Notes for EECS 126: Probability and Random Processes
UC Berkeley Fall 2019

Aditya Sengupta

February 5, 2020

Contents

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 1: Introduction

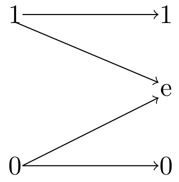
Lecturer: Shyam Parekh *29 August* *Aditya Sengupta*

Note: *LaTeX format adapted from template for lecture notes from CS 267, Applications of Parallel Computing, UC Berkeley EECS department.*

1.1 Applications of Probability

1.1.1 Capacity of a BEC

A binary erasure channel represents a transfer of bits over a channel, in which there is some probability that each bit is erased, i.e. it cannot be unambiguously read as 0 or 1.



Similarly, a BSC represents a channel in which a 0 can be read as a 1 or vice versa, instead of registering as an erased bit.

The key problem in a BEC or BSC is to maximize the capacity (the largest achievable rate of transmission) while the error still tends too 0.

1.1.2 Erdos-Renyi Random Graphs

An E-R random graph is a graph whose edges are given by a probability function $p(n)$ of the number of vertices n. $p(n)$ represents the probability of links being present for any given pair. Let $p(n) = \lambda \frac{\ln n}{n}$. If $\lambda > 1$, $G(n, p(n))$ is almost surely connected as $n \to \infty$, and if $\lambda < 1$, it is almost surely disconnected. “Almost surely” can be formalized as $\forall \epsilon > 0, \exists N \in \mathbb{R} \,\text{s.t.}\, n > N \implies$ the probability $p > 1 - \epsilon$.

1.1.3 DTMC and CTMC

A Markov chain is a system in which the current state is sufficient to describe the state evolution. This can be applied to fields such as queueing theory.

1.1.4 Estimation

A widely applicable problem is to estimate future states of a system given some system observation. One method to do this is maximum likelihood estimation, which chooses parameters θ to maximize the probability of the observed data D.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \mathbb{P}(D \mid \theta) \tag{1.1}$$

Another technique is max a posteriori estimation, which chooses the value that is most probable given observed data and a prior belief.

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \mathbb{P}(\theta \mid D) \tag{1.2}$$

1.1.5 Hypothesis Testing

Suppose there is a binary variable X that is observed at $\hat{X}$. Hypothesis testing seeks to minimize the proobability that $\hat{X} \neq X$. Specifically, we want to minimize $P(\hat{X} = 0 \mid X = 1)$ subject to $\mathbb{P}(\hat{X} = 1 \mid X = 0) \leq \beta$. This is resolved by Neyman-Pearson hypothesis testing.

1.1.6 Kalman Filtering

Kalman filtering is a method to obtain the optimal linear state estimate of a dynamic system. It combines a physical estimate of the system state with noisy measurements that are linear in the state to obtain the state estimate having the minimum mean-squared error at any given time.

Consider a discrete-time dynamic system with a state-transition rule

$$\vec{x}[k+1] = A[k]\vec{x}[k] + \vec{w}[k] \tag{1.3}$$

where $w[k]$ is Gaussian process noise described by a covariance matrix $Q[k]$. The impact of the white noise on the state prediction can be captured in a state covariance matrix P, with state-transition rule

$$P[k+1] = A[k]P[k]A[k]^T + Q[k] \tag{1.4}$$

A linear measurement of the system state is taken at each timestep k, given by

$$\vec{z}[k] = H[k]\vec{x}[k] + v[k] \tag{1.5}$$

where $v[k]$ is Gaussian measurement noise described by a covariance matrix $R[k]$.

The Kalman filter is an optimal state observer: given a series of measurements $\vec{z}[k]$, it reconstructs the $\vec{x}[k]$ that has the minimum mean-squared error (MMSE). For each timestep, the Kalman filter makes a state prediction and updates it based on the error between the predicted and actual measurements. Specifically, the error term in the measurement domain is converted to one in the state domain by multiplying by the *Kalman gain* matrix $K[k]$, which combines the process and measurement covariances to produce the optimal measurement-to-state gain.

1.2 Probabilistic Models

Probabilistic models are constructed by

- defining an experiment
- defining the sample space Ω of possible outcomes
- defining the probability law that assigns probabilities to subsets of Ω

For example, when flipping a coin, the sample space is {H, T}, and a probability law would have to assign probabilities to the subsets $\{\{\}, \{\text{H}\}, \{\text{T}\}, \{\text{H, T}\}\}$.

The set of events in the sample space form a $\sigma-$field, meaning they satisfy the field axioms. This is going to be relevant to this class never.

Suppose we randomly sample a real number in $[0, 1]$. The probability that any one element x is chosen is zero, because there are uncountably many reals in $[0, 1]$. The probability that a rational number is chosen is also 0, and it is possible to construct a set with uncountably many elements but zero probability, called the Cantor set. This is constructed by recursively removing the open middle third of each segment of the current interval, i.e. first deleting $\left(\frac{1}{3}, \frac{2}{3}\right)$, then deleting the middle third of $\left[0, \frac{1}{3}\right]$, i.e. $\left(\frac{1}{9}, \frac{2}{9}\right)$ and that of $\left[\frac{2}{3}, 1\right]$, i.e. $\left(\frac{7}{9}, \frac{8}{9}\right)$, and so on.

1.3 Probability axioms

A probability law must satisfy the following axioms:

- non-negativity: $\forall A \in \Omega, \mathbb{P}(A) \geq 0$
- additivity: $\forall A_i$ disjoint, $\mathbb{P}(\cup_i A_i) = \sum_i \mathbb{P}(A_i)$
- normalization: $P(\Omega) = 1$

1.4 Algebra of sets

Distributivity holds for sets:

$$S \cup (T \cap V) = (S \cup T) \cap (S \cup V) \tag{1.6}$$

$$S \cap (T \cup V) = (S \cap T) \cup (S \cap V) \tag{1.7}$$

$$\tag{1.8}$$

De Morgan's laws are also useful:

$$(\cup_n S_n)^c = \cap^n S_n^c \tag{1.9}$$

1.5 Properties of Probability Laws

1. $A \subset B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$; this can be shown using
2. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
3. $P(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(A^c \cap B) + \mathbb{P}(A^c \cap B^c \cap C)$.

1.6 Conditional Probability

Sometimes we know something about the result of an experiment that can change its probability. Consider a fair die in which you know the outcome is even. Then,

$$\mathbb{P}(6 \mid \text{outcome is even}) = \frac{1}{3} \tag{1.10}$$

More generally,

$$\mathbb{P}(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{1.11}$$

if $P(B) > 0$.

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 2: Conditional, Total Probability

Lecturer: Shyam Parekh *3 September* *Aditya Sengupta*

2.1 Conditional Probability

Consider a fair six-sided die. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. A simple example of a conditional probability is

$$\mathbb{P}(\text{outcome is 6} \mid \text{outcome is even}) \tag{2.1}$$

This condition restricts the sample space to $\Omega' = \{2, 4, 6\}$. The probability that an element chosen from this is 6 is simply $\frac{1}{3}$.

Assuming that $P(B) > 0$,

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{2.2}$$

Example 2.1. Consider two rolls of a fair 4-sided die. The outcomes exist in the sample space $\{1, 2, 3, 4\} \times \{1, 2, 3, 4\}$ (the Cartesian product of sample spaces). Each outcome in what I'm going to call Ω^2 is equally likely and has probability $\frac{1}{16}$. Then, consider the event $B = \min(x, y) = 2$, where $(x, y) \in \Omega^2$. There are five tuples in which this is the case, therefore $\mathbb{P}(B) = \frac{5}{16}$.

Consider

$$A = \{\max(x, y) = m \mid (x, y) \in \Omega^2\} \tag{2.3}$$

$$\mathbb{P}(A \mid B) = \begin{cases} \frac{2}{5} & m = 3, 4 \\ \frac{1}{5} & m = 2 \\ 0 & m = 1 \end{cases} \tag{2.4}$$

□

Example 2.2. Suppose a radar detection system is set up to detect airplanes. If an airplane is present, it is detected with $\mathbb{P}(D \mid P) = 0.99$. The probability of a false alarm, i.e. detecting an airplane when it is not there, is $\mathbb{P}(D \mid P^c) = 0.1$. The probability of a plane being present at all is $\mathbb{P}(P) = 0.05$.

We want to compute the probability of a false alarm and also no airplane being present (a joint probability, not a conditional one.)

$$\mathbb{P}(P^c \cup D) = \mathbb{P}(P^c) \cdot \mathbb{P}(D \mid P^c) = 0.95 \cdot 0.1 = 0.095 \tag{2.5}$$

Next, we want to compute the probability of an airplane passing undetected.

$$\mathbb{P}(P \cup D^c) = \mathbb{P}(P) \cdot \mathbb{P}(D^c \mid P) = \mathbb{P}(P) \cdot (1 - \mathbb{P}(D \mid P)) = 0.05 \cdot 0.01 = 0.0005 \tag{2.6}$$

□

Example 2.3. Suppose three cards are drawn randomly from a deck of 52 cards without replacement. We want to find the probability that none of the drawn cards are hearts. Let A_i be the event that the ith card is not a heart.

$$\begin{aligned} \mathbb{P}(A_1 \cap A_2 \cap A_3) &= \mathbb{P}(A_1) \cdot \mathbb{P}(A_2 \mid A_1) \cdot \mathbb{P}(A_3 \mid A_2 \cap A_1) && (2.7) \\ &= \frac{39}{52} \cdot \frac{38}{51} \cdot \frac{37}{50} && (2.8) \end{aligned}$$

□

2.2 Total Probability Theorem

Theorem 2.1. *Let $\{A_i \mid 1 \leq i \leq n\}$ be disjoint events. Then they form a partition of Ω, and*

$$\begin{aligned} \mathbb{P}(B) &= \sum_{i=1}^{n} \mathbb{P}(B \cap A_i) && (2.9) \\ &= \sum_{i=1}^{n} \mathbb{P}(A_i)\, \mathbb{P}(B \mid A_i) && (2.10) \end{aligned}$$

2.3 Bayes' Rule

Theorem 2.2. *Let A_i be disjoint events that form a partition of Ω. Assume $\forall 1 \leq i \leq n, \mathbb{P}(A_i) > 0$. Then, for any event B such that $\mathbb{P}(B) > 0$,*

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(A_i) \cdot \mathbb{P}(B \mid A_i)}{\mathbb{P}(B)} \tag{2.11}$$

Bayes' rule is useful for investigating cause-and-effect relationships. $\mathbb{P}(A_i \mid B)$ is called the posterior probability, which is calculable given the set of prior probabilities $\mathbb{P}(A_i)$. Combining this with the total probability theorem gives us

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(A_i) \cdot \mathbb{P}(B \mid A_i)}{\sum_{i=1}^{n} \mathbb{P}(A_i)\,\mathbb{P}(B \mid A_i)} \tag{2.12}$$

2.4 False Positives

Suppose there is a test for some event (e.g. someone having a rare disease) that comes up positive given that the event has happened with probability p_f, and comes up negative given the event has not happened with probability p_n. These are usually large, e.g. $p_f = p_n = 0.95$. The event has probability d, which is typically small, e.g. $d = 0.001$.

Given that the test is positive, what is the probability that the event has happened?

Suppose the event is E, and the detection event is D. We want to find $\mathbb{P}(E \mid D)$, which we can find using Bayes' rule:

$$\begin{aligned}\mathbb{P}(E \mid D) &= \frac{\mathbb{P}(D \mid E) \cdot \mathbb{P}(E)}{\mathbb{P}(E)\,\mathbb{P}(D \mid E) + \mathbb{P}(E^c)\,\mathbb{P}(D \mid E^c)} && (2.13)\\ &= \frac{p_f d}{p_f d + p_n(1-d)} && (2.14)\end{aligned}$$

For the proposed values of p_f, p_n, and d, $\mathbb{P}(E \mid D)$ comes out to 0.0187, which is much lower than what might be expected. Intuitively, this is because E^c is the majority of the sample space, but given E, D is most of the sample subspace. The probability of a false negative $(p_n(1-d))$ is much greater than that of a false positive, so the overall probability of a false positive is dominated by this.

2.5 Independence

Two events A, B are independent iff $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.

Remark 2.3. *If $\mathbb{P}(B) > 0, \mathbb{P}(A \mid B) = \mathbb{P}(A) \iff A, B$ are independent.*

Note that A, B disjoint does not imply they are independent.

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 3: Independence, Counting Review, Random Variables

Lecturer: Shyam Parekh *5 September* *Aditya Sengupta*

3.1 Independent Events

Example 3.1. Consider two rolls of a fair 4-sided die. Each of 16 outcomes is equally likely. Let the outcomes of the two individual rolls be a, b. For any i, j,

$$\mathbb{P}(a = i) = \frac{4}{16} \tag{3.1}$$

$$\mathbb{P}(b = j) = \frac{4}{16} \tag{3.2}$$

$$\mathbb{P}(a = i \wedge b = j) = \frac{1}{16} = \frac{4}{16} \cdot \frac{4}{16} \tag{3.3}$$

Therefore the events are independent.

Consider the events $A =$ the first roll is 1, and $S =$ the sum of two rolls is 5.

$$\mathbb{P}(A) = \frac{4}{16} \tag{3.4}$$

$$\mathbb{P}(B) = \mathbb{P}((1,4),(2,3),(3,2),(4,1)) = \frac{4}{16} \tag{3.5}$$

$$\mathbb{P}(A \cap B) = \frac{1}{16} = \frac{4}{16} \cdot \frac{4}{16} \tag{3.6}$$

Therefore the events are independent.

Let M, m represent the maximum and the minimum of two rolls.

$$\mathbb{P}(M = 2) = \mathbb{P}((1,2),(2,1),(2,2)) = \frac{3}{16} \tag{3.7}$$

$$\mathbb{P}(m = 2) = \mathbb{P}((2,2),(2,3),(3,2),(4,2),(2,4)) = \frac{5}{16} \tag{3.8}$$

$$\mathbb{P}(M = 2 \cap m = 2) = \mathbb{P}((2,2)) = \frac{1}{16} \neq \mathbb{P}(M = 2) \cdot \mathbb{P}(m = 2) \tag{3.9}$$

Therefore $M = 2$ and $m = 2$ are not independent.

□

3.1.1 Conditional Independence

Given an event C, the events A, B are conditionally independent if and only if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid C) \cdot \mathbb{P}(B \mid C) \tag{3.10}$$

Remark 3.1. *Unconditional independence does not imply conditional independence, and vice versa.*

3.1.2 Independence of several events

Theorem 3.2. *Events $A_i, 1 \leq i \leq n$ are independent iff $\forall S \subset \{1, \ldots, n\}$,*

$$\mathbb{P}(\cap_{i \in S} A_i) = \prod_{i \in S} \mathbb{P}(A_i) \tag{3.11}$$

It is necessary to have every subset included because of the following:

Lemma 3.3. *Pairwise independence does not imply independence.*

Example 3.2. Consider two independent fair coin tosses. Suppose H_i is the event that the ith toss comoes up heads, and suppose D is the event that two tosses have different outcomes. H_1 and H_2 are independent. Consider D given H_1:

$$\mathbb{P}(D \mid H_1) = \frac{\mathbb{P}(D \cap H_1)}{\mathbb{P}(H_1)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2} = \mathbb{P}(D) \tag{3.12}$$

Conditioning on H_1 does not change $\mathbb{P}(D)$, therefore D and H_1 are independent. Similarly, D and H_2 are independent.

However, consider both at once:

$$\mathbb{P}(D \cap H_1 \cap H_2) = 0 \neq \mathbb{P}(D) \cdot \mathbb{P}(H_1) \cdot \mathbb{P}(H_2) \tag{3.13}$$

Therefore, we have pairwise independence but not independence.

□

Similarly, joint independence does not imply independence.

Example 3.3. Consider two independent rolls of a fair six-sided die, whose outcomes are represented by d_1, d_2. Let A be the event $d_1 \in \{1, 2, 3\}$, B be that $d_1 \in \{3, 4, 5\}$, and C be that $d_1 + d_2 = 9$.

$$\mathbb{P}(A \cap B) = \frac{1}{6} \neq \mathbb{P}(A) \cdot \mathbb{P}(B) = \frac{1}{2} \cdot \frac{1}{2} \tag{3.14}$$

$$\tag{3.15}$$

Similarly, it can be verified that B, C or A, C are not pairwise independent. However,

$$\mathbb{P}(A \cap B \cap C) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{4}{36} = \frac{1}{36} \tag{3.16}$$

Therefore, A, B, C are *jointly independent.*

□

A sequence of independent trials is a set of trials in which the likelihood of each possible outcome does not change between each trial.

A sequence of independent trials is Bernoulli when there are only two possible outcomes of each trial.

3.2 Binomial Distribution

The binomial distribution models a sequence of Bernoulli trials, such as n tosses of a coin. Specifically, it gives the probability that k of the trials succeed, e.g. the probability of k heads.

$$\mathbb{P}(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.17}$$

where p is the probability of a success in each trial.

3.3 Counting Terminology and Results

1. There are $n!$ permutations of n objects.
2. There are ${}^kP_n = \frac{n!}{(n-k)!}$ $k-$permutations of n objects.
3. There are $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ ways to choose k objects out of n.
4. A partition of a set of n objects is a set of r groups each containing n_i objects such that $\sum_{i=1}^{r} n_i = n$. The number of ways to do this is $\binom{n}{n_1, \ldots, n_r} = \frac{n!}{n_1! n_2! \ldots n_r!}$

3.4 Random Variables

A random variable (RV) is a real-valued function $X : \Omega \to \mathbb{R}$ of the outcome of an experiment.

Theorem 3.4. *A function of a random variable is another random variable.*

Proof. `https://math.stackexchange.com/questions/1554011/proving-function-of-random-variable-is-a-random-variable` □

We can associate attributes to each random variable such as its mean and variance.

An RV can be conditioned on an event or on another RV. This allows us to define independence of RVs, or of an RV and an event.

3.4.1 Discrete RVs

A discrete RV is an RV to which only countable values can be associated. To each discrete RV, we can associate a probability mass function.

Some important discrete RVs are:

1. Bernoulli RV: with an associated parameter $0 \leq p \leq 1$, X takes on the values 0 or 1, corresponding to failure or success of a single binary experiment with probability of success p. The associated PMF is

$$p_x(k) = \begin{cases} p & k = 1 \\ 1-p & k = 0 \end{cases} \tag{3.18}$$

2. Binomial RV: with associated parameters $n \in \mathbb{Z}, 0 \leq p \leq 1$, X represents the total number of successful Bernoulli trials out of a total n. The associated PMF is

$$p_x(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.19}$$

3. Geometric RV: with an associated parameter $0 < p < 1$, X represents the number of failed trials until a successful one. It takes on any integer values greater than or equal to 0. The associated PMF is

$$p_x(k) = (1-p)^{k-1} p \tag{3.20}$$

   It can be verified that this sums to 1 as is required for a probability:

$$\sum_{k=1}^{\infty} (1-p)^{k-1} p = p \sum_{k=0}^{\infty} (1-p)^k = \frac{p}{1-(1-p)} = 1 \tag{3.21}$$

4. Poisson RV: with an associated parameter $\lambda > 0$, X represents a rapidly-decaying probability distribution for quantities like the number of discrete events in a continuous interval of time. Its PMF is

$$p_x(k) = \frac{e^{-\lambda} \lambda^k}{k!} \tag{3.22}$$

5. Uniform RV: with parameters $a < b$ representing the range, X represents a uniformly sampled random variable. Its PMF is

$$p_x(k) = \begin{cases} \frac{1}{b-a+1} & k \in [a,b] \cap \mathbb{Z} \\ 0 & k \notin [a,b] \cap \mathbb{Z} \end{cases} \tag{3.23}$$

Theorem 3.5. *Let $X_i, i = 1, \ldots, n$ be i.i.d(independent and identically distributed) Bernoulli RVs. Then $Y = \sum_{i=1}^{n} X_i$ is a binomial RV.*

3.5 Expectation

The expected value is also more commonly referred to as the average or mean. For an RV X with PMF $p_x(x)$,

$$\mathbb{E}[X] = \sum_{x} x p_x(x) \tag{3.24}$$

Technically, the condition $\sum_x |x| p_x(x) < \infty$ must be satisfied for the expectation to be well defined.

Example 3.4. Consider a PMF such that $x = 2^k$ with probability 2^{-k} for $k \in \mathbb{Z}, k > 1$ and $x = -2^{|k|}$ with probability $2^{-|k|}$ for $k \in \mathbb{Z}, k < -1$. We can verify that the probabilities sum to 1, but $\mathbb{E}[x]$ is not defined, because the above condition is not satisfied.

□

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 4: Expectation, Variance

Lecturer: Shyam Parekh *10 September* *Aditya Sengupta*

4.1 Definitions of expectation and variance

Previously, we saw that the expected value of a discrete RV is given by a weighted sum over its PMF:

$$\mathbb{E}[X] = \sum_x x p_x(x) \tag{4.1}$$

Any operation applied to X is also a random variable whose expectation and variance can be found. In particular, we refer to $\mathbb{E}[X^n]$ as the nth moment of X. More generally, we can say

$$\mathbb{E}[g(X)] = \sum_x g(x) p_x(x) \tag{4.2}$$

We can calculate the variance, which is a quantity representing the spread of a distribution, by

$$\text{var}(X) = \mathbb{E}[(X - E[X])^2] \geq 0 \tag{4.3}$$

The standard deviation of a distribution is the square root of its variance.

The variance is zero if and only if $X = E[X]$ with probability 1.

$$\sum_x (X - E[X])^2 p_X(x) = 0 \tag{4.4}$$

Suppose $Y = aX + b$, i.e. Y is linearly related to X. The expected value is

$$\mathbb{E}[Y] = a\,\mathbb{E}[X] + b \tag{4.5}$$

and the variance is

$$\begin{aligned} \text{var}(Y) &= \sum_x (aX + b - \mathbb{E}[aX + b])^2 p_x(X) && (4.6) \\ &= \sum_x (aX + b - a\,\mathbb{E}[X] - b)^2 p_x(X) && (4.7) \\ &= a^2 \sum_x (X - \mathbb{E}[X])^2 p_x(X) = a^2\,\text{var}(X) && (4.8) \end{aligned}$$

Therefore, linear scaling becomes quadratic under variance, and constant shifts do not affect the variance.

We can show that $\mathrm{var}(X) = \mathbb{E}[X^2] - E[X]^2$:

$$\begin{aligned} \mathrm{var}(X) &= \sum_x (X - \mathbb{E}[X])^2 p_x(X) && (4.9) \\ &= \sum_x (X^2 - 2X\,\mathbb{E}[X] + (E[X])^2) p_x(X) && (4.10) \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 && (4.11) \end{aligned}$$

4.2 Expectation and variance of common RVs

4.2.1 Bernoulli

With parameter p, the expectation of a Bernoulli RV is $\mathbb{E}[X] = p \cdot 1 + (1-p) \cdot 0 = p$, and the variance is $\mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1-p)$.

4.2.2 Binomial

With parameters n, p, the expectation of a binomial RV is $\mathbb{E}[X] = np$ and the variance is $\mathrm{var}[X] = np(1-p)$. This is obtained by sampling from a Bernoulli distribution n times [citation needed].

4.2.3 Geometric

$$\begin{aligned} \mathbb{E}[X] &= \sum_{k=1}^{\infty} k(1-p)^{k-1} p && (4.12) \\ &= -p \sum_{k=0}^{\infty} \frac{d}{dp} (1-p)^k && (4.13) \\ &= -p \frac{d}{dp} \frac{1}{1-(1-p)} = \frac{1}{p} && (4.14) \end{aligned}$$

Similarly, $\mathbb{E}[X^2] = \frac{2}{p^2} - \frac{1}{p}$, therefore $\mathrm{var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}$.

4.2.4 Poisson

$$\begin{aligned} \mathbb{E}[X] = \sum_{k=0}^{\infty} k \frac{e^{-\lambda}\lambda^k}{k!} &= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} && (4.15) \\ &= \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \lambda && (4.16) \end{aligned}$$

The second moment is

$$
\begin{aligned}
\mathbb{E}[X^2] &= \sum_{k=0}^{\infty} k^2 \frac{e^{-\lambda}\lambda^k}{k!} && (4.17)\\
&= \sum_{k=0}^{\infty} k(k-1)e^{-\lambda\lambda^k} k! + \sum_{k=0}^{\infty} k \frac{e^{-\lambda}\lambda^k}{k!} && (4.18)\\
&= \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{]\lambda^{k-2}}{(k-2)!} + \lambda && (4.19)\\
&= \lambda^2 + \lambda && (4.20)
\end{aligned}
$$

Therefore $\mathrm{var}[X] = \lambda$.

4.3 Joint PMFs

The joint PMF of two variables is the probability that they both have a certain fixed value,

$$
\mathbb{P}_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}(\{X = x\} \cap \{Y = y\}) \tag{4.21}
$$

From this, we can get the *marginal* PMFs by summing over all values of one of the variables,

$$
\begin{aligned}
\mathbb{P}_X(x) &= \sum_y \mathbb{P}_{X,Y}(x,y) && (4.22)\\
\mathbb{P}_Y(y) &= \sum_x \mathbb{P}_{X,Y}(x,y) && (4.23)\\
& && (4.24)
\end{aligned}
$$

We can extend the idea of the linear expectation of a linear function of random variables to these joint distributions,

$$
\begin{aligned}
\mathbb{E}[g(X,Y)] &= \sum_x \sum_y g(x,y)\, \mathbb{P}_{X,Y}(x,y) && (4.25)\\
\mathbb{E}[aX + bY + c] &= a\,\mathbb{E}[X] + b\,\mathbb{E}[Y] + c && (4.26)
\end{aligned}
$$

4.4 Conditional PMFs

We can condition a PMF on an event:

$$
\mathbb{P}_{X|A}(X = x \mid A) = \frac{\mathbb{P}(\{X = x\} \cap A)}{\mathbb{P}(A)} \tag{4.27}
$$

and we can verify that this is still a valid PMF:

$$\sum_x \mathbb{P}_{X|A}(x|A) = 1 \tag{4.28}$$

We can also condition on another RV,

$$\mathbb{P}_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{\mathbb{P}_{X,Y}(x, y)}{\mathbb{P}_Y(y)} \tag{4.29}$$

Again, we can verify that this is still a valid PMF,

$$\sum_x \mathbb{P}_{X|Y}(x|y) = 1 \tag{4.30}$$

4.5 Conditional Expectation

Suppose $\mathbb{P}(A) > 0$. Then

$$\mathbb{E}[X \mid A] = \sum_x x\, \mathbb{P}_{x|A}(x \mid A) \tag{4.31}$$

$$\mathbb{E}[X \mid Y = y] = \sum_x x\, \mathbb{P}(X \mid Y)(x \mid y) \tag{4.32}$$

4.6 Total Expectation Theorem

Theorem 4.1. *Let $\{A_i\}_{i=1}^n$ be disjoint events such that $\mathbb{P}(A_i) > 0$ for each i and such that they form a partition of Ω. Then*

$$\mathbb{E}[X] = \sum_n P(A_i)\, \mathbb{E}[X \mid A_i] \tag{4.33}$$

$$\mathbb{E}[X] = \sum_x x\, \mathbb{P}_X(x) = \sum_x x \sum_{i=1}^n \mathbb{P}(A_i)\, \mathbb{P}_{X|A_i}(x) \tag{4.34}$$

We can think of $X|Y$ as a random variable, which allows us to rewrite a marginal PMF as follows:

$$\mathbb{E}[X] = \sum_y \mathbb{P}_Y(y)\, \mathbb{E}[X \mid Y = y] = \mathbb{E}[\mathbb{E}[X|Y = y_i]] = \tag{4.35}$$

This is referred to as the *theorem of iterated expectations.*

Example 4.1. Suppose a message is being sent from Boston to New York with probability 0.5, to Chicago with probability 0.3, and to San Francisco with probability 0.2. Each of these destinations has an expectation of the transit time; say $\mathbb{E}[T \mid N] = 0.05, \mathbb{E}[T \mid C] = 0.1, \mathbb{E}[T \mid S] = 0.3$.

The overall expectation of the transit time can be found by multiplying pairwise the probability of each destination by its expected value,

$$\mathbb{E}[T] = 0.5 \cdot 0.05 + 0.3 \cdot 0.1 + 0.2 \cdot 0.3 = 0.115 \tag{4.36}$$

□

4.7 Geometric RV Expectation and Moments

We can use the idea of conditional expectation to examine thresholds on the geometric distribution.

$$\mathbb{E}[X \mid X = 1] = 1 \tag{4.37}$$

$$\mathbb{E}[X \mid X > 1] = 1 + \mathbb{E}[X] \tag{4.38}$$

By the total expectation theorem, we get that

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{P}(X = 1)\,\mathbb{E}[X \mid X = 1] + \mathbb{P}(X > 1)\,\mathbb{E}[X \mid X > 1] && (4.39) \\ &= p \cdot 1 + (1 - p) \cdot (1 + \mathbb{E}[X]) && (4.40) \end{aligned}$$

Therefore, $\mathbb{E}[X] = \frac{1}{p}$. Similarly, $\mathbb{E}[X^2 \mid X = 1] = 1$ and $\mathbb{E}[X^2 \mid X > 1] = \mathbb{E}[(1 + X)^2]$, so similarly we can get the above expression for $\mathbb{E}[X^2]$ and from that the variance.

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 5: RV independence, covariance

Lecturer: Shyam Parekh *12 September* *Aditya Sengupta*

5.1 Independence of random variables

A random variable X is independent of event A iff $\mathbb{P}(X = x \text{ and } A) = \mathbb{P}_X(x) \cdot \mathbb{P}(A)\ \forall x$. Similarly, we say that random variables X, Y are independent iff $\mathbb{P}_{X,Y}(X = x, Y = y) = \mathbb{P}_X(x) \cdot \mathbb{P}_Y(y)\ \forall x, y$.

Equivalently, X, Y are independent iff $\mathbb{P}_{X|Y}(x|y) = \mathbb{P}_X(x)\ \forall y \,\text{s.t.}\, \mathbb{P}_Y(y) > 0\ \forall x$.

Theorem 5.1. *If X, Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.*

Proof.

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_x \sum_y xy\, \mathbb{P}_{X,Y}(x,y) && (5.1)\\
&= \sum_x \sum_y xy\, \mathbb{P}_X(x)\, \mathbb{P}_Y(y) && (5.2)\\
&= \sum_x x\, \mathbb{P}_X(x) \cdot \sum_y y\, \mathbb{P}_Y(y) && (5.3)\\
&= \mathbb{E}[X]\,\mathbb{E}[Y] && (5.4)
\end{aligned}$$

□

5.2 Covariance

The covariance of two random variables is given by

$$\begin{aligned}
\operatorname{cov}(X,Y) &= \mathbb{E}[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])] && (5.5)\\
&= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] && (5.6)
\end{aligned}$$

The correlation coefficient of two random variables is given by

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}} \tag{5.7}$$

provided that $\operatorname{var}(X), \operatorname{var}(Y) > 0$.

If X, Y are independent, then $\operatorname{cov}(X,Y) = 0$. Also, if $\operatorname{var}(X), \operatorname{var}(Y) > 0$ but they are independent, then $\rho(X,Y) = 0$. In this case, we say that X, Y are uncorrelated.

Note that the covariance of two random variables being zeroo does not mean that they are independent.

Example 5.1. Randomly choose any of the four points on the unit circle with integer coordinates. The x and y coordinates are dependent, as we can see by calculating the marginal PMF for x (and similarly the one for y):

$$\mathbb{P}_X(x) = \begin{cases} \frac{1}{4} & x = -1 \\ \frac{1}{4} & x = 1 \\ \frac{1}{2} & x = 0 \end{cases} \tag{5.8}$$

Also, $\mathbb{E}[XY] = 0$, and $\mathbb{E}[X]\,\mathbb{E}[Y] = 0$, so the correlation coefficient is zero. However, X and Y are not independent. For example, if $X = \pm 1$, Y can only be 0.

□

Theorem 5.2.

$$|\rho(X,Y)| \leq 1 \tag{5.9}$$

Proof. We first show the Cauchy-Schwarz inequality holds on expected values:

$$(\mathbb{E}[XY])^2 \leq \mathbb{E}[X^2]\,\mathbb{E}[Y^2] \tag{5.10}$$

NOTE: I couldn't follow the proof in lecture, so this is one that I came up with.

Consider the vector space of random variables (they can be added, scaled, and have a zero element, and things like distributivity follow easily). Then, define an inner product on this: $\langle X, Y \rangle = \mathbb{E}[XY]$. This is symmetric by the symmetry of regular multiplication, linear by the linearity of expectation, and positive-definite by $\mathbb{E}[X^2] > 0 \ \forall X$ such that $\mathbb{P}_X(0) = 0$. `https://inst.eecs.berkeley.edu/~ee126/sp18/projection.pdf`. Since it's an inner product space, the inner product satisfies the C-S inequality.

Consider $\tilde{X} = X - \mathbb{E}[X], \tilde{Y} = Y - \mathbb{E}[Y]$.

$$\rho(X,Y)^2 = \frac{(\mathbb{E}[\tilde{X}\tilde{Y}])}{\mathbb{E}[\tilde{X}^2]\,\mathbb{E}[\tilde{Y}^2]} \leq 1 \tag{5.11}$$

$$\therefore |\rho(X,Y)| \leq 1 \tag{5.12}$$

□

If X, Y are independent, then their variances add linearly: $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$. This is because they can be shifted without affecting their variance to have zero mean, and the freshman's dream holds because the cross-term is zero by independence:

$$\text{var}(X+Y) = \text{var}(\tilde{X} + \tilde{Y}) = \mathbb{E}[(\tilde{X} + \tilde{Y}^2)] \tag{5.13}$$

where the second term vanishes because both variables are zero-mean. We expand this further,

$$\mathrm{var}(X+Y) = \mathbb{E}[\tilde{X}^2] + \mathbb{E}[\tilde{Y}^2] + 2\,\mathbb{E}[\tilde{X}\tilde{Y}] = \mathbb{E}[\tilde{X}^2] + \mathbb{E}[\tilde{Y}^2] = \mathrm{var}(X) + \mathrm{var}(Y) \tag{5.14}$$

5.3 Variance of Binomial RVs

Let $\{X_i\}_{i=1}^n$ be i.i.d. Bernoulli RVs. We can use the linearity of variances of independent variables to get

$$\mathrm{var}(X) = \mathrm{var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{var}(X_i) = np(1-p) \tag{5.15}$$

5.4 Entropy

Suppose X takes values x_i with probabilities p_i. The entropy of X is defined as

$$H(X) = -\sum_{i=1}^n p_i \ln(p_i) \tag{5.16}$$

Each component $-\log(p_i)$ is the self-information of $X = x_i$, and the entropy is the expected value of the self-information.

5.5 Continuous RVs

A random variable is continuous if there is a function $f_x \geq 0$ such that

$$p(x \in B) = \int_B f_x(x)dx \tag{5.17}$$

Here, f_x is referred to as a probability density function. f_x is Riemann integrable. For our purposes, f_x is piecewise continuous. The sets B over which we integrate are unions of finite or countably infinite intervals of $\mathbb{R}$.

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 6: Continuous Random Variables

Lecturer: Shyam Parekh *17 September* *Aditya Sengupta*

NOTE: this was a lecture by TAs, so there are slides with all this content. That meant I didn't go into as much detail as I would otherwise.

6.1 Probability Densities

In a continuous space, we describe a distribution with a probability density function (pdf). A valid probability density of a continuous random variable satisfies the conditions of

1. non-negativity: $\forall x, f(x) \geq 0$
2. summing to 1: $\int_D f(x) = 1$ (D is the domain of x)

We can get probabilities from densities by $\mathbb{P}(X \in B) = \int_B f_X(x)dx$ for any $B \subset D$. Specifically, for an interval in the reals,

$$\mathbb{P}(X \in [a, b]) = \mathbb{P}(a \leq X \leq B) = \int_a^b f_X(x)dx \tag{6.1}$$

Since the probability of attaining any particular point is zero, whether this interval is open or closed does not matter.

Example 6.1. Suppose we randomly sample a point in the unit sphere.

- The probability of picking the origin is 0.
- The probability density of picking the origin is $\frac{1}{\frac{4}{3}\pi r^3} = \frac{3}{4\pi}$.
- The probability of picking a point on the surface is 0, because the surface has no volume.
- The probability of picking a point within a radius of $\frac{1}{2}$ is $\frac{1}{8}$.

□

6.2 Cumulative Distribution Functions

The cumulative distribution function (cdf) of a random variable is $F_X(x) = \mathbb{P}(X \leq x)$. This is given explicitly by

$$F_X(x) = \int_{-\infty}^{x} f(t)dt \tag{6.2}$$

Since the pdf is always nonnegative, the cdf is always increasing.

Example 6.2. Let R be the distance from the origin to a point randomly sampled on a unit sphere.

- The cdf of R is $F_R(r) = \frac{3}{4\pi} \cdot \frac{4}{3}\pi r^3 = r^3$.
- The pdf of R is $\frac{d}{dr} r^3 = 3r^2$.
- The expectation of R is $\int_0^1 r \cdot 3r^2 dr = \frac{3}{4}$.

□

Note that the density can be greater than 1, as long as it is less than 1 when integrated over any interval. (??)

6.3 Continuous Distributions

6.3.1 Uniform Distribution

The density is uniform across a bounded interval. For $X \sim Unif(a, b)$,

$$f_X(x) = \frac{1}{b-a}, a < x < b \tag{6.3}$$

$$\mathbb{E}[X] = \frac{a+b}{2}, \text{var}(X) = \frac{(b-a)^2}{12} \tag{6.4}$$

6.3.2 Exponential Distribution

The density uniformly decays.

$$f_X(x) = \lambda e^{-\lambda x}, x > 0 \tag{6.5}$$

The cdf is $F_X(x) = 1 - e^{-\lambda x}$, the expectation is $\frac{1}{\lambda}$, and the variance is $\frac{1}{\lambda^2}$.

The exponential distribution is memoryless, meaning that $\mathbb{P}(X > x + a \mid X > x) = \mathbb{P}(X > a)$. The analogous discrete distribution is the geometric distribution. These are the only distributions in the discrete and continuous spaces with the memoryless property.

Intuitively, $Geo(p)$ approaches $Exp(\lambda)$ as $n \to \infty$. Recall the cdf of the geometric distribution is $F_X(n) = 1 - (1-p)^n$. As an ansatz, let $\delta = \frac{-\ln(1-p)}{\lambda}$; we have $e^{-\lambda\delta} = 1 - p$, so that $F_X(n) = F_Y(n\lambda)$. If we drive $\delta \to 0$, we have infinitely many trials per second while the expected number of trials stays constant. This approaches a continuous exponential distribution.

6.4 Gaussian Distribution

$$f_X(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \tag{6.6}$$

The cdf is given by the error function.

The sum of two independent Gaussians is Gaussian: if $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$ then $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. This is only true if they are independent.

A Gaussian multiplied by a constant is Gaussian: if $X \sim N(\mu, \sigma^2)$, then $aX \sim N(a\mu, a^2\sigma^2)$. These properties allow us to convert any Gaussian to the standard Gaussian $N(0, 1)$.

6.5 Joint PDFs

We can make them, and resolve them by double-integration. This gives us conditional probability densities:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \tag{6.7}$$

6.6 Independence

Similar to discrete RVs, there are three different definitions that are independent:

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) \tag{6.8}$$

and two others on the slides.

6.7 Conditional Expectation

$$\mathbb{E}[Y \mid X = x] = \int_{-\infty}^{\infty} y \ldots\ldots\ldots \tag{6.9}$$

6.8 Combining Discrete and Continuous RVs

You can do that.

Example 6.3. Let X be the outcome of a dice roll and let $Y = Exp(X)$. Then

$$f_{Y|X}(y \mid x) = xe^{-xy} \tag{6.10}$$

This allows us to define a marginal density,

$$f_Y(y) = \sum_{x=1}^{6} f(y \mid x)p_x(x) = \frac{1}{6}\left(f(y \mid x = 1) + \cdots + f(y \mid x = 6)\right) \tag{6.11}$$

□

6.9 Change of Variables

Example 6.4. Let $X \sim U[0, 1]$ and $Y = 2X$. We can't directly manipulate the density, but we can manipulate the cdf:

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(2X \leq Y) = \mathbb{P}(X \leq \frac{y}{2}) = F_X\left(\frac{y}{2}\right) \tag{6.12}$$

Then, we take a derivative:

$$f_Y(y) = \frac{d}{dy}F_X(\frac{y}{2}) = f_X\left(\frac{y}{2}\right) \cdot \frac{1}{2} \tag{6.13}$$

This is the pdf.

□

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 7: Order Statistics, MGFs

Lecturer: Shyam Parekh *19 September* *Aditya Sengupta*

7.1 Order Statistics

kth order statistics refers to the kth smallest random variable from a set of RVs $x_1, \dots, x_n$. In particular, we're interested in the smallest and the largest RVs.

Suppose $X_1, \dots, X_n$ are i.i.d. RVs with CDFs $F_X(x)$. Let

$$Y = \min_{1 \le k \le n} X_k \tag{7.1}$$

$$Z = \max_{1 \le k \le n} X_k \tag{7.2}$$

Then

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \le y) && (7.3)\\ &= \mathbb{P}(\min(X_1, \dots, X_n) \le y) && (7.4)\\ &= 1 - \mathbb{P}(\min(X_1, \dots, X_n) \ge Y) && (7.5)\\ &= 1 - (1 - F_X(y))^n && (7.6) \end{aligned}$$

Similarly,

$$\begin{aligned} F_Z(z) &= \mathbb{P}(Z \le z) && (7.7)\\ &= \mathbb{P}(\max(x_1, \dots, x_n) \le z) && (7.8)\\ &= \mathbb{P}(x_1 \le z)\,\mathbb{P}(x_2 \le z) \dots \mathbb{P}(x_n \le z) && (7.9)\\ &= (F_X(z))^n && (7.10) \end{aligned}$$

Example 7.1. The min of exponentially distributed i.i.d RVs can be found using the cdf, $F_X(x) = 1 - e^{-\lambda x}, x \ge 0$.

$$F_Y(y) = 1 - (1 - (1 - e^{-\lambda y}))^n = 1 - e^{-n\lambda y} \tag{7.11}$$

If x_i has parameter λ_i, we get

$$F_Y(y) = 1 - e^{y \sum_i \lambda_i} \tag{7.12}$$

That is, Y is exponentially distributed with parameter $\sum_i \lambda_i$.

□

7.2 Convolution

7.2.1 Discrete Case

Let X and Y be independent RVs with PMFs p_X, p_Y.

$$p_Z(z) = \mathbb{P}(X + Y = z) = \sum_x \mathbb{P}(X = x, Y = z - x) = \sum_x \mathbb{P}_x(x)\,\mathbb{P}_Y(z - x) \tag{7.13}$$

This is called convolution.

7.2.2 Continuous Case

Let X, Y be continuous RVs with PDFs f_X, f_Y. Let $Z = X + Y$ with PDF f_Z.

$$\mathbb{P}(X + Y \leq z \mid X = x) = \mathbb{P}(Y \leq z - x \mid X = x) = \mathbb{P}(Y \leq z - x) \tag{7.14}$$

By differentiating, we have

$$f_{Z|X}(z|x) = f_Y(z - x) \tag{7.15}$$

Using the multiplication rule, we find that the joint pdf on X and Z equals the unconditional pdf on X multiplies by the conditional pdf on Z given X:

$$f_{X,Z}(x, z) = f_X(x) f_{Z|X}(z|x) \tag{7.16}$$

Finally,

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_{X,Z}(x, z)dx && (7.17) \\ &= \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)dx && (7.18) \end{aligned}$$

This is the continuous convolution of X and Y.

Example 7.2. The convolution of two rectangles is a triangle.

□

Example 7.3. Let X, Y be independent normally distributed RVs with parameters (μ_x, σ_x^2) and (μ_y, σ_y^2). Then, the sum is also normally distributed with parameter $(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$.

□

7.3 Moment Generating Functions

For a random variable X, the moment generating function is

$$M_X(s) = \mathbb{E}[e^{sx}] \tag{7.19}$$

for a real number s.

Example 7.4. Consider a Poisson RV X with PMF $p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$. The MGF is

$$\begin{aligned} M_X(s) &= \sum_{x=0}^{\infty} e^{sx} \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{x=a}^{\infty} \frac{a^x}{x!} && (7.20) \\ &= e^{\lambda(e^s - 1)} && (7.21) \end{aligned}$$

where $a = \lambda e^s$.

□

Example 7.5. Consider an exponential RV with parameter λ; the MGF is

$$M(s) = \lambda \int_0^{\infty} e^{sx} e^{-\lambda x} dx = \frac{\lambda}{\lambda - s}, s < \lambda \tag{7.22}$$

□

The MGFs can give us the moments of a continuous random variable. Let X be a continuous RV; then

$$M(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x) dx \tag{7.23}$$

$$\left. \frac{d^n M(s)}{ds^n} \right|_{s=0} = \int x^n f(x) dx = \mathbb{E}[X^n] \tag{7.24}$$

An MGF uniquely specifies the pdf/pmf of a random variable.

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 8: Moment Generating Functions, Bounds

Lecturer: Shyam Parekh *24 September* *Aditya Sengupta*

8.1 MGF Examples

Example 8.1. Consider a Poisson(λ) distributed random variable. Its MGF is

$$\begin{aligned} M(s) &= \sum_{x=0}^{\infty} e^{sx}\frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{(e^s\lambda)^x}{x!} && (8.1)\\ &= e^{\lambda(e^s-1)} && (8.2)\end{aligned}$$

□

Example 8.2. Consider an Exponential(λ) random variable. Its MGF is

$$\begin{aligned} M(s) &= \lambda\int_0^{\infty} e^{sx}e^{-\lambda x}dx = \lambda\left.\frac{e^{s-\lambda}}{s-\lambda}\right|_0^{\infty} && (8.3)\\ &= \frac{\lambda}{\lambda-s}, s<\lambda && (8.4)\end{aligned}$$

□

Example 8.3. Consider $Y \sim \mathcal{N}(0,1)$.

$$M_Y(s) = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{y^2/2}e^{sy}dy \qquad (8.5)$$

Completing the square, we get

$$M_Y(s) = e^{s^2/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-y^2/2+sy-s^2/2}dy \tag{8.6}$$

$$= \frac{1}{\sqrt{2\pi}}e^{s^2/2}\int_{-\infty}^{\infty} e^{-(y-s)^2/2}dy \tag{8.7}$$

$$= e^{s^2/2} \tag{8.8}$$

□

8.2 Moments from MGF

Lemma 8.1. *The nth moment of a distribution is equal to the nth derivative of its MGF,*

$$\left.\frac{d^n M(s)}{ds^n}\right|_{s=0} = E[X^n] \tag{8.9}$$

Lemma 8.2. *If $M_X(s)$ is finite over $(-a, a)$ for some $a > 0$, then the MGF uniquely determines the CDF for the random variable X.*

Example 8.4. Consider a random variable whose MGF is

$$M(s) = \frac{1}{4}e^{-s} + \frac{1}{2} + \frac{1}{8}e^{4s} + \frac{1}{8}e^{5s} \tag{8.10}$$

Then, the PMF of X is the result of an inverse Laplace transform:

$$P(X = x) = \frac{1}{4}\delta[x+1] + \frac{1}{2}\delta[x] + \frac{1}{8}\delta[x-4] + \frac{1}{8}\delta[x-5] \tag{8.11}$$

□

Example 8.5. Suppose

$$M(s) = \frac{pe^s}{1-(1-p)e^s} \tag{8.12}$$

We recognize the denominator as the sum of an infinite geometric series, so we rewrite the MGF:

$$M(s) = pe^s\left(1 + (1-p)e^s + (1-p)^2e^{2s} + \dots\right) \tag{8.13}$$

Assuming $(1-p)e^s < 1$, we get

$$P(X = k) = p(1-p)^{k-1} \tag{8.14}$$

i.e. $X \sim \text{Geom}(p)$.

□

Theorem 8.3. *If $Z = \sum X_i$, then $M_Z(s) = \prod M_{x_i}(s)$.*

Example 8.6. The MGF of a Bernoulli(p) random variable is $M_{x_i}(s) = 1 - p + pe^s$, therefore the MGF of a Binomial(n, p) random variable is

$$M_X(s) = \prod_{i=1}^{n} M_{x_i}(s) = (1 - p + pe^s)^n \tag{8.15}$$

From this, we can get the moments of the binomial distribution:

$$\mathbb{E}[X] = \frac{d}{ds}M_X(s)\bigg|_{s=0} = n(1-p+pe^s)^{n-1}pe^s\big|_{s=0} = np \tag{8.16}$$

$$s\,\mathbb{E}[X^2] = n^2p^2 + np(1-p) \tag{8.17}$$

$$\tag{8.18}$$

□

Example 8.7. Let $X \sim \text{Poisson}(\lambda), Y \sim \text{Poisson}(\mu)$ and let $Z = X + Y$. The MGF is

$$M_Z(s) = M_X(s)M_Y(s) = e^{\lambda(e^s-1)}e^{\mu(e^s-1)} = e^{(\lambda+\mu)(e^s-1)} \tag{8.19}$$

i.e. $Z \sim \text{Poisson}(\lambda + \mu)$. This gives us the linearity of the Poisson distribution.

□

8.3 Bounds

8.3.1 Markov Inequality

Theorem 8.4. *If a random variable $X \geq 0$, then $\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$ $\forall a > 0$.*

Proof.

$$\mathbb{E}[X] = \sum_{x} x\,\mathbb{P}(X = x) \tag{8.20}$$

$$\geq \sum_{x \geq a} x\,\mathbb{P}(X = x) \tag{8.21}$$

$$\geq a \sum_{x \geq a} \mathbb{P}(X = x) \tag{8.22}$$

$$= a\,\mathbb{P}(X \geq a) \tag{8.23}$$

Therefore

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \tag{8.24}$$

The continuous analogue follows similarly.

Alternatively, let $Y_a = au(a)$. Since $Y_a \leq X$, $\mathbb{E}[Y_a] \leq E[X]$, and

$$\mathbb{E}[Y_a] = a\,\mathbb{P}(Y_a = a) = a\,\mathbb{P}(x \geq a) \leq E[X] \implies \mathbb{P}(X \geq a) \leq \frac{E[X]}{a}\ \forall a > 0 \tag{8.25}$$

□

8.3.2 Chebyshev's Inequality

Theorem 8.5. *If X is a random variable with mean μ and variance σ^2, then*

$$\mathbb{P}(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}\ \forall c > 0 \tag{8.26}$$

Proof. Given $c > 0$,

$$\mathbb{P}(|X - \mu| \geq c) \leq E[(x - \mu)^2 \geq c^2] \leq \mathbb{E}[(x - \mu)^2]/c^2 = \frac{\sigma^2}{c^2} \tag{8.27}$$

□

If we take $c = k\sigma$ for some k, we get the probability of being within k standard deviations,

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}\ , k > 0 \tag{8.28}$$

8.3.3 Chernoff Bounds

Theorem 8.6. *(Positive Chernoff Bound) Let $M(s)$ be the MGF of an RV X. For every a and every $s \geq 0$, $\mathbb{P}(X \geq a) \leq e^{-sa}M(s)$.*

Proof. Let $s > 0$. Then $\mathbb{P}(X \geq a) = \mathbb{P}(sX \geq sa) = \mathbb{P}(e^{sX} \geq e^{sa})$. We apply the Markov inequality to get

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[e^{sx}]}{e^{sa}} \implies \mathbb{P}(X \geq a) \leq e^{-sa}M(s) \tag{8.29}$$

□

Theorem 8.7. *(Negative Chernoff Bound) For every a and every $s \leq 0$, $\mathbb{P}(x \leq a) \leq e^{-sa}M(s)$.*

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 9: Bounds, CLT

Lecturer: Shyam Parekh *1 October* *Aditya Sengupta*

9.1 Refining the Chernoff Bound

The Chernoff bound can be refined by applying a Legendre transform:

$$\mathbb{P}(X \geq a) \leq e^{-\phi(a)} \; \forall a \tag{9.1}$$

where $\phi(a) = \max_{s \geq 0}(sa - \ln M(s))$.

Proof.

$$\mathbb{P}(X \geq a) \leq \min_{s \geq 0}(e^{-sa}M(s)) = \min_{s \geq 0} e^{-(sa - \ln M(s))} \tag{9.2}$$

$$\leq e^{-\max_{s \geq 0}(sa - \ln M(s))} = e^{-\phi(a)} \tag{9.3}$$

□

9.2 Jensen's Inequality

Theorem 9.1. *Let f be a twice-differentiable convex function (i.e. $f''(x) > 0$.) Then*

$$f(E[x]) \leq E[f(x)]. \tag{9.4}$$

Proof. Wasn't proved in class, but I did a supplementary writeup on this! □

9.3 The Weak Law of Large Numbers

Theorem 9.2. *Let $X_1, X_2, \ldots$ be i.i.d. RVs with mean μ and variance σ^2. Let M_n be their sample mean, $M_n = \frac{X_1 + \cdots + X_n}{n}$. As $n \to \infty$, the sample mean converges to the actual mean.*

This comes out of the Central Limit Theorem, which states that the sample mean doesn't just have these properties; its distribution is the (unique) normal distribution that has these properties. The direct proof you're about to see proves equality of arbitrary probability bounds on the sample mean, but not equality of the distributions.

Proof. By linearity of expectation, $\mathbb{E}[M_n] = \mu$ and because i.i.d. variances are additive, $\text{var}(M_n) = \sum_{i=1}^{n} \frac{\text{var}(X_i)}{n^2} = \frac{\sigma^2}{n}$.

By Chebyshev's inequality,

$$\mathbb{P}(|M_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \tag{9.5}$$

$$\lim_{n\to\infty} \mathbb{P}(|M_n - \mu| \geq \epsilon) = \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0 \tag{9.6}$$

□

This gives us the notion of convergence in probability.

Let $X_1, \ldots, X_n$ be RVs, not necessarily i.i.d. We say X_n converges to X in probability if for every $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0 \tag{9.7}$$

Example 9.1. Suppose some fraction p of voters support a candidate. Let M_n be the fraction of voters sampled who support the candidate. Then by a result from the homework:

$$\mathbb{P}(|M_n - p| \geq \epsilon) \leq \frac{p(1-p)}{n\epsilon^2} \tag{9.8}$$

Further, we note that $\max_{0\leq p\leq 1} p(1-p) = \frac{1}{4}$. Without knowing p, we therefore have

$$\mathbb{P}(|M_n - p| \geq \epsilon) \leq \frac{1}{4n\epsilon^2} \tag{9.9}$$

□

9.4 The Strong Law of Large Numbers

Theorem 9.3. *Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ. Let M_n be their sample mean. Then, M_n converges almost surely to μ, i.e.*

$$\mathbb{P}(\lim_{n\to\infty} M_n = \mu) = 1 \tag{9.10}$$

You say "almost sure" because there are still technically sequences of events that don't have the sample mean equal to the actual mean, just that those have probability tending to 0. For example, if you flip a fair coin an infinite number of times, the sample mean will converge to the actual mean, but you still have the possibility of a sequence of n heads that goes to zero as $n \to \infty$, so it'll "almost surely" converge.

Wikipedia: "In other words, the set of possible exceptions may be non-empty, but it has probability zero. The concept is precisely the same as the concept of "almost everywhere" in measure theory." (`https://en.wikipedia.org/wiki/Almost_surely`). If (X, Σ, μ) is a measure space, a property P is said to hold almost everywhere in X if there exists a set $N \in \Sigma$ with measure 0 ($\mu(N) = 0$), and all $x \in X \setminus N$ have the property P.

Example 9.2. Let $X_i \sim \text{Unif}[0, 1]$. Let $Y_n = \min\{X_1, \dots, X_n\}$. Then the Y_is are nonincreasing, and since it's always nonnegative we have a lower bound on $Y = \lim Y_i$.

$$\mathbb{P}(Y_i \geq \epsilon) = \mathbb{P}(X_1 \geq \epsilon, X_2 \geq \epsilon, \dots) = (1 - \epsilon)^n \tag{9.11}$$

$$\mathbb{P}(Y \geq \epsilon) = \lim_{n\to\infty} (1 - \epsilon)^n = 0 \tag{9.12}$$

Therefore $Y \to 0$ almost surely.

There's something to be said for convergence here: we can say the lim inf of Y_i is 0 because the sequence keeps decreasing and is bounded below by 0. I think you can formally prove this with subsequences. Here's a sketch: almost surely (yay I get why you need that now - there's some probability that the X_is are all bigger than the current min, but that probability goes to 0) you'll always have the minimum of the X_i decreasing so you can't pick an infinite subsequence with subsequential limit greater than that. So that's the lim inf and the lim sup because you also can't pick an infinite subsequence with subsequential limit less than that: by definition it's bounded below by 0. So lim sup and lim inf are equal, so the limit is equal and it's 0.

□

Theorem 9.4 (Central Limit Theorem). *Let $X_1, X_2, \dots$ be i.i.d. RVs with mean μ and variance σ^2. Let $S_n = \sum_{i=1}^n X_i$. Define*

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} \tag{9.13}$$

Then, in the limit $n \to \infty$, the cdf of Z_n converges (pointwise) to the cdf of the standard normal:

$$F_{Z_n}(z) \to \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-z^2/2} dz \tag{9.14}$$

A sequence of RVs X_i is said to converge to X in distribution (or weakly) if for every $a \in \mathbb{R}$ such that $F_X(a)$ is continuous,

$$\lim_{n\to\infty} F_{X_n}(a) = F_X(a) \tag{9.15}$$

or equivalently with epsilon-delta notation, $\forall \epsilon > 0, \exists N \text{ s.t. } n > N \implies \limsup_a |F_{X_n}(a) - F_X(a)| < \epsilon$.

We've used a number of different modes of convergence. Specifically, almost-sure convergence and rth moment convergence (i.e. that $\mathbb{E}[(X_n - X)^r] \to 0$) implies in-probability convergence, which in turn implies weak convergence.

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 10: Probability Convergence, Information Theory

Lecturer: Shyam Parekh *3 October* *Aditya Sengupta*

10.1 Convergence

Alternate proof of almost-sure convergence of the sequence of minima of $Unif([0,1])$ random variables to 0, with slightly more formalization.

We can demonstrate through an example that almost sure convergence implies in-probability convergence.

Example 10.1. Let

$$X = \begin{cases} 0 & \text{with probability} \frac{1}{2} \\ 1 & \text{with probability} \frac{1}{2} \end{cases} \tag{10.1}$$

and let $X_n = \left(1 + \frac{1}{n}\right) x$, i.e. $X_n - X = \frac{X}{n}$. Then

$$\mathbb{P}(\lim_{\text{n}\to\infty} |X_n - X| \geq \epsilon) = \mathbb{P}(\lim_{\text{n}\to\infty} \frac{X}{n} \geq \epsilon) = 0 \tag{10.2}$$

The dual of this gives us the result that we want,

$$\mathbb{P}(\lim_{\text{n}\to\infty} |X_n - X| < \epsilon) = 1 \implies X_n \to x\text{a.s.} \tag{10.3}$$

□

Something we've been using without proof for a while is this:

Lemma 10.1. *Almost sure convergence implies in-probability convergence.*

Proof. Let $X_n \to X$ almost surely. Then

$$\mathbb{P}\left(\lim_{\text{n}\to\infty} X_n \to X\right) = 1 \tag{10.4}$$

Let ω represent an arbitrary element of the sample space.

$$\mathbb{P}\left(\omega : |X_n(\omega) - X(\omega)| < \epsilon \ \forall n > some m(\omega)\right) = 1 \tag{10.5}$$

(say there exists an m such that $|a_n - a| < \epsilon$ for all $n > m$). Then define

$$A_m = \cup_{n \geq m}\{\omega \mid |X_n(\omega) - X(\omega)| \geq \epsilon\} := \cup_{n\geq m}\{|X_n - X| \geq \epsilon\} \tag{10.6}$$

We see that $A_1 \supseteq A_2 \supseteq A_3 \supseteq \dots$, so we claim that $A_\infty = \cap^{m\geq 1} A_m$ is the limiting set (I don't know what he means by this). Then by De Morgan's laws, we get

$$\begin{aligned} A_\infty^{\mathsf{c}} &= \cup_{m\geq 1} A_m^{\mathsf{c}} &(10.7)\\ &= \cup_{m\geq 1} \cap^{n\geq m} \{|x_n - x| < \epsilon\} &(10.8)\\ \therefore \mathbb{P}(A_\infty^{\mathsf{c}}) = 1 & &(10.9)\end{aligned}$$

We get that this has probability 1 from the statement that the set of ω such that the X_ns are ϵ-close to the limiting $X(\omega)$ for some n spans the sample space.

Therefore

$$\begin{aligned} \mathbb{P}(A_\infty) &= 0 &(10.10)\\ \mathbb{P}(|X_n - X| \geq \epsilon) \leq \lim_{\mathrm{n}\to\infty} \mathbb{P}(A_n) &= 0 &(10.11)\end{aligned}$$

Therefore $X_n \to X$ in probability. □

Note: just because $\mathbb{P}(A_\infty^{\mathsf{c}}) = 1$, we're not allowed to say A_∞^{c} spans the sample space; there may be (there are!) some points in the sample space with probability 0 that aren't in that set. They don't converge but they don't matter, which is why almost sure convergence makes sense.

We can show that in-probability convergence doesn't imply convergence of the expectation (rth moment convergence for $r = 1$.)

Example 10.2. Let X_n be 0 with probability $1 - \frac{1}{n}$ and otherwise n^2. Then

$$\lim_{\mathrm{n}\to\infty} \mathbb{P}(|X_n| \geq \epsilon) = \lim_{\mathrm{n}\to\infty} \frac{1}{n} = 0 \tag{10.12}$$

so $X_n \to 0$ in probability, but $\lim_{\mathrm{n}\to\infty} \mathbb{E}[X_n] = \lim_{\mathrm{n}\to\infty} n = \infty$.

□

The relationship is as follows: almost sure convergence (through the SLLN) and rth moment convergence both separately imply in-probability convergence, which in turn implies in-distribution convergence.

Example 10.3. Let X_i be Bernoulli with parameter $\frac{1}{i}$. Then $E[|X_n|^r] = \frac{1}{n} \to 0$ so rth moment convergence holds, so convergence in probability holds. However, almost sure convergence does not hold. For that, we would want to show that

$$\mathbb{P}(|X_n - 0| < \epsilon \ \forall n \geq m) \lim_{\text{n}\to\infty} \prod_{i=m}^{n} \left(1 - \frac{1}{i}\right) = \lim_{\text{n}\to\infty} \prod_{i=m}^{n} \frac{i-1}{i} = \lim_{\text{n}\to\infty} \frac{m-1}{n}$$

This is telescoping so it simplifies nicely. The limit goes to 0, which shows a.s. convergence doesn't hold. For a.s. convergence to hold, you need to show that some threshold m exists such that the probability of X_n being $\epsilon-$small is 1, but for any m this goes to 0.

□

10.2 Information Theory

Todo: TeX a block diagram. Information theory covers the flow of information from a source to a destination: source, to source encoder, to channel encoder, over a potentially noisy channel, through a channel decoder, then through a destination decoder and finally to the destination.

Theorem 10.2 (Source Coding Theorem)**.** *Consider N i.i.d. RVs each with entropy $H(X)$. This can be compressed into no more than $NH(x)$ bits with negligible risk of information loss as $N \to \infty$. Conversely, if they are compressed into fewer than $NH(x)$ bits, it is virtually certain that information will be lost.*

Theorem 10.3 (Channel Coding Theorem)**.** *Any rate below the channel capacity $\frac{\#\ \textit{message input bits}}{\#\ \textit{of bits transmitted}}$ is achievable. Conversely, any sequence of codes with $P_e(n) \to 0$ as $n \to \infty$ (where $P_e(n)$ is the max probability of error over possible input messages for the channel) has rate R less than the channel capacity.*

The two main channel modes we'll look at are binary erasure channels and binary symmetric channels. In a BEC, $c(b) = b$ with probability $1 - p$ (b is a bit, 0 or 1) and $c(b) = e$ with probability p (e is an error state). In a BSC, $c(b) = b$ with probability $1 - p$ and $c(b) = 1 - b$ with probability p.

Let's look at the achievable transmission rate for a BEC. We say R is an achievable transmission rate if there exist channel encoding and decoding functions (f_n, g_n) which encode messages of length $\lfloor nR \rfloor$ and decode the corresponding messages of length n; further, $P_e(n) \to 0$ as $n \to \infty$ is required. The capacity is $R = \sup\{r \mid r \text{ achievable}\}$.

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 11: Channel Coding Theorems

Lecturer: Shyam Parekh *8 October* *Aditya Sengupta*

11.1 Lyapunov's Inequality

Suppose $r > s \geq 1$. If $X_n \to X$ under the L_r norm, then $X_n \to X$ under the L_s norm. Lyapunov's inequality gives us a formal statement of this:

Theorem 11.1.

$$\left(\mathbb{E}\left[|X_n - X|^s\right]\right)^{1/s} \leq \left(\mathbb{E}\left[|X_n - X|^r\right]\right)^{1/r} \tag{11.1}$$

11.2 Binary Channels contd.

We previously defined the capacity of a channel. We claim that the BEC has capacity $1-p$, and the BSC has capacity $1 - H(p)$. First, we consider the BEC. This has an input set $\mathcal{X} = \{0, 1\}$ and an output set $\mathcal{Y} = \{0, 1, e\}$. Then, encoding is a function $f_n : \mathcal{X}^L \to \mathcal{X}^n$ and decoding is a function $\mathcal{Y}^n \to \mathcal{Y}^l$. The max probability of error is

$$p_e(n) = \max_{x \in \mathcal{X}^L} \mathbb{P}\left(g_n(\mathcal{Y}^{(n)}) \neq x \mid x^{(n)} = f_n(x)\right) \tag{11.2}$$

Let R be an achievable rate; then $L(n) = \lfloor nR \rfloor$ is the minimal number of bits you send ovwer the channel so that the message can be recovered. We will show that the capacity (the maximal achievable rate) cannot exceed $1-p$; suppose that we had feedback, i.e. the receiver could tell the sender when it received an erased bit. Then, to send 1 bit, an expected $\frac{1}{1-p}$ bits would have to be sent. Therefore $R = \frac{1}{1/(1-p)} = 1 - p$, i.e. even with the additional feedback information, we wouldn't be able to do any better than $1-p$.

Since we are claiming the capacity is $1-p$, we want to show that a rate $1 - p - \epsilon$ is achievable for any $\epsilon > 0$. We do this by writing a codebook. There are a total $2^{L(n)}$ possible messages, and we can associate a length n sequence of bits to be transmitted to each of these. However, we generate these sequences randomly, without requiring that they are unique. Since the probability of error is p, we can assume that $\lfloor n(1-p) \rfloor$ of the bits are transmitted; without loss of generality, suppose they're the first $\lfloor n(1-p) \rfloor$ bits. The decoding scheme looks for a match for these bits. If no unique match exists, it declares an error. Therefore, the probability of an error is

$$\mathbb{P}(\text{error}) = \mathbb{P}\left(\cup_{i=2}^{2^{L(n)}} (c_1 = c_i)\right) \leq \sum_{i=2}^{2^{L(n)}} 2^{\lfloor n(1-p) \rfloor} \tag{11.3}$$

$$\leq 2^{L(n)} 2^{-\lfloor n(1-p) \rfloor} \approx 2^{-n\epsilon} \tag{11.4}$$

Therefore the error goes to 0 as $n \to \infty$, so $R < 1 - p$ is achievable. Therefore the capacity of the BEC is $1-p$.

11.3 Huffman Coding

Huffman coding works on the principle that more likely symbols should be encoded in shorter bitstrings, and less likely symbols should be encoded in longer bitstrings. Suppose we want to encode sequences of the letters A, B, C, and D, and the frequencies associated to each are $55\%, 30\%, 10\%$, and 5%. We need to ensure that the codes are prefix-free, i.e. there is a unique way to decode a bitstring. Huffman coding is an optimal way to assign these, by building a binary tree using a priority queue. There's a lab on this.

11.4 Markov chains

A model has the "Markov property" if the effect of the past is summarized in the current state. For example, suppose you're betting on coin tosses. You have money given by F_n and you bet G_{n+1} on the next toss being heads. Then

$$F_{n+1} = F_n + G_{n+1} \left(1(x_{n+1} = H) - 1(x_{n+1} = T)\right) \tag{11.5}$$

and $F_{n+1} = F_n$ if $F_n = 0$.

As another example, suppose T_n is the time for the nth earthquake that is greater than 4 on the Richter scale. Then $T_{n+1} = T_n + Exp\left(\frac{1}{\lambda}\right)$; this is a Poisson process.

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 12: Discrete Time Markov Chains

Lecturer: Shyam Parekh *15 October* *Aditya Sengupta*

A discrete-time Markov chain describes a process whose state changes in discrete steps and cycles between a finite set of possible states. It also satisfies the Markov property, that the state only depends on the previous state:

$$\mathbb{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = x_{n-1}, \dots, X_1 = x_1) = \mathbb{P}(X_{n+1} = j \mid X_n = i) = p_{ij} \tag{12.1}$$

where $p_{ij} \geq 0$ and $\sum_j p_{ij} = 1$.

We also want the Markov chain to be time-independent.

We can make a probability transition matrix $P = \left[p_{ij}\right]$ to encode all of these probabilities. Let $\pi^{(n)} = \left[\mathbb{P}(X_n = 1) \dots \mathbb{P}(X_n = m)\right]$ represent the PMF of the state at time n. Then, the distribution at time $n+1$ can be found using the law of total probability to be

$$\pi^{(n+1)}(i) = \sum_j \pi^{(n)}(j) p_{ji} \tag{12.2}$$

and from this we get

$$\pi^{(n)} = \pi^{(0)} P^n \tag{12.3}$$

In general, if you want the probability of a particular sequence, you can multiply together a number of conditional probabilities, for example:

$$\mathbb{P}(X_0, X_1, X_2 \dots, X_n) = \mathbb{P}(X_0)\,\mathbb{P}(X_1 \mid X_0)\,\mathbb{P}(X_2 \mid X_1, X_0) \dots \mathbb{P}(X_n \mid X_{n-1}, X_{n-2}, \dots, X_1) \tag{12.4}$$

Then, the Markov property allows us to drop most of the conditioning:

$$\mathbb{P}(X_0, X_1, X_2 \dots, X_n) = \mathbb{P}(X_0)\,\mathbb{P}(X_1 \mid X_0)\,\mathbb{P}(X_2 \mid X_1) \dots \mathbb{P}(X_n \mid X_{n-1}) \tag{12.5}$$

12.1 n-step transitions

Let $r_{ij}(n)$ be the probability that $X_n = j$ given that $X_0 = i$. It is given by this recurrence relation:

$$r_{ij}(n) = \sum_{k \in S} r_{ik}(n-1) p_{kj} \tag{12.6}$$

Since $r_{ij}(1) = p_{ij}$, this just reduces to a matrix power.

We say state j is accessible from state i if $\exists n \in \mathbb{N}$ such that $r_{ij}(n) > 0$.

A state i is recurrent if for all j reachable from i, i is reachable from j, i.e. there is a nonzero probability of looping between i and j infinitely. If $A(i)$ is the set of reachable states from i, then i is recurrent iff $\forall j \in A(i), i \in A(j)$.

A state i is transient if it is not recurrent.

For any recurrent state i, all states $A(i)$ form a recurrent class.

Lemma 12.1. *Any Markov chain can be decomposed into one or more recurrent classes.*

Lemma 12.2. *A state in a recurrent class is not reachable fromo states in any other recurrent class (if it were, they would form one larger recurrent class.)*

Lemma 12.3. *Transient states are not reachable from a recurrent state. Moreover, at least one recurrent state is reachable from every transient state.*

A Markov chain is called irreducible if it only has one recurrent class.

For a non-irreducible Markov chain, we can identify the recurrent classes by an algorithm that I'll get off the slides later.

12.2 Periodicity

The periodicity of a state i in a Markov chain is

$$d(i) = \gcd\{n \geq | \, r_{ii}(n) > 0\} \tag{12.7}$$

Intuitively, this says that "all paths back to i take a multiple of $d(i)$ steps".

Theorem 12.4. *If i and j are in the same recurrent class, then $d(i) = d(j)$.*

We define a Markov chain as aperiodic if $d(i) = 1$ for all i; otherwise we say the chain is periodic with period d.

We say that a state $\pi_0(j)$ is stationary if choosing the initial state according to $\mathbb{P}(X_0 = j) = \pi_0(j)$ implies $\mathbb{P}(X_n = j) = \pi_0(j)$ for all n.

The balance equations give us stationarity:

$$\pi_0(j) = \sum_{k=1}^{m} \pi_0(k) p_{kj} \tag{12.8}$$

This can be written as $\pi_0 = \pi_0 P$; that is, π_0 is a left eigenvector of P with eigenvalue 1.

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 13: Discrete-Time Markov Chains II

Lecturer: Shyam Parekh *17 October* *Aditya Sengupta*

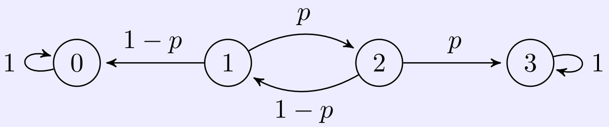
13.1 Hitting Times / First Passage Times

Let $\{X_n, n \geq 0\}$ be a DTMC over a finite set S. Consider a set $A \subset S$; let $T_A = \min\{n \geq 0 \mid x_n \in A\}$. We're interested in the expected value of the minimum number of steps required to reach the set A given that we start at some state i, i.e. $\mathbb{E}[T_A \mid X_0 = i]$. We write this more compactly as $\mathbb{E}_i[T_A]$. Let $\beta(i) = \mathbb{E}_i[T_A]$ for all $i \in S$. Then, in general, we can say that

$$\beta(i) = \begin{cases} 1 + \sum_j p_{ij}\beta_j & i \notin A \\ 0 & i \in A \end{cases} \tag{13.1}$$

These two cases are referred to as the first set equations (FSEs). The nontrivial case comes out of summing up $\sum_j p_{ij}(1 + \beta(j))$; for each state j that could be reached from i in one step, the expected time is 1 to get to j plus the expected time to reach the set A from the state j, which is $\beta(j)$. Further, we pull out the 1 from recognizing that the p_{ij}s sum to 1.

Example 13.1. This is called the Gambler's Ruin: consider a Markov chain in which with probability p, the state i moves to $i + 1$ unless you're at the end in either direction (at 0 or at some maximum, say 3), in which case you stay there with probability 1 (you don't have any money left, or you've got enough and don't want to risk more).



Consider the expected time to reach a success-or-failure state, i.e. $A = \{0, 3\}$.

$$\beta(0) = 0, \beta(3) = 0\beta(1) = 1 + p\beta(2), \beta(2) = 1 + (1 - p)\beta(1) \tag{13.2}$$

Solving the system of linear equations, we get

$$\beta(1) = \frac{1 + p}{1 - p + p^2}, \beta(2) = \frac{2 - p}{1 - p + p^2} \tag{13.3}$$

□

Given $A, B \subset S$ disjoint, we are interested in $P_i(T_A < T_B) = \mathbb{P}(T_A < T_B \mid X_0 = i)$. Let this be represented by α_i. Then

$$\alpha(i) = \begin{cases} p_{ij}\alpha(j) & i \notin A \cup B \\ 1 & i \in A \\ 0 & i \in B \end{cases} \tag{13.4}$$

Example 13.2. With the same setup as above, we see that $\alpha(0) = 0, \alpha(3) = 1, \alpha(1) = p\alpha(2), \alpha(2) = p + (1-p)\alpha(1)$. This gives us

$$\alpha(1) = \frac{p^2}{1-p+p^2}, \alpha(2) = \frac{p}{1-p+p^2} \tag{13.5}$$

□

13.2 Countably Infinite DTMCs

We say $\{X_n, n \geq 0\}$ is a countably infinite DTMC if its set of possible states $\mathcal{X}$ is countably infinite. Let T_x and T_x^+ be defined by

$$T_x = \min\{n \geq 0 \mid X_n = x\} \tag{13.6}$$

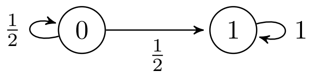
$$T_x^+ = \min\{n \geq 1 \mid X_n = x\} \tag{13.7}$$

For $x, y \in \mathcal{X}$, we defin $\rho_{x,y} = \mathbb{P}_x(T_y^+ < \infty)$ and $\rho_x = \rho_{x,x}$. We say x is recurrent if $\rho_x = 1$, and it is transient if $\rho_x < 1$.

Proposition 13.1. *Let $N_x = \sum_{n \geq 0} \mathbb{1}\{X_n = x\}$, i.e. the number of visits to x. If x is recurrent, then $N_x \to \infty$ almost surely and $\mathbb{E}_x[N_x] = \infty$; if x is transient, then $N_x < \infty$ almost surely and $\mathbb{E}_x[N_x] = \frac{\rho_x}{1-\rho_x} < \infty$.*

We get this result from the Markov chain being memoryless and following a geometric distribution; every time x is reached, there is a probability ρ_x that x will be reached again in finite time, so the expected number of visits for a transient state is given by the expectation of a Geometric(ρ_x) variable.

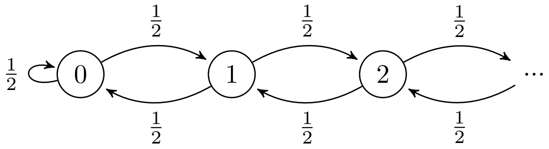
To illustrate this, consider the following finite DTMC:



Here, 0 is transient and 1 is recurrent. This corresponds to $\rho_0 = \frac{1}{2}$, and $\rho_1 = 1$.

13.3 Random Walks

Consider the following random walk:



This is a "random walk reflected at zero". To find the ρ coefficients, we use its recursive structure and the Markov property; for example,

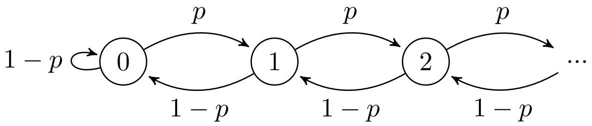
$$\rho_{2,0} = \rho_{2,1} \cdot \rho_{1,0} = \rho_{1,0}^2 \tag{13.8}$$

Also, we get the following linear relationships by examining the first few states:

$$\rho_0 = \frac{1}{2} + \frac{1}{2}\rho_{1,0} \tag{13.9}$$
$$\rho_{1,0} = \frac{1}{2} + \frac{1}{2}\rho_{2,0} \tag{13.10}$$

Solving all of these together, we get $\rho_{1,0} = \frac{1}{2} + \frac{1}{2}\rho_{1,0}^2 \implies \rho_{1,0} = 1$. Therefore $\rho_0 = 1$, so the zero state is recurrent and $N_0 \to \infty$ almost surely.

Next, consider a general random walk reflected at 0, which is almost the same but with probability of a forward transition p:



We claim that $p > \frac{1}{2}$ implies that all states are transient, and $p \leq \frac{1}{2}$ implies that all states are recurrent. Note that this is not a property of a finite Markov chain: if the chain terminated, all the states would be recurrent because they all satisfy $j \in A(i) \implies i \in A(j)$. This shows that there is something nontrivial about the countably-finite extension.

Proposition 13.2. *A finite DTMC must have at least one recurrent state.*

Proof. (mine) Suppose this is not the case; then every state in the set on which the DTMC is defined is transient, so each one is visited finitely many times. Index this by $\mathcal{X} = \{X_1, \ldots, X_n\}$ with a finite number of visits each $\{N_1, \ldots, N_n\}$. At time $1 + \sum_{i=1}^n N_i$, we must be at one of the states, but we've already exhausted all the visits, so we cannot be at any state. Therefore this is a contradiction and at least one state must be recurrent. □

State x communicates with state y if $\rho_{x,y} > 0$ and $\rho_{y,x} > 0$.

This is equivalent to the definition in which $y \in A(x), x \in A(y)$, but moe general, because it applies to the countably-infinite case as well.

A communicating class is a maximal set of states which communicate with each other.

Lemma 13.3. *States communicating is an equivalence relation, and so the communicating classes on a Markov chain are the equivalence classes under this relation.*

Proof.

- Symmetry: if x communicates with y, then $\rho_{x,y} > 0$ and $\rho_{y,x} = 0$ and so y communicates with x.
- Reflexivity: $\rho_x > 0$ by definition unless it transitions to a different state immediately with probability 1 and there is no path back to it. (The reference I looked up for this says reflexivity is "trivial", because of course it does.)
- Transitivity: if x communicates with y and y communicates with z, then the probability of moving $x \to y$ in finite time is nonzero, and the probability of moving $y \to z$ in finite time is nonzero, so the probability of moving $x \to z$ in finite time is at least their product and is therefore nonzero. Therefore $\rho_{xz} > 0$, so x communicates with z.

□

A DTMC is irreducible if it consists of only a single communicating class.

Theorem 13.4. *Recurrence and transience are class properties: if one element in a communicating class is recurrent or transient, all of the others are.*

State x is positive-recurrent if it is recurrent and $\mathbb{E}_x[T_x^+] < \infty$. State x is null-recurrent if it is recurrent and $\mathbb{E}_x[T_x^+] = \infty$.

Remark 13.5. *Positive- and null-recurrence are also class properties.*

Theorem 13.6 (Big Theorem)**.** *If a Markov chain is irreducible and positive recurrent, then the steady-state probability of each state gives the unique stationary distribution; further, it implies that*

$$\lim_{x \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{1}\{X_i = x\} = \pi(x) \tag{13.11}$$

This condition is called ergodicity. Finally, an irreducible positive-recurrent aperiodic Markov chain has $\pi_n(x) \to \pi(x)$: the distribution after n time trainsitions tends to the unique stationary distribution.

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 14: Discrete-Time Markov Chains III

Lecturer: Shyam Parekh *22 October* *Aditya Sengupta*

14.1 Ergodic Theorem

We'll start by restating the Ergodic Theorem.

Theorem 14.1 (Ergodic Theorem)**.** *Consider an irreducible DTMC.*

1. *If it is positive-recurrent, there exists a unique invariant distribution. This is an if-and-only-if: if there exists an invariant distribution, the DTMC must be positive-recurrent.*
2. *Ergodicity: if the DTMC is positive-recurrent, then the invariant distribution is almost surely just the fraction of time spent in each state as time goes to infinity:*

$$\lim_{\mathrm{n}\to\infty} \frac{1}{n}\sum_{i=0}^{n} \mathbb{1}\{X_i = x\} = \pi(x) \tag{14.1}$$

In words, the *__time average__* *tends to the* *__ensemble average__*.

3. *If the DTMC is positive-recurrent and aperiodic, then the distribution of states π_n tends to π as $n \to \infty$.*

Theorem 14.2. *Consider an irreducible positive-recurrent DTMC with invariant distribution π. Then*

$$\pi(x) = \frac{1}{E_x[T_x^+]} \tag{14.2}$$

Proof. Let $\tau_1, \tau_2, \ldots$ be the inter-visit intervals for x, i.e. the time intervals between consecutive visits to x (τ_1 is the time between the initial state and first visit, τ_2 is the time between the first visit and the second visit, etc.) The Markov property means the τ_is are i.i.d. By the strong law of large numbers, we can say that

$$\frac{1}{n}\sum_{i=1}^{n} \tau_i \to \mathbb{E}_x[T_x^+] \text{ almost surely} \tag{14.3}$$

The sample mean of the τ_is is given by the ratio of some large time t to the number of visits to x in that time, so we can rewrite the above convergence statement as

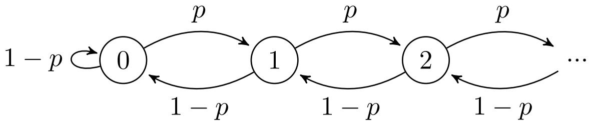
$$\frac{t}{\sum_{i=0}^{t-1} \mathbb{1}\{X_i = x\}} \to \mathbb{E}_x[T_x^+] \tag{14.4}$$

Rearranging, we get

$$\frac{1}{t}\sum_{i=0}^{t-1} \mathbb{1}\{X_i = x\} \to \frac{1}{\mathbb{E}_x[T_x^+]} \tag{14.5}$$

Therefore, by the ergodic theorem, we get $\pi(x) \to \frac{1}{\mathbb{E}_x[T_x^+]}$. □

14.2 General Random Walk with Reflection



We previously stated that $p < \frac{1}{2}$ implied this Markov chain was positive recurrent. By the ergodic theorem, we can say that means there exists a unique invariant distribution. We'll show on the homework that this is

$$\pi(k) = (1-\rho)\rho^k, k \geq 0 \tag{14.6}$$

where $\rho = \frac{p}{1-p}$.

Now, consider the case $p = \frac{1}{2}$, which we previously claimed is null-recurrent. We showed last time that this DTMC is recurrent, and we can solve the first step equations for recurrence:

$$\begin{aligned}\mathbb{E}_1[T_0^+] &= 1 + \frac{1}{2}\mathbb{E}_2[T_0^+] \\ &= 1 + \mathbb{E}_1[T_0^+] + (\text{something positive})\end{aligned}$$

The only way this can be satisfied is if $\mathbb{E}_1[T_0^+] = \infty$. We can extend this argument and say that each state x has $\mathbb{E}_x[T_x^+] = \infty$, so the Markov chain is null-recurrent. We can intuitively make sense of this as saying that since there are infinite recurrent states, the expected time we'll spend at any particular state goes to 0.

Finally, consider $p > \frac{1}{2}$. This is a transient DTMC. Let Y_ns be i.i.d with PMF

$$Y_n = \begin{cases} 1 & \text{with probability } p \\ -1 & \text{with probability } 1-p \end{cases} \tag{14.7}$$

Let $X_n = \max X_{n-1} + Y_n, 0$ for $n \geq 1$. Based on this, we can say $X_n \geq X_0 + \sum_{i=1}^n Y_i$, and so

$$\frac{1}{n}X_n \geq \frac{X_0}{n} + \frac{1}{n}\sum_{i=1}^n Y_i \tag{14.8}$$

The sample mean of the Y_is goes almost surely to the true mean by the strong law of large numbers, and because $p > \frac{1}{2}$ this is some positive number. Therefore, the probability that $X_n = 0$ goes to 0 almost surely. This implies 0 is transient, which in turn implies the DTMC is transient (just do this recursively, I think.)

14.3 Balance Equations

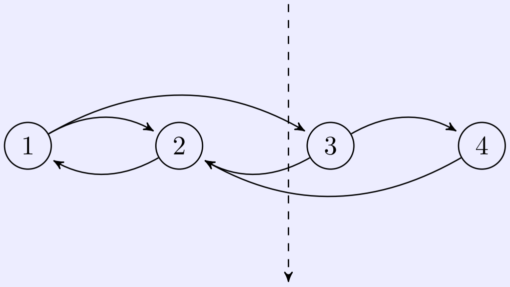
The global balance equations state that a distribution must be invariant under one time transition: $\pi = \pi P$. In particular, that means the probability of each state must be invariant under the time transition:

$$\pi(k) = \sum_i \pi(i) p_{ik} = \sum_{i \neq k} \pi(i) p_{ik} + \pi(k) p_{kk} \tag{14.9}$$

$$\pi(k)(1 - p_{kk}) = \pi(k) \sum_{i \neq k} p_{ki} = \sum_{i \neq k} \pi(i) p_{ik} \tag{14.10}$$

We don't have to consider the entirety of the DTMC at once, because global balance is quite a strong condition and we can stil comment on balance even if global balance is not satisfied. This leads to the idea of local balance: if you create a cut of a DTMC, then the probability mass flow in both directions must be equal.

Example 14.1.



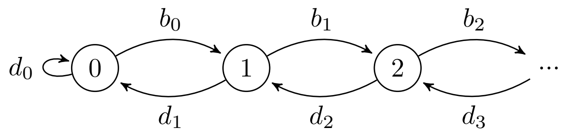
(sorry about the arrow, this is some jank LaTeXwork - basically it thinks this dotted line is being drawn between two hidden Markov nodes and so it's a directed edge.)

Place a cut between 2 and 3. Then local balance implies

$$\pi(1) p_{13} = \pi(3) p_{32} + \pi(4) p_{42} \tag{14.11}$$

□

The global balance equations imply all the local balance equations and vice versa. This gives us a tool for dealing with systems like the random walk reflected at zero. Suppose the forward transition probabilities i to $i+1$ are b_i and the backward ones i to $i+1$ are d_{i+1}.



We call this a birth-death DTMC.

Place a cut between 1 and 2; then $\pi(1)b_1 = \pi(2)d_2$, and in general

$$\pi(k)b_k = \pi(k+1)d_{k+1}, k \geq 0 \tag{14.12}$$

which gives us the recurrence relation

$$\pi(k) = \pi(0)\frac{b_0 \dots b_{k-1}}{d_1 \dots d_k} \tag{14.13}$$

Suppose $b_i = p$ and $d_i = 1 - p$ for all $i \geq 0$ and suppose $p < \frac{1}{2}$. Then this gives us

$$\pi(k) = \pi(0)\rho^k, k \geq 1 \tag{14.14}$$

where $\rho = \frac{p}{1-p}$. Therefore, $\pi(k) = (1-\rho)\rho^k$ (normalize to find $\pi(0)$.)

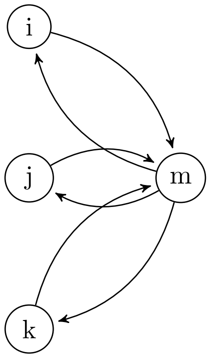
14.3.1 Detailed Balance Equations

The detailed balance equations state that

$$\sum_{x \neq m} \pi(x)p_{xm} = \pi(m)\sum_{x \neq m} p_{mx}. \tag{14.15}$$

I'm not sure I got all of this right?

This is useful in a case where we have one highly connected node:



The detailed balance equations imply the global balance equations, and for birth-death DTMCs, the balance equations imply the detailed balance equations. The DBEs are useful for when we look at reversibility.

14.4 Reversibility

Suppose $X_n, n \geq 0$ is an irreducible and positive-recurrent DTMC. Suppose the invariant distribution is π and $\pi_0 = \pi$.

If for all $n \geq 0$, the joint distribution of $X_0, \dots, X_n$ is the same as the joint distribution of $X_n, \dots, X_0$, then the DTMC is reversible.

Theorem 14.3. *The reverse of a DTMC has the Markov property.*

Proof.

$$P(X_k = i \mid X_{k+1} = j, X_{k+2} = i_{k+2}, \dots, X_n = i_n) = \frac{\mathbb{P}(X_k = i, X_{k+1} = j, X_{k+2} = i_{k+2} \dots, X_n = i_n)}{\mathbb{P}(X_{k+1} = j, X_{k+2} = i_{k+2}, \dots, X_n = i_n)} \tag{14.16}$$

$$= \frac{\pi(i) p_{ij} p_{j,i_{k+2}} \dots p_{i_{n-1},i_n}}{\pi(j) p_{j,i_{k+2}} \dots p_{i_{n-1},i_n}} \tag{14.17}$$

$$= \frac{\pi(i) p_{ij} \cancel{p_{j,i_{k+2}} \dots p_{i_{n-1},i_n}}}{\pi(j) \cancel{p_{j,i_{k+2}} \dots p_{i_{n-1},i_n}}} \tag{14.18}$$

$$= \frac{\pi(i) p_{ij}}{\pi(j)} \tag{14.19}$$

We see that the Markov property is satisfied because the dependence in everything from i_{k+2} onwards drops out. □

We can further recognize our result as $\mathbb{P}(X_k = j \mid X_{k-1} = i) = p_{ij}$. Therefore we get that for a reversible Markov chain, $\pi(i)p_{ij} = \pi(j)p_{ji}$, which is just the DBE on two states.

Theorem 14.4. *A DTMC is reversible if and only if the DBEs hold.*

Lemma 14.5. *If π satisfies the DBEs, then π is an invariant distribution.*

14.5 Poisson Processes

An arrival process is called a Poisson process with rate $\lambda > 0$ if it satisfies

1. time-homogeneity: $p(k, \tau)$ is the same for any τ.
2. independence: the number of arrivals in a given interval is independent of what happens outside that interval
3. $p(0, \tau) = 1 - \lambda\tau + O(\tau)$; $p(1, \tau) = \lambda\tau + O_1(\tau)$ and in general $p(k, \tau) = O_k(\tau)$.

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 15: Poisson Processes

Lecturer: Shyam Parekh *24 October* *Aditya Sengupta*

15.1 Deriving the distribution

A Poisson random variable can be approximated to a binomial one. Recall that a Poisson random variable has PMF $P_Z(k) = \frac{e^{-\lambda}\lambda^k}{k!}$, and a binomial random variable has PMF $P_s(k) = \binom{n}{k}p^k(1-p)^{n-k}$. If we let $np \to \lambda$ while also letting $n \to \infty$, we can show that $P_s(k) \to P_z(k)$.

A Poisson process, given by the distribution $p(k, \tau)$ (the probability of k arrivals in a time interval τ) has the following properties:

1. Time homogeneity: $p(k, \tau)$ is a pure function of τ.
2. Independence: the number of arrivals in any interval is independent of what happens outside the interval.
3. When $\tau \to 0$, $p(0, \tau) \approx 1 - \lambda\tau$, $p(1, \tau) \approx \lambda\tau$ and $p(k, \tau) \approx 0$ for $k > 1$.

Consider a time interval $[0, \tau]$, and divide this up into arbitrarily small windows of time δ, i.e. $\frac{\tau}{\delta} = n$. Then, the number of arrivals in $[0, \delta]$ (and by independence, any $\delta-$width) is given by the third property to be either 0 or 1 with probabilities $1 - \lambda\delta$ or $\lambda\delta$ respectively. This is the definition of a binomial process, as we are essentially drawing n Bernoulli($\lambda\delta$) random variables and summing them.

This tells us that the average number of arrivals in an interval τ is $np = n\delta\lambda$. Since $\delta = \frac{\tau}{n}$, this is $\lambda\tau$. Identifying this as a Poisson parameter, we can say that

$$p(k, \tau) = \frac{e^{-\lambda\tau}(\lambda\tau)^k}{k!} \; k = 0, 1, 2, \ldots \tag{15.1}$$

Let N_τ be the number of arrivals within time τ. Then $\mathbb{E}[N_\tau] = \lambda\tau$ and $\text{var}(N_\tau) = \lambda\tau$, because we know a Poisson distribution must satisfy these.

15.2 Time for first arrival

Let the time for the first arrival be T. If we start at time $t = 0$, then

$$\mathbb{P}(T \leq t) = 1 - \mathbb{P}(T > t) = 1 - p(0, t) = 1 - e^{-\lambda t} \tag{15.2}$$

Therefore $T \sim Exp(\lambda)$. This tells us that $\mathbb{E}[T] = \frac{1}{\lambda}$ and $\text{var}(\lambda) = \frac{1}{\lambda^2}$.

15.3 Arrivals in disjoint intervals

Suppose we have disjoint intervals with widths τ_1, τ_2 with N_1 and N_2 arrivals respectively. We know that Poisson random variables sum in distribution, so $N_1 + N_2 s \sim Pois(\lambda(\tau_1 + \tau_2))$.

15.4 Properties of Poisson Processes

1. For any $t > 0$, a PP after t is independent of the PP up until t.
2. For any $t > 0$, let T be the time of the first arrival after t. Then $P(T - t > s) = p(0, (t + s) - t) = p(0, s) = e^{-\lambda s}$.

This lets us say that the inter-arrival times T_k are i.i.d $Exp(\lambda)$.

15.5 kth Arrival Time

Let Y_k be the kth arrival time. $Y_k = \sum_{i=1}^{k} T_i$, and the T_is are i.i.d. so $\mathbb{E}[Y_k] = \frac{k}{\lambda}$ and $\text{var}(Y_k) = \frac{k}{\lambda^2}$. The pdf is

$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}, y \geq 0 \tag{15.3}$$

Proof. The cdf is

$$F_{Y_k}(y) = \mathbb{P}(Y_k \leq y) = \sum_{n=k}^{\infty} p(n, y) = 1 - \sum_{n=0}^{k-1} p(n, y) \tag{15.4}$$

Then, we take a derivative and it works. □

We call this the Erlang distribution of order k.

15.6 Splitting and Merging of Poisson Processes

Suppose for every arrival in a Poisson process, we categorize it in one of two "buckets": the left with probability p and the right with probability $1 - p$. Then the arrivals in the left bucket is a PP with parameter λp and the right bucket is a PP with parameter $\lambda(1 - p)$.

Merging works similarly: if we have two Poisson processes with parameters λ_1, λ_2, then their combination is Poisson with parameter $\lambda_1 + \lambda_2$.

If we consider a time interval δ, then to first order this can be verified:

$$p_{\lambda_1+\lambda_2}(0,\delta) = (1-\lambda_1\delta)(1-\lambda_2\delta) \approx 1-(\lambda_1+\lambda_2)\delta \tag{15.5}$$

$$p_{\lambda_1+\lambda_2}(1,\delta) = \lambda_1\delta(1-\lambda_2\delta) + \lambda_2\delta(1-\lambda_1\delta) \approx (\lambda_1+\lambda_2)\delta \tag{15.6}$$

If $T_a \sim Exp(\lambda_a)$ and $T_b \sim Exp(\lambda_b)$, then $Z = \min\{T_a, T_b\} \sim Exp(\lambda_a + \lambda_b)$. Therefore the inter-arrival time of the merged Poisson process with parameters λ_a, λ_b is $Exp(\lambda_a+\lambda_b)$. Additionally, we have $\mathbb{P}(\min\{T_a, T_b\} = T_a) = \frac{\lambda_a}{\lambda_a+\lambda_b}$.

15.7 Residual Life Paradox

Start a PP(λ) from $t = -\infty$. Let t^* be the current time, let u be the time of the last arrival, and let v be the time of the next arrival. We want to find the width of the interval, $L = (v - t^*) + (t^* - u)$.

Due to independence, $(t^* - u)$ is independent of $(v - t^*)$. By memorylessness, $v - t^*$ is $Exp(\lambda)$. Therefore

$$\mathbb{P}((t^* - u) > x) = \mathbb{P}(\text{no arrival in } [t^* - x, t^*]) = p(0, x) = e^{-\lambda x} \tag{15.7}$$

Therefore $t^* - u \sim Exp(\lambda)$, so L is an Erlang distribution of order 2. But we know this should be exponential.

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 16: Continuous-Time Markov Chains

Lecturer: Shyam Parekh *29 October* *Aditya Sengupta*

So far, we've discussed Markov chains with discrete jumps, i.e. chains that could be characterized by the transitions from $X_0 \to X_1 \to X_2$. However, some problems can only be modelled in continuous time. To account for this, we introduce the machinery of CTMCs.

Let X be a countable state space, and let π be a distribution on X. Then, we define the analogy to the state transition matrix P:

$$Q := \{Q(i,j) \mid i,j \in X\} \,\text{s. t.}\, Q(i,j) \geq 0 \forall i \neq j \tag{16.1}$$

We refer to this as the rate matrix or infinitesimal generator. Further, we can require of the rate matrix that $\sum_j Q(i,j) = 0$ for all i: the rates at which we go to every state is 0. This will make intuitive sense when we look at Q as the derivative of some analogous discrete-time matrix P. (I think.)

As a consequence of this, we get that the rate of a self-transition is given by

$$Q(i,i) = -\sum_{j \neq i} Q(i,j) \tag{16.2}$$

Further, we define $q(i) = -Q(i,i)$. Now, we can define a CTMC.

A continuous-time Markov chain over a countable state space X with initial distribution π and rate matrix Q is the process $\{X_t, t \geq 0\}$ such that

1. $\mathbb{P}(X_0 = i) = \pi(i)$: the initial distribution works the way it does for DTMCs.
2. $\mathbb{P}(X_{t+\epsilon} = j \mid X_t = i) = (1 - q(i)\epsilon) \cdot \mathbb{1}\{j = i\} + Q(i,j)\epsilon\mathbb{1}\{j \neq i\} + O(\epsilon^2)$: the infinitesimal-time transition probability can be linearly approximated using the transition rates. Further, the Markov property holds.

Note that $\sum_j \mathbb{P}(X_{t+\epsilon} = j \mid X_t = i) \approx 1$, so this is consistent with $Q(i,j)$ being the probability of jumping $i \to j$ in unit time.

16.1 Construction

Suppose at time t we're in state i. Let $\tau \sim Exp(q(i))$. At time $t+\tau$, the chain jumps to $j \neq i$ with probability

$$\Gamma(i,j) = \frac{Q(i,j)}{q(i)} \;\forall j \neq i \tag{16.3}$$

For $j \neq i$, note that

$$\mathbb{P}(X_{t+\epsilon} = j \mid X_t = i) = q(i)\epsilon\frac{Q(i,j)}{q(i)} + O(\epsilon^2) = \epsilon Q(i,j) + O(\epsilon^2) \tag{16.4}$$

If we sum over all of these but retain the higher-order terms, we get the probability of a transition in time ϵ is $q(i)e^{-q(i)\epsilon} \approx q(i)\epsilon$. I think this is true, but I can't really tell. This tells us that transition times are exponentially distributed, which is a characteristic of a Poisson process.

From this we see that a transition is the merging of Poisson processes with rates $\frac{Q(i,j)}{q(i)}$, determining whether i will transition to each possible j.

16.2 Embedded Markov Chain

The DTMC $\{X_n, n \geq 0\}$ with transition matrix Γ is called the embedded Markov chain of a CTMC.

Example 16.1. Consider a $PP(\lambda)$ that we model in a CTMC where the state is the number of arrivals in the last interval of time $\leq t$. This is modelled in the rate matrix

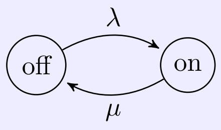
$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \dots \\ & -\lambda & \lambda & 0 & \dots \\ & & -\lambda & \lambda & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ & & \dots & -\lambda & \lambda \end{bmatrix}$$

and the associated discrete-time transition matrix,

$$\Gamma = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix}$$

□

Example 16.2. Consider the on-off source in the following CTMC:

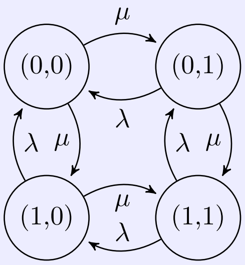


□

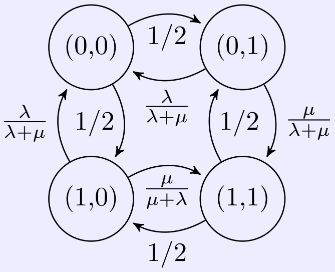
Example 16.3. Suppose we have two machines that both fail in time $Exp(\lambda)$ independently. If they fail, they're repaired in time $Exp(\mu)$ independently.

Let (i, j) be the state of this system, where a variable is 0 if its corresponding machine isn't working, and 1 if it is.

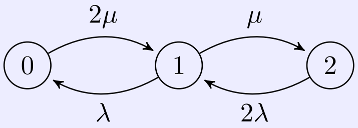
This is embedded into the following CTMC:



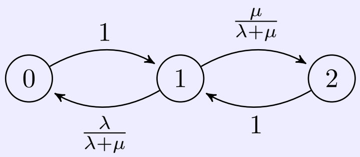
To compute the associated rate matrix, we note that if both machines are broken or are working, the odds are equally good that either one will break or will be fixed first, so the transitions going out from $(0, 0)$ and $(1, 1)$ all have probability $\frac{1}{2}$. The others have probabilities that we can get from comparing exponential variables as usual.


(0,0)
(0,1)
(1,0)
(1,1)
1/2
$\frac{\lambda}{\lambda+\mu}$
$\frac{\lambda}{\lambda+\mu}$
1/2
1/2
$\frac{\mu}{\lambda+\mu}$
$\frac{\mu}{\mu+\lambda}$
1/2

Another possible setup that is more easily extended is to set the number of machines working at any given time as the state.



and the embedded DTMC is


1
$\frac{\mu}{\lambda+\mu}$
0
1
2
$\frac{\lambda}{\lambda+\mu}$
1

□

EECS 126: Probability and Random Processes **Fall 2019**

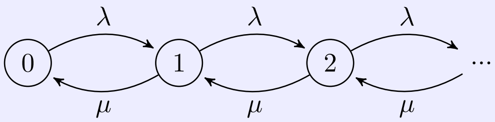
Lecture 17: Spooky Continuous-Time Markov Chains

Lecturer: Shyam Parekh *31 October* *Aditya Sengupta*

17.1 The M/M/1 Queue

Example 17.1. Consider a system with a single server for customers arriving according to a Poisson process with parameter λ, and in which the service time is $\tau_i \sim Exp(\mu)$.

The embedded DTMC is the birth-death Markov chain with $b_i = \lambda$ and $d_i = \mu$, where the state is the number of customers in the queue.



The associated rate matrix is

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

□

17.2 CTMC Properties

1. A CTMC is irreducible if and only if its embedded DTMC is irreducible.
2. A state is recurrent or transient in a CTMC if and only if it's recurrent or transient in the embedded DTMC.
3. Positive and null recurrence are not equivalent in the CTMC and the embedded DTMC. For example, Markov chains that are positive-recurrent in the continuous case over an infinite state space (i.e. have infinite expected transitions in finite time) are called explosive Markov chains, even though the embedded MC may be null-recurrent. `http://www.columbia.edu/~ks20/stochastic-I/stochastic-I-CTMC.pdf`
4. There is no notion of periodicity in a CTMC.

17.3 Throwing a Bunch of Theorems At Us

(his words) Let $P_t = \big[\mathbb{P}_t(i,j)\big] = \big[\mathbb{P}(X_t = j \mid X_0 = i)\big]$. This satisfies the property $P_{t+s} = P_t P_s = P_s P_t$. If we take a derivative in time, we should get the rate matrix multiplied by this transition-probability matrix.

Theorem 17.1.

$$\frac{d}{dt} P_t = Q P_t \tag{17.1}$$

This is the Kolmogorov backward equation (KBE).

Theorem 17.2.

$$\frac{d}{dt} P_t = P_t Q \tag{17.2}$$

This is the Kolmogorov forward equation (KFE).

I don't know how both of these are true in general at the same time, because that suggests any matrix commutes with its exponential which I'm pretty sure is not the case.

Here's a good derivation: `https://mast.queensu.ca/~stat455/lecturenotes/set5.pdf`.

So P_t is both a left and right eigenvector of the derivative operator, meaning it's an exponential with eigenvalue (rate) Q. This is the only way I can intuit going from the KBE and KFE to the next theorem, because he didn't do it.

Theorem 17.3.

$$P_t = e^{tQ} = \sum_{k=0}^{\infty} \frac{t^k Q^k}{k!} \tag{17.3}$$

This tells us that the distribution of a CTMC after t transitions is given by $\pi_t = \pi_0 e^{Qt}$.

17.4 Balance Equations

Theorem 17.4. *$\pi Q = 0$ if and only if $\pi P_t = \pi$; any π satisfying this is called the stationary or invariant distribution.*

Suppose $\pi Q = 0$. Then $\sum_i \pi(i) Q(i,j) = 0 \ \forall j$.

In general, the rate into state j should be the same as the rate out:

$$\sum_{i \neq j} \pi(i) Q(i,j) = \pi_j \sum_{k \neq j} Q(j,k) \tag{17.4}$$

Theorem 17.5 (Big Theorem)**.** *the name is still ridiculous.*

Consider an irreducible CTMC.

1. *If and only if it is positive-recurrent, there exists a unique invariant distribution.*
2. *If it is positive-recurrent, then* $\lim_{t\to\infty} \frac{1}{t}\int_0^t \mathbb{1}\{X_u = i\}du = \pi(i)$ *almost surely, i.e. it is ergodic.*
3. *If it is positive-recurrent, then* $P_t(i,j) \to \pi(j)$ *as* $t \to \infty$.
4. *If it is not positive-recurrent, then there exists no invariant distribution, and the fraction of time spent in any given state goes to 0:*

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t \mathbb{1}\{X_u = i\}du = 0 \ \forall i \tag{17.5}$$

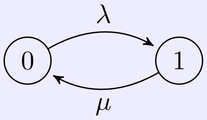
We can relate the invariant distributions of a CTMC and its embedded MC.

Theorem 17.6. *Consider a CTMC* $\{X_t, t \geq 0\}$ *and the embedded MC* $\{X_n, n \geq 0\}$. *Assume that both are irreducible positive-recurrent. Let* π, α *be the corresponding invariant distributions. Then*

$$\pi(i) = \frac{\alpha(i)/q(i)}{\sum_k \alpha(k)/q(k)} \tag{17.6}$$

where $q(i) = -Q(i,i)$.

Example 17.2. Consider the on-off source CTMC:



Solving the first-step equations (I think?), we get that $\pi(0) = \frac{\mu}{\lambda+\mu}$ and $\pi(1) = \frac{\lambda}{\lambda+\mu}$.
□

Example 17.3. For another example, we can return to the M/M/1 queue, with all of its birth coefficients equal to λ and all of its death coefficients equal to μ. If $\lambda > \mu$, the entire chain is a transient class; if $\lambda = \mu$, we get null recurence, and if $\lambda < \mu$ we get positive recurrence.

In the case $\lambda < \mu$, we get that there must exist an invariant distribution, which is given by $\pi(i) = (1-\rho)\rho^i$ where $\rho = \frac{\lambda}{\mu}$.
□

The rate matrix and the embedded DTMC matrix are related by $R = I + \frac{1}{q}Q$.

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 18: Simulated DTMCs, Erdős-Rényi Random Graphs

Lecturer: Shyam Parekh *7 November* *Aditya Sengupta*

18.1 Simulated DTMCs

18.1.1 CTMC Review

Recall that if π is an invariant for a CTMC and α is an invariant for its embedded DTMC, then

$$\pi(i) = \frac{\alpha(i)/q(i)}{\sum_k \alpha(k)/q(k)} \tag{18.1}$$

or conversely,

$$\alpha(i) = \frac{\pi(i)q(i)}{\sum_k \pi(k)q(k)} \tag{18.2}$$

As stated, the ergodic theorem held for countably infinite state spaces $\mathcal{X}$; a finite irreducible CTMC is always positive-recurrent. Further, if the chain is not irreducible, it can be decomposed into recurrent and transient classes. In this case, the invariant distribution is some linear combination of the invariant distributions of each recurrent class.

Recall that the balance equations for CTMCs said that we could find the invariant distribution by solving $\pi Q = 0$, which we get by equating flow-in and flow-out. Further, the detailed balance equations tell us that $\pi(i)Q(i,j) = \pi(j)Q(j,i)$.

18.1.2 Uniformization

We can convert a CTMC with rate matrix Q into its embedded DTMC by taking $q = \sup q(i)$ and $R = I + \frac{1}{q}Q$. The invariant distribution holds in both: $\pi R = \pi \iff \pi Q = 0$.

For example, for the on-off source, we have $Q = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}$ and so $R = \begin{bmatrix} 0 & 1 \\ \frac{\mu}{\lambda} & 1 - \frac{\mu}{\lambda} \end{bmatrix}$ and we can verify that these have the same invariant distribution $\pi = \frac{1}{\lambda+\mu}\begin{bmatrix} \lambda & \mu \end{bmatrix}$.

18.1.3 The PASTA property

Recall that the stationary distribution for the M/M/1 queue Markov chain was $\pi(i) = (1-\rho)\rho^i$ for $i \geq 0$ and $\rho = \frac{\lambda}{\mu}$. We can use this as an illustration of the PASTA property: Poisson Arrivals See Time Averages. By this property, we know that the probability that an arriving customer finds i customers in the queue is $(1-\rho)\rho^i$. Let D be the total delay in the system (waiting and service). Then, we can find the MGF of D:

$$M_D(s) = \mathbb{E}[e^{sD}] = \sum_{i=0}^{\infty} \rho^i (1-\rho) \left(\frac{\mu}{\mu - s}\right)^{i+1} = \frac{\mu - \lambda}{(\mu - \lambda) - s} \tag{18.3}$$

Therefore, the total delay is distributed by $Exp(\mu - \lambda)$.

18.1.4 Hitting times/probabilities in a CTMC

Let $A \subset \mathcal{X}$ and let $T_A = \inf\{t \mid X_t \in A\}$. We want to find $\mathbb{E}_i[T_A]$. Let $\beta(i) = \mathbb{E}_i[T_A]$.

$$\beta(i) = \begin{cases} \frac{1}{q(i)} + \sum_{j \neq i} \frac{Q(i,j)}{q(i)} \beta(j) & i \notin A \\ 0 & i \in A \end{cases} \tag{18.4}$$

and if we take $A, B \subset \mathcal{X}$ with $A \cap B = \varnothing$, we can find $\mathbb{P}_i(T_A < T_B)$:

$$\alpha(i) = \begin{cases} \sum_{j \neq i} \frac{Q(i,j)}{q(i)} \alpha(j) & i \notin A \cup B \\ 1 & i \in A \\ 0 & i \in B \end{cases} \tag{18.5}$$

18.2 Erdős-Rényi Random Graphs

18.2.1 Properties of E-R Random Graphs

An E-R random graph $G(n,p)$ is a random undirected graph with n vertices, where each of the $\binom{n}{2}$ edges exists with probability p.

We can think of an E-R random graph as a distribution over the set of all graphs. There are $2^{\binom{n}{2}}$ possible graphs, which we can enumerate. Let G_0 be a particular graph with m edges. Then

$$\mathbb{P}(G = G_0) = p^m (1-p)^{\binom{n}{2} - m} \tag{18.6}$$

The expected number of edges can be found using linearity of expectation: each edge exists with probability p, and there are $\binom{n}{2}$ of them, so the expected number of edges is $\binom{n}{2}p$.

The degree of an arbitrary vertex is given by

$$\mathbb{P}(D = d) = \binom{n-1}{d} p^d (1-p)^{n-1-d} \tag{18.7}$$

and $\mathbb{E}[D] = (n-1)p$.

When $p(n) = \frac{\lambda}{n}, \lambda > 0$, then

$$\mathbb{E}[D] = \frac{n-1}{n}\lambda \xrightarrow{n\to\infty} \lambda \tag{18.8}$$

The distribution of D turns out to be $Poisson(\lambda)$.

The probability that a given vertex is isolated is $(1-p)^{n-1}$.

Theorem 18.1 (Erdős-Rényi Theorem)**.** *Let $p(n) = \lambda\frac{\ln(n)}{n}$.*

- *If $\lambda < 1$, then $\mathbb{P}\{G(n,p)$ is connected$\} \xrightarrow{n\to\infty} 0$.*
- *If $\lambda > 1$, then $\mathbb{P}\{G(n,p)$ is connected$\} \xrightarrow{n\to\infty} 1$.*

Proof. For the case $\lambda < 1$, let X_n be the number of isolated nodes. We want to show that $\mathbb{P}(X_n > 0) \to 1$.

$$\begin{aligned} \mathbb{P}(X_n = 0) &= \mathbb{P}(|X_n - \mathbb{E}[X_n]| = E[X_n]) && (18.9)\\ &\le \mathbb{P}(|X_n - \mathbb{E}[X_n]| \ge \mathbb{E}[X_n]) && (18.10)\\ &\le \frac{\mathrm{var}(X_n)}{\mathbb{E}[X_n]^2} && (18.11) \end{aligned}$$

We can also look at X_n as the sum of indicator variables i, indicating whether i is isolated. Then by the variance of identically distributed RVs, we get

$$\mathrm{var}(X_n) = n\,\mathrm{var}(I_1) + n(n-1)\,\mathrm{cov}(I_1, I_2) \tag{18.12}$$

$I_1 \sim Bernoulli((1-p(n))^{n-1}) = Bernoulli(q(n))$, so $\mathrm{var}(I_1) = q(n)(1-q(n))$. □

18.2.2 Percolation Theory

Suppose we have a particle in a 2D grid starting on the top and trying to percolate to the bottom. It can either transition or not transition to an adjacent point with probability p. If $p < \frac{1}{2}$, there is no path to get to a particular point with probability 1, and if $p > \frac{1}{2}$ there is a path with probability 1.

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 19: MAP and MLE Estimates

Lecturer: Shyam Parekh *12 November* *Aditya Sengupta*

19.1 Applying Bayes' Rule

Bayes' rule lets us incorporate a prior probability into our analysis of events: we can condition the probability of something happening on some expectation of how likely we think it is. This is useful when we are trying to determine hidden parameters and only have some observed data that tells us indirectly about these parameters. The distribution that incorporates the prior as well as observed data is referred to as the posterior distribution. Suppose we have events C_i with prior probabilities p_i, and probabilities of the observed event/data S conditioned on C_i given by q_i. Then, we update $\mathbb{P}(C_i)$ based on seeing S as follows:

$$\begin{aligned}\mathbb{P}(C_i \mid S) = \frac{\mathbb{P}(C_i \cap S)}{\mathbb{P}(S)} &= \frac{\mathbb{P}(C_i)\,\mathbb{P}(S \mid C_i)}{\mathbb{P}(S)} &(19.1)\\ &= \frac{\mathbb{P}(C_i)\,\mathbb{P}(S \mid C_i)}{\sum_j \mathbb{P}(S \mid C_j)\,\mathbb{P}(C_j)} &(19.2)\\ &= \frac{p_i q_i}{\sum_j p_j q_j} &(19.3)\end{aligned}$$

19.2 MAP and MLE

19.2.1 Max A Posteriori Estimation

The MAP estimate of the correct C_i or a parameter associated with the C_is just has us maximize the numerator in the above application of Bayes' rule:

$$\arg\max_i \mathbb{P}(C_i \mid S) = \arg\max_i p_i q_i \tag{19.4}$$

This gives us the most likely cause for S given some priors.

19.2.2 Maximum Likelihood Estimation

MLE gives us the most likely causation:

$$\arg\max_i \mathbb{P}(C_i \mid S) = \arg\max_i q_i \tag{19.5}$$

Note that if the prior is uniform, i.e. $p_i = \frac{1}{n}$ for all i, then MAP and MLE are the same.

Example 19.1. Suppose we have two coins with probabilities of getting heads p_1 and p_2. We choose one of these coins, toss it n times and get k heads. The probability that we picked the first coin is

$$\mathbb{P}(\text{coin } 1 \mid k \text{ heads}) \propto \mathbb{P}(k \text{ heads} \mid \text{coin } 1)\,\mathbb{P}(\text{coin1}) \tag{19.6}$$

and the same for coin 2. Therefore, the updated $\mathbb{P}(\text{coin } 1) \propto p_1^k(1-p_1)^{n-k}$ and $\mathbb{P}(\text{coin } 2) \propto p_2^k(1-p_2)^{n-k}$ up to the same normalizing factor.

Here, MLE and MAP are the same because the prior probability of picking either coin is the same.

□

Let X and Y be discrete random variables.

$$MAP[X \mid Y = y] = \arg\max_x \mathbb{P}[X = x \mid Y = y] \tag{19.7}$$

$$MLE[X \mid Y = y] = \arg\max_x \mathbb{P}[Y = y \mid X = x] \tag{19.8}$$

Example 19.2. Romeo and Juliet are supposed to meet, but Juliet is late by some random time $X \sim Unif[0, \theta]$. Suppose Romeo thinks she'll be late by some time $Unif[0, 1]$, i.e. our prior is that $\theta = 1$. Then, the posterior distribution can be computed as follows. First, we write out the prior:

$$f_\theta(\theta) = \begin{cases} 1 & 0 \le \theta \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{19.9}$$

and the probability of the observation:

$$f_{X|\theta}(x|\theta) = \begin{cases} \frac{1}{\theta} & 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases} \tag{19.10}$$

Then, the posterior distribution is

$$
\begin{aligned}
f_{\theta|X}(\theta|x) &= \frac{f_\theta(\theta) f_{X|\theta}(x|\theta)}{\int f_\theta \theta' f_{X|\theta}(x|\theta') d\theta'} && (19.11)\\
&= \frac{1/\theta}{\int_x^1 \frac{1}{\theta'} d\theta'}, 0 \leq x \leq \theta \leq 1 && (19.12)\\
&= \begin{cases} \frac{1}{\theta |\ln x|} & 0 \leq x \leq \theta \leq 1 \\ 0 & \text{otherwise} \end{cases} && (19.13)
\end{aligned}
$$

□

The MAP posterior PDF is "optimistic" in x, as it decreases in θ faster than in x. Note that here the prior is uniform, so MAP and MLE are the same.

19.3 Binary Symmetric Channel Inference

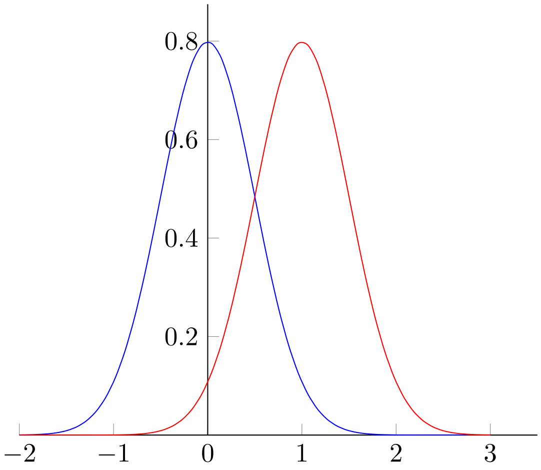
Suppose in a BSC with probability p of switching, we get an output 1. We want to estimate the input. The input $x = 1$ with probability α and $x = 0$ with probability $1 - \alpha$. Then

$$
\mathbb{P}(X = 0 \mid Y = 1) = \frac{\mathbb{P}(X = 0, Y = 1)}{\mathbb{P}(Y = 1)} = \frac{\mathbb{P}(X = 0, Y = 1)}{\mathbb{P}(Y = 1 \mid X = 0)\,\mathbb{P}(X = 0) + \mathbb{P}(Y = 1 \mid X = 1)\,\mathbb{P}(X = 1)} \tag{19.14}
$$

Substituting in, we get that $\mathbb{P}(X = 0 \mid Y = 1) \propto p(1 - \alpha)$ and $\mathbb{P}(X = 1 \mid Y = 1) \propto (1 - p)\alpha$.

19.4 Gaussian Channel Inference

Suppose we send $X \in \{0, 1\}$ and receive $Y = X + Z$ where $Z \sim N(0, \sigma^2)$ independent of X. With a prior of $X \sim Bernoulli(\alpha)$, we can infer from Y what X is. Let $f_i(y) = f_{Y|X}(y \mid x = 0) \sim N(i, \sigma^2)$ for $i = 0, 1$.



Then the posterior is

$$f_{X|Y}(x = 0 \mid y) = \frac{(1-\alpha)f_0(y)}{(1-\alpha)f_0(y) + \alpha f_1(y)} \tag{19.15}$$

$$f_{X|Y}(x = 1 \mid y) = \frac{\alpha f_1(y)}{(1-\alpha)f_0(y) + \alpha f_1(y)} \tag{19.16}$$

The MAP is 0 if $(1-\alpha)f_0(y) > \alpha f_1(y)$, and 1 otherwise. This is the same as an indicator function of $\mathbb{1}\{y \geq \frac{1}{2} + \sigma^2 \log\left(\frac{1-\alpha}{\alpha}\right)\}$.

The MLE is 0 if $f_0(y) > f_1(y)$ and 1 otherwise.

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 20: Hypothesis Testing, Randomization

Lecturer: Shyam Parekh *14 November* *Aditya Sengupta*

20.1 Binary Hypothesis Testing

Suppose $X \in \{0, 1\}$, and we observe some output Y. We want to determine $\hat{X}$ such that we maximize the probability of correct detection (PCD, or true positive) $= \mathbb{P}(\hat{X} = 1 \mid X = 1)$ subject to the probability of a false alarm (PFA, or false positive) $= \mathbb{P}(\hat{X} = 1 \mid X = 0)$.

If the solution is $PCD = R(\beta)$, then $R(\beta)$ is called the Receiver Operating Characteristic (ROC).

$$R(\beta) = \max PCD \mid PFA \leq \beta \tag{20.1}$$

Theorem 20.1 (Neyman-Pearson Theorem)**.** *The $\hat{X}$ that maximizes the PCD such that $PFA \leq \beta$ is given by*

$$\hat{X} = \begin{cases} 1 & L(Y) > \lambda \\ 1 w.p. \gamma & L(Y) = \lambda \\ 0 & L(Y) < \lambda \end{cases} \tag{20.2}$$

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

$$L(y) = \frac{f_{Y|X}(y|1)}{f_{Y|X}(y|0)} \tag{20.3}$$

If $L(Y)$ is large, then for an observed Y, $X = 1$ is more likely. As λ reduces, we choose $\hat{X} = 1$ more frequently. This leads to the PCD and PFA both increasing. We further constrain this by choosing λ such that $PFA = \beta$.

Example 20.1. Consideer the Gaussian channel where $Y = X + Z$ and $Z \sim \mathbb{N}(0, \sigma^2)$. The MLE estimate is

$$MLE[X \mid Y = y] = \arg\max_x f_{Y|X}(y|x) = \mathbb{1}\{y \geq 0.5\} \tag{20.4}$$

and the MAP estimate is

$$MAP[X \mid Y = y] = \arg max_x f_{X|Y}(x|y) = \mathbb{1}\{y \geq 0.5 + \sigma^2 \log\left(\frac{p_0}{p_1}\right)\} \tag{20.5}$$

For hypothesis testing, we choose the bound β on $\mathbb{P}(\hat{X} = 1 \mid X = 0)$. By the Neyman-Pearson theorem,

$$L(y) = \exp\left(\frac{2y-1}{2\sigma^2}\right) \tag{20.6}$$

Note that for any X, $\mathbb{P}[L(y) = \lambda] = 0$; $L(y)$ is strictly increasing in y. By the NP theorem, $\hat{X} = \mathbb{1}\{y \geq y_0\}$.

Suppose $\mathbb{P}(Y \geq y_0 \mid X = 0) = \beta$. This implies that $\mathbb{P}(\mathcal{N}(0,1) \geq \frac{y_0}{\sigma}) = \beta$; we can look this up by standard confidence intervals. Similarly,$PCD = \mathbb{P}[\hat{X} = 1 \mid X = 1] = \mathbb{P}[\mathbb{N}(1,\sigma^2) \geq y_0] = \mathbb{P}[\mathbb{N}(0, 1 \geq y(\beta) - \sigma^2]$. (I can't tell if that's a σ or σ^2.)

□

Example 20.2. Mean of Exponential Random Variables.

Two machines produce light bulbs. The lightbulbs have lifespans $\sim Exp(\lambda_0)$ or $Exp(\lambda_1)$ respectively. Suppose $\lambda_0 < \lambda_1$, i.e. machine 1 is defective. Suppose we observe lifespans $Y = (Y_1, \ldots, Y_n)$: which machine might have produced these?

$$\begin{aligned} L(y) &= \frac{\prod_{i=1}^n \lambda_1 \exp(-\lambda_1 y_1)}{\prod_{i=1}^n \lambda_0 \exp(-\lambda_0 y_0)} \quad (20.7) \\ &= \left(\frac{\lambda_1}{\lambda_0}\right)^n \exp\left(-(\lambda_1 - \lambda_0)\sum_{i=1}^n y_i\right) \quad (20.8) \end{aligned}$$

Since $\lambda_1 > \lambda_0$, $L(y)$ is decreasing in $\sum_{i=1}^n y_i$. Also, $\mathbb{P}(L(y) = \lambda) = 0$. Therefore

$$\hat{X} = \mathbb{1}\{\sum_{i=1}^n y_i \leq a\} \tag{20.9}$$

We choose a such that

$$\mathbb{P}(\sum_{i=1}^n Y_i \leq a \mid X = 0) = \beta \tag{20.10}$$

Applying the Central Limit Theorem, we get that this is equivalent to

$$\mathbb{P}(\mathbb{N}(0,1) \leq \lambda_0 \frac{a - n/\lambda_0}{\sqrt{n}}) = \beta \tag{20.11}$$

If we take $\beta = 0.05$, we get that $a = (n + 1.65\sqrt{n})\lambda_0^{-1}$. (I think he got the number wrong and it should be 1.96.)

□

20.2 Randomization

Let $X \in \{0, 1\}$ and $Y \in \{A, B, C\}$. Suppose the probabilities that $Y = A, B, C$ conditioned on $X = 0$ are 0.2, 0.5, 0.3 and conditioned on $X = 1$ are 0.2, 0.2, 0.6. This gives us

$$L(Y) = \begin{cases} 0.4 & Y = B \\ 1 & Y = A \\ 2 & Y = C \end{cases} \tag{20.12}$$

Suppose $\lambda = 2.1$; then the PCD is 0 because $L(Y) < \lambda$ everywhere and the PFA is also 0. If we set $\lambda = 2$, the PCD becomes 0.6γ and the PFA is 0.3γ. If $\gamma = 1$ then the PCD is $0.6 + 0.2\gamma$ and the PFA is $0.3 + 0.2\gamma$. I don't know how he got those numbers but it seems alright.

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 21: Hypothesis testing, linear least squares estimators

Lecturer: Shyam Parekh *19 November* *Aditya Sengupta*

21.1 Neyman-Pearson Hypothesis Testing is Optimal

Proof. (of Neyman-Pearson optimality) Consider another decision rule $\tilde{X}$ such that $\mathbb{P}(\tilde{X} = 1 \mid X = 0) \leq \beta$; we wannt to show that $\mathbb{P}(\tilde{X} = 1 \mid X = 1) \leq \mathbb{P}(\hat{X} = 1 \mid X = 1)$. Observe that

$$(\hat{X} - \tilde{X})(L(Y) - \lambda) \geq 0 \tag{21.1}$$

which we can verify by taking all three cases, $L(Y) > \lambda$, $L(Y) = \lambda$, and $L(Y) < \lambda$. Taking expectation on both sides, we get

$$\mathbb{E}[\hat{X}L(Y) \mid X = 0] - \mathbb{E}[\tilde{X}L(Y) \mid X = 0] \geq \lambda \left(\mathbb{E}[\hat{X} \mid X = 0] - \mathbb{E}[\tilde{X} \mid X = 0]\right) \tag{21.2}$$

Since $\hat{X}$ and $\tilde{X}$ can only take on the values 0 or 1, their expectation is equivalent to the probability that they are 1. Further, by the setup, we know that $\mathbb{P}(\hat{X} = 1 \mid X = 0) = \beta \geq \mathbb{P}(\tilde{X} = 1 \mid X = 0)$. Therefore the right hand side is some positive number.

This implies that

$$\mathbb{E}[\hat{X}L(Y) \mid X = 0] \geq \mathbb{E}[\tilde{X}L(Y) \mid Y = 0] \tag{21.3}$$

Further, we can go through algebra to show that $\mathbb{E}[g(Y)L(Y) \mid X = 0] = \mathbb{E}[g(Y) \mid X = 1]$. Therefore

$$\mathbb{E}[\hat{X}L(Y) \mid X = 0] = \mathbb{E}[\hat{X} \mid X = 1] = \mathbb{P}(\hat{X} = 1 \mid X = 1) \geq \mathbb{E}[\tilde{X} \mid X = 1] = \mathbb{P}(\tilde{X} = 1 \mid X = 1) \tag{21.4}$$

□

21.2 Estimation

21.2.1 Linear Algebra Setup

We want to develop geometric intuition for estimating a random variable. Consider the following space of random variables:

$$\mathcal{H} = \{X \mid X \in \mathbb{R}, \mathbb{E}[X^2] < \infty\} \tag{21.5}$$

Definitions of vector spaces, bases, spanning, linear independence, subspaces, inner product spaces.

An inner product space that is complete is called a Hilbert space. A space is complete if any Cauchy sequence in the space converges in the space, but we won't deal with those details.

Define $\langle X, Y \rangle = \mathbb{E}[XY]$. We can show this makes $\mathcal{H}$ into an inner product space. The only nontrivial piece of algebra required for this is the Cauchy-Schwartz inequality:

$$|\,\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]} \tag{21.6}$$

Note that over a finite sample space, $\mathbb{E}[XY] = \sum_{\omega \in \Omega} X(\omega)Y(\omega)P(\omega)$. Since Ω is finite, $\mathcal{H}$ is finite-dimensional, and has a basis $\{\mathbb{1}_\omega\}_{\omega \in \Omega}$.

21.3 Projections

We want to find $L[Y|X]$, the "best" linear (in the form $a + bX$) estimate of Y given X. Given $Y \in V$ and a subspace $U \subseteq V$, we want to find the closest $X \in U$ to Y. This is the orthogonal projection of Y on U.

The orthogonal projetion on a subspace U is $P : V \to U, y \to P_y = \arg\min_{x \in U} \|y - x\|$.

$P_y \in U$ and $y - P_y \in U^\perp$.

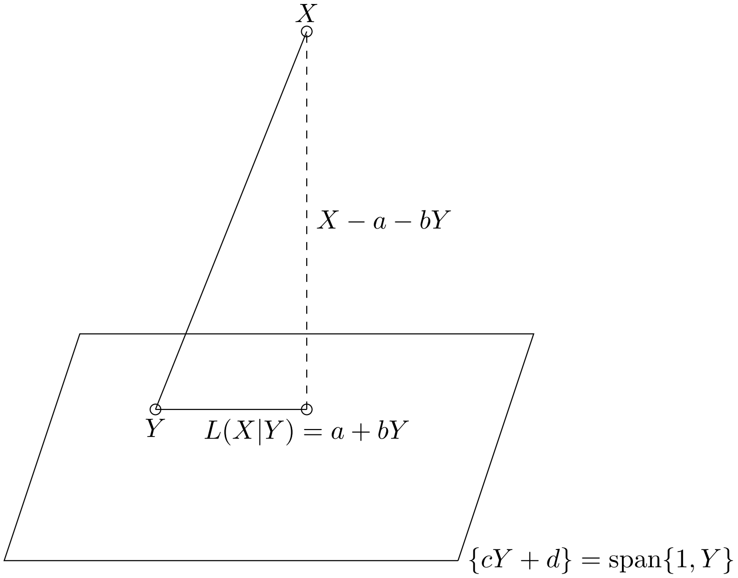
If U is finite-dimensional with an orthonormal basis $\{v_i\}_{i=1}^n$, then $P_y = \sum_{i=1}^n \langle y, v_i \rangle v_i$. Further, we can use the Gram-Schmidt process to construct an orthonormal basis out of any basis.

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 22: Computing LLSE, Minimum Mean-Squared Estimation

Lecturer: Shyam Parekh *21 November* *Aditya Sengupta*

Recall that the LLSE problem is to find $L(X|Y)$ given $X, Y \in \mathcal{H}$. This is an orthogonal projection:



Our task is to minimize $\mathbb{E}[(X - a - bY)^2]$ over all $a, b \in \mathbb{R}$.

The first method is algebraic, which is to take partial derivatives with respect to a and b of the expected value of the error squared, and set them both to zero. The second method is geometric: the error vector, denoted by a dotted line in the above diagram, is orthogonal to the space $\{cY + d\}$, meaning it is orthogonal to each basis element. That is, $\mathbb{E}[X - a - bY] = 0$ and $\mathbb{E}[(X - a - bY)Y] = 0$.

To illustrate these projections, imagine a unit cube, as represented in Figure 22.1.

Suppose we wanted to represent the vector $(1, 1, 1)$ in the $x - y$ plane basis. It is the sum of the projection onto this basis $(1, 1, 0)$ and an error term $(0, 0, 1)$. Similarly, we could represent it in the $x-$axis basis as $(1, 0, 0) + (0, 1, 1)$, or the $y-$axis basis as $(0, 1, 0) + (1, 0, 1)$. In each of these cases, we see that the error term is orthogonal to the pojection, like we would expect.

To bring this reasoning to random variabes, we make $\{1, Y\}$ into an orthonormal basis by applying Gram-Schmidt. We get $\{1, (Y - E[Y])/\sqrt{\text{var}\, Y}\}$. This gives us a closed form for the linear least-squares estimator:

Theorem 22.1. *Let $X, Y \in \mathcal{H}$ and let Y not be a constant. Then*

$$L[X|Y] = \mathbb{E}[X] + \frac{\text{cov}(X, Y)}{\text{var}(Y)}(Y - \mathbb{E}[Y]) \tag{22.1}$$

Further, the squared error of this estimator is

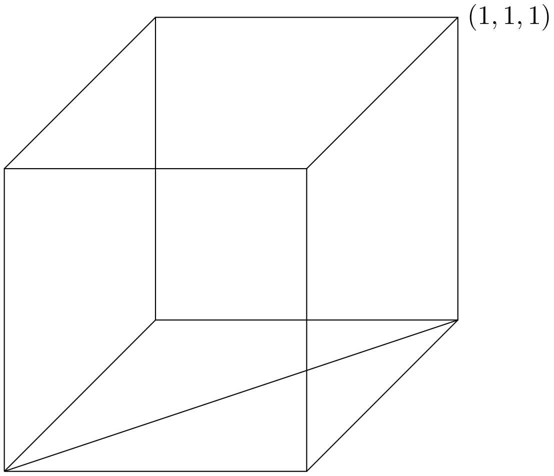


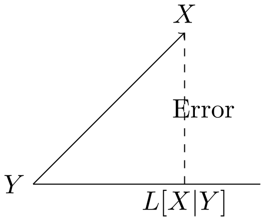
Figure 22.1: A unit cube, to represent projections

$$\mathbb{E}[(X - L[X|Y])^2] = \text{var}(X) - \frac{\text{cov}(X,Y)^2}{\text{var}(Y)} \tag{22.2}$$

Proof. The linear estimator formula comes out of taking the projection $\sum_{i=1}^{n} \langle x, u_i \rangle u_i$ to the orthonormal basis we found above. To compute the error term, we can reason geometrically. For simplicity, let X and Y be zero-mean (take $X \to X - \mathbb{E}[X]$ and $Y \to Y - \mathbb{E}[Y]$). Then, the linear estimator is

$$L[X|Y] = \frac{\text{cov}(X,Y)}{\text{var}(Y)} Y \tag{22.3}$$

Then, we draw the projection and note we can use Pythagoras' theorem:



The error term is therefore given by

$$\|\text{Error}\|^2 = \|X\|^2 - \|L(X|Y)\|^2 = \text{var}(X) - \frac{\text{cov}(X,Y)^2}{\text{var}(Y)} \tag{22.4}$$

□

Example 22.1. Let $Y = \alpha X + Z$, where X, Z are both zero-mean and independent. The true relation is given by $Y = \alpha X$ and there is an additional noise term that we denote Z. The optimal linear esatimator is

$$L[X|Y] = \frac{\operatorname{cov}(X,Y)}{\operatorname{var}(Y)} Y \tag{22.5}$$

Further, based on the form for Y, we know that $\operatorname{cov}(X,Y) = \alpha\,\mathbb{E}[X^2]$ and $\operatorname{var}(Y) = \alpha^2\,\mathbb{E}[X^2] + \operatorname{var}(Z)$. Therefore

$$L[X|Y] = \frac{\alpha\,\mathbb{E}[X^2]}{\alpha^2\,\mathbb{E}[X^2] + \mathbb{E}[Z^2]} = \frac{\alpha^{-1}}{1 + SNR^{-1}} \tag{22.6}$$

where SNR, the signal-to-noise ratio, is the ratio of the power (squared-magnitude, e.g. imagine X and Y are voltages, then their squares are proportional to power) of the signal to the noise. As the SNR goes to 0, the noise drowns out the signal and we get $L[X|Y] \to 0$; as the SNR goes to ∞, the signal dominates and we get the correct estimate of $L[X|Y] \to \frac{1}{\alpha} Y$.

□

Example 22.2. Let $X = \alpha Y + \beta Y^2$ where $Y \sim Unif[0,1]$. Then

$$L[X|Y] = \mathbb{E}[X] + \frac{\operatorname{cov}(X,Y)}{\operatorname{var}(Y)}(Y - \mathbb{E}[Y]) \tag{22.7}$$

The expected value of X is

$$\mathbb{E}[X] = \alpha\,\mathbb{E}[Y] + \beta\,\mathbb{E}[Y^2] = \frac{\alpha}{2} + \frac{\beta}{3} \tag{22.8}$$

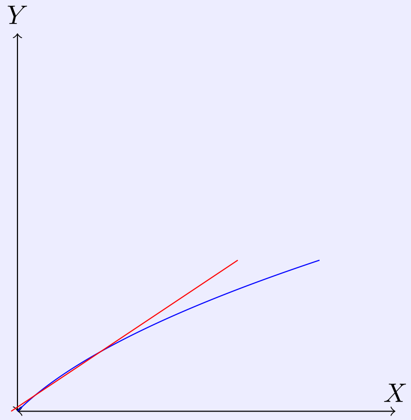
The covariance of X and Y is

$$\operatorname{cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}[\alpha Y^2 + \beta Y^3] - \left(\frac{\alpha}{2} + \frac{\beta}{3}\right)\frac{1}{2} = \frac{\alpha + \beta}{12} \tag{22.9}$$

and the variance of Y is $\frac{1}{12}$. Therefore, we get

$$L[X|Y] = -\frac{\beta}{6} + (\alpha + \beta)Y \tag{22.10}$$

This can be visualized as follows:



We see that locally, the linear estimator (red) approximates the true values (blue) but this linear approximation does not extend very well.

□

22.1 Linear Regression

Suppose we observe i.i.d $\{(X_i, Y_i)\}$. We want to find $g(Y) = a + bY$ such that the error $\mathbb{E}[|X - a - bY|^2]$ is minimized. We can do this by taking the partial derivatives of the total error of the observed points. For each k, we want

$$Err_k = \sum_{j=1}^{k} |X_j - a - bY_j|^2 \tag{22.11}$$

$$\frac{\partial Err_k}{\partial a} = 0, \frac{\partial Err_k}{\partial b} = 0 \tag{22.12}$$

After much algebra, we can show that

$$a + bY = \mathbb{E}_k[X] + \frac{\mathrm{cov}_k(X, Y)}{\mathrm{var}_k(Y)}(Y - \mathbb{E}_k[Y]) \tag{22.13}$$

where each subscript k indicates that it pertains to the value after k observations:

$$\mathbb{E}_k[X] = \frac{1}{k}\sum_{j=1}^{k} X_j \tag{22.14}$$

$$\mathbb{E}_k[Y] = \frac{1}{k}\sum_{j=1}^{k} Y_j \tag{22.15}$$

$$\text{cov}_k(X,Y) = \frac{1}{k}\sum_{j=1}^{k} X_j Y_j - \mathbb{E}_k[X]\,\mathbb{E}_k[Y] \tag{22.16}$$

$$\text{var}_k(Y) = \frac{1}{k}\sum_{j=1}^{k} Y_j^2 - (\mathbb{E}_k[Y])^2 \tag{22.17}$$

Essentially, this gives us a sample variance and covariance on the true values, that are given in general by:

$$M_n = \frac{1}{n}\sum_{i=1}^{n} Z_i \tag{22.18}$$

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - M - n)^2 \tag{22.19}$$

$$S_{u,v} = \frac{1}{n}\sum_{i=1}^{n}(U_i - M_n^u)(V_i - M_n^v) \tag{22.20}$$

These will converge to the true values with enough samples, in accordance with the strong law of large numbers.

Theorem 22.2. *Linear regression converges to the linear least-squares estimate of a random variable.*

22.2 Minimum mean-squared estimate (MMSE)

Suppose we know the joint distribution of X and Y. We want to find $g(Y)$ such that $\mathbb{E}[(X - g(Y))^2]$ is minimized.

Theorem 22.3. *The MMSE of X given Y is given by*

$$g(Y) = \mathbb{E}[X|Y] \tag{22.21}$$

Recall that $\mathbb{E}[X|Y]$ is a random variable that is a function of Y.

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 23: Minimum Mean-Squared Estimation, Gram-Schmidt Process

Lecturer: Shyam Parekh *26 November* *Aditya Sengupta*

23.1 The MMSE theorem

Recall that the MMSE problem is to find $g(Y)$ such that $\mathbb{E}[(X - g(Y))^2]$ is minimized, givenn the joint distribution of X and Y. Last time, we showed that the MMSE of X given Y is $\mathbb{E}[X|Y]$.

$$\mathbb{E}[X \mid Y = y] = \int x f_{X|Y}(x|y)dx = \int x \frac{f_{X,Y}(x,y)}{f_Y(y)}dx \tag{23.1}$$

Some properties of the random variable $\mathbb{E}[X|Y]$ are

1. Linearity: $\mathbb{E}[aX_1 + bX_2 \mid Y] = a\,\mathbb{E}[X_1 \mid Y] + b\,\mathbb{E}[X_2 \mid Y]$

2. Factoring: $\mathbb{E}[h(Y)X \mid Y] = h(Y)\,\mathbb{E}[X \mid Y]$

3. Iterated expectation: $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$

4. Independence: if X and Y are independent, $\mathbb{E}[X|Y] = E[X]$.

Lemma 23.1. *For any function φ,*

$$\mathbb{E}[(X - E[X|Y])\varphi(Y)] = 0 \tag{23.2}$$

and further, if $g(Y)$ is such that

$$\mathbb{E}[(X - g(Y))\varphi(Y)] = 0 \ \forall\varphi \tag{23.3}$$

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

Proof. We look at the following expression:

$$\mathbb{E}[\mathbb{E}[X|Y]\varphi(Y)] = \mathbb{E}[\mathbb{E}[\varphi(Y)X \mid Y]] = \mathbb{E}[X\varphi(Y)] \tag{23.4}$$

Therefore, by linearity,

$$\mathbb{E}[(X - \mathbb{E}[X|Y])\varphi(Y)] = \mathbb{E}[X\varphi(Y)] - \mathbb{E}[\mathbb{E}[X|Y]\varphi(Y)] = 0 \tag{23.5}$$

Further, for the general $g(Y)$ case, let $\varphi(Y) = g(Y) - \mathbb{E}[X|Y]$. That is,

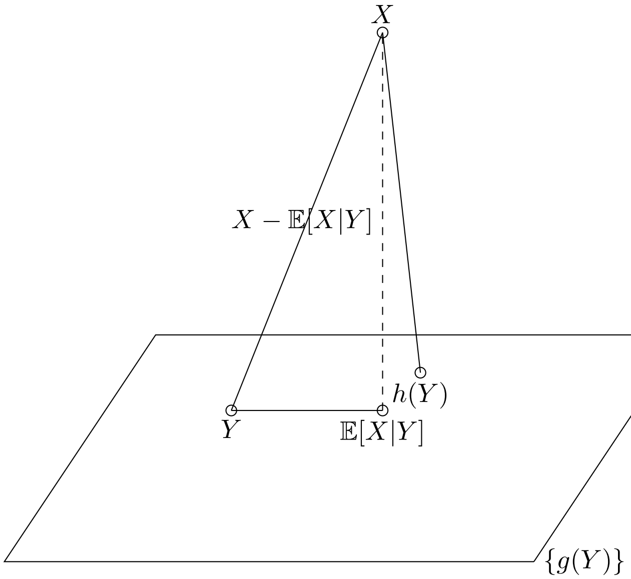
$$\begin{aligned} \mathbb{E}[(X - g(Y))\varphi(Y)] =& \mathbb{E}[(g(Y) - E[X|Y])^2] = 0 && (23.6)\\ &\implies \mathbb{E}[(g(Y) - \mathbb{E}[X|Y]) \cdot ((g(Y) - X) - (\mathbb{E}[X|Y] - X))] = 0 && (23.7) \end{aligned}$$

We recognize the first term as $\varphi(Y)$. Expanding this out, we note that the term $\varphi(Y)(\mathbb{E}[X|Y] - X)$ must be zero in expectation by the first part of the lemma, meaning the term $(g(Y) - X)\varphi(Y)$ must also be zero in expectation. Therefore $\mathbb{E}[\varphi(Y)] = 0$ which means its first and second moments are both zero, so $\varphi(Y) = 0$ and $g(Y) = \mathbb{E}[X|Y]$. □

Now, we can prove the MMSE theorem.

Proof. (of the MMSE theorem)

Suppose there were some projection $h(Y)$ such that the distance-squared were less than what we claim is the minimal value.



We can expand out the expectation with the greater value:

$$\mathbb{E}[(X - h(Y))^2] = \mathbb{E}[(X - \mathbb{E}[X|Y])^2] + \mathbb{E}[(\mathbb{E}[X|Y] - h(Y))^2] + 2\,\mathbb{E}[(X - \mathbb{E}[X|Y])(\mathbb{E}[X|Y] - h(Y))] \tag{23.8}$$

The cross term is zero by the lemma, because it is some function of Y multiplied by $X - \mathbb{E}[X|Y]$, and the term $\mathbb{E}[(\mathbb{E}[X|Y] - h(Y))^2]$ must be nonnegative. Therefore

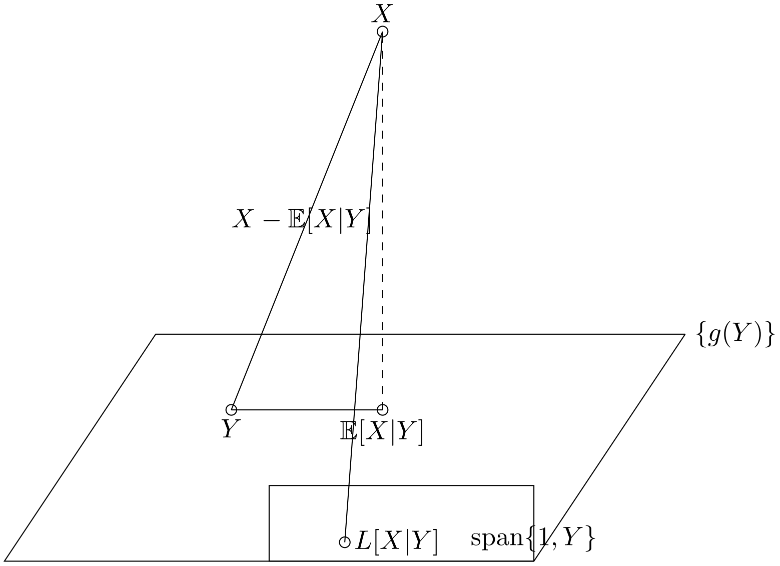
$$\mathbb{E}[(X - h(Y))^2] \geq \mathbb{E}[(X - \mathbb{E}[X|Y])^2] \tag{23.9}$$

□

23.2 Relating MMSE and LLSE

In general, the LLSE and MMSE are not equal, and the error of the LLSE is greater than or equal to that of the MMSE.

NOTE: tikz to be edited



Example 23.1. Let $Y \sim Unif[-1, 1]$ and let $X = Y^2$. The MMSE is

$$E[X|Y] = Y^2 \tag{23.10}$$

whereas the LLSE is

$$\text{Ł}[X|Y] = \mathbb{E}[X] = \frac{1}{3} \tag{23.11}$$

These are clearly different.

□

Example 23.2. Suppose X, Y, Z are i.i.d. and we have access to their sum.

$$\mathbb{E}[X|X+Y+Z] = \frac{1}{3}(X+Y+Z) \tag{23.12}$$

By symmetry, this is also the same as $\mathbb{E}[Y|X+Y+Z]$ and $\mathbb{E}[Z|X+Y+Z]$. In this case, the LLSE and MMSE are the same.

□

23.3 Jointly Gaussian RVs, LLSE of vectors

Suppose we have random variables X and Y given by

$$\begin{bmatrix} X \\ Y \end{bmatrix} = A_{2\times k} Z_{k\times 1} + \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \tag{23.13}$$

where the Z_is are i.i.d. $\mathcal{N}(0,1)$. In this case, the LLSE and MMSE are the same.

Before we can analyze this further, we have to develop the machinery of finding the LLSE of vectors. Let X and Y be random vectors, and suppose $\Sigma_Y = E[(Y - \mathbb{E}[Y])(Y - E[Y])^T]$. Then

$$L[X|Y] = \mathbb{E}[X] + \text{cov}(X,Y)\Sigma_Y^{-1}(Y - \mathbb{E}[Y]) \tag{23.14}$$

where the covariance of the two vectors is given by $\text{cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T]$.

The error of the linear estimator is

$$\mathbb{E}[\|X - L[X|Y]\|^2] = tr\left(\Sigma_x - \text{cov}(X,Y)\Sigma_y^{-1}\,\text{cov}(Y,X)\right) \tag{23.15}$$

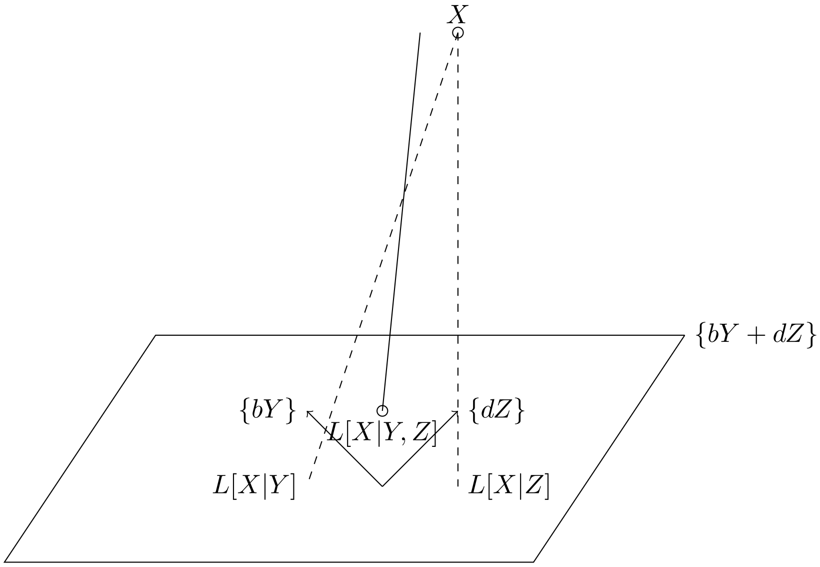
23.4 LLSE Updates

Suppose we have a linear estimate $L[X|Y]$ and then we get a second observation Z that is orthogonal to Y. Then

$$L[X|Y,Z] = L[X|Y] + L[X|Z] \tag{23.16}$$

We need to show that $X - (L[X|Y] + L[X|Z])$ is orthogonal to Y and Z. This can be done geometrically.

NOTE: tikz to be edited.


X
$\{bY + dZ\}$
$\{bY\}$
$L[X|Y, Z]$
$\{dZ\}$
$L[X|Y]$
$L[X|Z]$

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 24: Jointly Gaussian Random Variables, Kalman Filter

Lecturer: Shyam Parekh *3 December* *Aditya Sengupta*

We know that given a finite basis $\{u_i\}_{i=1}^n$ of a subspace U, we can find an orthonormal basis of U. This allows us to define jointly Gaussian random variables. $Y_1, \ldots, Y_n$ are jointly Gaussian RVs if

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = A_{n \times k} Z_{k \times 1} + \mu_{Y\,n \times 1} \tag{24.1}$$

where each Z_i in $Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_k \end{bmatrix}$ is i.i.d. $\mathcal{N}(0,1)$. Y has covariance matrix $\Sigma_Y = AA^T$. We denote this by $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$.

Theorem 24.1. *The joint density of $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$ is*

$$f_Y(y) = \frac{1}{\sqrt{2\pi|\Sigma_Y|}} \exp\left(-\frac{1}{2}(y-\mu_Y)^T \Sigma_Y^{-1}(y-\mu_Y)\right) \tag{24.2}$$

We assume that Σ_Y is invertible.

Example 24.1. Let $Y_i \sim \mathcal{N}(0, \Sigma_i^2)$ for $i = 1, 2$.

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \tag{24.3}$$

The joint density is

$$\begin{aligned} f_Y(y_1, y_2) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2}\begin{bmatrix} y_1 & y_2 \end{bmatrix} \begin{bmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}\right) && (24.4) \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2}\left(\frac{y_1^2}{\sigma_1^2} + \frac{y_2^2}{\sigma_2^2}\right)\right) && (24.5) \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2}\frac{y_1^2}{\sigma_1^2}\right) \cdot \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{1}{2}\frac{y_2^2}{\sigma_2^2}\right) && (24.6) \end{aligned}$$

This satisfies the definition of being jointly Gaussian. However, if we say that $f_Y(y_1, y_2)$ is defined this way if $y_1 y_2 > 0$ and the density is 0 otherwise, they are no longer jointly Gaussian.

□

Theorem 24.2. *Jointly Gaussian RVs are independent iff they are uncorrelated.*

The components being uncorrelated implies the components are independent. This is easy to verify when all $\mathrm{cov}(Y_i, Y_j) = 0$ for all $i \neq j$.

The level curves of the joint density of jointly Gaussian RVs are ellipses.

Theorem 24.3. *For jointly Gaussian RVs, the LLSE and MMSE are the same.*

Proof. $X - L[X|Y]$ and Y are uncorrelated, so they are independent. This means $X - L[X|Y]$ and $\phi(Y)$ are independent for any function ϕ. This means $X - L[X|Y]$ is orthogonal to the space of functions of Y. Therefore $X - L[X|Y] = \mathbb{E}[X|Y]$ so the LLSE is the MMSE. □

EECS 126: Probability and Random Processes **Fall 2019**

Lecture 25: The Kalman Filter

Lecturer: Shyam Parekh *5 December* *Aditya Sengupta*

Suppose we have a system with state $\{X_i\}_{i=1}^n$ and output $\{Y_i\}_{i=1}^n$. In general both are vectors, and typically Y will have lower dimensionality than X. We have some model that gives us how the states change over time:

$$X_n = AX_{n-1} + V_n \tag{25.1}$$

where V_n is zero-mean process noise with a known covariance matrix Q. We measure Y according to some known translation, given by a matrix C, and we also have some zero-mean measurement noise W_n with known covariance R:

$$Y_n = CX_n + W_n \tag{25.2}$$

The estimation problem here is to estimate X_n, the state, given the observed outputs Y_n. Since we assume everything is linear, we denote this by $L[X_n \mid Y_1, \dots, Y_n]$.

For the scalar case, we assume $|a| < 1$; for the vector case, we can either assume $|\,|A|\,| < 1$ or add a control term. In the scalar case, we can also rescale such that $c = 1$.

Recall that if X, Y, Z are zero-mean, then $L[X \mid Y, Z] = L[X \mid Y] + L[X \mid Z - L[Z \mid Y]]$. We can intuitively understand this as an update satep: we first have the estimate $L[X \mid Y]$, then when we observe Z, we can update this by adding the corrective term.

25.1 Deriving the Kalman Filter

Let $\hat{x}_{n|n} = L[x_n \mid y_1, \dots, y_n]$. From the above result, we can write a recurrence relation for this:

$$\hat{x}_{n|n} = L[x_n \mid y_1, \dots, y_{n-1}] + L[x_n \mid y_n - L[y_n \mid y_1, \dots, y_{n-1}]] = \hat{x}_{n|n-1} + K_n\tilde{y_n} \tag{25.3}$$

The first term is a prediction: assuming we know $\hat{x}_{n-1|n-1}$, we can just multiply by A to get our prediction $\hat{x}_{n|n-1}$. The second is an update: we multiply the error (or more optimistically innovation) term $y_n - L[y_n \mid y_1, \dots, y_{n-1}]$ by some "magical" gain to give us the linear least-square estimate of the state. Intuitively, we make a guess as to what'll happen to the state, and from there a guess as to what measurement to expect. When we get an actual measurement, we have a term denoting how far off we are, so we multiply this by the Kalman gain matrix K_n to translate that to a correction in the state domain.

$$\hat{x}_{n|n-1} = a\hat{x}_{n-1|n-1} \tag{25.4}$$

$$\tilde{y}_n = y_n - L[y_n \mid y_1, \dots, y_{n-1}] = y_n - CL[x_n \mid y_1, \dots, y_{n-1}] = y_n - C\hat{x}_{n|n-1} \tag{25.5}$$

(In the scalar case, we can drop the C.) Therefore, we get the overall equation

$$\hat{x}_{n|n} = A\hat{x}_{n-1|n-1} + K_n(y_n - CA\hat{x}_{n-1|n-1}) \tag{25.6}$$

This is all predicated on our ability to derive the magic scalar/matrix K_n. The website has a derivation of the scalar case, so I'm going to derive the vector case.

Along with the states $\hat{x}_{n|n}$, we maintain estimates of the covariance of the state, $P_{n|n}$. During a 'predict' step, the state is propagated forward in time according to $\hat{x}_{n|n-1} = A\hat{x}_{n-1|n-1} + V_n$ and the covariance is updated according to $P_{n|n-1} = AP_{n-1|n-1}A^T + Q$. This is important for finding the steady-state covariance and therefore for the steady-state Kalman gain.

$$\hat{x}_{n|n-1} = A\hat{x}_{n-1|n-1} + V_n \tag{25.7}$$

$$P_{n|n-1} = AP_{n-1|n-1}A^T + Q \tag{25.8}$$

During an 'update' step, the state is updated according to $\hat{x}_{n|n} = \hat{x}_{n|n-1} + K_n(y_n - C\hat{x}_{n|n-1}) = (I - KC)\hat{x}_{n|n-1} + Ky_n$. Recall that measurement covariance is given by a matrix R. The covariance is updated according to $P_{n|n} = (I - K_nC)P_{n|n-1}(I - K_nC)^T + K_nRK_n^T$. The Kalman gain is the matrix that minimizes $P_{n|n}$.

$$\hat{x}_{n|n} = (I - KC)\hat{x}_{n|n-1} + Ky_n + W_n \tag{25.9}$$

$$P_{n|n} = (I - K_nC)P_{n|n-1}(I - K_nC)^T + K_nRK_n^T \tag{25.10}$$

To find the gain, we set the following:

$$\frac{\partial \mathrm{Tr} P_{n|n}}{\partial K_n} = 0 \tag{25.11}$$

$$\frac{\partial \mathrm{Tr} P_{n|n-1}}{\partial K_n} - 2\frac{\partial \mathrm{Tr} K_nCP_{n|n-1}}{\partial K_n} + \frac{\partial \mathrm{Tr} K_nCP_{n|n-1}C^TK_n^T}{\partial K_n} + \frac{\partial \mathrm{Tr} K_nRK_n^T}{\partial K_n} = 0 \tag{25.12}$$

The first term is a constant with respect to K_n, so we drop it. The other three can be evaluated using trace differentiation properties: $\frac{\partial \mathrm{Tr} ABA^T}{\partial A} = 2AB$ for symmetric A, and $\frac{\partial \mathrm{Tr} AC}{\partial A} = C^T$. We get

$$-2P_{n|n-1}C^T + 2K_nCP_{n|n-1}C^T + 2K_nR = 0 \tag{25.13}$$

This gives us the Kalman gain:

$$K_n = P_{n|n-1}C^T[CP_{n|n-1}C^T + R]^{-1}. \tag{25.14}$$

Further, we can put the system into a steady state. That is, we can find the K that keeps the covariance constant, so that we can precompute the gain and compute MMSE state estimates online without having to track the covariance. We do this by requiring $P_{n+1|n} = P_{n|n-1}$ (we do the update before the prediction just for algebraic simplicity).

$$P_{n|n-1} = AP_{n|n-1}A^T + Q - AP_{n|n-1}C^T(CP_{n|n-1}C^T + R)^{-1}CP_{n|n-1}A^T \tag{25.15}$$

This is a specific form of something called an algebraic Ricatti equation, which can be analytically solved. In practice, you can get this solution just by running the filter until a steady state is reached.