Random Processes in Systems

Electrical Engineering 226A
Fall 2017

Professor Venkat Anantharam
Scribe: Sinho Chewi

Contents

Lecture 1

August 24

1.1 Probability Framework

$(\Omega, \mathcal{F}, \mathbb{P})$.

- Ω is a set called the **sample space**. Its points are called **outcomes**.
- $\mathcal{F}$ is a collection of subsets of Ω callled **events** (often it is *not possible* to let Ω be the family of all subsets of Ω if one wants to satisfy the axioms of the theory).
- $\mathbb{P} : \mathcal{F} \to [0, 1]$, where $\mathbb{P}(A)$ for $A \in \mathcal{F}$ is called the **probability** of the event A.

Example 1.1. Take $\Omega = \{1, 2, \ldots, 6\}$, $\mathcal{F}$ is all subsets of Ω, and $\mathbb{P}(A) = |A|/6$ (where $|A|$ is the cardinality of A). This is the probability model for a single roll of a fair six-sided die.

Example 1.2. Take $\Omega = \{1, \ldots, 6\}^{\mathbb{Z}}$, the set of all two-sided infinite sequences with values from $\{1, \ldots, 6\}$. $\mathcal{F}$ is the subsets of Ω that can be constructed from subsets based on finitely many time indices by complementation, countable unions, and countable intersections. If

$$A = \{\omega \in \Omega : \omega_k = a_k,\ \omega_{k+1} = a_{k+1}, \ldots, \omega_{k+\ell} = a_{k+\ell}\},$$

where $a_k \in \{1, \ldots, 6\}$, $a_{k+1} \in \{1, \ldots, 6\}$, $\ldots$, $a_{k+\ell} \in \{1, \ldots, 6\}$, then

$$\mathbb{P}(A) = \left(\frac{1}{6}\right)^{\ell+1}.$$

It turns out from the axioms that this will uniquely specify $\mathbb{P}(A)$ for all $A \in \mathcal{F}$.

$\mathcal{F}$ captures the idea of information. For example, we could consider the experiment of rolling two dice, where we only care about the first die. Then, take $\Omega = \{(i, j) : 1 \le i, j \le 6\}$, which has size 36, $\mathcal{F}$ to be all subsets of the form $A \times \{1, \ldots, 6\}$ where $A \subseteq \{1, \ldots, 6\}$, and

$$\mathbb{P}(A) = \frac{|A|}{36}, \qquad \text{for } A \in \mathcal{F}.$$

There are two sets of axioms.

Axioms for $\mathcal{F}$:

1. $A \in \mathcal{F} \implies A^{\mathrm{c}} \in \mathcal{F}$ ($A^{\mathrm{c}} = \Omega \setminus A$ is the **complement** of A)
2. $\varnothing \in \mathcal{F}$

3. $A \in \mathcal{F}$, $B \in \mathcal{F} \implies A \cup B \in \mathcal{F}$

4. If $A_1, A_2, A_3, \ldots$ is a countable collection of events, i.e., $A_1 \in \mathcal{F}$, $A_2 \in \mathcal{F}$, etc., then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Axioms for probability:

1. $\mathbb{P}(\varnothing) = 0$.

1' $\mathbb{P}(\Omega) = 1$.

2. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ if $A \cap B = \varnothing$.

2' If $A_1, A_2, A_3, \ldots$ is a sequence of mutually disjoint events ($A_i \cap A_j = \varnothing$ if $i \neq j$), then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Example 1.3. Take $\Omega = [0,1]$, $\mathcal{F}$ is all subsets of Ω that can be constructed from intervals of the type $[a,b]$ or $(a,b]$ or $[a,b)$ or (a,b) by the operations of complementation, countable unions, and countable intersections, and $\mathbb{P}(A)$ is the length of A. This models picking a point uniformly at random on $[0,1]$.

The key thing that makes this theory tick is that when probabilities are replaced by conditional probabilities, they still obey the axioms of probability. Given an event B with $\mathbb{P}(B) > 0$, $\mathbb{P}(A \mid B)$ is defined to be $\mathbb{P}(A \cap B)/\mathbb{P}(B)$. Then, $\mathbb{P}(\varnothing \mid B) = 0$; $\mathbb{P}(\Omega \mid B) = 1$; if A_1, A_2 are disjoint, then

$$\mathbb{P}(A_1 \cap A_2 \mid B) = \mathbb{P}(A_1 \mid B) + \mathbb{P}(A_2 \mid B);$$

similarly for countable disjoint unions.

Lecture 2

August 29

2.1 Review of Probability

Reading from the book:

- Chapter 0
- Chapter 1, Sections 1-6 (starts on Thursday)

Last lecture: $(\Omega, \mathcal{F}, \mathbb{P})$.

Upshot: $\mathbb{P}(A)$ is the probability of *event* A, e.g. $\mathbb{P}(\text{roll of the die} \geq 5)$ or $\mathbb{P}(\text{disease is cancer})$.

2.1.1 Conditional Probability

If B is an event with $\mathbb{P}(B) > 0$, define

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

for all events A. This is called the conditional probability of A given B. Then, $A \mapsto \mathbb{P}(A \mid B)$ satisfies the axioms of probability.

Bayes rule:

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

(if $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$). This is used for inference.

Example 2.1. Suppose we have a transmitter and a receiver. The link is either good or bad. Suppose that we know (from the model) $\mathbb{P}(\text{channel bad})$ and $\mathbb{P}(\text{channel good})$ and the two add up to 1, and suppose we know $\mathbb{P}(\text{reception is good} \mid \text{channel good})$, $\mathbb{P}(\text{reception is good} \mid \text{channel bad})$, and also $\mathbb{P}(\text{reception is good})$. Then,

$$\mathbb{P}(\text{channel bad} \mid \text{reception is good}) = \frac{\mathbb{P}(\text{reception is good} \mid \text{channel bad})\mathbb{P}(\text{channel bad})}{\mathbb{P}(\text{reception is good})}.$$

2.1.2 Total Probability Formula

Suppose $A_1, A_2, \dots, A_K$ form a partition of the sample space. Then, for any event B,

$$\mathbb{P}(B) = \sum_{k=1}^{K} \mathbb{P}(B \mid A_k)\mathbb{P}(A_k).$$

More generally, if $A_1, A_2, A_3, \dots$ is a countably infinite partition of Ω, then for all events B,

$$\mathbb{P}(B) = \sum_{k=1}^{\infty} \mathbb{P}(B \mid A_k)\mathbb{P}(A_k).$$

We could have figured out $\mathbb{P}$(reception is good) in 2.1 from the Total Probability Theorem.

2.1.3 Inclusion-Exclusion Formula

For all events A, B,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

More generally,

$$\begin{aligned}\mathbb{P}\left(\bigcup_{k=1}^{K} A_k\right) &= \sum_{k=1}^{K} \mathbb{P}(A_k) - \sum_{1 \le i_1 < i_2 \le K} \mathbb{P}(A_{i_1} \cap A_{i_2}) + \sum_{1 \le i_1 < i_2 < i_3 \le K} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}) \\ &\quad - \sum_{1 \le i_1 < i_2 < i_3 < i_4 \le K} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3} \cap A_{i_4}) + \cdots - (-1)^K \mathbb{P}(A_1 \cap \cdots \cap A_K).\end{aligned}$$

We can immediately write down

$$\mathbb{P}(A \cup B \mid C) = \mathbb{P}(A \mid C) + \mathbb{P}(B \mid C) - \mathbb{P}(A \cap B \mid C)$$

since conditional probabilities satisfy Kolmogorov's axioms.

2.1.4 Monotonic Continuity of Probability

1. If $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$ is an increasing sequence of events, then $\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \lim_{k\to\infty} \mathbb{P}(A_k)$. From the countable additivity of probability,

$$\begin{aligned}\text{LHS} = \mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) &= \mathbb{P}(A_1) + \mathbb{P}(A_2 \setminus A_1) + \mathbb{P}(A_3 \setminus A_2) + \cdots \\ &= \lim_{k\to\infty} \mathbb{P}(A_1) + \mathbb{P}(A_2 \setminus A_1) + \cdots + \mathbb{P}(A_k \setminus A_{k-1}) = \lim_{k\to\infty} \mathbb{P}(A_k),\end{aligned}$$

   because $\mathbb{P}(A_k) = \mathbb{P}(A_1) + \mathbb{P}(A_2 \setminus A_1) + \cdots + \mathbb{P}(A_k \setminus A_{k-1})$.

2. Similarly, if $B_1 \supseteq B_2 \supseteq \cdots$ is a sequence of events, then $\mathbb{P}(\bigcap_{k=1}^{\infty} B_k) = \lim_{k\to\infty} \mathbb{P}(B_k)$ (this has the same kind of proof).

2.1.5 Independence

We say events A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. What this definition is trying to capture is that $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ (if $\mathbb{P}(B) > 0$). Note that this is equivalent to $\mathbb{P}(B \mid A) = \mathbb{P}(B)$ (if $\mathbb{P}(A) > 0$). For multiple events $A_1, \dots, A_k$, we say they are independent if for every subset $S \subseteq \{1, \dots, k\}$, $\mathbb{P}(\bigcap_{k\in S} A_k) = \prod_{k\in S} \mathbb{P}(A_k)$. It is *not sufficient* to just assume that $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$ for all $1 \le i < j \le k$.

2.1.6 Borel-Cantelli Lemmas

Given a sequence of real numbers $(x_n,\ n \ge 1)$,

$$\limsup_{n\to\infty} x_n = \lim_{k\to\infty} \left(\sup_{m\ge k} x_m\right),$$

$$\liminf_{n\to\infty} x_n = \lim_{k\to\infty}\left(\inf_{m\geq k} x_m\right).$$

Consider the sequence $(1-1/n,\ n\geq 1)$; there is no maximum, but the supremum is 1. $\lim_{n\to\infty} x_n$ exists iff $\limsup_{n\to\infty} x_n = \liminf_{n\to\infty} x_n$. For example, in the sequence $1, -1, 1, -1, \dots$ has $\limsup_{n\to\infty} x_n = 1$ and $\liminf_{n\to\infty} x_n = -1$. As another example, $\limsup_{n\to\infty} \sin n = 1$ and $\liminf_{n\to\infty} \sin n = -1$.

Given a sequence of events $A_1, A_2, A_3, \dots$, consider $\bigcap_{k\geq 1}(\bigcup_{m\geq k} A_m)$. This is the event that the events $(A_\ell,\ \ell \geq 1)$ occur *infinitely often*. This is denoted $\limsup_{n\to\infty} A_n$. $\bigcup_{k\geq 1}(\bigcap_{m\geq k} A_m)$ is the event that eventually the events A_m keep happening.

Lemma 2.2 (First Borel-Cantelli Lemma)**.** *If $(A_n,\ n\geq 1)$ is a sequence of events and*

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty,$$

then $\mathbb{P}(\limsup_{n\to\infty} A_n) = 0$, i.e. $\mathbb{P}(A_n$ occur infinitely often$) = 0$.

Proof.

$$\begin{aligned}\mathbb{P}\left(\limsup_{n\to\infty} A_n\right) &= \lim_{k\to\infty} \mathbb{P}\left(\bigcup_{m\geq k} A_m\right) && \text{by monotonic continuity of probability}\\ &\leq \lim_{k\to\infty} \sum_{m=k}^{\infty} \mathbb{P}(A_m) && \text{by union bound}\end{aligned}$$

but the RHS $\to 0$ as $k\to\infty$ by the assumption that $\sum_{\ell=1}^{\infty} \mathbb{P}(A_\ell) < \infty$. □

Lemma 2.3 (Second Borel-Cantelli Lemma)**.** *If $A_1, A_2, \dots$ are independent events, and*

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty,$$

then $\mathbb{P}(\limsup_{n\to\infty} A_n) = 1$.

Lecture 3

August 31

3.1 Random Variables

Definition 3.1. A **real-valued random variable** is a real-valued function on the sample space with the restriction that for all $x \in \mathbb{R}$, $\{\omega \in \Omega : X(\omega) \leq x\}$ is an event. We write this event as $\{X \leq x\}$.

From the axioms of a probability model, this turns out to be enough to make $\{X \in A\}$ an event for any subset $A \subseteq \mathbb{R}$ that can be constructed from intervals by complementation, countable unions, and countable intersections. The function $x \in \mathbb{R} \mapsto \mathbb{P}(X \leq x)$ is denoted F_X. $F_X(x)$ is the **cumulative distribution function of X at x**. F_X is non-decreasing and it increases from 0 to 1 (usually: sometimes we will find it useful to allow $\{X = \infty\}$ and $\{X = -\infty\}$ to have positive probability; when we do this, we are talk about "extended random variables" $(\Omega, \mathcal{F}, \mathbb{P}) \xrightarrow{X} \mathbb{R} \cup \{\infty, -\infty\}$). We will also consider random variables taking values in abstract spaces: $(\Omega, \mathcal{F}, \mathbb{P}) \xrightarrow{X} (E, \mathcal{E})$, where E is some set of values (e.g. $\{\text{red}, \text{blue}, \text{green}\}$) and $\mathcal{E}$ is the allowed collection of subsets of E satisfying the axioms of Kolmogorov which apply to $\mathcal{F}$.

To do calculations with X, it will then be enough to know the **distribution** of X. This is the map taking $B \in \mathcal{E} \mapsto \mathbb{P}(X \in B)$.

Note: $\{X \in B\}$ means $\{\omega \in \Omega : X(\omega) \in B\}$.

A real-valued random variable is called **discrete** if there is a countable subset of real numbers that carries all of the probability, i.e. there is a subset $\{x_1, x_2, \ldots\} \subseteq \mathbb{R}$, countable, such that $\mathbb{P}(X \in \{x_1, x_2, \ldots\}) = 1$. We would calculate with such X via the $\mathbb{P}(X = x_k)$, $k = 1, 2, \ldots$, called the **probability mass function**.

It is called **absolutely continuous** if there is a non-negative function $f_X : \mathbb{R} \to [0, \infty)$ such that

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

for all $x \in \mathbb{R}$.

A random variable is **singular continuous** if its CDF (F_X) is continuous but its derivative is the zero function. For example, let F_1 be the function which increases from 0 to 1/2 on the interval $[0, 1/3]$, stays flat, and then increases from 1/2 to 1 on $[1/3, 2/3]$. Let F_2 be the function which increases from 0 to 1/4 on $[0, 1/9]$, stays flat, increases from 1/4 to 1/2 on $[2/9, 1/3]$, stays flat, increases from 1/2 to 3/4 on $[2/3, 7/9]$, stays flat, and increases from 3/4 to 1 on $[8/9, 1]$. Do this iteratively. Then, $\lim_{n \to \infty} F_n$ is a continuous function on $\mathbb{R}$ with zero derivative "everywhere".

The most general real-valued random variable will have its distribution expressible as the sum of three parts, one of each type. We will only consider distributions that are built up from a discrete and absolutely continuous part.

3.1.1 Standard Distributions

Bernoulli(p): $\mathbb{P}(X=0)=1-p$, $\mathbb{P}(X=1)=p$, where $p\in[0,1]$.

Binomial(n,p): $\mathbb{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ where $n\geq 1$, $p\in[0,1]$.

Recall: If $X_1,X_2,\ldots,X_n$ are *independent* Bernoulli(p) random variables, then $X_1+\cdots+X_n$ will be Binomial(n,p).

Given n real-valued random variables $X_1,\ldots,X_n$ on $(\Omega,\mathcal{F},\mathbb{P})$ which are jointly defined, their joint CDF $F_{X_1,\ldots,X_n}:\mathbb{R}^n\to[0,1]$ is defined by $F_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\triangleq\mathbb{P}(X_1\leq x_1,X_2\leq x_2,\ldots,X_n\leq x_n)$. We will say that $X_1,\ldots,X_n$ are **independent** if for all valid subsets $A_1\subseteq\mathbb{R},A_2\subseteq\mathbb{R},\ldots,A_n\subseteq\mathbb{R}$, the events $\{X_1\in A_1\},\ldots,\{X_n\in A_n\}$ are mutually independent. It turns out that this is equivalent to requiring that

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n)=F_{X_1}(x_1)\cdots F_{X_n}(x_n)$$

for all $(x_1,\ldots,x_n)\in\mathbb{R}^n$. If $X_1,\ldots,X_n$ are each *discrete*, we can discuss them through their joint PMF and independence is equivalent to $\mathbb{P}(X_1=y_1,\ldots,X_n=y_n)=\prod_{i=1}^n\mathbb{P}(X_i=y_i)$. We say that $(X_1,\ldots,X_n)$ are **jointly continuous** if we can write $F_{X_1,\ldots,X_n}(x_1,\ldots,x_n)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f_{X_1,\ldots,X_n}(y_1,\ldots,y_n)\,\mathrm{d}y_1\cdots\mathrm{d}y_n$ for some $f_X(y_1,\ldots,y_n)\geq 0$. Here, independence is equivalent to $f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)=\prod_{i=1}^n f_{X_i}(x_i)$.

Poisson distribution:

$$\mathbb{P}(X=n)=\frac{\lambda^n}{n!}\mathrm{e}^{-\lambda},\qquad n=0,1,2,\ldots$$

where $\lambda>0$ (if $\lambda=0$, $\mathbb{P}(X=0)=1$).

Geometric distribution: $\mathbb{P}(X=k)=p(1-p)^k$, $k=0,1,2,\ldots$, where $p\in(0,1]$.

Exponential distribution:

$$f_X(x)=\begin{cases}\lambda\mathrm{e}^{-\lambda x}, & t\geq 0\\ 0, & t<0\end{cases}$$

for $\lambda>0$.

Gaussian:

$$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-(x-m)^2/(2\sigma^2)},$$

where $m\in\mathbb{R}$, $\sigma>0$ (later we will allow $\sigma=0$ but this has to be done through the characteristic function).

The **characteristic function** of a real-valued random variable X is by definition $\theta\mapsto\mathbb{E}[\mathrm{e}^{\mathrm{i}\theta X}]$, where $\mathrm{i}\triangleq\sqrt{-1}$, denoted Φ_X for now. We will see that this is well-defined for all θ, for all random variables X.

Example 3.2. If X has density f_X, $\Phi_X(\theta)=\int_{-\infty}^{\infty}\mathrm{e}^{\mathrm{i}\theta x}f_X(x)\,\mathrm{d}x$.

The **moment generating function** is $M_X(t)\triangleq\mathbb{E}[\mathrm{e}^{tX}]$. This can sometimes be undefined for all $t\neq 0$.

3.2 Expectation

You have been told that given a random variable X, for any function $h(X)$ of X:

$$\mathbb{E}[h(X)]=\begin{cases}\sum_i h(x_i)\mathbb{P}(X=x_i), & \text{if } X \text{ is discrete}\\ \int_{-\infty}^{\infty}h(x)f_X(x)\,\mathrm{d}x, & \text{if } X \text{ is continuous}\end{cases}$$

The truth is that this could potentially have the value $\infty - \infty$, and we do not know what this means. This kind of difficulty also arises even when we sum series.

Example 3.3. What is $1 + (-1) + 1 + (-1) + 1 + (-1) + \cdots$?

To deal with this, what we will mean by $\mathbb{E}[X]$ for X a random variable is $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$ if at least one of $\mathbb{E}[X^+]$ and $\mathbb{E}[X^-]$ is finite, and undefined otherwise. Here, $\mathbb{E}[X^+] \triangleq \mathbb{E}[\max(X, 0)]$ and $\mathbb{E}[X^-] \triangleq \mathbb{E}[\max(-X, 0)]$. To define the right hand sides, one only needs to define expectation for non-negative random variables. This is done by approximating them from below by a sequence of simple random variables. A **simple random variable** is piecewise constant with finitely many pieces. (See the handout.)

Lecture 4

September 5

4.1 Discrete-Time Finite State Markov Chains

4.1.1 Basic Definition

A discrete-time stochastic process $\dots, X_{-2}, X_{-1}, X_0, X_1, X_2, \dots$ with each RV taking only finitely many values can be specified, for the purpose of calculations, via its "finite-dimensional distributions", i.e. the numbers $\mathbb{P}(X_k = i_k, X_{k+1} = i_{k+1}, \dots, X_{k+\ell} = i_{k+\ell})$ for all k, all $\ell \geq 0$, and all $(i_k, \dots, i_{k+\ell})$.

Let $\mathcal{S}$ be a finite set. We will say that an $\mathcal{S}$-valued random sequence $(X_n,\ n \in \mathbb{Z})$ is a **Markov chain** if $\mathbb{P}(X_{n+1} = j \mid X_n = i, A) = \mathbb{P}(X_{n+1} = j \mid X_n = i)$ for all $n \in \mathbb{Z}$, for all $i, j \in \mathcal{S}$, for all events A that depend only on the past at time n (i.e. that are functions of $(X_k,\ k \leq n)$, i.e. $\mathbb{1}_A = g(X_k,\ k \leq n)$ for some function $g(\cdot)$). Equivalently, we require

$$\mathbb{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_{n-d} = i_{n-d}) = \mathbb{P}(X_{n+1} = j \mid X_n = i)$$

for all n, all $d \geq 0$, all $i, i_{n-1}, \dots, i_{n-d}, j \in \mathcal{S}$. We will assume in addition that $\mathbb{P}(X_{n+1} = j \mid X_n = i)$ does not depend on n (time homogeneity assumption). We will write $p(i,j)$ for $\mathbb{P}(X_{n+1} = j \mid X_n = i)$ (sometimes people write $p(j \mid i)$ instead). Let

$$\underbrace{P}_{|\mathcal{S}| \times |\mathcal{S}| \text{ matrix}} \triangleq \big[p(i,j)\big] .$$

P is a square matrix with non-negative entries, each of whose rows sums to 1, i.e.

$$P \underbrace{\mathbf{1}}_{\text{column vector of all 1s}} = \mathbf{1}.$$

Such a matrix is called a **stochastic matrix**.

P can be *drawn* via a transition diagram.

4.1.2 Chapman-Kolmogorov Equations

To fully specify the joint distributions, one needs more than just P. Typically, one discusses the one-sided case, i.e. $(X_0, X_1, X_2, \dots)$, specifies the initial distribution (a row vector $(1 \times \mathcal{S})$, listing $[\mathbb{P}(X_0 = i),\ i \in \mathcal{S}]$, and P. This is enough to compute, for all $k \geq 0$, $\ell \geq 0$, $i_k, i_{k+1}, \dots, i_{k+\ell} \in \mathcal{S}$,

$$\mathbb{P}(X_k = i_k, X_{k+1} = i_{k+1}, \dots, X_{k+\ell} = i_{k+\ell}).$$

Why? The reason is the Chapman-Kolmogorov equations, which say that the ℓ-step transition probabilities are given by the ℓth power of P (for $\ell \geq 0$; $P^0 \triangleq I$) which manifest as $P^{m+n} = P^m P^n$ for all $m \geq 0$, $n \geq 0$.

In the discrete-time setting, the content of the statement is in the first assertion. Let us check the equations. For $\ell = 0$, $\mathbb{P}(X_n = j \mid X_n = i) = \delta_{i,j}$, denoted $p^{(0)}(i,j)$, is the **Kronecker delta function** defined by:

$$\delta_{i,j} = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}$$

This says that the 0-step transition probability matrix is the identity matrix. For $\ell = 1$,

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

by definition. These probabilities are denoted $p^{(1)}(i,j)$. Let us do $\ell = 2$.

$$\begin{aligned} p^{(2)}(i,j) \triangleq \mathbb{P}(X_{n+2} = j \mid X_n = i) &= \sum_{k \in \mathcal{S}} \mathbb{P}(X_{n+2} = j, X_{n+1} = k \mid X_n = i) \\ &= \sum_{k \in \mathcal{S}} \mathbb{P}(X_{n+2} = j \mid X_{n+1} = k, X_n = i)\mathbb{P}(X_{n+1} = k \mid X_n = i) \\ &= \sum_{k \in \mathcal{S}} \mathbb{P}(X_{n+2} = j \mid X_{n+1} = k)\mathbb{P}(X_{n+1} = k \mid X_n = i) \\ &= \sum_{k \in \mathcal{S}} p(k,j)p(i,k) \\ &= \sum_{k \in \mathcal{S}} p(i,k)p(k,j) \\ &= P^2(i,j). \end{aligned}$$

So, $[p^{(2)}(i,j)] = P^2$. Similarly,

$$[p^{(\ell)}(i,j)] = P[p^{(\ell-1)}(i,j)] = P^\ell$$

by induction.

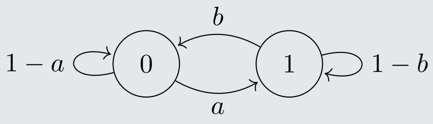
4.1.3 Recurrence & Transience

Let P ($|\mathcal{S}| \times |\mathcal{S}|$) be a stochastic matrix. A probability distribution on $\mathcal{S}$ viewed as a row vector is called a **stationary distribution** of P if $\pi P = \pi$. We will show that every stochastic matrix P admits a stationary distribution.

For any $y \in S$, we define $T_y = \inf\{n \geq 1 : X_n = y\}$. Note that $T_y \in \{1, 2, \ldots\} \cup \{\infty\}$.

Definition 4.1. If $\mathbb{P}(T_y < \infty \mid X_0 = y) = 1$, we call the state y **recurrent**, whereas if the quantity $\mathbb{P}(T_y < \infty \mid X_0 = y) < 1$, we call the state y **transient**.

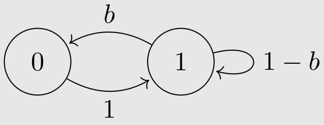
Example 4.2 (2-State MC)**.** If $0 < a < 1$, $0 < b < 1$:



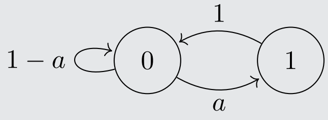
Here,

$$\pi = \begin{bmatrix} \dfrac{b}{a+b} & \dfrac{a}{a+b} \end{bmatrix}.$$

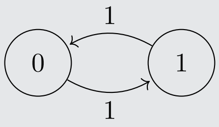
If $a = 1$, $0 < b < 1$:



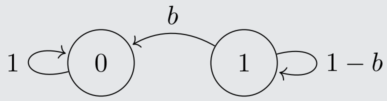
If $0 < a < 1$, $b = 1$:



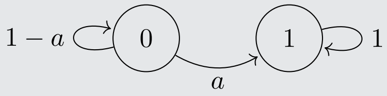
If $a = b = 1$:



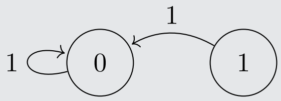
If $a = 0$, $0 < b < 1$:



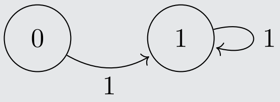
If $0 < a < 1$, $b = 0$:



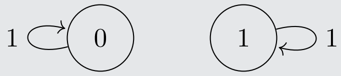
If $a = 0$, $b = 1$:



If $a = 1$, $b = 0$:



If $a = b = 0$:



In all 9 cases,

$$P = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}.$$

Let us focus on $a > 0$, $b > 0$ (the first cases). You can compute

$$P^n = \frac{1}{a+b}\begin{bmatrix} b + ac^n & a - ac^n \\ b - bc^n & a + bc^n \end{bmatrix}$$

where $c = 1 - a - b$. You want to think of this as

$$\begin{bmatrix} \dfrac{b}{a+b} & \dfrac{a}{a+b} \\ \dfrac{b}{a+b} & \dfrac{a}{a+b} \end{bmatrix} + c^n \begin{bmatrix} \dfrac{a}{a+b} & -\dfrac{a}{a+b} \\ -\dfrac{b}{a+b} & \dfrac{b}{a+b} \end{bmatrix}.$$

You want to observe that $|c| < 1$ (except when $a = 1$, $b = 1$). If $a = 1$, $b = 1$, then $c = -1$.

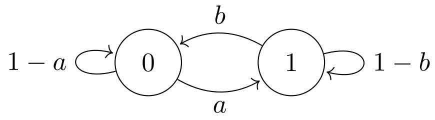
We will revisit this while doing the general theory. Think of this in linear algebra.

Lecture 5

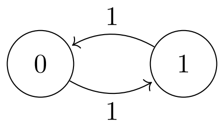
September 7

5.1 Stationary Distributions

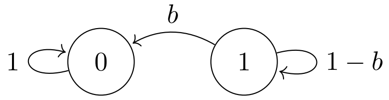
If $0 < a < 1$, $0 < b < 1$:



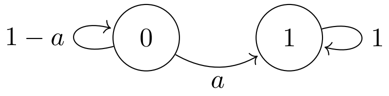
There is a single aperiodic irreducible class. There is a similar picture as long as $a > 0$, $b > 0$, except when $a = 1$ and $b = 1$:



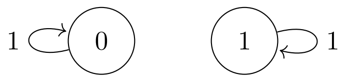
which exhibits *periodicity*. Then there is the case



(the case when $b = 1$ is similar) and



(the $a = 1$ case is similar). These two cases exhibit *transience*. Finally, there is the case



which has multiple irreducible classes.

Every finite Markov chain admits a stationary probability distribution, i.e. there is a row vector π with non-negative entries summing to 1 such that $\pi P = \pi$. We already know that $P\mathbf{1} = \mathbf{1}$, so we know there is a row vector η, $\eta \neq 0$, such that $\eta P = \eta$. We will write $\eta = \eta_+ - \eta_-$ where $\eta_+ \geq 0$, $\eta_- \geq 0$ with disjoint supports.

$$\overbrace{(\eta_+ - \eta_-)P}^{\eta P} = \eta_+ P - \eta_- P = \overbrace{\eta_+ - \eta_-}^{\eta}$$

so $\eta_+ P \geq \eta_+$ and $\eta_- P \geq \eta_-$ (because P has non-negative entries). But, $\eta_+ P\mathbf{1} = \eta_+\mathbf{1}$ (because $P\mathbf{1} = \mathbf{1}$) so $\eta_+ P = \eta_+$ and $\eta_- P = \eta_-$, so there is a stationary probability distribution.

A finite state DTMC is called **irreducible** if for every pair of states $i, j \in \mathcal{S}$, where $\mathcal{S}$ is the state space, i communicates with j, i.e. $p^{(n)}(i,j) > 0$ for some $n \geq 0$ (so by convention we say $i \to i$ even if there is no self-loop at i).

Our argument (just given) actually shows that for an irreducible Markov chain there is a unique stationary probability distribution.

Proof: If $\pi_1 P = \pi_1$ and $\pi_2 P = \pi_2$ and $\pi_1 \neq \pi_2$, where $\pi_1\mathbf{1} = 1$ and $\pi_2\mathbf{1} = 1$, and π_1 and π_2 have all entries non-negative, then write $\pi_1 - \pi_2$ as $\beta_+ - \beta_-$, where β_+, β_- are non-negative and have disjoint supports. Then, $(\beta_+ - \beta_-)P = \beta_+ P - \beta_- P = \beta_+ - \beta_-$ so $\beta_+ P \geq \beta_+$ and $\beta_- P \geq \beta_-$, so $\beta_+ P = \beta_+$, $\beta_- P = \beta_-$, which is impossible since β_+ and β_- must have support a proper subset of the state space.

Every eigenvalue of a stochastic matrix must have absolute value at most 1.

Proof: Let λ be an eigenvalue and $x \in \mathbb{C}^{|\mathcal{S}|}$ be an associated column eigenvector. Observe that

$$\max_{1\leq j\leq |\mathcal{S}|} |x(j)| \geq \max_{1\leq j\leq |\mathcal{S}|} |(Px)(j)|$$

because $(Px)(j) = \sum_j p(i,j)x(j)$ and $p(i,j) \geq 0$, $\sum_{j=1}^{|\mathcal{S}|} p(i,j) = 1$, using the principle

$$|\alpha a + (1-\alpha)b| \leq \alpha|a| + (1-\alpha)|b|$$

for $\alpha \in [0,1]$. Since $Px = \lambda x$, this means $|\lambda| \leq 1$.

5.2 Classification of States

$T_y \triangleq \inf\{n \geq 1 : X_n = y\}$ is the first "return" time of state y. Note that $\mathbb{P}(T_y = \infty) > 0$ can happen (for some initial distributions). Also, $g_y \triangleq \mathbb{P}(T_y < \infty \mid X_0 = y)$.

Definition: y is called **recurrent** if $g_y = 1$, and y is called **transient** if $g_y < 1$.

Notation. $\rho_{x,y} = \mathbb{P}(T_y < \infty \mid X_0 = x)$.

Define a subset $B \subseteq \mathcal{S}$ to be irreducible if for all $i, j \in B$, i communicates with j. (This definition is consistent with the earlier definition of what it means for the DTMC to be irreducible.)

Define a set of states $B \subseteq \mathcal{S}$ to be **closed** if for all $i \in B$, for all $j \in \mathcal{S} \setminus B$, i does not communicate with j.

Theorem 5.1. *Every finite state Markov chain with state space $\mathcal{S}$ has $\mathcal{S}$ comprised of disjoint subsets $S = T \cup R_1 \cup R_2 \cup \cdots \cup R_k$ (a partition into disjoint subsets) where T is entirely comprised of transient states and each R_i is closed and irreducible.*

We will also show: Every state in any closed irreducible subset of states is recurrent. Actually, T has more detail.

We will prove:

- Every state $i \in \mathcal{S}$ for which there is some state $j \in \mathcal{S}$ with $i \to j$ but $j \not\to i$ must be a transient state.

We will define $N_y = \sum_{k=1}^{\infty} \mathbb{1}_{\{X_k = y\}}$ (the number of visits to state y). Note $\mathbb{P}(N_y = \infty) > 0$ is possible. We will show

$$\mathbb{E}[N_y \mid X_0 = x] = \frac{\rho_{x,y}}{1 - \rho_{y,y}}$$

$$= \rho_{x,y} + \rho_{x,y}\rho_{y,y} + \rho_{x,y}\rho_{y,y}^2 + \cdots$$

(where $\rho_{y,y} = g_y$).

Lecture 6

September 12

6.1 Stopping Times

Let $T_y = \inf\{n \geq 1 : X_n = y\} \in \{1, 2, \ldots\}$ and $g_y \triangleq \mathbb{P}(T_y < \infty \mid X_0 = y)$. If $g_y < 1$, y is called **transient**. If $g_y = 1$, y is called **recurrent**.

Definition 6.1. A random variable T is called a **random time** if it takes values in $\{0, 1, 2, \ldots\} \cup \{\infty\}$. It is called a **stopping time** of the sequence $\{X_0, X_1, X_2, \ldots\}$ if the event $\{T = n\}$ is determined by $(X_0, \ldots, X_n)$ for all $n = 0, 1, 2, \ldots$.

Example 6.2. Each return time T_y, $y \in \mathcal{S}$, is a stopping time in a DTMC.

Example 6.3. Let $T = \inf\{n \geq 0 : (X_{n-2}, X_{n-1}, X_n) = (a, b, c)\}$ for some $a, b, c \in \mathcal{S}$. T is a stopping time.

Example 6.4. The following is not a stopping time (in general): $V = \inf\{n \geq 1 : X_{n+1} = y\}$. In such cases where whether or not the next state is y is determined by the current state, V would be a stopping time.

6.2 Strong Markov Property

Theorem 6.5 (Strong Markov Property)**.** *One version: Let T be a stopping time with $\mathbb{P}(T < \infty) = 1$. Then,*

$$\mathbb{P}(X_{T+1} = y \mid X_T = x) = p(x, y)$$

(the transition probability). More generally, if A is any event determined by $(X_0, \ldots, X_T)$, then

$$\mathbb{P}(X_{T+1} = y \mid X_T = x, A) = p(x, y).$$

Proof. We will first prove that if T is a stopping time, then for each $n \geq 0$,

$$\mathbb{P}(X_{T+1} = y \mid X_T = x, T = n) = p(x, y).$$

Then,

$$\text{LHS} = \frac{\mathbb{P}(X_{T+1} = y, X_n = x, T = n)}{\mathbb{P}(X_T = x, T = n)}.$$

The numerator is $\sum_{x_0,x_1,\dots,x_{n-1}} \mathbb{P}(X_{n+1} = y, X_n = x, X_{n-1} = x_{n-1}, \dots, X_0 = x_0)$, where the summation is over all $(x_0, x_1, \dots, x_{n-1})$ such that $\{X_0 = x, \dots, X_{n-1} = x_{n-1}, X_n = x\}$ implies $\{T = n\}$. We can then write

$$\begin{aligned}\text{numerator} = &\sum_{(x_0,x_1,\dots,x_{n-1})} \mathbb{P}(X_{n+1} = y \mid X_n = x, X_{n-1} = x_{n-1}, \dots, X_0 = x_0)\\ &\quad \times \mathbb{P}(X_n = x, X_{n+1} = x_{n-1}, \dots, X_0 = x_0)\\ &= p(x,y) \cdot \text{denominator}.\end{aligned}$$

Similarly, one can show $\mathbb{P}(X_{T+1} = y \mid X_T = x, T = n, A) = p(x, y)$ when T is a stopping time and A is determined by $(X_0, \dots, X_T)$. To set $\mathbb{P}(X_{T+1} = y \mid X_T = x) = p(x, y)$ for T a stopping time with $\mathbb{P}(T < \infty) = 1$, write it as

$$\sum_{n=0}^{\infty} \mathbb{P}(X_{T+1} = y, T = n \mid X_T = x) = \sum_{n=0}^{\infty} \underbrace{\mathbb{P}(X_{T+1} = y \mid X_T = x, T = n)}_{p(x,y)} \mathbb{P}(T = n \mid X_T = x). \qquad \square$$

Recall

$$\begin{aligned}\rho_{x,y} &\triangleq \mathbb{P}(T_y < \infty \mid X_0 = x) \qquad (\text{so } g_y = \rho_{y,y}),\\ N(y) &\triangleq \sum_{m=1}^{\infty} \mathbb{1}\{X_m = y\}.\end{aligned}$$

The key tool is

$$\mathbb{E}[N(y) \mid X_0 = x] = \frac{\rho_{x,y}}{1 - \rho_{y,y}}, \tag{6.1}$$

also written as $\mathbb{E}_x[N(y)]$ in the book. To see (6.1), define the stopping times $T_y^{(k+1)} \triangleq \inf\{n > T_y^{(k)} : X_n = y\}$ for $k = 1, 2, \dots,$ with $T_y^{(1)} \triangleq T_y$. Each $T_y^{(k)} \in \{1, 2, \dots\} \cup \{\infty\}$. One can prove using the Strong Markov Property 6.5 that $\mathbb{P}(T_y^{(k)} < \infty \mid X_0 = x) = \rho_{x,y}\rho_{y,y}^{k-1}$, for $k \geq 1$. One can show (do it)

$$\mathbb{E}[N(y) \mid X_0 = x] = \sum_{k=1}^{\infty} \mathbb{P}(T_y^{(k)} < \infty \mid X_0 = x) = \sum_{k=1}^{\infty} \rho_{x,y}\rho_{y,y}^{k-1} = \frac{\rho_{x,y}}{1 - \rho_{y,y}}.$$

Also, $\mathbb{E}[N(y) \mid X_0 = x] = \sum_{n=1}^{\infty} p^{(n)}(x, y)$ (this comes from

$$\mathbb{E}[N(y) \mid X_0 = x] = \mathbb{E}\Big[\sum_{n=1}^{\infty} \mathbb{1}\{X_n = y\} \;\Big|\; X_0 = x\Big],$$

the definition of $N(y)$).

6.3 Classification of States

We say $x \in \mathcal{S}$ **communicates with** $y \in \mathcal{S}$ (written $x \to y$) if $p^{(n)}(x, y) > 0$ for some $n \geq 0$ ($n = 0$ takes care of $x \to x$). We will say x **and** y **communicate with each other** if $x \to y$ and $y \to x$ (we will write $x \leftrightarrow y$).

Fact: $x \leftrightarrow y$ is a binary relation on the set of pairs of states which is reflexive (i.e. $x \leftrightarrow x$), symmetric ($x \leftrightarrow y \implies y \leftrightarrow x$), and transitive ($x \leftrightarrow y$ and $y \leftrightarrow z \implies x \leftrightarrow z$). The equivalence classes of $\mathcal{S}$ (the state space) associated to the equivalence relation $\leftrightarrow$ are called the communicating classes of P (or of the Markov chain with transition probability matrix P).

Theorem 6.6. *Transience and recurrence are* ***class properties*** *(i.e. if $x \in \mathcal{S}$ is recurrent, and $x \leftrightarrow y$, then y is recurrent; the same is true for transient states).*

From

$$\mathbb{E}[N(y) \mid X_0 = x] = \frac{\rho_{x,y}}{1 - \rho_{y,y}}$$

it holds that

$$\mathbb{E}[N(y) \mid X_0 = y] = \frac{\rho_{y,y}}{1 - \rho_{y,y}}.$$

Proof of 6.6. Suppose x is recurrent, i.e., $g_x = 1$, i.e., $\rho_{x,x} = 1$. Equivalently, $\mathbb{E}[N(x) \mid X_0 = x] = \infty$, i.e., $\sum_{n=1}^{\infty} p^{(n)}(x,x) = \infty$. If $x \leftrightarrow y$, then by definition there is some $\ell \geq 0$ and $m \geq 0$ such that $p^{(\ell)}(x,y) > 0$ and $p^{(m)}(y,x) > 0$. By the Markov Property (or by Chapman-Kolmogorov), $p^{(\ell+n+m)}(y,y) \geq p^{(m)}(y,x)p^{(n)}(x,x)p^{(\ell)}(x,y)$. This implies that $\sum_{k=1}^{\infty} p^{(k)}(y,y) = \infty$, so $g_y = 1$, i.e., y is recurrent. □

Theorem 6.7. *Every closed irreducible subset of states has at least one recurrent state. (In fact, such a subset is a communicating class and so all of its states are recurrent.)*

Proof. We know enough to write this proof, see the book. □

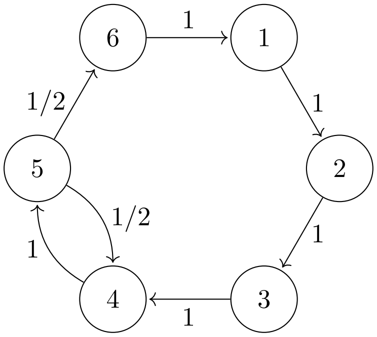
Theorem (Classification): We can write a partition of $\mathcal{S}$ as $\mathcal{S} = T \cup R_1 \cup \cdots \cup R_k$, where $\mathcal{S}$ is the state space, T is the transient states, and $R_1, \ldots, R_k$ are closed irreducible subsets.

To *compute* the probabilities of getting eventually trapped in the various closed irreducible communicating classes starting from an arbitrary initial distribution, first crush all states in each such class to one state. This leaves a Markov chain with $|T| + k$ states. The $|T| \times |T|$ block will be strictly substochastic.

Lecture 7

September 14

7.1 Period of a State in a DTMC (Finite State)



Here:

$$p^{(n)}(1,1) \begin{cases} > 0, & \text{if } n \in \{6, 8, 10, \dots\} \\ = 0, & \text{otherwise} \end{cases}$$

Intuitively, the “period” of state 1 should be 2.

Definition 7.1. Given $x \in \mathcal{S}$, where $\mathcal{S}$ is the state space, consider $\{n > 0 : p^{(n)}(x,x) > 0\}$. Take the greatest common divisor of this set. This is called the **period** of x.

Theorem 7.2. *Given any set of non-negative integers $\{n_1, n_2, n_3, \dots\}$, either finite or countably infinite, $0 \le n_1 < n_2 < n_3 < \cdots$, let d be their GCD. Then, one can find integers $a_1, \dots, a_k$ (not necessarily positive) and $n_{i_1}, \dots, n_{i_k}$ in the given set of integers such that $a_1 n_{i_1} + a_2 n_{i_2} + \cdots + a_k n_{i_k} = d$.*

This is called a Bezout identity. This result is proved using the Euclidean division algorithm. The Bezout identity tells us that $b_1 m_1 + \cdots + b_\ell m_\ell = c_1 r_1 + \cdots + c_p r_p + d$ for some $m_1, \dots, m_\ell, r_1, \dots, r_p$ and positive integer coefficients $b_1, \dots, b_\ell, c_1, \dots, c_p$ with $p^{(m_i)}(x,x) > 0$ for $1 \le i \le \ell$, $p^{(r_j)}(x,x) > 0$ for $1 \le j \le p$, where d is the period of x. So, there is some $s > 0$ such that $p^{(s)}(x,x) > 0$ and $p^{(s+d)}(x,x) > 0$ (take $s = c_1 r_1 + \cdots + c_p r_p$). Hence, $p^{(2s)}(x,x) > 0$, $p^{(2s+d)}(x,x) > 0$, $p^{(2s+2d)}(x,x) > 0$. Finally, we see that for $n \ge s^2$ divisible by d, $p^{(n)}(x,x) > 0$.

7.1.1 Euclidean Division Algorithm to Find the GCD of Two Numbers

$$\gcd(45, 10) = 5$$

$$45 = 4 \cdot 10 + 5 \qquad\qquad 1 \cdot 45 - 4 \cdot 10 = 5$$

$$\begin{aligned}
10 &= 2\cdot\underline{5}+0 \\
\gcd(393,21) &= 3 \\
393 &= 18\cdot 21+15 && 3\cdot 393-56\cdot 21=3 \\
21 &= 1\cdot 15+6 && -2\cdot 21+3\cdot 15=3 \\
15 &= 2\cdot 6+3 && 15-2\cdot 6=3 \\
6 &= 2\cdot\underline{3}+0
\end{aligned}$$

This gives the Bezout identity.

7.1.2 Periodic Markov Chains

Theorem 7.3. *The period is a class property, i.e., if x and y belong to the same communicating class (i.e., $x \leftrightarrow y$), then they have the same period.*

Proof. Suppose that $p^{(k)}(x,y)>0$ and $p^{(\ell)}(x,y)>0$. If $\operatorname{period}(x)=d$, we know $p^{(md)}(x,x)>0$ for all $m\geq m_0$ for some m_0 large enough. Also, d divides $k+\ell$, say $k+\ell=jd$. So, $p^{(md)}(y,y)>0$ for all $m\geq m_0+j$. So, $\operatorname{period}(y)$ divides $\operatorname{period}(x)$ but the same argument can also be run in reverse. □

To understand an irreducible Markov chain with period $d>1$, you should convince yourself that P^d can be written in block diagonal form with d blocks, each block corresponding to an irreducible aperiodic Markov chain.

Example 7.4. For $d=3$, by reordering the states P can be made to look like $\begin{bmatrix} 0 & P^{1,2} & 0 \\ 0 & 0 & P^{2,3} \\ P^{3,1} & 0 & 0 \end{bmatrix}$ so

$$P^3 = \begin{bmatrix} P^{1,2}P^{2,3}P^{3,1} & 0 & 0 \\ 0 & P^{2,3}P^{3,1}P^{1,2} & 0 \\ 0 & 0 & P^{3,1}P^{1,2}P^{2,3} \end{bmatrix}.$$

If

$$\begin{aligned}
\pi_1 P^{1,2} &= \pi_2, \\
\pi_2 P^{2,3} &= \pi_3, \\
\pi_3 P^{3,1} &= \pi_1,
\end{aligned}$$

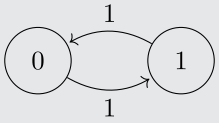
then

$$\begin{aligned}
\pi_1 P^{1,2}P^{2,3}P^{3,1} &= \pi_1 \\
\pi_2 P^{2,3}P^{3,1}P^{1,2} &= \pi_2 \\
\pi_3 P^{3,1}P^{1,2}P^{2,3} &= \pi_3
\end{aligned}$$

and:

$$\begin{aligned}
\begin{bmatrix}\pi_1/3 & \pi_2/3 & \pi_3/3\end{bmatrix} P &= \begin{bmatrix}\pi_1/3 & \pi_2/3 & \pi_3/3\end{bmatrix}, \\
\begin{bmatrix}\pi_1/3 & \mathrm{e}^{2\pi\mathrm{i}/3}\pi_2/3 & \mathrm{e}^{4\pi\mathrm{i}/3}\pi_3/3\end{bmatrix} P &= \mathrm{e}^{4\pi\mathrm{i}/3}\begin{bmatrix}\pi_1/3 & \mathrm{e}^{2\pi\mathrm{i}/3}\pi_2/3 & \mathrm{e}^{4\pi\mathrm{i}/3}\pi_3/3\end{bmatrix}, \\
\begin{bmatrix}\pi_1/3 & \mathrm{e}^{4\pi\mathrm{i}/3}\pi_2/3 & \mathrm{e}^{2\pi\mathrm{i}/3}\pi_3/3\end{bmatrix} P &= \mathrm{e}^{2\pi\mathrm{i}/3}\begin{bmatrix}\pi_1/3 & \mathrm{e}^{4\pi\mathrm{i}/3}\pi_2/3 & \mathrm{e}^{2\pi\mathrm{i}/3}\pi_3/3\end{bmatrix}.
\end{aligned}$$

Example 7.5. For the two-state chain



we have

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad P^2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Example 7.6. For the six-state chain above,

$$P = \begin{array}{c} 1 \\ 3 \\ 5 \\ 2 \\ 4 \\ 6 \end{array} \left[\begin{array}{ccc|ccc} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ \hline 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

7.2 Perron-Frobenius Theorem

To analyze periodicity, we will use the Perron-Frobenius Theorem for irreducible stochastic matrices. We will state this after discussing the Jordan canonical form. Any square matrix $A \in \mathbb{C}^{n\times n}$ can be put in this form. High-level conceptual things to recognize:

- Eigenvectors are not enough. One also needs "generalized eigenvectors" to understand how a matrix acts on vectors.

Example 7.7. Take

$$A \in \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

Then,

$$\det(\lambda I - A) = \det \begin{bmatrix} \lambda - 1 & -1 \\ 0 & \lambda - 1 \end{bmatrix} = \underbrace{(\lambda-1)^2}_{\text{characteristic polynomial}}$$

so there is a single eigenvalue equal to 1 with algebraic multiplicity 2. The **algebraic multiplicity** of an eigenvalue is the number of times it shows up as a root of the characteristic polynomial. Then,

$$\underbrace{\mathcal{N}(I-A)}_{\text{null space of } \lambda I - A \text{ for } \lambda=1} = \mathcal{N}\left(\begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix}\right) = \operatorname{span}\left\{\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right\}.$$

But, $\mathcal{N}((I-A)^2) = \mathbb{R}^2$ because $(I-A)^2$ is the zero matrix. In this example,

$$(I-A)\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 0 \\ x_2 \end{bmatrix} = \begin{bmatrix} -x_2 \\ 0 \end{bmatrix} \in \mathcal{N}(I-A).$$

This is called "**nilpotency**" and $I - A$ is "**nilpotent**".

Theorem 7.8. *Every square matrix $A \in \mathbb{C}^{n\times n}$ can be written as*

$$M^{-1}AM = \textit{block diagonal},$$

where M is a non-singular $n \times n$ matrix (change of basis) and each block diagonal is a matrix which looks like $\begin{bmatrix} \lambda & 1 & 0 & 0 & 0 \\ 0 & \lambda & 1 & 0 & 0 \\ 0 & 0 & \lambda & 1 & 0 \\ 0 & 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & 0 & \lambda \end{bmatrix}$*, where λ is an eigenvalue of A.*

The dimension of the eigenspace of an eigenvalue λ is called its **geometric multiplicity** g_λ. *Always*, $g_\lambda \leq a_\lambda$, where a_λ is the algebraic multiplicity. *Always*: The number of Jordan blocks for λ is g_λ, and the sum of their dimensions is a_λ.

Theorem 7.9 (Perron-Frobenius Theorem)**.** *The Perron-Frobenius Theorem for aperiodic irreducible stochastic matrices P says: 1 is an eigenvalue of algebraic and geometric multiplicity 1, and its eigenspace is the span of a vector with strictly positive entries. All other eigenvalues have absolute value strictly less than 1.*

For an irreducible stochastic matrix of period d, each dth root of unity is an eigenvalue of algebraic and geometric multiplicity 1. All other eigenvalues have absolute value strictly less than 1.

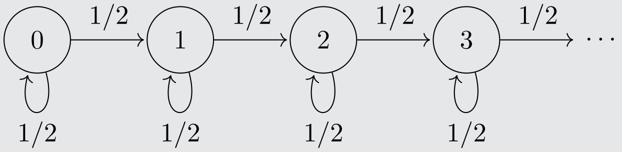
Lecture 8

September 19

8.1 Countable-State Discrete-Time Markov Chains with Infinite State Space

If we enumerate the state space somehow, the transition probabilities are described by $p(i, j)$ as in the finite state case and we can draw a transition diagram similarly.

Example 8.1. Here is an infinite chain.

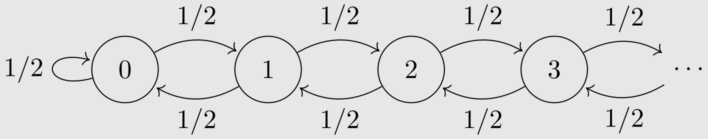


Every state is *transient* (this is impossible for a finite-state chain) because

$$g_y \triangleq \mathbb{P}(T_y < \infty \mid X_0 = y) = \frac{1}{2}$$

for $y = 0, 1, 2, \ldots$. Here, $T_y \triangleq \inf\{n \geq 1 : X_n = y\}$ as before. Each state is a communicating class.

Example 8.2. Another example:



Here, every state is recurrent, i.e.,

$$\mathbb{P}(T_y < \infty \mid X_0 = y) = 1. \tag{8.1}$$

The chain is irreducible, but there is no stationary probability distribution. This can be verified by checking that if π were a stationary distribution, then we would have $\pi(i) = \pi(i + 1)$ for all $i \geq 0$, and this is incompatible with $\pi(i) \geq 0$ for all i and $\sum_{i \in \mathbb{N}} \pi(i) = 1$. This cannot happen in a finite-state chain. Let us check (8.1). We write the first-step equations. Let $\alpha_k = \mathbb{P}(T_0 < \infty \mid X_0 = k)$ for $k = 0, 1, \ldots$.

So, $g_0 \triangleq \mathbb{P}(T_0 < \infty \mid X_0 = 0) = \alpha_0$.

$$\alpha_0 = \frac{1}{2} + \frac{1}{2}\alpha_1,$$
$$\alpha_1 = \frac{1}{2} + \frac{1}{2}\alpha_2.$$

because

$$\begin{aligned}\mathbb{P}(T_0 < \infty \mid X_0 = 1) &= \mathbb{P}(T_0 < \infty, X_1 = 0 \mid X_0 = 1) + \mathbb{P}(T_0 < \infty, X_1 = 2 \mid X_0 = 1)\\ &= \underbrace{\mathbb{P}(T_0 < \infty \mid X_1 = 0, X_0 = 1)}_{1} \cdot \frac{1}{2} + \underbrace{\mathbb{P}(T_0 < \infty \mid X_1 = 2, X_0 = 1)}_{\alpha_2} \cdot \frac{1}{2}.\end{aligned}$$

Also, $\alpha_2 = \alpha_1^2$. So,

$$\alpha_1 = \frac{1}{2} + \frac{1}{2}\alpha_1^2 \implies \alpha_1 = 1,$$

and so $\alpha_0 = 1$. Since $\alpha_0 = 1$, state 0 is recurrent. Then, the same logic as for finite-state chains shows that every state is recurrent. Formally,

$$\begin{aligned}\mathbb{P}(T_0 < \infty \mid X_0 = 2) &= \mathbb{P}(T_0 < \infty, T_1 < \infty \mid X_0 = 2)\\ &= \mathbb{P}(T_1 < \infty \mid X_0 = 2)\mathbb{P}(T_0 < \infty \mid T_1 < \infty, X_0 = 2)\\ &= \mathbb{P}(T_1 < \infty \mid X_0 = 2)\mathbb{P}(T_0 < \infty \mid X_0 = 1)\end{aligned}$$

and also $\mathbb{P}(T_1 < \infty \mid X_0 = 2) = \mathbb{P}(T_0 < \infty \mid X_0 = 1)$.

In 8.2, the *reason* why it is possible to have every state recurrent but no stationary distribution is that the expected time to return from a state to itself is ∞ (for each $y = 0, 1, 2, \dots$), i.e., $\mathbb{E}[T_y \mid X_0 = y] = \infty$. In an irreducible finite-state Markov chain (so every state is recurrent), we must have $\mathbb{E}[T_y \mid X_0 = y] < \infty$. See the first part of Handout 6. It gives a procedure for each subset A of the state space $\mathcal{S}$ of a finite-state DTMC for computing $(k_i^A,\ i \in \mathcal{S})$, where $k_i^A = \mathbb{E}[\text{time to hit } A \mid X_0 = i]$. Apply this procedure to a modified MC where y is replaced by y_{initial} and y_{final}. Each $p(x, y)$ becomes $p(x, y_{\text{final}})$, each $p(y, z)$ becomes $p(y_{\text{initial}}, z)$, and $p(y_{\text{final}}, y_{\text{final}}) = 1$. Take $A = \{y_{\text{final}}\}$ and compute $k^{y_{\text{final}}}$.

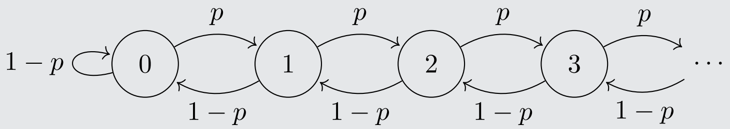
Let $m_k = \mathbb{E}[T_0 \mid X_0 = k]$, where $T_y = \inf\{n \geq 1 : X_n = y\}$. We want to show $m_0 = \infty$.

$$\begin{aligned}m_0 &= \frac{1}{2} + \frac{1}{2}(1 + m_1)\\ m_1 &= \frac{1}{2} + \frac{1}{2}(1 + m_2)\\ m_2 &= 2m_1\end{aligned}$$

The mean time to get to 0 starting in state 2 is the mean time to get to 1 starting in state 2, plus the mean time to get to 0 starting in state 1. Thus, we have the equation $m_1 = 1 + m_1$, so $m_1 = \infty$. Then,

$$m_0 = 1 + \frac{1}{2}m_1 = \infty.$$

Example 8.3. Third example:



with $0 < p < 1/2$. You will find $g_y = 1$ for each y (recurrent) and $\mathbb{E}[T_y < \infty \mid X_0 = y] < \infty$.

Definition 8.4. We say a state $y \in \mathcal{S}$, where $\mathcal{S}$ is the state space, is **positive-recurrent** if it is recurrent (i.e., $\mathbb{P}(T_y < \infty \mid X_0 = y) = 1$) and if $\mathbb{E}[T_y \mid X_0 = y] < \infty$. We say $y \in \mathcal{S}$ is **null-recurrent** if it is recurrent but $\mathbb{E}[T_y \mid X_0 = y] = \infty$.

We will prove that positive-recurrence and null-recurrence are class properties (there will be a handout). (The proof that recurrence is a class property is the same as for finite-state chains.) So, we will just consider positive-recurrence. Let $f_{i,j}^{(n)} \triangleq \mathbb{P}(T_j = n \mid X_0 = i)$, for $n = 1, 2, \dots$, where $T_j = \inf\{n \geq 1 : X_n = j\}$. Define $F_{i,j}(s) \triangleq \sum_{n=1}^{\infty} f_{i,j}^{(n)} s^n$, which is well-defined for $|s| < 1$. Let $P_{i,j}(s) \triangleq \sum_{n=0}^{\infty} p_{i,j}^{(n)} s^n$. One can show

$$\begin{aligned} P_{i,i}(s) &= 1 + F_{i,i}(s) P_{i,i}(s), \\ P_{i,j}(s) &= F_{i,j}(s) P_{j,j}(s), \qquad \text{if } j \neq i. \end{aligned}$$

The first equation says $p_{i,i}^{(0)} = 1$ and $p_{i,i}^{(n)} = \sum_{k=1}^{n} f_{i,i}^{(k)} p_{i,i}^{(n-k)}$. The second equation says

$$p_{i,j}^{(n)} = \sum_{k=1}^{n} f_{i,j}^{(k)} p_{j,j}^{(n-k)}.$$

It turns out

$$\mathbb{E}[T_y \mid X_0 = j] = F'_{j,j}(1),$$

where the expression on the right is the derivative from the left. It turns out that j is positive-recurrent iff there is a positive constant $K < \infty$ such that $P'_{j,j}(s) \leq K(P_{j,j}(s))^2$ for all s sufficiently close to 1.

Lecture 9

September 21

9.1 Stationary Distributions in Countable-State Markov Chains

Handout 9 shows that for an irreducible recurrent countable-state DTMC, given any $y \in \mathcal{S}$ (where $\mathcal{S}$ is the state space), because y is recurrent, $\mathbb{P}(T_y < \infty \mid X_0 = y) = 1$. We can consider

$$\mu(x) = \sum_{n \geq 1} \mathbb{P}(X_n = x, T_y \geq n \mid X_0 = y).$$

The handout shows that $\sum_{z \in \mathcal{S}} \mu(z) p(z, x) = \mu(x)$ for all $x \in \mathcal{S}$ and $\sum_{x \in \mathcal{S}} \mu(x) = \mathbb{E}[T_y \mid X_0 = y]$ (which is finite only if y is positive-recurrent). The general picture for a countable-state DTMC will have recurrent communicating classes that are closed and either finite and positive-recurrent, infinite and positive-recurrent, or infinite and null-recurrent. The transient classes can lead to other transient classes or to recurrent classes. The chain can "go to ∞" purely via transient classes. In fact, there are many ways of going to ∞.

For finite-state irreducible Markov chains we saw that there is a unique stationary distribution, call it π, and $p_{x,y}^{(n)} \to \pi(y)$ as $n \to \infty$ if the chain is aperiodic (for all $x, y \in \mathcal{S}$). If the chain is periodic,

$$\frac{1}{n} \sum_{k=1}^{n} p^{(k)}(x, y) \to \pi(y),$$

as $n \to \infty$, for all $x, y \in \mathcal{S}$. For example, when

$$P = \begin{bmatrix} 0 & P_{1,2} \\ P_{2,1} & 0 \end{bmatrix}$$

then there is a stationary distribution

$$\begin{bmatrix} \pi_1 & \pi_2 \end{bmatrix} \begin{bmatrix} 0 & P_{1,2} \\ P_{2,1} & 0 \end{bmatrix} = \begin{bmatrix} \pi_1 & \pi_2 \end{bmatrix}.$$

Similar statements are true for countable-state irreducible positive-recurrent DTMCs, i.e.:

1. there is a unique stationary distribution, call it π;
2. if the chain is aperiodic, $\sum_{x \in \mathcal{S}} p_0(x) p^{(n)}(x, y) \to \pi(y)$ as $n \to \infty$ for all initial distributions p_0;
3. in the periodic case,

$$\frac{1}{n} \sum_{x \in \mathcal{S}} \sum_{k=1}^{n} p_0(x) p^{(k)}(x, y) \to \pi(y)$$

as $n \to \infty$, for all p_0.

This is proved in Handout 9.

9.2 Coupling

2 is proved in the book using the idea of coupling. Coupling is a technique for proving statements about probabilities from pointwise statements.

General Idea: Suppose X and Y are real-valued random variables about whose *distributions* I want to prove something. I will construct $\tilde{X}$ and $\tilde{Y}$ on the same sample space (i.e., so that I can talk about their joint distribution) in such a way such that $\tilde{X} \overset{\mathsf{d}}{=} X$ and $\tilde{Y} \overset{\mathsf{d}}{=} Y$ (i.e., $\tilde{X}$ and X have the same distribution) and then prove a pointwise statement about $(\tilde{X}, \tilde{Y})$.

Example 9.1. Suppose X and Y are (non-negative integer-valued) geometric random variables,

$$\mathbb{P}(X = k) = (1-a)a^k, \qquad k = 0, 1, 2, \dots$$
$$\mathbb{P}(Y = k) = (1-b)b^k, \qquad k = 0, 1, 2, \dots$$

where $0 < a < 1$ and $0 < b < 1$. Assume $a < b$. I want to prove that for any increasing function $f : \{0, 1, 2, \dots\} \to \mathbb{R}$, $\mathbb{E}[f(X)] \leq \mathbb{E}[f(Y)]$. The direct approach is to show that for all increasing f as above, $\sum_{k=0}^{\infty} f(k)(1-a)a^k \leq \sum_{k=0}^{\infty} f(k)(1-b)b^k$. Is this even true? After all, $a < b$ then $1 - a > 1 - b$. However, $f(k) = f(0) + \sum_{\ell=1}^{k} g(\ell)$ where $g(\ell) = f(\ell) - f(\ell - 1) \geq 0$ for all ℓ. Then

$$\begin{aligned}
\text{LHS} &= f(0) + \sum_{k=0}^{\infty} \sum_{\ell=1}^{k} g(\ell)(1-a)a^k \\
&= f(0) + \sum_{\ell=1}^{\infty} g(\ell) \sum_{k=\ell-1}^{\infty} (1-a)a^k \\
&= f(0) + \sum_{\ell=1}^{\infty} g(\ell) a^{\ell-1} \\
&\leq f(0) + \sum_{\ell=1}^{\infty} g(\ell) b^{\ell-1} \\
&= \text{RHS}.
\end{aligned}$$

A coupling proof can be given instead by constructing a jointly defined pair of random variables $(\tilde{X}, \tilde{Y})$ such that $\tilde{X} \sim \text{Geometric}(a)$, $\tilde{Y} \sim \text{Geometric}(b)$, and $\tilde{X} \leq \tilde{Y}$ pointwise. This implies that

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

because $X \overset{\mathsf{d}}{=} \tilde{X}$, because $\tilde{X} \leq \tilde{Y}$ and f is increasing, and because $Y \overset{\mathsf{d}}{=} \tilde{Y}$.

Consider two partitions of $[0, 1]$, one with heights $1 - a$, $1 - a + (1-a)a$, $1 - a + (1-a)a + (1-a)a^2$, ... and the other with heights $1 - b$, $1 - b + (1-b)b$, $1 - b + (1-b)b + (1-b)b^2$, ... Then, pick a point uniformly at random on the interval. This defines the coupled pair $(\tilde{X}, \tilde{Y})$.

Theorem 9.2. *Given an irreducible aperiodic positive-recurrent DTMC (countable-state) with (unique) stationary distribution π, then for all $x, y \in \mathcal{S}$, $p^{(n)}(x, y) \to \pi(y)$ as $n \to \infty$.*

Proof by Coupling. We will construct a Markov chain on $\mathcal{S} \times \mathcal{S}$, call it $((\hat{X}_n, \hat{Y}_n),\ n \geq 0)$, initialized in

$$\hat{p}_0(x', y') = \begin{cases} 0, & \text{if } x' \neq x \\ \pi(y), & \text{if } x' = x \end{cases}$$

with transition probabilities

$$\begin{aligned} \hat{p}\big((x_1, y_1), (x_2, y_2)\big) &= p(x_1, x_2)p(y_1, y_2) \qquad \text{if } x_1 \neq y_1, \\ \hat{p}\big((x, x), (x', x')\big) &= p(x, x'), \end{aligned}$$

i.e., once we get to the diagonal, we stay put. What we need to show is that if $V = \inf\{n \geq 0 : \hat{X}_n = \hat{Y}_n\}$, then $\mathbb{P}(V < \infty) = 1$, where the last expression is computed under the dynamics of the coupled Markov chain. The book proves that $\mathbb{P}(V < \infty) = 1$ by considering the chain on $\mathcal{S} \times \mathcal{S}$ with transition probabilities $\tilde{p}((x_1, y_1), (x_2, y_2)) = p(x_1, x_2)p(y_1, y_2)$.

Given the original chain is positive-recurrent, $\pi(x) > 0$ for all $x \in \mathcal{S}$. This implies $\pi(x_1)\pi(y_1) > 0$ for all $(x_1, y_1) \in \mathcal{S} \times \mathcal{S}$. Also note that $(x_1, y_1) \mapsto \pi(x_1)\pi(y_1)$ is the stationary distribution for the product chain defined by $\tilde{p}$. This implies that each state (x, y) in the $\tilde{p}$-chain is positive-recurrent. □

In fact, for every $z \in \mathcal{S}$, if $V_{(z,z)} \triangleq \inf\{n \geq 0 : \hat{X}_n = z, \hat{Y}_n = z\}$, then $\mathbb{P}(V_{(z,z)} < \infty) = 1$.

Lecture 10

September 26

10.1 Simple Random Walk in $\mathbb{Z}^d$

For $d = 1$, the transitions are

$$p(i, i+1) = \frac{1}{2}$$
$$p(i, i-1) = \frac{1}{2}$$

for all $i \in \mathbb{Z}$. For $d = 2$,

$$p\big((i,j),(i+u, j+v)\big) = \frac{1}{4}$$

for each $(u,v) \in \{\pm 1\} \times \{\pm 1\}$. Similarly in dimension d, the state space is $\mathbb{Z}^d$, and

$$p\big((i_1, \dots, i_d),(i_1+u_1, \dots, i_d+u_d)\big) = \frac{1}{2^d}$$

for each $(u_1, \dots, u_d) \in \{\pm 1\}^d$. For each $d \geq 1$, the SRW is an irreducible DTMC with period 2. If there were a stationary probability distribution, it would give equal probability to all states which is impossible. So, positive recurrence is ruled out.

Recall:

$$\begin{aligned} T_y &= \inf\{n \geq 1 : X_n = y\}, \\ g_y &\triangleq \mathbb{P}(T_y < \infty \mid X_0 = y), \\ \rho_{x,y} &\triangleq \mathbb{P}(T_y < \infty \mid X_0 = x), \\ N_y &= \sum_{n=1}^{\infty} \mathbb{1}\{X_n = y\}, \\ \mathbb{E}[N_y \mid X_0 = x] &= \frac{\rho_{x,y}}{1-\rho_{y,y}} = \rho_{x,y} + \rho_{x,y}\rho_{y,y} + \rho_{x,y}\rho_{y,y}^2 + \cdots . \end{aligned}$$

The last formula tells us that $g_y = 1 \iff \mathbb{E}[N_y \mid X_0 = y] = \infty$. Also, $\mathbb{E}[N_y \mid X_0 = x] = \sum_{n=1}^{\infty} p^{(n)}(x,y)$.

Consider SRW on $\mathbb{Z}$ (i.e., $d = 1$).

$$p^{(n)}(0,0) = \begin{cases} 0, & \text{if } n \text{ is odd} \\ \dbinom{2m}{m}\dfrac{1}{2^{2m}}, & \text{if } n = 2m \end{cases}$$

A basic fact well worth knowing is **Stirling's approximation** for the factorial.

$$n! = \left(\frac{n}{\mathrm{e}}\right)^n \sqrt{2\pi n}\mathrm{e}^{\alpha_n}$$

for some

$$\frac{1}{12n+1} \leq \alpha_n \leq \frac{1}{12n}.$$

Some intuition for the main part of this formula comes from

$$\begin{aligned}\log n! &= \log 1 + \log 2 + \cdots + \log n \\ &\underbrace{\approx}_{\text{approximately}} \int_1^n \log x \,\mathrm{d}x \\ &= x\log x - x\Big|_1^n \\ &\approx n \log\frac{n}{\mathrm{e}} = \log\left(\frac{n}{e}\right)^n.\end{aligned}$$

So,

$$\binom{2m}{m}\frac{1}{2^{2m}} = \frac{2m!}{m!m!}\frac{1}{2^{2m}} = \frac{(2m/\mathrm{e})^{2m}\sqrt{4\pi m}\mathrm{e}^{\alpha_{2m}}}{(m/\mathrm{e})^m(m/\mathrm{e})^m(2\pi m)\mathrm{e}^{2\alpha_m}}\frac{1}{2^{2m}} = \frac{1}{\sqrt{\pi m}}\mathrm{e}^{\alpha_{2m}-2\alpha_m}.$$

So, for $d = 1$,

$$\sum_{n=1}^{\infty} p^{(n)}(0,0) = \sum_{m=1}^{\infty}\frac{1}{\sqrt{\pi m}}\mathrm{e}^{\alpha_{2m}-2\alpha_m} = \infty.$$

For $d = 2$,

$$\sum_{n=1}^{\infty} p^{(n)}\big((0,0),(0,0)\big) = \sum_{m=1}^{\infty}\frac{1}{\pi m}\mathrm{e}^{2\alpha_{2m}-4\alpha_m} = \infty.$$

For $d \geq 3$,

$$\sum_{n=1}^{\infty} p^{(n)}\big((\underbrace{0,\ldots,0}_{d}),(\underbrace{0,\ldots,0}_{d})\big) \leq C\sum_{m=1}^{\infty}\frac{1}{m^{d/2}} < \infty.$$

10.2 Branching Processes

Another important example of countable-state DTMC examples are **branching processes**.

To define the branching process we will define an *array* of i.i.d. non-negative integer-valued random variables.

$$\begin{matrix} Y_{0,1} & Y_{0,2} & Y_{0,3} & \cdots \\ Y_{1,1} & Y_{1,2} & Y_{1,3} & \cdots \\ Y_{2,1} & Y_{2,2} & Y_{3,3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{matrix}$$

We will also have X_0 non-negative integer-valued, independent of all the $Y_{i,j}$. X_0 denotes the initial number of people alive.

$$\begin{aligned}X_1 &= Y_{0,1} + Y_{0,2} + \cdots + Y_{0,X_0} \\ X_2 &= Y_{1,1} + Y_{1,2} + \cdots + Y_{1,X_1}\end{aligned}$$

$$\vdots$$

$$X_{n+1} = Y_{n,1} + Y_{n,2} + \cdots + Y_{n,X_n}$$

(If $X_n = 0$, then $X_{n+1} = 0$.) $Y_{n,j}$ represents the number of children produced by the jth person alive at time $n \geq 0$, $j \geq 1$, and X_n is the number of people alive at time n. Let $\mu \triangleq \mathbb{E}[Y_{n,j}]$, which is the same for all $n \geq 0$, $j \geq 1$ (this is the mean number of progeny in one step).

$$\begin{aligned}
\mathbb{E}[X_{n+1}] &= \sum_{k=0}^{\infty} \mathbb{E}[X_{n+1} \mid X_n = k]\mathbb{P}(X_n = k) \\
&= \sum_{k=0}^{\infty} \mathbb{E}\Big[\sum_{j=1}^{k} Y_{n,j} \;\Big|\; X_n = k\Big]\mathbb{P}(X_n = k) \\
&= \sum_{k=0}^{\infty} \mu k \mathbb{P}(X_n = k) = \mu\, \mathbb{E}[X_n].
\end{aligned}$$

This already tells us that if $\mu < 1$ (and $\mathbb{E}[X_0] < \infty$), then $\mathbb{E}[X_n] = \mu^n\, \mathbb{E}[X_0] \to 0$ as $n \to \infty$. This implies $\mathbb{P}(X_n \geq 1) \leq \mathbb{E}[X_n] \to 0$ as $n \to \infty$. So, because X_n is non-negative integer-valued, $\mathbb{P}(X_n = 0) \to 1$ as $n \to \infty$. This is because the LHS of the inequality is $\mathbb{P}(X_n = 1) + \mathbb{P}(X_n = 2) + \mathbb{P}(X_n = 3) + \cdots$ and the RHS of the inequality is $\mathbb{P}(X_n = 1) + 2\mathbb{P}(X_n = 2) + 3\mathbb{P}(X_n = 3) + \cdots$ and $\mathbb{P}(X_n = 0) = 1 - \mathbb{P}(X_n \geq 1)$. Also if $\mu > 1$, if $\mathbb{E}[X_0] > 0$, then $\mathbb{E}[X_n] = \mu^n\, \mathbb{E}[X_0] \to \infty$ as $n \to \infty$.

- $\boxed{\mu < 1}$: **subcritical**
- $\boxed{\mu > 1}$: **supercritical**
- $\boxed{\mu = 1}$: **critical**

Even though $\mathbb{E}[X_n] \to \infty$ when $\mu > 1$ (where $\mathbb{E}[X_0] > 0$), we might have $\lim_{n\to\infty} \mathbb{P}(X_n = 0) > 0$. The limit exists since $\mathbb{P}(X_{n+1} = 0) \geq \mathbb{P}(X_n = 0)$. This limit is called the **extinction probability**. We can in fact figure out the extinction probability because of a beautiful recursion for the generating function of X_n. Define $G_n(s) \triangleq \mathbb{E}[s^{X_n}]$ where s is a dummy variable or a complex variable of absolute value less than 1. Note that the generating function can be used to compute the moments, e.g.

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2} G_n(s)\Big|_{s=1} = \mathbb{E}[X(X-1)].$$

Let $G(s)$ denote $\mathbb{E}[s^{Y_{n,j}}]$ (which is the same for any $n \geq 0$, $j \geq 1$). Then,

$$\begin{aligned}
G_{n+1}(s) = \mathbb{E}[s^{X_{n+1}}] &= \sum_{k=0}^{\infty} \mathbb{E}[s^{X_{n+1}} \mid X_n = k]\mathbb{P}(X_n = k) \\
&= \sum_{k=0}^{\infty} \mathbb{E}[s^{Y_{n,1}+\cdots+Y_{n,k}} \mid X_n = k]\mathbb{P}(X_n = k) \\
&= \sum_{k=0}^{\infty} \mathbb{E}[s^{Y}]^k \mathbb{P}(X_n = k) \\
&= \sum_{k=0}^{\infty} G(s)^k \mathbb{P}(X_n = k) \\
&= G_n\big(G(s)\big).
\end{aligned}$$

This uses:

- We can drop the conditioning.

- $Y_{n,1}, \dots, Y_{n,k}$ are i.i.d.

So,

$$G_{n+1}(s) = G_0\big(\underbrace{G(G(\cdots G(G(s)))))}_{n+1 \text{ times}}\big)$$

where $G_0(s) = \mathbb{E}[s^{X_0}]$. To make life easy, assume that $X_0 = 1$ with probability 1, so $G_0(s) = s$. So,

$$G_{n+1}(s) = G^{(n+1)}(s) = G\big(G \circ \cdots \circ G(s)\big) = G\big(G_n(s)\big),$$

where $G^{(n)}$ denotes G concatenated with itself n times. Also, $\mathbb{P}(X_n = 0) = G_n(0)$ because

$$G_n(0) = \mathbb{P}(X_n = 0) + s\mathbb{P}(X_n = 1) + s^2\mathbb{P}(X_n = 2) + \cdots .$$

So, if $\alpha \triangleq \lim_{n\to\infty} \mathbb{P}(X_n = 0)$, we learn that $\alpha = G(\alpha)$ (from setting $s = 0$ and letting $n \to \infty$ in $G_{n+1}(s) = G(G_n(s)))$.

Example 10.1. Say:

$$\begin{aligned}
\mathbb{P}(Y_{n,j} = 0) &= u_0\\
\mathbb{P}(Y_{n,j} = 1) &= u_1\\
\mathbb{P}(Y_{n,j} = 2) &= u_2\\
\mathbb{P}(Y_{n,j} = 3) &= u_3
\end{aligned}$$

Then, $G(s) = u_0 + u_1 s + u_2 s^2 + u_3 s^3$.

In general,

$$G'(s) = \frac{\mathrm{d}}{\mathrm{d}s}\,\mathbb{E}[s^Y] = \mathbb{E}[Y s^{Y-1}]$$

where $Y \overset{\mathsf{d}}{=} Y_{n,j}$.

$$\begin{aligned}
G(s) &= \sum_{k=0}^{\infty} \mathbb{P}(Y = k)s^k,\\
G'(s) &= \sum_{k=0}^{\infty} k\mathbb{P}(Y = k)s^{k-1}\\
&\ge 0 \qquad \text{for } s \in [0,1],\\
G''(s) &= \sum_{k=0}^{\infty} k(k-1)\mathbb{P}(Y = k)s^{k-2}\\
&\ge 0 \qquad \text{for } s \in [0,1].
\end{aligned}$$

Also, $G'(1) = \mu$.

Conclusion: $\alpha < 1$ when $\mu > 1$ and given by the lower fixed point in the intersection of the diagonal and the function G.

Lecture 11

September 28

11.1 Poisson Process

A **continuous-time stochastic process** is a family of random variables $(X(t),\ t \in \mathbb{R})$ or $(X(t),\ t \geq 0)$. There are two viewpoints you could have.

- For each t, $X(t)$ is a random variable.
- The entire function $(t \mapsto X(t))$ is a random function of time. This is called the **sample path**.

$$(\Omega, \mathcal{F}, \mathbb{P}) \xrightarrow{X(t)} \mathbb{R}$$

One should know that specifying the finite-dimensional distributions still leaves some freedom in creating a stochastic process in the sense that the sample path behavior is not completely specified. One often prescribes the sample paths and checks that the FDDs are correct.

A **Poisson process** $(N(t),\ t \geq 0)$ is a non-negative integer-valued process whose sample paths (with probability 1) increase by jumps of size 1 and which has the FDDs given by the properties:

1. $\mathbb{P}(N(0) = 0) = 1$;
2. $$\mathbb{P}\big(N(t) - N(s) = k\big) = \frac{(\lambda(t-s))^k}{k!}\mathrm{e}^{-\lambda(t-s)}, \qquad k \geq 0,\ 0 \leq s < t$$
   (here $\lambda \geq 0$);
3. independent increments.

Let S_k be the kth arrival time for $k \geq 1$, $S_0 = 0$, and T_m is the mth interarrival time, $m \geq 1$. $N(t)$ is the number of arrivals up to time $t \geq 0$, so $N(t) = \sum_{k=1}^{\infty} \mathbb{1}\{S_k \leq t\}$. For $k \geq 1$,

$$\begin{aligned} S_k &= \sup\{s > 0 : N(s) < k\} \\ &= \inf\{s > 0 : N(s) \geq k\}. \end{aligned}$$

It turns out that, by virtue of requiring that $(N(t),\ t \geq 0)$ is a Poisson process of *rate* λ, $T_1, T_2, T_3, \ldots$ are independent and exponentially distributed with parameter λ, i.e., $\mathbb{P}(T_m > t) = \mathrm{e}^{-\lambda t}$ for all $t > 0$.

$$\begin{aligned} \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda} &= \lambda \sum_{l=0}^{\infty} \frac{\lambda^l}{l!}\mathrm{e}^{-\lambda} \\ &= \lambda, \\ \mathbb{E}[T_m] &= \int_0^{\infty} t\lambda \mathrm{e}^{-\lambda t}\,\mathrm{d}t = \int_0^{\infty} \mathrm{e}^{-\lambda t}\,\mathrm{d}t = \frac{1}{\lambda}. \end{aligned}$$

Another way to compute this is to compute the characteristic function

$$\Phi_{T_m}(\theta) = \mathbb{E}[\mathrm{e}^{\mathrm{i}\theta T_m}] = \mathrm{i}\int_0^\infty \mathrm{e}^{\mathrm{i}\theta t}\lambda \mathrm{e}^{-\lambda t}\,\mathrm{d}t = \frac{\lambda}{\lambda - \mathrm{i}\theta},$$

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\Phi_{I_m}(\theta) = \frac{\mathrm{i}\lambda}{(\lambda-\mathrm{i}\theta)^2} = \frac{\mathrm{i}}{\lambda} \qquad \text{at } \theta = 0.$$

To check the two views of the Poisson process, consider first T_1.

$$\mathbb{P}(T_1 > t) = \mathbb{P}\big(N(t) = 0\big) = \frac{(\lambda t)^0}{0!}\mathrm{e}^{-\lambda t} = \mathrm{e}^{-\lambda t}.$$

For the more general case, recall that the sum of n i.i.d. exponentials, Exponential(λ), has a *gamma* distribution, $\Gamma(n,\lambda)$. The density of this is $\dfrac{\lambda^n t^{n-1}}{(n-1)!}\mathrm{e}^{-\lambda t}$. To verify this, note that it is true for $n = 1$. For $n = 2$, we have to compute the convolution of two exponential densities.

$$\int_0^t \lambda \mathrm{e}^{-\lambda s}\lambda \mathrm{e}^{-\lambda(t-s)}\,\mathrm{d}s = \lambda^2 \mathrm{e}^{-\lambda t}\int_0^t \mathrm{d}s = \lambda^2 t\mathrm{e}^{-\lambda t}$$

which is consistent with the general formula. By induction,

$$\int_0^t \lambda \mathrm{e}^{-\lambda(t-s)}\frac{\lambda^k s^{k-1}}{(k-1)!}\mathrm{e}^{-\lambda s}\,\mathrm{d}s = \frac{\lambda^{k+1}}{(k-1)!}\mathrm{e}^{-\lambda t}\int_0^t s^{k-1}\,\mathrm{d}s = \frac{\lambda^{k+1}t^k\mathrm{e}^{-\lambda t}}{k!}.$$

Formally, note

$$\mathbb{P}\big(N(t) < n\big) = \sum_{k=0}^{n-1}\frac{(\lambda t)^k}{k!}\mathrm{e}^{-\lambda t}.$$

We have to check that this equals the probability that a $\Gamma(n,\lambda)$ random variable exceeds t.

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\Big(\mathrm{e}^{-\lambda t} + \lambda t\mathrm{e}^{-\lambda t} + \frac{(\lambda t)^2}{2!}\mathrm{e}^{-\lambda t} + \cdots + \frac{(\lambda t)^{n-1}}{(n-1)!}\mathrm{e}^{-\lambda t}\Big)\\ &\quad = -\lambda \mathrm{e}^{-\lambda t} + \lambda \mathrm{e}^{-\lambda t} - \lambda^2 t\mathrm{e}^{-\lambda t} + \lambda^2 t\mathrm{e}^{-\lambda t} - \frac{\lambda^3 t^2}{2!}\mathrm{e}^{-\lambda t} + \cdots - \frac{\lambda^n t^{n-1}}{(n-1)!}\mathrm{e}^{-\lambda t}\end{aligned}$$

and the last term is the only one which survives.

11.1.1 Memorylessness Property for the Exponential Distribution

Recall the **memorylessness property** of an exponential random variable: $\mathbb{P}(T > t+s \mid T > s) = \mathbb{P}(T > t)$ when $T \sim$ Exponential(λ).

Proof:

$$\text{LHS} = \frac{\mathbb{P}(T > t+s, T > s)}{\mathbb{P}(T > s)} = \frac{\mathbb{P}(T > t+s)}{\mathbb{P}(T > s)} = \frac{\mathrm{e}^{-\lambda(t+s)}}{\mathrm{e}^{-\lambda s}} = \mathrm{e}^{-\lambda t} = \mathbb{P}(T > t).$$

11.1.2 Properties of the Poisson Process

Let $A_1 < A_2 < \cdots < A_k$ be the arrival times in (s,t) of the points in (s,t).

$$\mathbb{P}\big(A_1 \in (t_1, t_1 + \mathrm{d}t_1), \ldots, A_k \in (t_k, t_k + \mathrm{d}t_k) \mid N(t) - N(s) = k\big)$$

has the distribution of k uniform order statistics on the interval (s,t).

Theorem 11.1 (Superposition)**.** *If $(N_1(t),\ t \geq 0)$ is a Poisson process of rate λ_1, and $(N_2(t),\ t \geq 0)$ is a Poisson process of rate λ_2 and $(N_1(t),\ t \geq 0) \perp\!\!\!\perp (N_2(t),\ t \geq 0)$, let $N(t) \triangleq N_1(t) + N_2(t)$. Then, $(N(t),\ t \geq 0)$ is a Poisson process of rate $\lambda_1 + \lambda_2$.*

Proof. 1. $\mathbb{P}(N(0) = 0) = \mathbb{P}(N_1(0) = 0, N_2(0) = 0) = 1$.

2. For $0 \leq s \leq t$,

$$\begin{aligned}
\mathbb{P}\big(N(t) - N(s) = k\big) &= \sum_{l=0}^{k} \mathbb{P}\big(N_1(t) - N_1(s) = l, N_2(t) - N_2(s) = k - l\big) \\
&= \sum_{l=0}^{k} \frac{(\lambda_1(t-s))^l}{l!} \mathrm{e}^{-\lambda_1(t-s)} \frac{(\lambda_2(t-s))^{k-l}}{(k-l)!} \mathrm{e}^{-\lambda_2(t-s)} \\
&= \frac{1}{k!} \sum_{l=0}^{k} \binom{k}{l} \big(\lambda_1(t-s)\big)^l \big(\lambda_2(t-s)\big)^{k-l} \mathrm{e}^{-(\lambda_1+\lambda_2)(t-s)} \\
&= \frac{((\lambda_1 + \lambda_2)(t-s))^k}{k!} \mathrm{e}^{-(\lambda_1+\lambda_2)(t-s)}.
\end{aligned}$$

3. N has independent increments because

$$\begin{aligned}
N(t_1) - N(s_1) &= N_1(t_1) - N_1(s_1) + N_2(t_1) - N_2(s_1), \\
N(t_2) - N(s_2) &= N_1(t_2) - N_1(s_2) + N_2(t_2) - N_2(s_2).
\end{aligned}$$

□

This is consistent with the interarrival time viewpoint.

Example 11.2. If $T_1^{(1)} \sim \text{Exponential}(\lambda_1)$ and $T_1^{(2)} \sim \text{Exponential}(\lambda_2)$, $T_1^{(1)} \perp\!\!\!\perp T_1^{(2)}$, then

$$\min(T_1^{(1)}, T_1^{(2)}) \sim \text{Exponential}(\lambda_1 + \lambda_2).$$

Proof.

$$\begin{aligned}
\mathbb{P}\big(\min(T_1^{(1)}, T_1^{(2)}) > t\big) &= \mathbb{P}(T_1^{(1)} > t, T_1^{(2)} > t) \\
&= \mathbb{P}(T_1^{(1)} > t)\mathbb{P}(T_1^{(2)} > t) \\
&= \mathrm{e}^{-\lambda_1 t}\mathrm{e}^{-\lambda_2 t} = \mathrm{e}^{-(\lambda_1+\lambda_2)t}.
\end{aligned}$$

We can also show

$$\mathbb{P}(T_1^{(1)} < T_1^{(2)}) = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

and also if $I \in \{1, 2\}$ is defined by $I = 1$ iff $T_1^{(1)} < T_1^{(2)}$, $I = 2$ iff $T_1^{(1)} > T_1^{(2)}$, then

$$\mathbb{P}\big(I = 1 \bigm| \min(T_1^{(1)}, T_1^{(2)}) \in (t, t + \mathrm{d}t)\big) = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

because

$$\begin{aligned}
\frac{\mathbb{P}(I = 1, \min(T_1^{(1)}, T_1^{(2)}) \in (t, t+\mathrm{d}t))}{\mathbb{P}(\min(T_1^{(1)}, T_1^{(2)}) \in (t, t+\mathrm{d}t))} &= \frac{\mathbb{P}(T_1^{(1)} \in (t, t+\mathrm{d}t), T_1^{(2)} > t)}{\mathbb{P}(\min(T_1^{(1)}, T_1^{(2)}) \in (t, t+\mathrm{d}t))} = \frac{(\lambda_1 \mathrm{e}^{-\lambda_1 t}\,\mathrm{d}t)\mathrm{e}^{-\lambda_2 t}}{(\lambda_1 + \lambda_2)\mathrm{e}^{-(\lambda_1+\lambda_2)t}\,\mathrm{d}t} \\
&= \frac{\lambda_1}{\lambda_1 + \lambda_2}.
\end{aligned}$$

Theorem 11.3 (Thinning)**.** *Given $(N(t),\ t \geq 0)$, a Poisson process of rate λ, color each point red or blue independently, red with probability p, blue with probability $1-p$. Let $(N_{\mathrm{B}}(t),\ t \geq 0)$ and $(N_{\mathrm{R}}(t),\ t \geq 0)$ be the processes of red and blue points respectively. Then, these are independent Poisson processes of rates $\lambda(1-p)$ and λp respectively.*

Lecture 12

October 3

12.1 More on Poisson Processes

The Poisson process with rate λ, $\lambda > 0$, is a stochastic process $(N(t),\ t \geq 0)$ whose sample paths are piecewise constant and increase by jumps of size 1, right-continuous, such that:

1. $N(0) = 0$;
2. For $0 \leq s < t$,

   $$\mathbb{P}\big(N(t) - N(s) = k\big) = \frac{(\lambda(t-s))^k}{k!}\mathrm{e}^{-\lambda(t-s)}$$

   so $\mathbb{E}[N(t) - N(s)] = \lambda(s-t)$;
3. The increments are independent, i.e.,

   $$\big(\underbrace{N(t_1) - N(s_1)}_{N((s_1,t_1])}, \underbrace{N(t_2) - N(s_2)}_{N((s_2,t_2])}, \dots, \underbrace{N(t_k) - N(s_k)}_{N((s_k,t_k])}\big)$$

   is independent.

Also,

$$\begin{aligned} T_0 = S_0 &= 0, \\ T_1 = S_1 &= \inf\{t : N(t) = 1\}, \\ S_2 &= \inf\{t : N(t) = 2\}, \\ &\vdots \\ S_k &= \inf\{t : N(t) = k\}, \end{aligned}$$

with $T_1 = S_1$, $T_2 = S_2 - S_1$, $\dots$, $T_k = S_k - S_{k-1}$.

We saw *last time*: $T_1 \sim \text{Exponential}(\lambda)$. *Proof*:

$$\mathbb{P}(T_1 > t) = \mathbb{P}\big(N(t) = 0\big) = \mathrm{e}^{-\lambda t}.$$

Also, $S_k \sim \text{Gamma}(k, \lambda)$, with

$$f_{S_k}(t) = \frac{\lambda^k t^{k-1}}{(k-1)!}\mathrm{e}^{-\lambda t}.$$

This is enough to conclude that $T_1, T_2, T_3, \dots$ are i.i.d. Exponential(λ) because $T_1, T_2, T_3, \dots$ are independent. For example,

$$\begin{aligned}
&\mathbb{P}\big(T_2 > t \bigm| T_1 \in (u, u+\mathrm{d}u)\big) \\
&\quad= \frac{\mathbb{P}(T_2 > t, T_1 \in (u, u+\mathrm{d}u))}{\mathbb{P}(T_1 \in (u, u+\mathrm{d}u))} \\
&\quad= \frac{\mathbb{P}(N(u) = 0, N(u, u+\mathrm{d}u) = 1, N(t+u) = 1)}{\mathbb{P}(N(u) = 0, N(u+\mathrm{d}u) = 1)} \\
&\quad= \frac{\mathbb{P}(N(u) = 0)\mathbb{P}(N(u+\mathrm{d}u) = 1 \mid N(u) = 0)\mathbb{P}(N(t+u) = 1 \mid N(u) = 0, N(u+\mathrm{d}u) = 1)}{\mathbb{P}(N(u) = 0)\mathbb{P}(N(u+\mathrm{d}u) = 1 \mid N(u) = 1)} \\
&\quad= \frac{\mathrm{e}^{-\lambda u} \cdot \lambda\,\mathrm{d}u \cdot \mathrm{e}^{-\lambda(t+u-u)}}{\mathrm{e}^{-\lambda u} \cdot \lambda\,\mathrm{d}u} = \mathrm{e}^{-\lambda t}.
\end{aligned}$$

If $(N_1(t),\ t \ge 0) \perp\!\!\!\perp (N_2(t),\ t \ge 0)$ are Poisson processes of rates λ_1 and λ_2 respectively, then $(N(t),\ t \ge 0)$, where $N(t) = N_1(t) + N_2(t)$, is a Poisson process of rate $\lambda_1 + \lambda_2$. Further, if we let $(T_k^{(1)},\ k = 1, 2, \dots)$, $(T_k^{(2)},\ k = 1, 2, \dots)$ be the interarrival times of the respective processes and $(T_k,\ k = 1, 2, \dots)$ be the interarrival times of $(N(t),\ t \ge 0)$, so $S_k^{(1)} = T_1^{(1)} + \cdots + T_k^{(1)}$, $S_k^{(2)} = T_1^{(2)} + \cdots + T_k^{(2)}$, $S_k = T_1 + \cdots + T_k$, so $T_1 = \min(T_1^{(1)}, T_1^{(2)})$, S_2 is the second smallest among $(S_1^{(1)}, S_2^{(1)}, S_1^{(2)}, S_2^{(2)})$, and S_3 similarly, let:

$$I_k = \begin{cases} 1, & \text{if } S_k \text{ is a point from the first process} \\ 2, & \text{if } S_k \text{ is a point from the second process} \end{cases}$$

Then, $(I_1, I_2, I_3, \dots)$ are i.i.d. Bernoulli with

$$\mathbb{P}(I_k = 1) = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

For $0 \le u_1 < u_2 < \cdots < u_{k-1}$, conditional on $S_1 \in (u_1, u_1 + \mathrm{d}u_1)$, …, $S_{k-1} \in (u_{k-1}, u_{k-1} + \mathrm{d}u_{k-1})$, $I_1 = i_1$, …, $I_{k-1} = i_{k-1}$, then $(N_1(t - u_{k-1}),\ t \ge u_{k-1})$, $(N_2(t - u_{k-1}),\ t \ge u_{k-1})$ are *independent* Poisson processes of rates λ_1 and λ_2 respectively.

12.1.1 Conditioning on a Later Arrival

Given a Poisson process of rate λ, and $0 \le v_1 < v_2 < \cdots < v_k \le t$, let us compute

$$\begin{aligned}
&\mathbb{P}\big(T_1 \in (v_1, v_1 + \mathrm{d}v_1), T_2 \in (v_2 - v_1, v_2 - v_1 + \mathrm{d}v_2), \dots, T_k \in (v_k - v_{k-1}, v_k - v_{k-1} + \mathrm{d}v_k) \bigm| N(t) = k\big) \\
&\quad= \frac{\mathbb{P}(T_1 \in (v_1, v_1 + \mathrm{d}v_1), \dots, T_k \in (v_k - v_{k-1}, v_k - v_{k-1} + \mathrm{d}v_k), N(t) = k)}{\mathbb{P}(N(t) = k)} \\
&\quad= \frac{\mathbb{P}(T_1 \in (v_1, v_1 + \mathrm{d}v_1), \dots, T_k \in (v_k - v_{k-1}, v_k - v_{k-1} + \mathrm{d}v_k), T_{k+1} > t - v_k)}{\mathbb{P}(N(t) = k)} \\
&\quad= \frac{\lambda\mathrm{e}^{-\lambda_1 v_1}\lambda\mathrm{e}^{-\lambda(v_2 - v_1)} \dots \lambda\mathrm{e}^{-\lambda(v_k - v_{k-1})}\lambda\mathrm{e}^{-\lambda(t - v_k)}\,\mathrm{d}v_1 \cdots \mathrm{d}v_k}{(\lambda t)^k \mathrm{e}^{-\lambda t}/k!} \\
&\quad= \frac{k!}{t^k}\mathrm{d}v_1 \cdots \mathrm{d}v_k.
\end{aligned}$$

What this tells us is that if we define $U_1, U_2, \dots, U_k$, independent, each uniformly distributed on $[0, t]$, and consider their order statistics $(U_{[1]}, U_{[2]}, \dots, U_{[k]})$, i.e., $\{U_{[1]}, \dots, U_{[k]}\} = \{U_1, \dots, U_k\}$ and

$$U_{[1]} \le U_{[2]} \le \cdots \le U_{[k]},$$

then $(U_{[1]}, U_{[2]}, \dots, U_{[k]}) \overset{\mathsf{d}}{=} \mathcal{L}(S_1, S_2, \dots, S_k \mid N(t) = k)$.

12.1.2 Poisson Law of Rare Events

Let $X_{n,1}, X_{n,2}, \dots, X_{n,n}$ be i.i.d. Bernoulli random variables with $\mathbb{P}(X_{n,k} = 1) = p_n$. So,

$$\mathbb{E}[\underbrace{X_{n,1} + \cdots + X_{n,n}}_{V_n}] = np_n.$$

Consider, as $n \to \infty$, $np_n \to \lambda$. Then,

$$\lim_{n\to\infty} \mathbb{P}(V_n = k) = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda} \qquad \text{for all } k \geq 0.$$

In fact, $\mathbb{E}[z^{V_n}] \to \mathrm{e}^{\lambda(1-z)}$. This is the **Poisson law for rare events**. If $X \sim \text{Bernoulli}(p)$,

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

Also, if $Y \sim \text{Poisson}(\lambda)$, then

$$\mathbb{E}[z^Y] = \sum_{k=0}^{\infty} z^k \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda} = \mathrm{e}^{-\lambda}\mathrm{e}^{\lambda z} = \mathrm{e}^{\lambda(1-z)}.$$

$V_n = X_{n,1} + \cdots + X_{n,n}$ so

$$\mathbb{E}[z^{V_n}] = \mathbb{E}[z^{X_{n,1}}]^n = \big(1 + p_n(1-z)\big)^n \to \mathrm{e}^{\lambda(1-z)} \qquad \text{as } n \to \infty$$

since $np_n \to \lambda$. Recall that

$$\mathrm{e} = \lim_{n\to\infty}\Big(1 + \frac{1}{n}\Big)^n.$$

In fact, with this notation, if $X_{n,k} \sim \text{Bernoulli}(p_{n,k})$ and $(X_{n,1}, \dots, X_{n,n})$ are independent, one can show that $|\mathbb{P}(V_n \in A) - \mathbb{P}(Z_n \in A)| \leq \sum_{k=1}^n p_{n,k}^2$, where A is any subset of $\{0, 1, \dots\}$, and $Z_n \sim \text{Poisson}(\sum_{k=1}^n p_{n,k})$.

12.2 Continuous-Time Markov Chains

A **countable-state continuous-time Markov chain** $(X_t,\ t \geq 0)$ or $(X_t,\ -\infty < t < \infty)$, is an $\mathcal{S}$-valued process, where $\mathcal{S}$ is the state space (countable set), satisfying the property that for all t, for all k, for all $s_0 < s_1 < \cdots < s_k < s < t$, for all $x_0, x_1, \dots, x_k, x, y$,

$$\mathbb{P}(X_t = y \mid X_s = x, X_{s_0} = x_0, \dots, X_{s_k} = x_k) = \mathbb{P}(X_t = y \mid X_s = x).$$

Lecture 13

October 5

13.1 Continuous-Time Markov Chains

$\mathcal{S}$ is a countable state space. $(X_t,\ t \geq 0)$ is a continuous-time $\mathcal{S}$-valued stochastic process. If

$$\mathbb{P}(X_t = y \mid X_s = x, X_{s_0} = x_0, \ldots, X_{s_k} = x_k) = \mathbb{P}(X_t = y \mid X_s = x)$$

for all $s_0 < s_1 < \cdots < s_k < s < t$, for all $x_0, x_1, \ldots, x_k, x, y \in \mathcal{S}$, then we say that $(X_t,\ t \geq 0)$ is a **continuous-time Markov chain**. We will also assume that $\mathbb{P}(X_t = x \mid X_s = y)$ depends on x, y, and $t - s$ only (i.e., we assume time homogeneity). Let us write $p_t(x,y)$ for $\mathbb{P}(X_t = y \mid X_0 = x)$. Then, $P_t = \big[p_t(x,y)\big]$ is a stochastic matrix if $\mathcal{S}$ is finite. More generally, $\sum_{y \in \mathcal{S}} p_t(x,y) = 1$ for all x, for all $t \geq 0$. Assume $P_0 = I$.

We have the **Chapman-Kolmogorov equations**, i.e., $P_t P_s = P_{t+s}$. *Proof*:

$$\begin{aligned}
p_{t+s}(x,y) &= \mathbb{P}(X_{t+s} = y \mid X_0 = s) \\
&= \sum_{z \in \mathcal{S}} \mathbb{P}(X_{t+s} = y, X_t = z \mid X_0 = x) \\
&= \sum_{z \in \mathcal{S}} \mathbb{P}(X_{t+s} = y \mid X_t = z, X_0 = x)\mathbb{P}(X_t = z \mid X_0 = x) \\
&= \sum_{z \in \mathcal{S}} p_t(x,z) p_s(z,y).
\end{aligned}$$

The Markov processes we are interested in will have the additional property that for all $y \neq x$,

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} p_\varepsilon(x,y)$$

exists (we denote it $q(x,y)$) and we will assume that $\sum_{y \neq x} q(x,y) < \infty$ for all $x \in \mathcal{S}$. We will denote this sum by λ_x. We will also set $q(x,x) = -\lambda_x$. $Q \triangleq \big[q(x,y)\big]$ is called the **rate matrix** or **generator matrix** of our Markov process. It turns out that one can write

$$\frac{\mathrm{d}}{\mathrm{d}t} P_t = P_t Q \qquad\qquad \textbf{Kolmogorov forward equation}$$

and (we will assume that we can write)

$$\frac{\mathrm{d}}{\mathrm{d}t} P_t = Q P_t \qquad\qquad \textbf{Kolmogorov backward equation}.$$

Note that $p_{t+\varepsilon}(x,y) = \sum_{z \in \mathcal{S}} p_t(x,z) p_\varepsilon(z,y)$ from the Chapman-Kolmogorov equations. So,

$$p_{t+\varepsilon}(x,y) - p_t(x,y) = \sum_{z \neq y} p_t(x,z) p_\varepsilon(z,y) + p_t(x,y)\big(p_\varepsilon(y,y) - 1\big).$$

Divide by ε, let $\varepsilon \to 0$.

$$\frac{\mathrm{d}}{\mathrm{d}t} p_t(x,y) = \sum_{z \neq y} p_t(x,z) q(z,y) - p_t(x,y)\lambda_y$$

because $1 - p_\varepsilon(y,y) = \sum_{z\neq y} p_\varepsilon(y,z)$, so $\varepsilon^{-1}\sum_{z\neq y} p_\varepsilon(y,z) \to \sum_{z\neq y} q(y,z) = \lambda_y$, and thus

$$\begin{aligned} &= \sum_{z\in\mathcal{S}} p_t(x,z)q(z,y) \\ &= (P_tQ)(x,y). \end{aligned}$$

For the backward equation,

$$\begin{aligned} p_{t+\varepsilon}(x,y) &= \sum_{z\in\mathcal{S}} p_\varepsilon(x,z)p_t(z,y), \\ p_{t+\varepsilon}(x,y) - p_t(x,y) &= \sum_{z\neq x} p_\varepsilon(x,z)p_t(z,y) + \big(p_\varepsilon(x,x) - 1\big)p_t(x,y). \end{aligned}$$

Divide by ε, let $\varepsilon \to 0$, and it is still true that

$$\frac{p_\varepsilon(x,z)}{\varepsilon} \to q(x,z) \qquad \text{and} \qquad \frac{p_\varepsilon(x,x)-1}{\varepsilon} \to -\lambda_x$$

but we can have summability issues since $p_\varepsilon(z,y)$ for fixed y can be not summable over z.

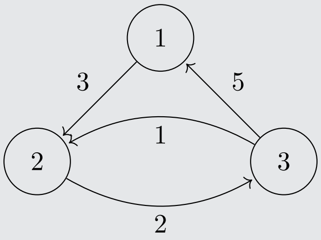
In the finite state space case, the equation

$$\frac{\mathrm{d}}{\mathrm{d}t}P_t = P_tQ \qquad \text{or} \qquad \frac{\mathrm{d}}{\mathrm{d}t}P_t = QP_t,$$

with the initial condition $P_0 = I$ has the solution $P_t = \mathrm{e}^{Qt}$ where e^{Qt} is notation for $\displaystyle\sum_{n=0}^{\infty}\frac{(Qt)^n}{n!}$ which will be summable for all t. We will also use the notation e^{Qt} for P_t when the state space is countably infinite, but one needs some care to interpret this.

First, by convention one describes a continuous-time Markov chain via a "rate diagram".

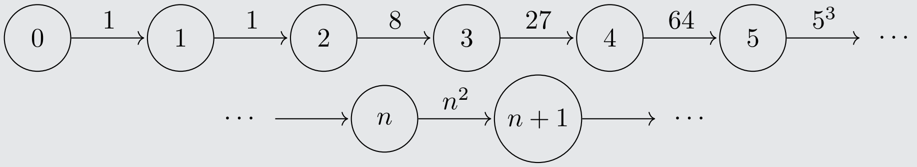
Example 13.1. Consider:



means

$$Q = \begin{bmatrix} -3 & 3 & 0 \\ 0 & -2 & 2 \\ 5 & 1 & -6 \end{bmatrix}.$$

Example 13.2. When $\mathcal{S}$ is infinite, we could have, e.g.,



Starting from 0, the mean time to "escape to ∞" will be

$$1+\sum_{n=1}^{\infty}\frac{1}{n^3}<\infty.$$

In this example, the chain stays in k for an exponentially distributed time of rate k^3. This means

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

cannot be solved for all t. If we start at $P_0 = I$, in finite time we could be at none of the states.

A basic intuition to have: If a positive random variable T has the property that

$$\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\mathbb{P}(T \le t+\varepsilon \mid T > t) = \lambda,$$

then T must be exponentially distributed with mean $1/\lambda$.

$$\mathbb{P}(T \le t+\varepsilon \mid T>t) = \frac{\mathbb{P}(T\le t+\varepsilon, T>t)}{\mathbb{P}(T>t)} = \frac{\mathbb{P}(T>t)-\mathbb{P}(T>t+\varepsilon)}{\mathbb{P}(T>t)}.$$

Let $\mathbb{P}(T>t)$ be denoted $g(t)$. Then, the assumption reads

$$\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\frac{g(t)-g(t+\varepsilon)}{g(t)} = \lambda,$$

i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}t}g(t) = -\lambda g(t),$$

so $g(t) = \mathrm{e}^{-\lambda t}$ because $g(0) = 1$.

From this picture, we learn that the time spent in each state x before we making a jump is exponentially distributed with rate λ_x. Further, we claim the jump, when made, is to state $y \neq x$ with probability $q(x,y)/\lambda_x$. This is because if we run a race between independent exponential random variables each of parameter $q(x,y)$ (for $y \neq x$), the jump will occur at time $\sim$ Exponential(λ_x), and the yth variable wins with probability $q(x,y)/\lambda_x$. If $T_1 \sim$ Exponential(λ_1), $T_2 \sim$ Exponential(λ_2), $T_1 \perp\!\!\!\perp T_2$, then

$$\{\min(T_1,T_2) > t\} = \{T_1 > t, T_2 > t\}$$

and

$$\mathbb{P}\big(T_1 \text{ wins} \bigm| \min(T_1,T_2)\in(t,t+\mathrm{d}t)\big) = \frac{\lambda_1}{\lambda_1+\lambda_2}.$$

So for us a continuous-time Markov process will be prescribed by giving Q (the rate matrix), i.e., by giving $q(x,y)$ for $y \neq x$ such that $\sum_{y\neq x} q(x,y) = \lambda_x < \infty$ ($q(x,x) \triangleq -\lambda_x$). We will assume "non-explosivity" from any initial condition and we will think of the process as being constructed via exponentially distributed random variables and $\mathcal{S}$-valued random variables as follows. Suppose the initial distribution is $(p_0(x),\ x\in\mathcal{S})$.

- Draw $x_0 \in \mathcal{S}$ according to the initial distribution.

- Start an exponential distribution of rate λ_{x_0}. When it expires, jump to state $y \neq x_0$ according to the realization of an $\mathcal{S}$-valued coin toss which equals $y \neq x_0$ with probability $q(x_0, y)/\lambda_{x_0}$.

- Suppose we jumped to x_1. Repeat the algorithm. Start an exponential random variable with rate λ_{x_1}. Draw a realization of an $\mathcal{S}$-valued coin with probabilities $q(x_1, y)/\lambda_{x_1}$ for $y \neq x_1$. Each time and "coin" is independent of all previous ones.

An even simpler picture is possible when $\sup_{x \in \mathcal{S}} \lambda_x < \lambda < \infty$. Create a single Poisson process of rate λ. When in state x and if a point of the Poisson process occurs, stay put with probability $1 - \lambda_x/\lambda$ and jump to $y \neq x$ with probability $q(x, y)/\lambda_x$ if you jump.

13.1.1 Definitions Parallel to the Discrete-Time Case

State x communicates with state y if $p_t(x, y) > 0$ for some $t > 0$ (and x communicates with itself by default).

New twist: If $p_t(x, y) > 0$ for some $t > 0$, then $p_t(x, y) > 0$ for all $t > 0$. Reason: Both are equivalent to the existence of a path from x to y in the rate diagram.

Lecture 14

October 10

14.1 Construction of Continuous-Time Markov Chains

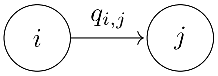
A continuous-time Markov chain $(X_t,\ t \geq 0)$ is an $\mathcal{S}$-valued process, where $\mathcal{S}$ is a countable set (finite or countably infinite). $q_{i,j}$ is the "rate" (non-negative) of transitions from state i to state j. $\lambda_i = \sum_{j \neq i} q_{i,j}$ (assume $0 \leq \lambda_i < \infty$). $q_{i,i} \triangleq -\lambda_i = -\sum_{j \neq i} q_{i,j}$. Assume $Q = \left[q_{i,j}\right]$ results in a "non-explosive" Markov chain.

Mental picture: When state i is entered, a "fresh" exponentially distributed random variable of parameter λ_i is generated (independent of everything that happened so far). Jump out of i when this random variable expires to state j with probability $q_{i,j}/\lambda_i$, where the choice of the state to jump to is independent of everything so far.

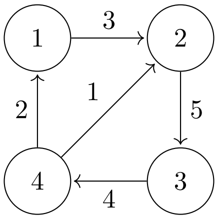
Another construction: When in state i, generate for each $j \neq i$ a "fresh" Exponential($q_{i,j}$) random variable. Thesea are independent and independent of the past. The one that expires earliest determines where you jump.

Another construction (if $\lambda_i < \lambda < \infty$ for all i): Run a Poisson process of rate λ. If in state i and a point of the Poisson process appears, jump to $j \neq i$ with probability $q_{i,j}/\lambda$, stay put with probability $1 - \lambda_i/\lambda$.

In the rate diagram (which has no self-loops), a directed edge is marked with $q_{i,j}$ if $q_{i,j} > 0$.



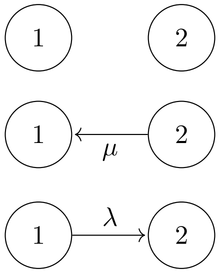
For example:



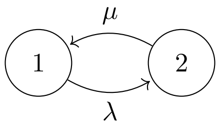
corresponds to the rate matrix

$$Q = \begin{bmatrix} -3 & 3 & 0 & 0 \\ 0 & -5 & 5 & 0 \\ 0 & 0 & -4 & 4 \\ 2 & 1 & 0 & -3 \end{bmatrix}.$$

14.2 Two-State Continuous-Time Markov Chains



The most general case is:



The transition matrix is

$$Q = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}.$$

Recall: $P_t = \mathrm{e}^{Qt}$ (from $\dot{P}_t = P_t Q$ and $\dot{P}_t = QP_t$) where e^{Qt} means $\displaystyle\sum_{k=0}^{\infty} \frac{(Qt)^k}{k!}$. Here,

$$P_t = \begin{bmatrix} \dfrac{\mu}{\mu+\lambda} + \dfrac{\lambda}{\mu+\lambda}\mathrm{e}^{-(\mu+\lambda)t} & \dfrac{\lambda}{\mu+\lambda} - \dfrac{\lambda}{\mu+\lambda}\mathrm{e}^{-(\mu+\lambda)t} \\ \dfrac{\mu}{\mu+\lambda} - \dfrac{\mu}{\mu+\lambda}\mathrm{e}^{-(\mu+\lambda)t} & \dfrac{\lambda}{\mu+\lambda} + \dfrac{\mu}{\mu+\lambda}\mathrm{e}^{-(\mu+\lambda)t} \end{bmatrix} \xrightarrow{\text{as } t\to\infty} \begin{bmatrix} \dfrac{\mu}{\mu+\lambda} & \dfrac{\lambda}{\mu+\lambda} \\ \dfrac{\mu}{\mu+\lambda} & \dfrac{\lambda}{\mu+\lambda} \end{bmatrix}$$

and

$$\begin{bmatrix} \dfrac{\mu}{\mu+\lambda} & \dfrac{\lambda}{\mu+\lambda} \end{bmatrix} \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix} = \begin{bmatrix} 0 & 0 \end{bmatrix}.$$

14.3 Parallels with Discrete-Time Markov Chains

14.3.1 Stationary Distributions

$P_t \triangleq \big[p_{i,j}(t)\big]$, where $p_{i,j}(t) \triangleq \mathbb{P}(X_t = j \mid X_0 = i)$. $\pi P = \pi$ in discrete time is replaced by $\pi Q = 0$ in continuous time. In other words, $\sum_{i\in\mathcal{S}} \pi_i q_{i,j} = 0$ for all j. $\pi Q = 0$ is equivalent to $\pi P_t = \pi$ for all $t > 0$. Given $\pi P_t = \pi$ for all $t > 0$,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\pi P_t = 0 &\implies \pi P_t Q = 0 \\ &\implies \pi Q = 0. \end{aligned}$$

Conversely, if $\pi Q = 0$, then $(\pi \dot{P_t}) = \pi Q P_t = 0$ so $\pi P_t = \pi P_0 = \pi$.

14.3.2 Classification of States

In discrete time, we say i communicates with j if we can get from i to j in the transition probability diagram. (i communicates with i by definition.) Equivalently, $p_{i,j}^{(n)} > 0$ for some $n \geq 0$. Notation: $i \to j$.

In continuous time, we will say i **communicates with** j if we can get from i to j in the rate diagram. (i communicates with i by definition.) Equivalently, $p_{i,j}(t) > 0$ for some $t > 0$, or equivalently, $p_{i,j}(t) > 0$ for all $t > 0$. Notation: $i \to j$.

As in discrete time, write $i \leftrightarrow j$ if $i \to j$ and $j \to i$. This defines an equivalence relation on $\mathcal{S}$. The equivalence classes are called the **communicating classes**. The continuous-time Markov chain is **irreducible** if $\mathcal{S}$ is a single communicating class.

The analog of $T_y \triangleq \inf\{n \geq 1 : X_n = y\}$ in discrete time is

$$T_i = \inf\{t > 0 : X_t = i, X_s \neq i \text{ for some } 0 < s < t\}.$$

In continuous time, if

$$\mathbb{P}(T_i < \infty \mid X_0 = i) \begin{cases} < 1, & \text{we say state } i \text{ is } \textbf{transient}, \\ = 1, & \text{we say state } i \text{ is } \textbf{recurrent}. \end{cases}$$

If state i is recurrent, we further look at $\mathbb{E}[T_i \mid X_0 = i]$. If this is finite, we say state i is **positive-recurrent**, and if this is ∞ we say state i is **null-recurrent**. One can show that transience, null recurrence, and positive recurrence are class properties. If $\sup_{i \in \mathcal{S}} \lambda_i = \infty$, technically one needs to write proofs for these claims. The structure of the proofs is *identical.* If $\sup_{i \in \mathcal{S}} \lambda_i < \lambda < \infty$, these statements are corollaries of the discrete-time statements. To see this, having picked $\lambda > \sup_{i \in \mathcal{S}} \lambda_i$, define a discrete-time Markov chain transition probability matrix (or array) associated to the so-called "**jump chain**" whose diagonal entries are $(1 - \lambda_i)/\lambda$ and the off-diagonal entries $j \neq i$ are $q_{i,j}/\lambda$.

14.3.3 Hitting Times

In a continuous-time MC, given a subset $A \subseteq S$, define $V_A = \inf\{t > 0 : X_t \in A\}$ (the first **hitting time of** A). V_i is different from T_i because T_i was the first *return* time, which required leaving i before coming back. Let $h_i^A \triangleq \mathbb{P}(V_A < \infty \mid X_0 = i)$. So, $h_i^A = 1$ for all $i \in A$. The equations satisfied by the $(h_i^A,\ i \in \mathcal{S})$ for fixed A are

$$\begin{aligned} h_i^A &= 1 \text{ for all } i \in A, \\ \sum_{j \in \mathcal{S}} q_{i,j} h_j^A &= 0 \text{ for all } i \notin A. \end{aligned}$$

These can be got from an infinitesimal first-step equation. Fix $\delta > 0$ small.

$$\begin{aligned} &\mathbb{P}(V^A < \infty \mid X_0 = i) \\ &\quad = \sum_{j \in \mathcal{S}} \mathbb{P}(V^A < \infty, X_\delta = j \mid X_0 = i) \\ &\quad = \sum_{j \neq i} \mathbb{P}(V^A < \infty \mid X_\delta = j) \underbrace{\mathbb{P}(X_\delta = j \mid X_0 = i)}_{q_{i,j}\delta + o(\delta)} + \mathbb{P}(V^A < \infty \mid X_\delta = i) \underbrace{\mathbb{P}(X_\delta = i \mid X_0 = i)}_{1 - \lambda_i \delta + o(\delta)} \end{aligned}$$

Little-o of delta, $o(\delta)$, is any function of δ such that when divided by δ, it goes to 0 as $\delta \to 0$. Then divide by δ and let $\delta \to 0$.

Lecture 15

October 12

15.1 Asymptotic Notation

Asymptotic notation is widely used and *standard*.

Purpose: Avoid keeping track of functions you do not care about except in so far as their asymptotics.

General situation: x is converging to x_0 and we are given functions f and g.

- $f(x) = o(g(x))$ as $x \to x_0$ if
$$\frac{f(x)}{g(x)} \to 0 \qquad \text{as } x \to x_0.$$
- $f(x) = O(g(x))$ as $x \to x_0$ if
$$\limsup_{x\to x_0} \frac{f(x)}{g(x)} < \infty$$
(used when $g(x) \geq 0$).
- $f(x) = \omega(g(x))$ as $x \to x_0$ if
$$\frac{f(x)}{g(x)} \to \infty \qquad \text{as } x \to x_0$$
(again used when $g(x) \geq 0$).
- $f(x) = \Omega(g(x))$ as $x \to x_0$ if
$$\liminf_{x\to x_0} \frac{f(x)}{g(x)} \geq K > 0$$
(again $g(x) \geq 0$ is assumed).
- $f(x) = \Theta(g(x))$ as $x \to x_0$ means $f(x) = \Omega(g(x))$ and $f(x) = O(g(x))$.

For a Poisson process of rate $\lambda > 0$, for $s_1 < s_2 < \cdots < s_k < t < t + \varepsilon$,

$$\begin{aligned}
\mathbb{P}\big(N(t+\varepsilon) - N(t) = 1 \bigm| N(s_1), N(s_2), \dots, N(s_k)\big) &= \lambda\varepsilon e^{-\lambda\varepsilon} = \lambda\varepsilon\big(1 - O(\varepsilon)\big) = \lambda\varepsilon + o(\varepsilon), \\
\mathbb{P}\big(N(t+\varepsilon) - N(t) \geq 2 \bigm| N(s_1), N(s_2), \dots, N(s_k)\big) &= 1 - e^{-\lambda\varepsilon} - \lambda\varepsilon e^{-\lambda\varepsilon} \\
&= 1 - \big(1 - \lambda\varepsilon + o(\varepsilon)\big) - \lambda\varepsilon\big(1 - O(\varepsilon)\big) \\
&= o(\varepsilon).
\end{aligned}$$

If $(X(t),\ t \ge 0)$ is an $\mathcal{S}$-valued CTMC with rate matrix Q, under our assumptions, then given the times $s_1 < s_2 < \cdots < s_k < t < t + \varepsilon$, $x_1, x_2, \ldots, x_k, x, y \in \mathcal{S}$,

$$\mathbb{P}\big(X(t+\varepsilon) = y \bigm| X(t) = x, X(s_1) = x_1, \ldots, X(s_k) = x_k\big) = \begin{cases} q_{x,y}\varepsilon + o(\varepsilon), & \text{for } y \ne x, \\ 1 - \lambda_x \varepsilon + o(\varepsilon), & \text{for } y = x. \end{cases}$$

15.2 CTMC Results

15.2.1 Hitting Times

Recall that if $A \subseteq S$ then $V^A \triangleq \inf\{t > 0 : X_t \in A\}$. We can write the analog of the first-step equations for $(\mathbb{E}_i[V^A],\ i \in \mathcal{S})$. Here, $\mathbb{E}_i[V^A] = \mathbb{E}_i[V^A \mid X_0 = i]$. These are:

$$\begin{aligned} \mathbb{E}_i[V^A] &= 0, && \text{for } i \in A, \\ -\sum_{j \in \mathcal{S}} q_{i,j}\, \mathbb{E}_j[V^A] &= 1, && \text{for } i \notin A. \end{aligned}$$

These can be proved by writing for $i \notin A$,

$$\mathbb{E}[V^A \mid X_0 = i] = \varepsilon + \big(1 - \lambda_i \varepsilon + o(\varepsilon)\big)\, \mathbb{E}[V^A \mid X_0 = i] + \sum_{j \notin A} \frac{q_{i,j}}{\lambda_i} \big(\lambda_i \varepsilon + o(\varepsilon)\big)\, \mathbb{E}[V^A \mid X_0 = j].$$

Divide by sides by ε and let $\varepsilon \to 0$. This gives the desired formula.

15.2.2 Stationarity

A probability distribution π on $\mathcal{S}$ is called **stationary** for the CTMC if $\pi Q = \pi$ (where π is a row vector). We saw that this is equivalent to $\pi P_t = \pi$ for all $t \ge 0$.

Let $T_i = \inf\{t > 0 : X_t = i, X_s \ne i \text{ for some } 0 < s < t\}$. Then, $\mathbb{P}_i(T_i < \infty) = 1$ if and only if i is recurrent and $\mathbb{P}_i(T_i < \infty) < 1$ if and only if i is transient. If i is recurrent and:

- $\mathbb{E}_i[T_i] = \infty$, then i is null-recurrent;
- $\mathbb{E}_i[T_i] < \infty$, then i is positive-recurrent.

Let $m_i \triangleq \mathbb{E}_i[T_i]$ and let $\mu_j^{(i)} \triangleq \mathbb{E}_i[\int_0^{T_i} \mathbb{1}\{X_s = j\}\, \mathrm{d}s]$. An irreducible positive-recurrent CTMC has a unique stationary distribution, call it π, and for any fixed $i \in \mathcal{S}$, for each $j \in \mathcal{S}$, then

$$\pi_j = \frac{\mu_j^{(i)}}{m_i}.$$

Note that $m_i = \sum_{j \in \mathcal{S}} \mu_j^{(i)}$, so $\mu_i^{(i)} = 1/\lambda_i$.

15.2.3 Jump Chain

To any CTMC, with the rate matrix Q of the kind we consider, there is an associated DTMC called the **jump chain** with transition probabilities $q_{i,j}/\lambda_i$, $j \ne i$, which has zero diagonal terms. An irreducible positive-recurrent CTMC has an irreducible positive-recurrent jump chain (this needs the condition $\sum_{k \in \mathcal{S}} \pi_k \lambda_k < \infty$). The stationary distribution of the jump chain is $\pi_i \lambda_i / (\sum_{j \in \mathcal{S}} \pi_j \lambda_j)$. To check this,

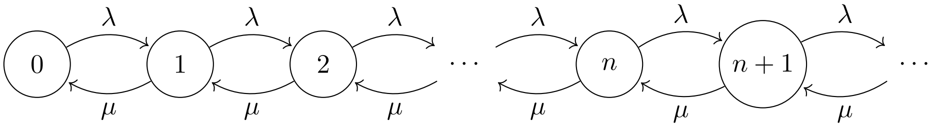
$$\sum_{i \ne j} \Big(\frac{\pi_i \lambda_i}{\sum_{k \in \mathcal{S}} \pi_k \lambda_k}\Big) \frac{q_{i,j}}{\lambda_i} = \frac{\pi_j \lambda_j}{\sum_{k \in \mathcal{S}} \pi_k \lambda_k}$$

using $\pi Q = 0$.

The analog of $\mathbb{E}_i[N_i]$ in discrete time, where $N_i \triangleq \sum_{n=1}^{\infty} \mathbb{1}\{X_n = i\}$, is $\mathbb{E}_i[\int_0^{\infty} \mathbb{1}\{X_s = i\}\, \mathrm{d}s] = \int_0^{\infty} p_{i,i}(t)\, \mathrm{d}t$.

15.3 Examples

15.3.1 Continuous-Time Birth-Death Process



The chain is irreducible. Also, it is transient if and only if $\lambda > \mu$, and null-recurrent if and only if $\lambda = \mu$. To verify this, take $A = \{0\}$ and $h_i^A \triangleq \mathbb{P}_i(V^A < \infty)$.

$$h_0^{\{0\}} = 1,$$
$$\sum_{j \in \mathcal{S}} q_{i,j} h_j^{\{0\}} = 0, \qquad \text{for all } i \neq 0.$$

Assume $\mu = \lambda$. For $i = 1$,

$$\lambda \cdot 1 + \lambda \cdot h_2^{\{0\}} - 2\lambda h_1^{\{0\}} = 0.$$

For $i = 2$,

$$\lambda h_1^{\{0\}} + \lambda h_3^{\{0\}} - 2\lambda h_2^{\{0\}} = 0.$$

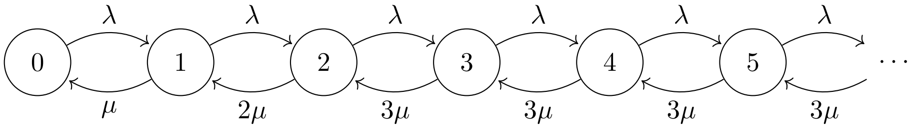
The other equations are similar. To solve this, set all $h_i^{\{0\}}$ to be 1, i.e., we have recurrence. Similarly, the equations for $\mathbb{E}_i[V^A]$ for $A = \{0\}$ will show that we have null recurrence. If $\lambda < \mu$, we have positive recurrence. Here, the stationary distribution is

$$\pi_i = \left(\frac{\lambda}{\mu}\right)^i \frac{1}{1 - \lambda/\mu}, \qquad i = 0, 1, 2, \ldots .$$

15.3.2 Queueing Theory

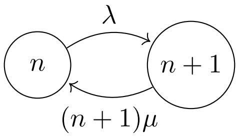
The continuous-time birth-death process is the **M/M/**1 queue model, which means memoryless arrivals, memoryless service, and one server. There is a buffer with infinite capacity. We see arrivals at the times of a Poisson(λ) process. There is one server which serves at the times of a Poisson process of rate μ, and the arrival process is independent of the service process. X_t is the number of the packets in the queue.

The **M/M/**k queue has k servers. For example, consider $k = 3$.



Now, $\lambda < 3\mu$ if and only if the chain is positive-recurrent.

In the extreme case we have the **M/M/**∞ queue.



The chain is always positive-recurrent. The stationary distribution is Poisson.

$$\pi_k = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k \mathrm{e}^{-\lambda/\mu}.$$

15.4 Time Reversal

Any stationary Markov chain can be run backwards in time. In discrete time, the time-reversed chain has transition probabilities

$$p^{(\mathrm{R})}_{i,j} = \frac{\pi_j p_{j,i}}{\pi_i}$$

where $p_{j,i}$ are the forward transition probabilities and π is the stationary distribution. Similarly, in continuous time,

$$q^{(\mathrm{R})}_{i,j} = \frac{\pi_j q_{j,i}}{\pi_i}$$

(even for $j = i$).

For the M/M/1 queue,

$$q^{(\mathrm{R})}_{n,n+1} = \frac{\pi_{n+1} q_{n+1,n}}{\pi_n} = \frac{(\lambda/\mu)^{n+1}\mu/(1-\lambda/\mu)}{(\lambda/\mu)^n/(1-\lambda/\mu)} = \lambda = q_{n,n+1}.$$

Similarly, $q^{(\mathrm{R})}_{n+1,n} = \mu$.

In stationarity in an M/M/1 queue, one has the amazing result that the process of departures is also a Poisson process with rate λ and $(A_s,\ s \geq t) \perp\!\!\!\perp X_t \perp\!\!\!\perp (D_s,\ s \leq t)$. (This is Burke's Theorem.)

Lecture 16

October 19

16.1 Conditional Expectation

B is an event and B^{c} is the complement of B. Given any event A, $\mathbb{P}(A \mid B) = \mathbb{E}[\mathbb{1}_A \mid B]$ is the conditional probability of A given B. Similarly, $\mathbb{P}(A \mid B^{\mathsf{c}}) = \mathbb{E}[\mathbb{1}_A \mid B^{\mathsf{c}}]$ is the conditional probability of A given B^{c}. We can keep track of both via the random variable $\mathbb{P}(A \mid B)\,\mathbb{1}_B + \mathbb{P}(A \mid B^{\mathsf{c}})\,\mathbb{1}_{B^{\mathsf{c}}}$. This is a random variable that is a deterministic function of $\mathbb{1}_B$.

Definition 16.1. Given any random variable Y and a RV X jointly defined with it, we create a new random variable denoted $\mathbb{E}(X \mid Y)$ which is supposed to have the property that

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

holds for all $h(Y)$ for which the expectation makes sense.

Fact: $\mathbb{E}(X \mid Y)$ exists and is unique almost surely (i.e., any two choices differ only on an event of probability zero).

Example 16.2. $\mathbb{E}(\mathbb{1}_A \mid \mathbb{1}_B) = \mathbb{P}(A \mid B)\,\mathbb{1}_B + \mathbb{P}(A \mid B^{\mathsf{c}})\,\mathbb{1}_{B^{\mathsf{c}}}$. To prove this, we need to show that for any deterministic function of $\mathbb{1}_B$, i.e., $a\,\mathbb{1}_B + b\,\mathbb{1}_{B^{\mathsf{c}}}$,

$$\mathbb{E}\big[\underbrace{(a\,\mathbb{1}_B + b\,\mathbb{1}_{B^{\mathsf{c}}})}_{h(Y)}\underbrace{\big(\mathbb{P}(A \mid B)\,\mathbb{1}_B + \mathbb{P}(A \mid B^{\mathsf{c}})\,\mathbb{1}_{B^{\mathsf{c}}}\big)}_{\mathbb{E}(X|Y)}\big] \overset{?}{=} \mathbb{E}\big[\underbrace{(a\,\mathbb{1}_B + b\,\mathbb{1}_{B^{\mathsf{c}}})}_{h(Y)}\underbrace{\mathbb{1}_A}_{X}\big].$$

Here, $X = \mathbb{1}_A$ and $Y = \mathbb{1}_B$. Then,

$$\text{LHS} = a\mathbb{P}(A \mid B)\mathbb{P}(B) + b\mathbb{P}(A \mid B^{\mathsf{c}})\mathbb{P}(B^{\mathsf{c}}) = a\mathbb{P}(A \cap B) + b\mathbb{P}(A \cap B^{\mathsf{c}}) = \text{RHS}.$$

Example 16.3. If Y is discrete, taking on values in the set $\{y_1, \dots, y_K\}$, then

$$\mathbb{E}(X \mid Y) = \sum_{k=1}^{K} \mathbb{E}[X \mid Y = y_k]\,\mathbb{1}_{\{Y = y_k\}}$$

where

$$\mathbb{E}[X \mid Y = y_k] = \frac{\mathbb{E}[X\,\mathbb{1}_{\{Y=y_k\}}]}{\mathbb{P}(Y = y_k)}.$$

The most general function of Y looks like $\sum_{k=1}^K a_k \mathbb{1}_{\{Y=y_k\}}$. We need to check:

$$\mathbb{E}\Big[\underbrace{\Big(\sum_{k=1}^K a_k \mathbb{1}_{\{Y=y_k\}}\Big)}_{h(Y)}\underbrace{\Big(\sum_{k=1}^K \mathbb{E}[X \mid Y=y_k]\,\mathbb{1}_{\{Y=y_k\}}\Big)}_{\mathbb{E}(X|Y)}\Big] = \mathbb{E}\Big[\underbrace{\sum_{k=1}^K a_k \mathbb{1}_{\{Y=y_k\}}}_{h(Y)}\underbrace{X}_{X}\Big].$$

Then,

$$\begin{aligned}\text{LHS} &= \sum_{k=1}^K a_k\,\mathbb{E}[X \mid Y=y_k]\mathbb{P}(Y=y_k)\\ &= \sum_{k=1}^K a_k\,\mathbb{E}[X\,\mathbb{1}_{\{Y=y_k\}}] = \mathbb{E}\Big[X\sum_{k=1}^K a_k\,\mathbb{1}_{\{Y=y_k\}}\Big] = \text{RHS}.\end{aligned}$$

Example 16.4. If X and Y have joint density $f_{X,Y}$, then $\mathbb{E}(X \mid Y) = \int_{-\infty}^{\infty} x f_{X|Y}(x \mid Y)\,\mathrm{d}x$. From a first course in probability:

$$\begin{aligned}\mathbb{E}[X \mid Y=y] &= \lim_{\varepsilon\to 0}\mathbb{E}\Big[X\;\Big|\;Y \in \Big(y-\frac{\varepsilon}{2}, y+\frac{\varepsilon}{2}\Big)\Big]\\ &= \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\,\mathrm{d}x.\end{aligned}$$

We need to check that for all $h(Y)$,

$$\mathbb{E}\Big[h(Y)\int_{-\infty}^{\infty} x f_{X|Y}(x \mid Y)\,\mathrm{d}x\Big] \overset{?}{=} \mathbb{E}[h(Y)X].$$

Then,

$$\begin{aligned}\text{LHS} &= \int_{-\infty}^{\infty}\Big(h(y)\int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\,\mathrm{d}x\Big) f_Y(y)\,\mathrm{d}y\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x h(y) f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \mathbb{E}[Xh(Y)] = \text{RHS}.\end{aligned}$$

16.1.1 MMSE

When second moments exist, one can show $\mathbb{E}[(X-\mathbb{E}(X \mid Y))^2] \le \mathbb{E}[(X-h(Y))^2]$ for all functions of Y, so $\mathbb{E}(X \mid Y)$ is the MMSE (minimum mean square estimate) of X given Y. To prove this, write the RHS as

$$\begin{aligned}&\mathbb{E}\big[\big(X-\mathbb{E}(X \mid Y)+\mathbb{E}(X \mid Y)-h(Y)\big)^2\big]\\ &\quad= \underbrace{\mathbb{E}\big[\big(X-\mathbb{E}(X \mid Y)\big)^2\big]}_{\text{LHS}} + \underbrace{\mathbb{E}\big[\big(\mathbb{E}(X \mid Y)-h(Y)\big)^2\big]}_{\ge 0}\\ &\qquad+ \underbrace{2\,\mathbb{E}\big[\big(\mathbb{E}(X \mid Y)-h(Y)\big)X\big] - 2\,\mathbb{E}\big[\big(\mathbb{E}(X \mid Y)-h(Y)\big)\,\mathbb{E}(X \mid Y)\big]}_{=0}\end{aligned}$$

from the definition of $\mathbb{E}(X \mid Y)$ as that function of Y for which for all functions of Y, say $h(Y)$,

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

16.1.2 Properties of Conditional Expectation

Theorem 16.5. *If X and Y are independent, then $\mathbb{E}(X \mid Y) = \mathbb{E}[X]$.*

Proof. For any function of Y, call it $h(Y)$, we need to show $\mathbb{E}[h(Y)\,\mathbb{E}[X]] \overset{?}{=} \mathbb{E}[h(Y)X]$. Then, the LHS is $\mathbb{E}[h(Y)]\,\mathbb{E}[X]$ and the RHS is $\mathbb{E}[h(Y)]\,\mathbb{E}[X]$ so LHS = RHS. □

Recall that $\mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2]$.

Theorem 16.6. $\mathbb{E}(aX_1 + bX_2 \mid Y) = a\,\mathbb{E}(X_1 \mid Y) + b\,\mathbb{E}(X_2 \mid Y)$.

Proof. Pick any function of Y, call it $h(Y)$. Compute

$$\begin{aligned}\mathbb{E}\big[h(Y)\big(a\,\mathbb{E}(X_1 \mid Y) + b\,\mathbb{E}(X_2 \mid Y)\big)\big] &= a\,\mathbb{E}[h(Y)\,\mathbb{E}(X_1 \mid Y)] + b\,\mathbb{E}[h(Y)\,\mathbb{E}(X_2 \mid Y)] \\ &= a\,\mathbb{E}[h(Y)X_1] + b\,\mathbb{E}[h(Y)X_2] \\ &= \mathbb{E}[h(Y)(aX_1 + bX_2)].\end{aligned}$$

□

Theorem 16.7. *Given any function of Y, say $h(Y)$, $\mathbb{E}(h(Y)X \mid Y) = h(Y)\,\mathbb{E}(X \mid Y)$.*

Proof. To prove this, take any function of Y, say $g(Y)$. Then,

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

using the definition of $\mathbb{E}(X \mid Y)$, where $g(Y)$ is the test function of Y, and $h(Y)\,\mathbb{E}(X \mid Y)$ is the proposed expression for $\mathbb{E}(h(Y)X \mid Y)$. □

Similar to $\mathbb{E}(X \mid Y)$, we can define $\mathbb{E}(X \mid Y_1, Y_2, \ldots, Y_L)$ as that function of $(Y_1, \ldots, Y_L)$ having the property that for every function of $(Y_1, \ldots, Y_L)$, call it $h(Y_1, \ldots, Y_L)$, we have

$$\mathbb{E}[h(Y_1, \ldots, Y_L)\,\mathbb{E}(X \mid Y_1, \ldots, Y_L)] = \mathbb{E}[h(Y_1, \ldots, Y_L)X].$$

We can even make sense of $\mathbb{E}(X \mid (Y_t,\ t \geq 0))$.

Theorem 16.8 (Successive Projection Property)**.** $\mathbb{E}(\mathbb{E}(X \mid Y, Z) \mid Y) = \mathbb{E}(X \mid Y)$.

Proof. We want to show that the LHS works as $\mathbb{E}(X \mid Y)$. So, take a test function $h(Y)$. Then, $\mathbb{E}[h(Y)\,\mathbb{E}(X \mid Y, Z)] = \mathbb{E}[h(Y)X]$ using the definition of $\mathbb{E}(X \mid Y, Z)$ and the fact that a function of Y can be thought of as a function of (Y, Z). □

This can be used to help in calculations as follows. You are interested in $\mathbb{E}(X \mid Y)$, but calculating it directly may be hard. Sometimes you can find Z such that computing $\mathbb{E}(X \mid Y, Z)$ and $\mathbb{E}(\mathbb{E}(X \mid Y, Z) \mid Y)$ is easy.

Example 16.9. X and Y are uniformly distributed and independent on $[0, 1]$.

$$\begin{aligned}\mathbb{E}(X \mid \min(X,Y)) &= \mathbb{E}\Big(\mathbb{E}(X \mid Y, \min(X,Y)) \;\Big|\; \min(X,Y)\Big) \\ &= \mathbb{E}\Big(\frac{1+Y}{2}\,\mathbb{1}\{Y = \min(X,Y)\} + \min(X,Y)\,\mathbb{1}\{Y > \min(X,Y)\} \;\Big|\; \min(X,Y)\Big) \\ &= \mathbb{E}\Big(\min(X,Y) + \frac{1-\min(X,Y)}{2}\,\mathbb{1}\{Y = \min(X,Y)\} \;\Big|\; \min(X,Y)\Big)\end{aligned}$$

$$= \frac{1}{4} + \frac{3}{4}\min(X, Y).$$

16.2 Convergence Concepts

Example 16.10. Consider the following sequence of functions. For each $m \geq 1$, and each $0 \leq k \leq m$, consider a function which is 1 on $k/(m+1) < t \leq (k+1)/(m+1)$ and 0 elsewhere. Let

$$n = \frac{m(m+1)}{2} + k.$$

Lecture 17

October 24

17.1 Convergence Concepts

Issue: A random variable X is a *function* on $(\Omega, \mathcal{F}, \mathbb{P})$, i.e., $X(\omega) \in \mathbb{R}$ for each $\omega \in \mathbb{R}$. So, given a sequence $(X_1, X_2, \dots)$ of random variables, the asymptotic behavior of $(X_1(\omega), X_2(\omega), \dots)$ might be different for different ω.

17.1.1 Almost Sure Convergence

Definition 17.1. We say $(X_n,\ n \geq 1)$ **converges almost surely** (or **almost everywhere**) to the extended real-valued random variable X if $\mathbb{P}(\{\omega : X_n(\omega) \text{ converges to } X(\omega)\}) = 1$. One writes

$$X_n \xrightarrow{\text{a.s.}} X.$$

Example 17.2. Suppose you start with \$1. Toss a fair coin: if heads, your wealth triples, and if tails your wealth becomes 0. Repeat (with independent coin tosses). Let $X_0 = 1$, X_n is the wealth after n coin tosses. Here, $X_n \xrightarrow{\text{a.s.}} 0$, but

$$\mathbb{E}[X_n] = \left(\frac{3}{2}\right)^n \to \infty \qquad \text{as } n \to \infty.$$

17.1.2 Convergence in Probability

Definition 17.3 (Convergence in Probability)**.** Given $(X_n,\ n \geq 1)$, a sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, we will say X_n **converges in probability to** X if, for every $\varepsilon > 0$, $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$. We write

$$X_n \xrightarrow{\mathbb{P}} X.$$

In 17.2, $X_n \xrightarrow{\mathbb{P}} 0$ also. Let us check. Fix $\varepsilon > 0$. Compute

$$\begin{aligned}
\mathbb{P}(|X_n - 0| > \varepsilon) &= \mathbb{P}(|X_n| > \varepsilon) \\
&= \mathbb{P}(X_n \neq 0) \qquad \text{if } \varepsilon < 1 \\
&= \frac{1}{2^n} \to 0 \qquad \text{as } n \to \infty
\end{aligned}$$

in this example.

Example 17.4 (An Example where $X_n \xrightarrow{\mathbb{P}} 0$ but $X_n \not\xrightarrow{\text{a.s.}} 0$)**.** Pick a point uniformly at random on $(0,1]$, call it U. For each n, write it as $m(m+1)/2+k$ for $m \geq 0$, $1 \leq k \leq m+1$. Let:

$$X_n = \begin{cases} 1, & \dfrac{k-1}{m+1} < U \leq \dfrac{k}{m+1} \\ 0, & \text{elsewhere} \end{cases}$$

Here, $X_n \xrightarrow{\mathbb{P}} 0$ but $\mathbb{P}(\limsup_{n\to\infty} X_n = 1) = 1$, $\mathbb{P}(\liminf_{n\to\infty} X_n = 0) = 1$, so $X_n \not\xrightarrow{\text{a.s.}} 0$.

However, we have:

Theorem 17.5. $X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{\mathbb{P}} X$.

Proof. Assume $X_n \xrightarrow{\text{a.s.}} X$. Fix $\varepsilon > 0$. Define the event $A_n \triangleq \{|X_n - X| > \varepsilon\}$. Consider

$$\limsup_{n\to\infty} A_n \triangleq \bigcap_{m\geq 1}\Big(\bigcup_{k\geq m} A_k\Big).$$

Since $X_n \xrightarrow{\text{a.s.}} X$, $\mathbb{P}(\limsup_{n\to\infty} A_n) = 0$. But, $\mathbb{P}(\limsup_{n\to\infty} A_n) \geq \limsup_{n\to\infty} \mathbb{P}(A_n)$. To see this, $\limsup_{n\to\infty}\mathbb{P}(A_n) = \inf_{m\geq 1}\sup_{k\geq m}\mathbb{P}(A_k)$. But,

$$\begin{aligned}\mathbb{P}\Big(\bigcup_{k\geq m} A_k\Big) &\geq \sup_{k\geq m}\mathbb{P}(A_k) \\ &\geq \limsup_{n\to\infty}\mathbb{P}(A_n)\end{aligned}$$

and $(\bigcup_{k\geq m} A_k,\ m \geq 1)$ is a decreasing sequence of events, so

$$\mathbb{P}\Big(\limsup_{n\to\infty} A_n\Big) = \lim_{m\to\infty}\mathbb{P}\Big(\bigcup_{k\geq m} A_k\Big) \geq \limsup_{n\to\infty}\mathbb{P}(A_n).$$

Hence, $\lim_{n\to\infty}\mathbb{P}(A_n)$ exists and equals 0, i.e., $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$. Since this holds for each $\varepsilon > 0$, we have $X_n \xrightarrow{\mathbb{P}} X$. □

Note: Fix $\varepsilon > 0$ and let $A_n = \{|X_n - X| > \varepsilon\}$. $X_n \xrightarrow{\text{a.s.}} X \implies \mathbb{1}_{A_n} \xrightarrow{\text{a.s.}} 0$.

17.1.3 Convergence in Distribution

A third important notion is **convergence in distribution**. This is a property of the sequence of CDFs $(F_{X_n},\ n \geq 1)$, where $F_{X_n}(x) = \mathbb{P}(X_n \leq x)$. We would like to capture the notion of $F_{X_n}(x)$ converging to $F_X(x)$ at each x. However, this has difficulties because it is *not* consistent with the convergence of constants.

Example 17.6. Let

$$X_n = 1 + \frac{1}{n}$$

(constant RVs). Let $X = 1$. Then, $F_{X_n}(1) = 0$ for each n, but $F_X(1) = 1$.

Definition 17.7. We say X_n converges in distribution to X if $F_{X_n}(x) \to F_X(x)$ as $n \to \infty$ for each x

at which $F_X(x)$ is continuous. We write

$$X_n \xrightarrow{\mathsf{d}} X.$$

This turns out to be equivalent to $\Phi_{X_n}(\theta) \to \Phi_X(\theta)$ for all $\theta \in \mathbb{R}$, where $\Phi_X(\theta) \triangleq \mathbb{E}[\mathrm{e}^{\mathrm{i}\theta X}]$ and $\mathrm{i} = \sqrt{-1}$.

Since $X_n \xrightarrow{\mathsf{d}} X$ does not even require that the X_n are jointly defined, we can talk about $X_n \xrightarrow{\mathsf{d}} X$ even if the question "Does $X_n \xrightarrow{\mathbb{P}} X$" does not make sense. More generally, $X_n \xrightarrow{\mathsf{d}} X \not\Longrightarrow X_n \xrightarrow{\mathbb{P}} X$ even when $(X_n,\ n \geq 1)$ are jointly defined.

Example 17.8. Pick $U \sim \text{Uniform}[0,1]$. Let $X_n = U$ for n even and $X_n = 1 - U$ for n odd. Hence, X_n does not converge either in probability or a.s.

But, we have:

Proposition 17.9. $X_n \xrightarrow{\mathbb{P}} X \implies X_n \xrightarrow{\mathsf{d}} X$.

Proof. Assume that we are given $(X_n,\ n \geq 1)$ such that $X_n \xrightarrow{\mathbb{P}} X$.

$$\begin{aligned} F_X(x+\varepsilon) &= \mathbb{P}(X \leq x+\varepsilon) \\ &\geq \mathbb{P}(X_n \leq x) - \mathbb{P}(|X_n - X| > \varepsilon). \end{aligned}$$

Since $X_n \xrightarrow{\mathbb{P}} X$, $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$. Also,

$$\begin{aligned} F_{X_n}(x) &= \mathbb{P}(X_n \leq x) \\ &\geq \mathbb{P}(X \leq x - \varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon). \end{aligned}$$

So,

$$\begin{aligned} F_X(x+\varepsilon) &\geq \limsup_{n\to\infty} F_{X_n}(x), \\ \liminf_{n\to\infty} F_{X_n}(x) &\geq F_X(x-\varepsilon). \end{aligned}$$

So, if x is a continuity point of X, then $\lim_{n\to\infty} F_{X_n}(x)$ exists and equals $F_X(x)$. □

$$X_n \xrightarrow{\text{a.s.}} X \implies X_n \xrightarrow{\mathbb{P}} X \implies X_n \xrightarrow{\mathsf{d}} X$$

$$X_n \xrightarrow{\text{a.s.}} X \not\Longleftarrow X_n \xrightarrow{\mathbb{P}} X \not\Longleftarrow X_n \xrightarrow{\mathsf{d}} X$$

17.1.4 Convergence in L^p

Definition 17.10. For $\infty > p \geq 1$, $\|X - Y\|_p \triangleq \mathbb{E}[|X-Y|^p]^{1/p}$, called the ***p*-norm distance** (**L^p distance**).

The most important case is $p = 2$.

Definition 17.11. We say $(X_n,\ n \geq 1)$ **converges in mean square** to X (write $X_n \xrightarrow{\text{m.s.}} X$) if

$$\mathbb{E}[|X_n - X|^2]^{1/2} \to 0 \qquad \text{as } n \to \infty.$$

We have

$$X_n \xrightarrow{\text{m.s.}} X \implies X_n \xrightarrow{\mathbb{P}} X$$

but

$$X_n \xrightarrow{\text{a.s.}} X \not\Longrightarrow X_n \xrightarrow{\text{m.s.}} X \not\Longrightarrow X_n \xrightarrow{\text{a.s.}} X.$$

Example 17.12. Triple your wealth or go bankrupt with equal probability. $\mathbb{P}(X_n = (3/2)^n) = 1/2^n$, $\mathbb{P}(X_n = 0) = 1 - 1/2^n$. Then, $\mathbb{E}[X_n^2] = (3/2)^{2n}/2^n \to \infty$, but $X_n \xrightarrow{\text{a.s.}} 0$ and also $X_n \xrightarrow{\mathbb{P}} X$.

To show $X_n \xrightarrow{\text{m.s.}} X \implies X_n \xrightarrow{\mathbb{P}} X$,

$$\begin{aligned}\mathbb{P}(|X_n - X| > \varepsilon) &\leq \frac{\mathbb{E}[|X_n - X|^2\, \mathbb{1}\{|X_n - X| > \varepsilon\}]}{\varepsilon^2} \\ &\leq \frac{\mathbb{E}[|X_n - X|^2]}{\varepsilon^2}.\end{aligned}$$

Lecture 18

October 26

18.1 Limit Theorems

Theorem 18.1 (Weak Law of Large Numbers)**.** *If $X_1, X_2, \dots$ are i.i.d. real-valued, $\mathbb{E}[|X_1|] < \infty$, then*

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow{\mathbb{P}} \mathbb{E}[X_1].$$

This is a corollary of the Strong Law of Large Numbers, which says that (under the same assumptions),

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow{\text{a.s.}} \mathbb{E}[X_1].$$

There is a proof by Etemadi.

Theorem 18.2 (Central Limit Theorem)**.** *If $X_1, X_2, \dots$ are i.i.d. real-valued, $\mathbb{E}[X_1] = a$, $\operatorname{var} X_1 = \sigma^2 < \infty$, then*

$$\frac{X_1 + \cdots + X_n - na}{\sqrt{n}\sigma} \xrightarrow{\mathsf{d}} \mathcal{N}(0,1),$$

where $\mathcal{N}(0,1)$ is the standard Gaussian density $\dfrac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}$ on $\mathbb{R}$.

This is proved in the handout assuming $\mathbb{E}[|X_1|^3] < \infty$.

18.2 Results that Allow the Exchange of a.s. Limits & Expectation

Recall the example where X_n is your wealth after n i.i.d. coin tosses when $X_0 = 1$, heads triples your wealth, and tails makes you bankrupt. $X_n \xrightarrow{\text{a.s.}} 0$ but $\mathbb{E}[X_n] \to \infty$ as $n \to \infty$. In fact,

$$\mathbb{E}[X_n] = \frac{3^n}{2^n}.$$

Then, $\mathbb{E}[\text{a.s.-}\lim_{n\to\infty} X_n] \ne \lim_{n\to\infty} \mathbb{E}[X_n]$.

Theorem 18.3 (Monotone Convergence Theorem)**.** *If $(X_n,\ n \ge 0)$ are non-negative and increase almost surely to their limit X, then $\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X]$.*

Lemma 18.4 (Fatou's Lemma)**.** *If $(X_n,\ n \geq 1)$ are non-negative, then*

$$\mathbb{E}\big[\liminf_{n\to\infty} X_n\big] \leq \liminf_{n\to\infty} \mathbb{E}[X_n].$$

Proof. $\inf_{k\geq m} X_k \leq X_\ell$ for all $\ell \geq m$, so $\mathbb{E}[\inf_{k\geq m} X_k] \leq \mathbb{E}[X_\ell]$ for all $\ell \geq m$. So,

$$\mathbb{E}\big[\inf_{k\geq m} X_k\big] \leq \inf_{\ell\geq m} \mathbb{E}[X_\ell]$$

for each $m \geq 1$, so by the MCT 18.3, the LHS converges to $\mathbb{E}[\liminf_{n\to\infty} X_n]$ and the RHS converges to $\liminf_{n\to\infty} \mathbb{E}[X_n]$. □

Here, $\liminf_{n\to\infty} X_n$ denotes the a.s. limit of $\inf_{k\geq m} X_k$ as $m \to \infty$.

Theorem 18.5 (Dominated Convergence Theorem)**.** *If $(X_n,\ n \geq 0)$ satisfies $X_n \xrightarrow{a.s.} X$ and there is some Y with $\mathbb{E}[|Y|] < \infty$ such that $|X_n| \leq Y$ for all large enough n, then $\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X]$. In fact, $\lim_{n\to\infty} \mathbb{E}[|X_n - X|] = 0$.*

In fact, one knows that "**uniform integrability**" is a necessary and sufficient condition to allow the interchange of expectation and a.s. limit. Something like the DCT 18.5 is more useful.

18.3 Martingales

Definition 18.6. Given a sequence of random variables $(X_0, X_1, X_2, \dots)$, a sequence of random variables $(M_0, M_1, M_2, \dots)$ is called a **martingale with respect to** $(X_0, X_1, X_2, \dots)$ if

1. for all $n \geq 0$, $\mathbb{E}[|M_n|] < \infty$ (not necessarily uniformly in n);
2. M_n is a deterministic function of $(X_0, \dots, X_n)$ for each n, i.e., $\mathbb{E}(M_n \mid X_0, X_1, \dots, X_n) = M_n$;
3. $\mathbb{E}(M_{n+1} \mid X_0, \dots, X_n) = M_n$ for all $n \geq 0$.

"Meaning": $(X_0, X_1, \dots, X_k)$ represents our "knowledge" at time k.

We will say $(M_n,\ n \geq 0)$ is a **martingale** if it is a martingale with respect to $(M_n,\ n \geq 0)$.

1. $\mathbb{E}[|M_n|] < \infty$.
2. M_n is automatically a function of $(M_0, \dots, M_n)$.
3. $\mathbb{E}(M_{n+1} \mid M_0, \dots, M_n) = M_n$.

Theorem 18.7 (Doob's Martingale Convergence Theorem)**.** *If $(M_n,\ n \geq 0)$ is a martingale with respect to $(X_n,\ n \geq 0)$ and there exists $K < \infty$ such that $\mathbb{E}[|M_n|] < K$ for all $n \geq 0$, then $(M_n,\ n \geq 0)$ converges a.s. as $n \to \infty$ to a limit, call it Z.*

(Of course, if there is uniform integrability, $\mathbb{E}[M_n] \to \mathbb{E}[Z]$ as $n \to \infty$. In fact, we have $\mathbb{E}[|M_n - Z|] \to 0$ as $n \to \infty$.)

Example 18.8. Suppose $X_1, X_2, X_3, \dots$ are i.i.d., $X_0 \perp\!\!\!\perp (X_1, X_2, \dots)$. Let $S_0 = X_0$ and also let $S_n = X_0 + X_1 + X_2 + \dots + X_n$, $n \geq 1$. If $\mathbb{E}[X_1] = a$, let $M_n \triangleq S_n - na$ for $n \geq 0$. Then, $(M_n,\ n \geq 0)$ is a martingale with respect to $(X_n,\ n \geq 0)$. Check $\mathbb{E}(S_{n+1} - (n+1)a \mid X_0, \dots, X_n) = S_n - na$.

Example 18.9. In a branching process, we have random variables

$$\begin{matrix} X_{0,1} & X_{0,2} & \cdots \\ \vdots & \vdots & \\ X_{n,1} & X_{n,2} & \cdots \\ \vdots & \vdots & \end{matrix}$$

where $X_{t,j}$ for $t \geq 0$, $j \geq 1$, i.i.d., non-negative integer-valued, represents the number of children of the jth individual alive at time t. Y_0 is the number of individuals at time 0, and:

$$\begin{aligned} Y_1 &= X_{0,1} + \cdots + X_{0,Y_0}, \\ &\vdots \\ Y_{n+1} &= X_{n,1} + \cdots + X_{n,Y_n}. \end{aligned}$$

Then, $\mathbb{E}(Y_{n+1} \mid Y_0, \ldots, Y_n) = \mu Y_n$, where $\mu = \mathbb{E}[X_{t,j}]$. Consequently, $\mathbb{E}[Y_{n+1}] = \mu\, \mathbb{E}[Y_n]$ so

$$\mathbb{E}[Y_n] = \mu^n\, \mathbb{E}[Y_0].$$

In the branching process context,

$$\Big(\frac{Y_n}{\mu^n},\ n \geq 0\Big)$$

is a martingale. Note that

$$\mathbb{E}\Big[\frac{Y_n}{\mu^n}\Big] = \frac{\mathbb{E}[Y_n]}{\mu^n} = \mathbb{E}[Y_0]$$

and the bound is uniform in n. By the Martingale Convergence Theorem 18.7, $\dfrac{Y_n}{\mu^n}$ converges a.s. as $n \to \infty$. Of course, this is true if $\mu < 1$. Before, we had

$$\begin{aligned} G_{Y_{n+1}}(s) = \mathbb{E}(s^{Y_{n+1}} \mid Y_n, \ldots, Y_0) &= \mathbb{E}(s^{X_{n,1}+\cdots+X_{n,Y_n}} \mid Y_n, \ldots, Y_0) \\ &= \sum_{k=0}^{\infty} G_X(s)^k \mathbb{P}(Y_n = k) = G_{Y_n}\big(G_X(s)\big). \end{aligned}$$

Also, $\mathbb{P}(Y_n = 0) = G_{Y_n}(0)$, so $\mathbb{P}(Y_n = 0) \to 1$ and so

$$\frac{Y_n}{\mu^n} \to 0 \qquad \text{a.s.}$$

Example 18.10. Given two probability distributions P and Q on any $(\Omega, \mathcal{F})$ such that $P(A) = 0$ implies $Q(A) = 0$ for all events A, one can define a random variable denoted $\dfrac{\mathrm{d}Q}{\mathrm{d}P}$ such that for any event A,

$$\mathbb{E}_P\Big[\frac{\mathrm{d}Q}{\mathrm{d}P}\mathbb{1}_A\Big] = Q(A),$$

where $\mathbb{E}_P$ denotes expectation with respect to P. $\dfrac{\mathrm{d}Q}{\mathrm{d}P}$ is called the **likelihood ratio** of Q with respect to P.

Concretely, suppose P and Q are on a finite set $\{x_1, \ldots, x_n\}$ (we need $P(x_j) = 0 \implies Q(x_j) = 0$).

Then,

$$\frac{\mathrm{d}Q}{\mathrm{d}P}(x_j) = \frac{Q(x_j)}{P(x_j)}.$$

Obviously, for any $A \subseteq \{x_1, \ldots, x_n\}$,

$$Q(A) = \sum_{x_j \in A} Q(x_j) = \sum_{x_j \in A} \frac{Q(x_j)}{P(x_j)} P(x_j).$$

Then,

$$\mathbb{E}_P\Big(\frac{Q(X_0, \ldots, X_{n+1})}{P(X_0, \ldots, X_{n+1})} \;\Big|\; X_0, \ldots, X_n\Big) = \frac{Q(X_0, \ldots, X_n)}{P(X_0, \ldots, X_n)}.$$

where $X_0, X_1, \ldots$ are taking values in a finite set, and we assume

$$P(x_0, \ldots, x_n) = 0 \implies Q(x_0, \ldots, x_n) = 0.$$

Lecture 19

October 31

19.1 Martingales

$(M_n,\ n \geq 0)$ is a **martingale with respect to** $(X_n,\ n \geq 0)$ if

(i) $\mathbb{E}[|M_n|] < \infty$ for all $n \geq 0$;

(ii) M_n is "**adapted** to" $(X_0, \dots, X_n)$ (i.e., a deterministic function of them)

(iii) $\mathbb{E}(M_{n+1} \mid X_0, \dots, X_n) = M_n$.

We say $(M_n,\ n \geq 0)$ is a **martingale** if it is a martingale with respect to itself.

If $(M_n,\ n \geq 0)$ is a martingale with respect to $(X_n,\ n \geq 0)$, then

$$\begin{aligned}\mathbb{E}(M_{n+1} \mid M_0, \dots, M_n) &= \mathbb{E}\big(\mathbb{E}(M_{n+1} \mid X_0, \dots, X_n, M_0, \dots, M_n) \bigm| M_0, \dots, M_n\big) \\ &= \mathbb{E}\big(\mathbb{E}(M_{n+1} \mid X_0, \dots, X_n) \bigm| M_0, \dots, M_n\big) \\ &= \mathbb{E}(M_n \mid M_0, \dots, M_n) = M_n\end{aligned}$$

which means $(M_n,\ n \geq 0)$ is a martingale.

We can have $(M_n,\ n \geq 0)$ be a martingale with respect to itself, but not with respect to $(X_n,\ n \geq 0)$.

Martingale Convergence Theorem: If $(M_n,\ n \geq 0)$ is a martingale and there exists $M < \infty$ such that $\mathbb{E}[|M_n|] \leq M < \infty$ for all $n \geq 0$, then $M_n \xrightarrow{\text{a.s.}} Z$ as $n \to \infty$ for some random variable Z. If $(M_n,\ n \geq 0)$ are dominated, i.e., $|M_n| \leq W$ for some random variable W, with $\mathbb{E}[W] < \infty$, then $\mathbb{E}[|M_n - Z|] \to 0$ as $n \to \infty$.

19.1.1 Branching Process

In the branching process,

$$\frac{X_n}{\mu^n} \xrightarrow{\text{a.s.}} Z \qquad \text{as } n \to \infty$$

where X_n is the number of individuals alive at time n in a branching process with mean number of children per stage μ.

19.1.2 Likelihood Ratio

Consider two possible probability models for observations $X_0, X_1, X_2, \dots$ drawn from a finite set $\mathcal{X}$, either $\mathbb{P}(X_0 = x_0, \dots, X_n = x_n) = p(x_1, \dots, x_n)$ or $\mathbb{P}(X_0 = x_0, \dots, X_n = x_n) = q(x_0, \dots, x_n)$ for $n \geq 0$, $(x_0, \dots, x_n) \in \mathcal{X}^{n+1}$. Then,

$$\ell(x_0, \dots, x_n) \triangleq \frac{q(x_0, \dots, x_n)}{p(x_0, \dots, x_n)}$$

is called the **likelihood ratio** at stage n. (We will assume that $p(x_0, \dots, x_n) = 0 \implies q(x_0, \dots, x_n) = 0$ for all n, for all $(x_0, \dots, x_n)$; we say Q is **absolutely continuous with respect to** P.) Let $(L_n,\ n \geq 0)$ be the sequence of random variables

$$L(X_0, \dots, X_n) \triangleq \ell(X_0, \dots, X_n) = \frac{q(X_0, \dots, X_n)}{p(X_0, \dots, X_n)}.$$

Then,

$$\begin{aligned}
\mathbb{E}_P(L_{n+1} \mid X_0, \dots, X_n) &= \mathbb{E}_P\Big(\frac{q(X_0, \dots, X_n)q(X_{n+1} \mid X_0, \dots, X_n)}{p(X_0, \dots, X_n)p(X_{n+1} \mid X_0, \dots, X_n)} \Bigm| X_0, \dots, X_n\Big) \\
&= L_n\, \mathbb{E}_P\Big(\frac{q(X_{n+1} \mid X_0, \dots, X_n)}{p(X_{n+1} \mid X_0, \dots, X_n)} \Bigm| X_0, \dots, X_n\Big) \\
&= L_n \sum_{x \in \mathcal{X}} \frac{q(x \mid X_0, \dots, X_n)}{p(x \mid X_0, \dots, X_n)} p(x \mid X_0, \dots, X_n) \\
&= L_n.
\end{aligned}$$

Notice $\mathbb{E}_P[L_{n+1}] = \mathbb{E}_P[\mathbb{E}_P(L_{n+1} \mid X_0, \dots, X_n)] = \mathbb{E}[L_n]$, so the expectation stays the same (true for any martingale). Here,

$$\begin{aligned}
L_0(X_0) &= \frac{q(X_0)}{p(X_0)}, \\
\mathbb{E}_P[L_0] &= \sum_{x \in \mathcal{X}} \frac{q(x)}{p(x)} p(x) = 1.
\end{aligned}$$

Hence, we can use the Martingale Convergence Theorem 18.7. So, $L_n \xrightarrow{\text{a.s.}} Z$ as $n \to \infty$ (for some random variable Z). Typically, even though $\mathbb{E}[L_n] = 1$ for all n, Z will be 0 a.s., e.g., for the i.i.d. case.

Example 19.1. If $\mathcal{X} = \{0, 1\}$, then

$$\begin{aligned}
p(x_0, \dots, x_n) &= \alpha^{k(x_0, \dots, x_n)}(1-\alpha)^{(n+1)-k(x_0, \dots, x_n)} \\
q(x_0, \dots, x_n) &= \beta^{k(x_0, \dots, x_n)}(1-\beta)^{(n+1)-k(x_0, \dots, x_n)}.
\end{aligned}$$

where $k(x_0, \dots, x_n)$ is the number of 1s in $(x_0, \dots, x_n)$. Then,

$$\begin{aligned}
L_n &= \Big(\frac{\beta}{\alpha}\Big)^{k(X_0, \dots, X_n)} \Big(\frac{1-\beta}{1-\alpha}\Big)^{(n-1)-k(X_0, \dots, X_n)} \\
&= \Big\{\Big(\frac{\beta}{\alpha}\Big)^{k(X_0, \dots, X_n)/(n+1)} \Big(\frac{1-\beta}{1-\alpha}\Big)^{1-k(X_0, \dots, X_n)/(n+1)}\Big\}^{n+1} \\
&= \mathrm{e}^{(k(X_0, \dots, X_{n+1})/(n+1))\log(\beta/\alpha) + (1-k(X_0, \dots, X_n)/(n+1))\log((1-\beta)/(1-\alpha))}.
\end{aligned}$$

It turns out that if $\alpha \neq \beta$, then

$$\alpha \log \frac{\beta}{\alpha} + (1-\alpha) \log \frac{1-\beta}{1-\alpha} < 0$$

(in fact,

$$\alpha \log \frac{\alpha}{\beta} + (1-\alpha) \log \frac{1-\alpha}{1-\beta}$$

is called the **relative entropy** of α with respect to β and is positive if $\alpha < \beta$).

19.1.3 Supermartingales & Submartingales

A **supermartingale** $(M_n,\ 0)$ **with respect to** $(X_n,\ n \geq 0)$ is a sequence of random variables satisfying

1. $\mathbb{E}[|M_n|] < \infty$;
2. M_n is adapted to $(X_0, \dots, X_n)$ for all $n \geq 0$;
3. $\mathbb{E}(M_{n+1} \mid X_0, \dots, X_n) \leq M_n$.

If we instead in 3 we have $\mathbb{E}(M_{n+1} \mid X_0, \dots, X_n) \geq M_n$, we say that $(M_n,\ n \geq 0)$ is a **submartingale with respect to** $(X_n,\ n \geq 0)$.

There is a so-called Submartingale Convergence Theorem. It says that if $(V_n,\ n \geq 0)$ is a submartingale (i.e., with respect to itself) and $\mathbb{E}[|V_n|] \leq M < \infty$ for all n, then $V_n \xrightarrow{\text{a.s.}} Z$ for some Z with $\mathbb{E}[|Z|] < \infty$.

Theorem 19.2. *Let $(Y_n,\ n \geq 0)$ be a supermartingale with respect to $(X_n,\ n \geq 0)$. Let $(H_n,\ n \geq 0)$ be an arbitrary non-negative sequence of random variables such that H_n is a deterministic function of $(X_0, \dots, X_n)$. Fix W_0, a RV dependent on X_0, and define $W_n = W_0 + \sum_{m=1}^n H_m(Y_m - Y_{m-1})$ for $n \geq 1$. Then, $(W_n,\ n \geq 0)$ is also a supermartingale.*

Proof.

$$\begin{aligned}\mathbb{E}(W_{n+1} \mid X_0, \dots, X_n) &= \mathbb{E}\big(W_n + H_{n+1}(Y_{n+1} - Y_n) \bigm| X_0, \dots, X_n\big) \\ &= W_n + H_{n+1}\,\mathbb{E}(Y_{n+1} - Y_n \mid X_0, \dots, X_n) \leq W_n.\end{aligned}$$ □

Consequence. Let T be a stopping time of $(X_n,\ n \geq 0)$ (i.e., for each k the indicator $\mathbb{1}\{T = k\}$ is a function of $(X_0, \dots, X_k)$). If $(Y_n,\ n \geq 0)$ is a supermartingale with respect to $(X_n,\ n \geq 0)$, let $(Y_{T \wedge n},\ n \geq 0)$ denote the stopped supermartingale.

$$Y_{T \wedge n} = \begin{cases} Y_T, & \text{if } T = n \\ Y_n, & \text{if } T > n \end{cases}$$

Proposition 19.3. *$(Y_{T \wedge n},\ n \geq 0)$ is also a supermartingale with respect to $(X_n,\ n \geq 0)$.*

Proof. Taking $H_m = \mathbb{1}\{T \geq m\}$ (which is a function of $(X_0, \dots, X_{m-1})$), this is a consequence of 19.2 $(W_0 = Y_0)$. □

This gives a proof of "**Wald's Identity**".

Proposition 19.4 (Wald's Identity)**.** *If $X_1, X_2, \dots$ are i.i.d. with $\mathbb{E}[|X_1|] < \infty$ and $\mathbb{E}[X_1] = \mu$, let $S_n = S_0 + X_1 + \dots + X_n$ where $S_0 \perp\!\!\!\perp (X_1, X_2, \dots)$. Then, $\mathbb{E}[S_{T \wedge n} - \mu(T \wedge n)] = \mathbb{E}[S_0]$ for every stopping time T.*

Proof. $(S_0, S_1 - \mu, S_2 - 2\mu, \dots)$ is a martingale so here, it and its negative are supermartingales. Hence, the expectation of its stopped version equals the initial expectation. □

19.1.4 Markov Chains

$(X_n,\ n \geq 0)$ is a discrete-time Markov chain with state space $\mathcal{S}$, transition probability matrix P. We looked at stopping times like $\inf\{n \geq 0 : X_n \in A\}$ for $A \subseteq S$. Let $h(x) \triangleq \mathbb{P}(V_A < \infty \mid X_0 = x)$. We can consider $(h(X_n),\ n \geq 0)$. This is a supermartingale with respect to $(X_n,\ n \geq 0)$. This comes from $h(x) = \sum_{y \in \mathcal{S}} p(x, y)h(y)$ for $x \notin A$ and $1 = h(x) \geq \sum_{y \in \mathcal{S}} p(x, y)h(y)$ for $x \in A$. This gives

$$\mathbb{E}\big(h(X_{n+1}) \bigm| X_0, \dots, X_n\big) = \mathbb{E}\big(h(X_{n+1}) \bigm| X_n\big) \leq h(X_n)$$

from the Markov property. Also, $(W_n,\ n \geq 0) \triangleq (h(X_{n \wedge V_A}),\ n \geq 0)$ is a martingale with respect to $(X_n,\ n \geq 0)$.

Let $u(x) \triangleq \mathbb{E}[V_A \mid X_0 = x]$.

Claim: $(u(X_n) + n,\ n \geq 0)$ is a submartingale with respect to $(X_n,\ n \geq 0)$. This comes from

$$u(x) \leq 1 + \sum_{y \in \mathcal{S}} p(x, y) u(y).$$

19.1.5 Azuma-Hoeffding Inequality

Theorem 19.5 (Azuma-Hoeffding Inequality)**.** *Let $(M_n,\ n \geq 0)$ be a martingale. Suppose*

$$|M_{n+1} - M_n| \leq c_{n+1}$$

for all $n \geq 0$ for some constants $c_1, c_2, \dots$. Then, $\mathbb{P}(|M_n - M_0| \geq \lambda) \leq 2\mathrm{e}^{-\lambda^2/(2\sum_{k=1}^{n} c_k^2)}$.

Lecture 20

November 2

Lecturer: Professor Thomas Courtade

20.1 Azuma's Inequality

$(M_n)_{n\geq 0}$ is a **martingale** with respect to $(X_n)_{n\geq 0}$ if $\mathbb{E}(M_{n+1} \mid X_0, \dots, X_n) = M_n$. Instead of writing $X_0, \dots, X_n$, one can write $\mathcal{F}_n$ for the "**filtration**".

An example is when $X_i \in \{-1, +1\}$, taking on each value with probability $1/2$, and then $M_n = \sum_{i\leq n} X_i$.

Theorem 20.1 (Azuma's Inequality)**.** *If $(M_n)_{n\geq 0}$ is a martingale with $|M_k - M_{k-1}| \leq c_k$ for all k, then for any integer $n \geq 1$ and $t > 0$,*

$$\mathbb{P}(M_n - M_0 \geq t) \leq \exp\Big(-\frac{t^2}{2\sum_{k=1}^n c_k^2}\Big).$$

Proof. Assume each c_k is 1 and $(M_n)_{n\geq 0}$ is a martingale with respect to $(X_n)_{n\geq 0}$, and $M_0 = 0$. We would like to show

$$\mathbb{P}(M_n \geq \lambda\sqrt{n}) \leq \exp\Big(-\frac{\lambda^2}{2}\Big), \qquad \lambda > 0.$$

Let $Y_i = M_i - M_{i-1}$. By the bounded differences condition, $|Y_i| \leq 1$ for all i. Then,

$$\mathbb{E}(Y_i \mid X_{i-1}, \dots, X_0) = \mathbb{E}(M_i - M_{i-1} \mid X_{i-1}, \dots, X_0) = M_{i-1} - M_{i-1} = 0.$$

By 20.2,

$$\mathbb{E}(\mathrm{e}^{\alpha Y_i} \mid X_{i-1}, \dots, X_0) \leq h\big(\mathbb{E}(Y_i \mid X_{i-1}, \dots, X_0)\big) = \underbrace{\frac{\mathrm{e}^{\alpha} + \mathrm{e}^{-\alpha}}{2}}_{\cosh\alpha} \leq \mathrm{e}^{\alpha^2/2}.$$

Thus,

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{\alpha M_n}] &= \mathbb{E}\Big[\prod_{i=1}^n \mathrm{e}^{\alpha Y_i}\Big] \\ &= \mathbb{E}\Big[\mathbb{E}\Big(\prod_{i=1}^n \mathrm{e}^{\alpha Y_i} \;\Big|\; X_{n-1}, \dots, X_0\Big)\Big]\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\Big[\prod_{i=1}^{n-1} \mathrm{e}^{\alpha Y_i}\, \mathbb{E}(\mathrm{e}^{\alpha Y_n} \mid X_{n-1}, \dots, X_0)\Big] \\
&\le \mathrm{e}^{\alpha^2/2}\, \mathbb{E}\Big[\prod_{i=1}^{n-1} \mathrm{e}^{\alpha Y_i}\Big] \le \mathrm{e}^{n\alpha^2/2}.
\end{aligned}
$$

So,

$$
\begin{aligned}
\mathbb{P}(M_n \ge \lambda\sqrt{n}) = \mathbb{P}(\mathrm{e}^{\alpha M_n} \ge \mathrm{e}^{\alpha\lambda\sqrt{n}}) &\le \frac{\mathbb{E}[\mathrm{e}^{\alpha M_n}]}{\mathrm{e}^{\alpha\lambda\sqrt{n}}} \\
&\le \mathrm{e}^{n\alpha^2/2 - \alpha\lambda\sqrt{n}} \\
&= \mathrm{e}^{-\lambda^2/2}
\end{aligned}
$$

by taking $\alpha = \lambda/\sqrt{n}$. □

Lemma 20.2. *For $y \in [-1, 1]$,*

$$
\mathrm{e}^{\alpha y} \le \underbrace{\frac{\mathrm{e}^{\alpha} + \mathrm{e}^{-\alpha}}{2} + \frac{\mathrm{e}^{\alpha} - \mathrm{e}^{-\alpha}}{2} y}_{h(y)}, \qquad \alpha > 0.
$$

If $|Y| \le 1$, then $\mathbb{E}[\mathrm{e}^{\alpha Y}] \le h(\mathbb{E}[Y])$.

Corollary 20.3. *If $(M_n)_{n\ge 0}$ is a martingale, then*

$$
\mathbb{P}(|M_n - M_0| \ge \lambda\sqrt{n}) \le 2\exp\Big(-\frac{\lambda^2}{2}\Big).
$$

Proof. If $(M_n)_{n\ge0}$ is a martingale, so is $(-M_n)_{n\ge0}$. □

Example 20.4. Let $X_0 = 0$, $(X_k)_{k\ge1}$ are i.i.d. Bernoulli(1/2) taking values ±1. Let $M_n = \sum_{k=0}^{n} X_k$.

$$
\frac{M_n}{\sqrt{n}} \to \mathcal{N}(0,1) \qquad \text{in distribution.}
$$

Then,

$$
\begin{aligned}
\mathbb{P}(M_n \ge \sqrt{2(1+\varepsilon) n \ln n}) &\le \exp\Big(-\frac{2(1+\varepsilon)}{2}\ln n\Big) \\
&= n^{-(1+\varepsilon)}.
\end{aligned}
$$

By the Borel-Cantelli Lemma 2.2,

$$
\mathbb{P}\Big(\frac{M_n}{\sqrt{n}} \ge \sqrt{2(1+\varepsilon)\ln n} \text{ i.o.}\Big) = 0.
$$

The LIL says

$$
\limsup_{n\to\infty} \frac{M_n}{\sqrt{n \ln\ln n}} = \sqrt{2} \qquad \text{a.s.}
$$

Example 20.5. Let $G = (V, E)$ be a graph, where V is the set of vertices and E is the set of edges. Let $\chi(G)$ be the smallest number of colors needed to color the vertices so that no edge is monochromatic. $\mathcal{G}(n, p)$ is the Erdos-Renyi random graph ensemble. If $G \sim \mathcal{G}(n, p)$ means I take a "random" graph on n vertices, picking each edge to be present with probability p independent of all others.

Vertex-Exposure Martingale: Let f be any graph-theoretic function (e.g., $f = \chi$). Let $\mathcal{G}(n, p)$ be the underlying probability space. Define $M_i(H) = \mathbb{E}[f(G) \mid \text{for } x, y \leq i, \{x, y\} \in G \iff \{x, y\} \in H]$. I claim that $(M_n)_{n\geq 0}$ is a martingale with respect to $(G_i)_{i\geq 0}$, where G_i is the subgraph G induces on the first i vertices. Note that $M_n(H) = f(H)$ and $M_0(H) = \mathbb{E}_{G\sim\mathcal{G}(n,p)}[f(G)]$. Why is this a martingale? (This actually follows from the Doob Martingale.)

$$\begin{aligned} M_{i-1}(H) &= \frac{1}{N_{i-1}} \sum_{G : G_{i-1} = H_{i-1}} f(G) \\ &= \frac{1}{N_{i-1}} \sum_{\tilde{H}_i : (\tilde{H}_i)_{i-1} = H_{i-1}} \sum_{G : G_i = \tilde{H}_i} f(G) \\ &= \mathbb{E}(M_i \mid G_{i-1}, \dots, G_0)(H). \end{aligned}$$

In general, if $f(X_1, X_2, \dots, X_n)$ is a function of random variables, then

$$B_i = \mathbb{E}\big(f(X) \bigm| X_1, \dots, X_i\big)$$

forms a martingale.

$$\mathbb{E}(B_i \mid X_1, \dots, X_{i-1}) = \mathbb{E}\big(f(X_1, \dots, X_n) \bigm| X_1, \dots, X_{i-1}\big) = B_{i-1}.$$

Theorem 20.6. *Let $G \sim \mathcal{G}(n, p)$. Then,*

$$\mathbb{P}(|\chi(G) - \mathbb{E}[\chi(G)]| \geq \lambda\sqrt{n}) \leq 2\exp\Big(-\frac{\lambda^2}{2}\Big).$$

Proof. $|M_i - M_{i-1}| \leq 1$. □

Lecture 21

November 7

Lecturer: Professor Thomas Courtade

21.1 Renewal Processes

21.1.1 Limit Theorems

Theorem 21.1. *For a renewal process N with average interarrival time $\mu < \infty$,*

$$\frac{N(t)}{t} \to \frac{1}{\mu}, \qquad \text{with probability } 1.$$

For discrete-time Markov chains, the number of arrivals to state j, $N_j(n)$, for time n, satisfies

$$\frac{N_j(n)}{n} \to \frac{1}{\mu_{j,j}}, \qquad \text{with probability 1.}$$

Proof of 21.1. The SLLN says

$$\frac{S_n}{n} \to \mu \qquad \text{with probability 1.}$$

From

$$\frac{N(t)+1}{N(t)+1} \frac{N(t)}{S_{N(t)+1}} \leq \frac{N(t)}{t} \leq \frac{N(t)}{S_{N(t)}}$$

and by the SLLN, both sides of the bound converge to $1/\mu$ with probability 1, since $N(t) \to \infty$ with probability 1. □

Note. This actually holds even if $\mu = \infty$, just truncate. The new renewal process is $\tilde{N}$, where

$$\tilde{X}_n = \min(B, X_n).$$

Then,

$$\frac{N(t)}{t} \leq \frac{\tilde{N}(t)}{t} \to \frac{1}{\tilde{\mu}}, \qquad \text{with probability 1.}$$

Take $B \to \infty$.

In particular,

$$\mathbb{P}\Big(\Big|\frac{N(t)}{t} - \frac{1}{\mu}\Big| > \varepsilon\Big) \to 0 \qquad \text{as } t \to \infty.$$

If $\mathbb{E}[X_i] = \mu$ and $\operatorname{var} X_i = \sigma^2$, then

$$\lim_{t\to\infty} \mathbb{P}\Big(\frac{N(t) - t/\mu}{\sigma\mu^{-3/2}\sqrt{t}} < \alpha\Big) = \Phi(\alpha).$$

21.1.2 Renewal Reward Process

In a Markov chain, imagine that whenever I enter state X_i, I collect the reward $r(X_i)$, then

$$\frac{1}{n}\sum_{i=1}^{n} r(X_i) \to \mathbb{E}_\pi[r(X)],$$

provided π exists and is unique.

R is a non-negative process that only depends on the "location within the current arrival period and the length of the period". Define $R_n = \int_{S_{n-1}}^{S_n} R(\tau)\,\mathrm{d}\tau$, the reward accumulated during the nth renewal period. Note: $R_1, R_2, R_3, \ldots$ are i.i.d. random variables.

Theorem 21.2. *Let $R(t)$, $t \geq 0$, be a renewal reward process with average interarrival time $\mu < \infty$.*

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t R(\tau)\,\mathrm{d}\tau = \frac{\mathbb{E}[R_1]}{\mu} \qquad \textit{with probability } 1.$$

Proof.

$$\frac{N(t)}{N(t)}\frac{\sum_{i=1}^{N(t)} R_i}{t} \leq \frac{\int_0^t R(\tau)\,\mathrm{d}\tau}{t} \leq \frac{\sum_{i=1}^{N(t)+1} R_i}{t}\frac{N(t)+1}{N(t)+1}$$

and we know

$$\begin{aligned} \frac{N(t)+1}{t} &\to \frac{1}{\mu}, && \text{with probability 1,} \\ \frac{\sum_{i=1}^{N(t)+1} R_t}{N(t)+1} &\to \mathbb{E}[R_1], && \text{with probability 1.} \end{aligned}$$

□

21.1.3 Residual Life Process

In the residual life process,

$$\mathbb{E}[R_n \mid X_n = x] = \frac{x^2}{2} \implies \mathbb{E}[R_n] = \frac{\mathbb{E}[X_n^2]}{2},$$

so

$$\lim_{t\to\infty} \frac{1}{t}\int_0^\tau R(\tau)\,\mathrm{d}\tau = \frac{\mathbb{E}[X_1]^2}{2\,\mathbb{E}[X_1]} = \frac{1}{2}\frac{\operatorname{var} X_1}{\mathbb{E}[X_1]} + \frac{1}{2}\,\mathbb{E}[X_1] \qquad \text{with probability 1.}$$

21.1.4 Time Average Duration

In the time average duration process,

$$\mathbb{E}[R_n \mid X_n = x] = x^2 \implies \mathbb{E}[R_n] = \mathbb{E}[X_n^2],$$

so

$$\lim_{t\to\infty} \frac{1}{t} \int_0^{\tau} R(\tau)\,\mathrm{d}\tau = \frac{\mathbb{E}[X_1]^2}{\mathbb{E}[X_1]} = \frac{\operatorname{var} X_1}{\mathbb{E}[X_1]} + \mathbb{E}[X_1] \qquad \text{with probability 1.}$$

21.1.5 Elementary Renewal Theorem

Wald's Equality: Let $X_1, X_2, \ldots$ be a sequence of random variables, each having mean μ. If N is a stopping time, with $\mathbb{E}[N] < \infty$, then $\mathbb{E}[\sum_{i=1}^N X_i] = \mathbb{E}[X]\,\mathbb{E}[N] = \mu\,\mathbb{E}[N]$.

We have seen that

$$\frac{N(t)}{t} \to \frac{1}{\mu}, \qquad \text{with probability 1.}$$

Theorem 21.3 (Elementary Renewal Theorem)**.**

$$\frac{\mathbb{E}[N(t)]}{t} \to \frac{1}{\mu}.$$

Proof.

$$\begin{aligned} \mathbb{E}[S_{N(t)+1}] &= \mathbb{E}\Big[\sum_{i=1}^{N(t)+1} X_i\Big] \\ &= \mu(\mathbb{E}[N(t)] + 1), \\ \frac{\mathbb{E}[N(t)]}{t} &= \frac{1}{t\mu}\,\mathbb{E}[S_{N(t)+1}] - \frac{1}{t} \\ &\geq \frac{1}{\mu} - \frac{1}{t}, \\ \liminf_{t\to\infty} \frac{\mathbb{E}[N(t)]}{t} &\geq \frac{1}{\mu}. \end{aligned}$$

Let $\tilde{X}_i = \min(B, X_i)$, $\tilde{S}_{N(t)+1} \leq t + B$.

$$\begin{aligned} \frac{\mathbb{E}[N(t)]}{t} \leq \frac{\mathbb{E}[\tilde{N}(t)]}{t} &\leq \frac{1 + B/t}{\tilde{\mu}_B} - \frac{1}{t}, \\ \limsup_{t\to\infty} \frac{\mathbb{E}[N(t)]}{t} &\leq \frac{1}{\tilde{\mu}_B} \xrightarrow[B\to\infty]{} \frac{1}{\mu}. \end{aligned}$$

□

Lecture 22

November 9

Lecturer: Professor Thomas Courtade

22.1 Renewal Theory

For a sequence of interarrival times $(X_i)_{i\in\mathbb{N}}$, i.i.d. with mean $\mu < \infty$, we saw last time that

$$\frac{N(t)}{t} \to \frac{1}{\mu}, \qquad \text{a.s.,} \tag{SL}$$

$$\frac{\mathbb{E}[N(t)]}{t} \to \frac{1}{\mu}, \tag{ERT}$$

$$\frac{1}{t}\int_0^t R(\tau)\,\mathrm{d}\tau \to \frac{\mathbb{E}[R_1]}{\mu}, \qquad \text{a.s.} \tag{RRT}$$

Claim: $N(t)+1$ is a stopping time for $(X_i)_{i\in\mathbb{N}}$.

$$\begin{aligned} \{N(t)+1 = n\} &= \{N(t) = n-1\} \\ &= \begin{matrix} (n-1)\text{th arrival came before time } t \\ n\text{th arrival came after time } t \end{matrix} \\ &= \Big\{\sum_{i=1}^{n-1} X_i \le t\Big\} \cup \Big\{\sum_{i=1}^{n} X_i > t\Big\}. \end{aligned}$$

Example 22.1 (Alternating Renewal Process)**.** Let $(X_i)_{i\in\mathbb{N}}$ be i.i.d. with mean μ_X, $(T_i)_{i\in\mathbb{N}}$ be i.i.d. with mean μ_T.

Q: What is the fraction of time spent in the "X" state?

$$\frac{1}{t}\int_0^t R(\tau)\,\mathrm{d}\tau \xrightarrow{\text{a.s.}} \frac{\mathbb{E}[R_1]}{\mu} = \frac{\mu_X}{\mu_X + \mu_Y}.$$

22.2 Application to Queueing Theory

Consider a G/G/1 queue: there is a general arrival process (renewal process) with general service time distribution and 1 server. Suppose customers arrive at rate λ. Then, the number of arrivals satisfies

$$\frac{N(t)}{t} \to \lambda \qquad \text{with probability 1.}$$

Theorem 22.2. *If the queue starts with $k \ge 1$ customers in line, μ is the service rate, and $\lambda < \mu$, then the queue will empty at some point with probability* 1.

Proof. Suppose that the queue never empties. Then, N_{s} is a renewal process. The number of people served at time t satisfies

$$\frac{N_{\mathrm{s}}(t)}{t} \xrightarrow{\text{a.s.}} \mu > \lambda \xleftarrow{\text{a.s.}} \frac{k + N_{\mathrm{a}}(t)}{t}$$

but this is a contradiction. □

In fact, if $\lambda < \mu$, then the fraction of time that the server is busy is λ/μ.

Theorem 22.3 (Little's Theorem)**.** *Consider a G/G/1 queue with arrival rate λ. $A(t)$ is the number of arrivals up to time t, $D(t)$ is the number of departures up to time t, and $L(t) = A(t) - D(t)$, the number of customers in the queue. Define*

$$\bar{L} = \lim_{t\to\infty} \frac{1}{t}\int_0^t L(\tau)\,\mathrm{d}\tau,$$

which exists with probability 1*, as the time average number of people in the queue, and*

$$\bar{W} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} W_i$$

which exists with probability 1*. Then, $\bar{L} = \lambda\bar{W}$.*

$\bar{L}$ is the average number of customers in the queue, λ is the arrival rate, and $\bar{W}$ is the average waiting time.

Example 22.4. I receive 100 emails per day and the average size of my inbox is 8000 emails. Q: How long should you wait for a response after emailing me?

$$8000 \text{ emails} = 100\,\frac{\text{emails}}{\text{day}} \times \bar{W} \implies \bar{W} = 80 \text{ days}.$$

Example 22.5. There are 5 births per day at the hospital. Most (90%) women stay 2 days. Some (10%) women stay 7 days. The average stay is $0.9 \cdot 2 + 0.1 \cdot 7 = 2.5$ days. Q: How many mothers are in the ward on average? $5 \cdot 2.5 = 12.5$ mothers.

22.3 Gaussian Random Vectors

A **standard normal random variable** $\mathcal{N}(0,1)$ is defined by its density

$$f_W(w) = \frac{1}{\sqrt{2\pi}}\exp\Bigl(-\frac{w^2}{2}\Bigr).$$

More generally, the $\mathcal{N}(\mu,\sigma^2)$ distribution, where μ is the mean and σ^2 is the variance, has density

$$f_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\Bigl(-\frac{(z-\mu)^2}{2\sigma^2}\Bigr).$$

A finite collection of "jointly" Gaussian random variables is called a **Gaussian random vector**. A collection $Z_1, Z_2, \dots, Z_n$ is jointly Gaussian if they can be expressed as $Z_j = \sum_{l=1}^{m} \alpha_{j,l} W_l + \mu_j$ for constants α, μ, and

$W_1, W_2, \dots, W_m$ are i.i.d. $\mathcal{N}(0,1)$. So, $Z = AW + \mu$, where $W = \begin{bmatrix} W_1 & W_2 & \cdots & W_m \end{bmatrix}^\mathsf{T}$. What does this imply for PDFs of JG RVs?

$$f_W(w) = \prod_{i=1}^{m} f_{W_i}(w_i) = \frac{1}{(2\pi)^{m/2}} \exp\Bigl(-\frac{w^\mathsf{T} w}{2}\Bigr)$$

If $y = Ax + b$, where A is invertible,

$$f_Y(y) = \frac{1}{|A|} f_X\bigl(A^{-1}(y-b)\bigr).$$

So,

$$f_Z(z) = \frac{1}{(2\pi)^{n/2}\sqrt{|AA^\mathsf{T}|}} \exp\Bigl(-\frac{1}{2}(z-\mu)^\mathsf{T}(AA^\mathsf{T})^{-1}(z-\mu)\Bigr)$$

What is AA^T? $\operatorname{cov} Z = \mathbb{E}[(Z-\mu)(Z-\mu)^\mathsf{T}] = A\,\mathbb{E}[WW^\mathsf{T}]A^\mathsf{T} = AA^\mathsf{T}$ since $\mathbb{E}[WW^\mathsf{T}] = I$. An n-dimensional Gaussian random vector with mean μ, covariance Σ has density

$$f_Z(z) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\Bigl(-\frac{(z-\mu)^\mathsf{T}\Sigma^{-1}(z-\mu)}{2}\Bigr).$$

Suppose we have a JG vector $\begin{bmatrix} X^\mathsf{T} & Y^\mathsf{T} \end{bmatrix}^\mathsf{T}$ with covariance

$$\begin{bmatrix} \Sigma_X & \Sigma_{X,Y} \\ \Sigma_{X,Y}^\mathsf{T} & \Sigma_Y \end{bmatrix}.$$

Assume that the vector is zero-mean. I claim that $X = \Sigma_{X,Y}\Sigma_Y^{-1}Y + V$, where V is Gaussian, independent of Y, and $\operatorname{cov} V = \Sigma_X - \Sigma_{X,Y}\Sigma_Y^{-1}\Sigma_{Y,X}$. Consequently, $\mathbb{E}(X \mid Y) = \Sigma_{X,Y}\Sigma_Y^{-1}Y$.

Lecture 23

November 14

Lecturer: Professor Thomas Courtade

23.1 Jointly Gaussian Random Variables

The density of the standard Gaussian is

$$f_W(w) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{w^2}{2}\Big).$$

Jointly Gaussian RVs are defined to be any affine combination of independent $\mathcal{N}(0,1)$ random variables. If $Z \sim \mathcal{N}(\mu, \Sigma)$, $\Sigma \succ 0$, then

$$f_Z(z) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\Big(-\frac{(z-\mu)^\mathsf{T}\Sigma^{-1}(z-\mu)}{2}\Big).$$

Suppose $\begin{bmatrix} X \\ Y \end{bmatrix}$ is a Gaussian random vector (zero-mean) with covariance matrix

$$\mathbb{E}\Big[\begin{bmatrix} X \\ Y \end{bmatrix}\begin{bmatrix} X \\ Y \end{bmatrix}^\mathsf{T}\Big] = \mathbb{E}\Big[\begin{bmatrix} X \\ Y \end{bmatrix}\begin{bmatrix} X^\mathsf{T} & Y^\mathsf{T} \end{bmatrix}\Big] = \mathbb{E}\Big[\begin{bmatrix} XX^\mathsf{T} & XY^\mathsf{T} \\ YX^\mathsf{T} & YY^\mathsf{T} \end{bmatrix}\Big] = \begin{bmatrix} \Sigma_X & \Sigma_{X,Y} \\ \Sigma_{Y,X} & \Sigma_Y \end{bmatrix}. \tag{23.1}$$

I claim: $X = \Sigma_{X,Y}\Sigma_Y^{-1}Y + V$, where $\operatorname{cov} V = \Sigma_X - \Sigma_{X,Y}\Sigma_Y^{-1}\Sigma_{Y,X}$, and V is Gaussian, independent of Y. $\Sigma_{X,Y}\Sigma_Y^{-1}Y$ is the predictable part of X from Y. V is called the **innovation**.

Suppose I observe $X_1, X_2, X_3, \dots, X_n$, and I want to predict X_{n+1}. In other words, solve the problem $\min_f \mathbb{E}[(X_{n+1} - f(X_1, \dots, X_n))^2]$. The optimal solution is $f(X_1, \dots, X_n) = \mathbb{E}(X_{n+1} \mid X_1, \dots, X_n)$.

Note: In the joint Gaussian case, $\mathbb{E}(X \mid Y) = \Sigma_{X,Y}\Sigma_Y^{-1}Y$ (i.e., the conditional expectation is a *linear* function when dealing with Gaussians).

To prove this, just show that the covariance matrix (23.1) holds for the proposted relation.

$$\begin{aligned} \mathbb{E}[XY^\mathsf{T}] &= \mathbb{E}[(\Sigma_{X,Y}\Sigma_Y^{-1}Y + V)Y^\mathsf{T}] \\ &= \Sigma_{X,Y}\Sigma_Y^{-1}\underbrace{\mathbb{E}[YY^\mathsf{T}]}_{\Sigma_Y} \\ &= \Sigma_{X,Y}. \end{aligned}$$

Clearly, $\mathbb{E}[YY^\mathsf{T}] = \Sigma_Y$. Also,

$$\mathbb{E}[XX^\mathsf{T}] = \mathbb{E}[(\Sigma_{X,Y}\Sigma_Y^{-1}Y + V)(\Sigma_{X,Y}\Sigma_Y^{-1}Y + V)^\mathsf{T}]$$

$$
\begin{aligned}
&= \mathbb{E}[(\Sigma_{X,Y}\Sigma_Y^{-1}Y)(\Sigma_{X,Y}\Sigma_Y^{-1}Y)^{\mathsf{T}}] + \mathbb{E}[VV^{\mathsf{T}}] \\
&= \Sigma_{X,Y}\Sigma_Y^{-1}I\Sigma_{Y,X} + \Sigma_X - \Sigma_{X,Y}\Sigma_Y^{-1}\Sigma_{Y,X} \\
&= \Sigma_X.
\end{aligned}
$$

23.2 Gaussian Processes, Second-Order Processes

A stochastic process $\{X_t\}_{t\in\mathcal{T}}$ is a **Gaussian process** if every finite collection of samples is jointly Gaussian.

A random process is a **second-order process** if $\mathbb{E}[X_t^2] < \infty$ for all t.

Note: Gaussian processes are second-order processes.

Suppose $\{X_t\}_{t\in\mathcal{T}}$ and $\{Y_t\}_{t\in\mathcal{T}}$ are SOPs. Then, $\mathbb{E}[X_tY_t] \le \sqrt{\mathbb{E}[X_t^2]\,\mathbb{E}[Y_t^2]} < \infty$ by Cauchy-Schwarz. Also, $\mathbb{E}[(X_t+Y_t)^2] = \mathbb{E}[X_t^2] + \mathbb{E}[Y_t^2] + 2\,\mathbb{E}[X_tY_t] < \infty$. Thus, SOPs can be viewed as elements of a vector space.

23.2.1 Covariance Function

Let $\{X_t\}_{t\in\mathcal{T}}$ be a zero-mean SOP. $K_X(t_1,t_2) = \mathbb{E}[X_{t_1}X_{t_2}]$. [Note: $K_X(t_1,t_2) < \infty$.]

Note: If we are dealing with a Gaussian process, then (the mean and) $K_X(\cdot,\cdot)$ completely specify the distribution of the process.

Example 23.1. If $X_t = \alpha\sin(\omega_0 t + \Theta)$, where $\Theta \sim \text{Uniform}[0, 2\pi)$, then

$$
\begin{aligned}
K_X(t_1,t_2) &= \mathbb{E}[\alpha\sin(\omega_0 t_1 + \Theta)\alpha\sin(\omega_0 t_2 + \Theta)] \\
&= \frac{\alpha^2}{2}\cos\big(\omega_0(t_1 - t_2)\big).
\end{aligned}
$$

Example 23.2 (DT "White Gaussian Noise" Process)**.** Let $(X_n)_{n\in\mathbb{N}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Then

$$K_X(k_1,k_2) = \delta(k_1 - k_2).$$

Example 23.3 (DT Gauss-Markov Process)**.** Let $X_0 = 0$, and $X_{n+1} = \alpha X_n + W_n$, where $W_n \sim \mathcal{N}(0,1)$ is independent of $X_{\le n}$ and $W_1, \dots, W_{n-1}$. Assume $|\alpha| < 1$. Then, for $k \ge 0$,

$$K_X(n+k,n) = \mathbb{E}[X_{n+k}X_n] = \mathbb{E}[(\alpha X_{n+k-1} + W_{n+k-1})X_n] = \alpha\,\mathbb{E}[X_{n+k-1}X_n] = \cdots = \alpha^k\,\mathbb{E}[X_n^2].$$

Also,

$$\mathbb{E}[X_n^2] = \frac{1-\alpha^{2n}}{1-\alpha^2}$$

because $\mathbb{E}[X_1^2] = 1$ and $\mathbb{E}[X_{n+1}^2] = \alpha^2\,\mathbb{E}[X_n^2] + 1$. So,

$$
\begin{aligned}
K_X(n+k,n) &= \alpha^{|k|}\frac{1-\alpha^{2n}}{1-\alpha^2} \\
&\approx \frac{\alpha^{|k|}}{1-\alpha^2}, \qquad \text{for } n \text{ big}.
\end{aligned}
$$

23.2.2 Stationarity

A process $\{X_t\}_{t\in\mathcal{T}}$ is **stationary** if

$$F_{X_{t_1},X_{t_2},\ldots,X_{t_k}}(x_1,x_2,\ldots,x_k) = F_{X_{t_1+\tau},X_{t_2+\tau},\ldots,X_{t_k+\tau}}(x_1,\ldots,x_k)$$

for all $k \geq 1$, $t_1,\ldots,t_k$, and $\tau \in \mathbb{R}$, i.e., the joint statistics are invariant to shifts in time.

Theorem 23.4. *A Gaussian process is stationary iff $K_X(t+\tau,t) = K_X(\tau,0)$ for all t, τ.*

In general, a zero-mean process is **wide-sense stationary** if $K_X(t+\tau,t) = K_X(\tau,0)$ for all t, τ.

Note: For a WSS process, $K_X(t+\tau,t) = K_X(t,t+\tau) = K_X(\tau,0) = K_X(0,-\tau) = K_X(-\tau,0)$. So, usually, we just write $K_X(\tau) \equiv K_X(0,\tau)$. By the above reasoning, this is a *symmetric function*.

Lecture 24

November 16

Lecturer: Professor Anant Sahai

24.1 Stationary & WSS Processes

The autocorrelation of X is $K_{X,X}(t_1, t_2) = \operatorname{cov}(X(t_1), X(t_2))$. For a WSS process, $K_{X,X}(\tau) = K_{X,X}(t, t+\tau)$ for all t.

24.1.1 Linear Time-Invariant (LTI) Systems

Linear time-invariant (LTI) systems are defined by their impulse responses.

$$\xrightarrow{x} \boxed{H \text{ (LTI)}} \xrightarrow{y}$$

Continuous Time	Discrete Time	Finite Cyclic Time
Dirac delta	Kronecker delta	$\begin{bmatrix}1\\0\\0\\\vdots\\0\end{bmatrix}$
time delay D_τ	discrete-time delay	cyclic time shift
linear system	linear system	matrix
LTI system	LTI system	circulant matrix
$\mathrm{e}^{\mathrm{i}\omega t}$	$\mathrm{e}^{\mathrm{i}\omega t}$	eigenbasis for the circulant matrices
impulse response	impulse response	first column
Gaussian white noise	i.i.d. $\mathcal{N}(0,1)$	i.i.d. $\mathcal{N}(0,1)$ random vector, $\mathcal{N}(0,I)$

A circulant matrix is of the form

$$\begin{bmatrix} h_0 & h_{n-1} & \cdots & \\ h_1 & h_0 & \cdots & \\ \vdots & h_1 & \ddots & \\ \vdots & \vdots & & \\ h_{n-1} & h_{n-2} & \cdots & h_0 \end{bmatrix}.$$

The eigenbasis for the circulant matrices consists of vectors with $h[j] = \mathrm{e}^{\mathrm{i}2\pi kj/n}$.

24.1.2 Power Spectral Density

The **power spectral density** $S_{X,X}$ is the FT of $K_{X,X}$.

$$\begin{aligned} S_{X,X}(f) &= \int_{-\infty}^{\infty} K_{X,X}(\tau)\mathrm{e}^{-\mathrm{i}2\pi f\tau}\,\mathrm{d}\tau, \\ K_{X,X}(\tau) &= \int_{-\infty}^{\infty} S_{X,X}(f)\mathrm{e}^{+\mathrm{i}2\pi f\tau}\,\mathrm{d}f. \end{aligned}$$

So, $K_{X,X}(0) = \int_{-\infty}^{\infty} S_{X,X}(f)\,\mathrm{d}f$.

White Gaussian noise is the “process” whose $S_{X,X}(f) = 1$ for all f.

What happens if the input signals are random?

$$\xrightarrow{X}\boxed{H \text{ (LTI)}}\xrightarrow{Y}$$

What is $S_{Y,Y}$?

Aside: If v is a vector of i.i.d. $\mathcal{N}(0,1)$ random variables, $y = Cv$, then

$$\Sigma_{y,y} = CC^* = \begin{bmatrix} h_0 & h_{n-1} & h_{n-2} & \cdots & h_1 \\ \vdots & \ddots & & & \\ h_{n-1} & \cdots & & & \end{bmatrix} \begin{bmatrix} h_0^* & h_1^* & \cdots \\ h_{n-1}^* & \ddots & \\ \vdots & & \\ h_1^* & \cdots & \end{bmatrix}$$

and in the frequency domain,

$$\begin{aligned} y_{\mathrm{F}} &= C_{\mathrm{F}} v_{\mathrm{F}} \\ &= \begin{bmatrix} H(0) & & & 0 \\ & H(1) & & \\ & & \ddots & \\ 0 & & & H(n-1) \end{bmatrix} v \end{aligned}$$

so

$$(\Sigma_{Y,Y})_{\mathrm{F}} = \begin{bmatrix} |H(0)|^2 & \\ & \ddots \end{bmatrix}.$$

By analogy, $S_{Y,Y}(f) = S_{X,X}(f)|H(f)|^2$.

Lecture 25

November 21

Lecturer: Soham-Rajesh Phade

25.1 Wiener Filter

25.1.1 MMSE & LLSE

Recall the MMSE estimator. If X, Y are real-valued jointly Gaussian RVs, zero mean, where Y is the observation, then $\hat{X}_{\mathrm{MMSE}}(Y)$ is the minimum mean squared error estimator which minimizes $\mathbb{E}[(X-\hat{X}(Y))^2]$. Then, $\hat{X}_{\mathrm{MMSE}}(Y) = \mathbb{E}(X \mid Y)$ *because*

$$\begin{aligned}
&\mathbb{E}\big[\big(X - \hat{X}(Y)\big)^2\big] \\
&\qquad = \mathbb{E}\big[\big\{\big(X - \hat{X}_{\mathrm{MMSE}}(Y)\big) + \big(\hat{X}_{\mathrm{MMSE}}(Y) - \hat{X}(Y)\big)\big\}^2\big] \\
&\qquad = \mathbb{E}\big[\big(X - \hat{X}_{\mathrm{MMSE}}(Y)\big)^2\big] + \mathbb{E}\big[\big(\hat{X}_{\mathrm{MMSE}}(Y) - \hat{X}(Y)\big)^2\big] + 2\underbrace{\mathbb{E}\big[\big(X - \hat{X}_{\mathrm{MMSE}}(Y)\big)\big(\hat{X}_{\mathrm{MMSE}}(Y) - \hat{X}(Y)\big)\big]}_{0}.
\end{aligned}$$

The last term is 0 by the **Orthogonality Principle**: $\mathbb{E}[(X - \hat{X}(Y))h(Y)] = 0$.

For the jointly Gaussian case, if $\mathbb{E}(X \mid Y) = aY$, then $\mathbb{E}[(X - aY)Y] = 0$ gives

$$a = \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}.$$

For $X = \begin{bmatrix} X_1 & \cdots & X_m \end{bmatrix}^{\mathsf{T}}$ and $Y = \begin{bmatrix} Y_1 & \cdots & Y_n \end{bmatrix}^{\mathsf{T}}$, then

$$\begin{aligned}
X &= PY + V \\
&= \Sigma_{X,Y}\Sigma_Y^{-1}Y + V
\end{aligned}$$

where $V \perp\!\!\!\perp Y$, $\Sigma_{X,Y} = \mathbb{E}[XY^{\mathsf{T}}]$, and $\Sigma_Y = \mathbb{E}[YY^{\mathsf{T}}]$.

LLSE estimator: $\hat{X}_{\mathrm{LLSE}}(Y)$ is a linear function of Y that minimizes

$$\mathbb{E}\big[\big(X - \hat{X}(Y)\big)^2\big] = \mathbb{E}\big[\big\{\big(X - \hat{X}_{\mathrm{LLSE}}(Y)\big) + \big(\hat{X}_{\mathrm{LLSE}}(Y) - \hat{X}(Y)\big)\big\}^2\big]$$

which gives the **Orthogonality Principle for the LLSE**:

$$\mathbb{E}\big[\big(X - \hat{X}_{\mathrm{LLSE}}(Y)\big)aY\big] = 0, \qquad \forall a.$$

Note that the error depends only on $\mathbb{E}[X^2]$, $\mathbb{E}[Y^2]$, and $\mathbb{E}[XY]$, or in the vector case, on $\Sigma_{X,Y}$, Σ_X, and Σ_Y. Then, $\hat{X}_{\mathrm{LLSE}}(Y) = bY$, where

$$b = \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}.$$

In the vector case, $\hat{X}(Y) = \Sigma_{X,Y}\Sigma_Y^{-1}Y$.

25.1.2 Wiener Filter

Let $(S(u),\ u \in \mathbb{R})$ be the signal and $(X(u),\ u \in \mathbb{R})$ be the observation. Assume that these two processes are jointly WSS and zero mean, $\mathbb{E}[X(u)] = \mathbb{E}[S(u)] = 0$ for all u. Thus,

$$\begin{aligned}
\mathbb{E}[X(t+\tau)X(t)] &= K_{X,X}(\tau),\\
\mathbb{E}[S(t+\tau)S(t)] &= K_{S,S}(\tau),\\
\mathbb{E}[X(t+\tau)S(t)] &= K_{X,S}(\tau).
\end{aligned}$$

Let $\hat{S}(t) = L[S(t) \mid X(u),\ -\infty < u < \infty]$. Then,

$$\begin{aligned}
\hat{S}(t_1) &= \int_{-\infty}^{\infty} h(u,t_1)X(t_1-u)\,\mathrm{d}u,\\
\hat{S}(t_2) &= \int_{-\infty}^{\infty} h(u,t_2)X(t_2-u)\,\mathrm{d}u.
\end{aligned}$$

The jointly WSS assumption gives $h(u,t_1) = h(u,t_2)$. The problem is to minimize $\mathbb{E}[(S(t) - \hat{S}(t))^2]$.

Orthogonality Principle: $\mathbb{E}[(S(t) - \hat{S}(t))X(u)] = 0$ for all u. So,

$$\int_{-\infty}^{\infty} h(v)\,\mathbb{E}[X(t-v)X(u)]\,\mathrm{d}v = \mathbb{E}[S(t)X(u)], \qquad \forall u.$$

Let $\tau = t-u$, so $\int_{-\infty}^{\infty} h(v)K_{X,X}(\tau - v)\,\mathrm{d}v = K_{S,X}(\tau)$ for all τ. Take the Fourier transform:

$$H(f)S_{X,X}(f) = S_{S,X}(f), \qquad \forall f,$$

so

$$H(f) = \frac{S_{S,X}(f)}{S_{X,X}(f)}.$$

This is the **Wiener filter**. Here, H is the Fourier transform of h.

Example 25.1. Suppose $X = S + V$, where $(V(t))_{t\in\mathbb{R}} \perp\!\!\!\perp (S(t))_{t\in\mathbb{R}}$. Then,

$$\begin{aligned}
K_{S,X}(\tau) &= \mathbb{E}[S(t+\tau)X(t)]\\
&= \mathbb{E}\big[S(t+\tau)\big(S(t)+V(t)\big)\big]\\
&= K_{S,S}(\tau) + 0,\\
K_{X,X}(\tau) &= \mathbb{E}[X(t+\tau)X(t)]\\
&= \mathbb{E}\big[\big(S(t+\tau)+V(t+\tau)\big)\big(S(t)+V(t)\big)\big]\\
&= K_{S,S}(\tau) + K_{V,V}(\tau).
\end{aligned}$$

So,

$$H(f) = \frac{S_{S,S}(f)}{S_{S,S}(f) + S_{V,V}(f)}.$$

The Wiener filter we just derived is a non-causal filter.

25.2 Hypothesis Testing

H_0 is the **null hypothesis** and H_1 is the **alternate hypothesis**. For example, the source may be $\{0,1\}$ and the observation (real-valued) may have some added noise, $X = S + V$.

The observation X is sampled from $f(\cdot \mid 0)$ if H_0 is true, and from $f(\cdot \mid 1)$ if H_1 is true.

25.2.1 Bayesian Formulation

In the Bayesian formulation, we also have $\pi_0 = \mathbb{P}(H_0 \text{ is true})$ and $\pi_1 = \mathbb{P}(H_1 \text{ is true})$. So, $\pi_0 + \pi_1 = 1$.

The **cost function** gives $C(1 \mid 0)$, the cost of predicting H_1 when H_0 is true, and $C(0 \mid 1)$, the cost of predicting H_0 when H_1 is true.

Performance criterion: minimize expected cost.

Decision rule:

$$\Delta(x) = \begin{cases} 0, & \text{declare } H_0 \text{ true} \\ 1, & \text{declare } H_1 \text{ true} \end{cases}$$

Threshold decision rule:

$$L(x) = \frac{f(x \mid 1)}{f(x \mid 0)}$$

is the likelihood function. The threshold decision rule is:

$$\Delta(x) = \begin{cases} 0, & \text{if } L(x) < \Lambda \\ 1, & \text{otherwise} \end{cases}$$

Proposition 25.2. *The optimal decision rule is the threshold decision rule with cutoff*

$$\Lambda = \frac{c(1 \mid 0)\pi(0)}{c(0 \mid 1)\pi(1)}.$$

Proof. The expected cost is

$$\int_{-\infty}^{\infty} [c(1 \mid 0)\pi(0)\,\mathbb{1}\{\Delta(x) = 1\} f(x \mid 0) + c(0 \mid 1)\pi(1)\,\mathbb{1}\{\Delta(x) = 0\} f(x \mid 1)]\,\mathrm{d}x$$
$$= \int_{-\infty}^{\infty} [c(1 \mid 0)\pi(0)\,\mathbb{1}\{\Delta(x) = 1\} + c(0 \mid 1)\pi(1)\,\mathbb{1}\{\Delta(x) = 0\} L(x)] f(x \mid 0)\,\mathrm{d}x.$$

To optimize this, $\Delta(x) = 1$ if $c(0 \mid 1)\pi(1)L(x) > c(1 \mid 0)\pi(0)$. □

25.2.2 Neyman-Pearson Formulation

The **false error** is $\mathbb{P}(H_1 \text{ true} \mid H_0 \text{ true})$ (**type I**). The **missed detection** is $\mathbb{P}(H_0 \text{ true} \mid H_1 \text{ true})$ (**type II**). We require the false error to be $\le \varepsilon$. Minimize the missed detection.

Randomized threshold rule: There is a cutoff Λ and a randomization $0 \le \alpha \le 1$.

$$\Delta^*(x) = \begin{cases} 0, & \text{if } L(x) < \Lambda, \\ \begin{cases} 0 & \text{with probability } 1 - \alpha \\ 1 & \text{with probability } \alpha \end{cases}, & \text{if } L(x) = \Lambda \\ 1, & \text{if } L(x) > \Lambda \end{cases}$$

Proposition 25.3. *The randomized threshold rule is optimal.*

Proof. Select Λ, α so that $\mathbb{P}_0(L(X) > \Lambda) + \alpha\mathbb{P}_0(L(X) = \Lambda) = \varepsilon$ (this uniquely determines Λ, α). Let Δ be some other decision rule. If Δ is valid, then $\int_{\{x:\Delta(x)=1\}} f(x \mid 0)\,\mathrm{d}x \le \varepsilon$. Now,

$$\big(\Delta^*(x) - \Delta(x)\big)\big(f(x \mid 1) - \Lambda f(x \mid 0)\big) \ge 0.$$

Integrate with respect to x.

$$\begin{aligned}\int_{-\infty}^{\infty} \mathbb{1}\{\Delta^*(x) = 1\} f(x \mid 1)\,\mathrm{d}x - \int_{-\infty}^{\infty} \mathbb{1}\{\Delta(x) = 1\} f(x \mid 1)\,\mathrm{d}x \\ \ge \Lambda\Big(\int_{-\infty}^{\infty} \mathbb{1}\{\Delta^*(x) = 1\} f(x \mid 0)\,\mathrm{d}x - \int_{-\infty}^{\infty} \mathbb{1}\{\Delta(x) = 1\} f(x \mid 0)\,\mathrm{d}x\Big)\end{aligned}$$

so

$$\mathbb{P}_1\big(\Delta^*(X) = 1\big) - \mathbb{P}_1\big(\Delta(X) = 1\big) \ge \Lambda\big[\underbrace{\mathbb{P}_0\big(\Delta^*(X) = 1\big)}_{=\varepsilon} - \underbrace{\mathbb{P}_0\big(\Delta(X) = 1\big)}_{\le\varepsilon}\big] \ge 0.$$

Then, $\mathbb{P}_1(\Delta^*(X) = 1) \ge \mathbb{P}_1(\Delta(X) = 1)$, so $\mathbb{P}_1(\Delta^*(X) = 0) \le \mathbb{P}_1(\Delta(X) = 0)$. □

Lecture 26

November 28

Lecturer: Professor Anant Sahai

26.1 Hypothesis Testing & Matched Filtering

Observe X, test the hypotheses H_0 and H_1. Under H_0, the density of X is $f(\cdot \mid 0)$; under H_1, the density of X is $f(\cdot \mid 1)$. The likelihood ratio is

$$L(x) = \frac{f(x \mid 1)}{f(x \mid 0)}.$$

Example 26.1. Suppose under H_0, $X \sim \text{Uniform}[0, 2]$, and under H_1, $X \sim \text{Uniform}[1, 3]$.

Question: In the Bayesian setting where

$$\pi_0 = \frac{1}{3},$$
$$\pi_1 = \frac{2}{3},$$

we observe $X = 1.2$. What is $\mathbb{P}(0 \mid X = 1.2)$? Answer: 1/3.

Threshold (randomized): If $L(X) < \Lambda$, say 0. If $L(X) > \Lambda$, say 1. If $L(X) = \Lambda$, guess by $(\alpha, 1 - \alpha)$.

26.1.1 Gaussian Examples

Example 26.2 (1-Dimensional)**.** Test $H_0 : X \sim \mathcal{N}(0, \sigma^2)$ vs. $H_1 : X \sim \mathcal{N}(m, \sigma^2)$.

$$\begin{aligned} L(x) = \frac{\mathrm{e}^{-(x-m)^2/(2\sigma^2)}}{\mathrm{e}^{-x^2/(2\sigma^2)}} = \mathrm{e}^{-((x-m)^2-x^2)/(2\sigma^2)} = \mathrm{e}^{-(-2xm+m^2)/(2\sigma^2)} &= \mathrm{e}^{-(m/\sigma^2)(-x+m/2)} \\ &= \mathrm{e}^{(m/\sigma^2)(x-m/2)}. \end{aligned}$$

Example 26.3 (k-Dimensional)**.** Test $H_0 : X \sim \mathcal{N}(0, \sigma^2 I_k)$ vs. $H_1 : X \sim \mathcal{N}(m, \sigma^2 I_k)$.

$$L(x) = \frac{\mathrm{e}^{-(x-m)^\mathsf{T}(x-m)/(2\sigma^2)}}{\mathrm{e}^{-x^\mathsf{T}x/(2\sigma^2)}} = \mathrm{e}^{-(x^\mathsf{T}x-2x^\mathsf{T}m+m^\mathsf{T}m-x^\mathsf{T}x)/(2\sigma^2)} = \mathrm{e}^{(m^\mathsf{T}/\sigma^2)(x-m/2)}.$$

Only the information contained in $m^\mathsf{T}X$ is useful. Change coordinates to $\tilde{X}$. Under H_0, $\tilde{X}[0] \sim \mathcal{N}(0, \sigma^2)$, and under H_1, $\tilde{X}[0] \sim \mathcal{N}(\|m\|_2, \sigma^2)$. For $j > 0$, under H_0 and H_1, $\tilde{X}[j] \sim \mathcal{N}(0, \sigma^2)$.

Example 26.4 (k-Dimensional (Non-White))**.** Test $H_0 : X \sim \mathcal{N}(0, C)$ vs. $H_1 : X \sim \mathcal{N}(\mu, C)$. Change coordinates from X to X'. If $X' = HX$, then

$$\begin{aligned} \mathbb{E}[X'(X')^\mathsf{T}] &= \mathbb{E}[HXX^\mathsf{T}H^\mathsf{T}] \\ &= HCH^\mathsf{T} = I_k. \end{aligned}$$

The new mean is $m' = Hm$. Look at $(m')^\mathsf{T}X' = m^\mathsf{T}H^\mathsf{T}HX$.

Example 26.5 (∞-Dimensional Case)**.** The signal is either 0 or f (e.g., a pulse), and the noise process is $(V(t))_{t\in\mathbb{R}}$. Observe $(X(t))_{t\in\mathbb{R}}$.

(Easy Case) Start with the core case when $(V(t))_{t\in\mathbb{R}}$ is white noise. The goal is to pass $(X(t))_{t\in\mathbb{R}}$ through a filter H to generate $(Y(t))_{t\in\mathbb{R}}$. Pick a time t_0 and look at $Y(t_0)$. We want to choose a hypothesis test based on $\langle f/\|f\|_2, X\rangle$.

(Hard Case) V is not white noise, so it has some $R_{V,V}$. The approach here is to first use a whitening filter, and then a matched filter.