# Stochastic Processes

## STAT-150 Lecture Series, Jim Pitman

(version 1.18, updated 11/27/2019)

**Fall 2019**

**University of California, Berkeley**

This document is intended to capture lecture material into a referential and well-prepared resource for the cohort of students taking the STAT-150 Stochastic Processes course taught by Jim Pitman. Many thanks to all the student contributors and a special thank-you to John-Michael Laurel and Daniel Suryakusuma.

# Contents

# LECTURE 1

## Stochastic Processes and Markov Chains

A *stochastic process* is a collection of random variables indexed by some *parameter set* $\mathcal{I}$. We shall use the following notation

$$(X_i, i \in \mathcal{I}) = (X_i)_{i \in \mathcal{I}}$$

The parameter set commonly represents a set of times, but can extend to e.g. space or space-time. Underlying (1) there is always a *probability measure* $\mathbb{P}$ on some outcome space $\Omega$ with $\mathbb{P}$ a function of subsets $\mathcal{F}$ of $\Omega$, ranging over a suitable collection of subsets $\mathcal{F}$ called *events*.

In the *canonical* setup, $\Omega$ is a *product space*

$$\Omega = \prod_{i \in \mathcal{I}} \mathcal{S}_i$$

where $\mathcal{S}_i$ is a space of values of $X_i$,and the $X_i$ are just coordinate maps on this product space, and $\mathbb{P}$ is a probability measure on the product space, with $\mathcal{F}$ the product $\sigma$-field. So $\omega = (x_i, i \in \mathcal{I}) \in \Omega$ and $X_i(\omega) = x_i$. But

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

could be any *probability space*, and each $X_i$ a random variable defined as a function on that space. All of the italicized terms here are standard. Their definitions can be found in `Wikipedia`. It's prudent to highlight the equivalence in notation between Pitman and Durret.

| Pitman | Durrett | description |
|---|---|---|
| $\mathbb{P}$ | $P$ | probability measure |
| $P$ | $p$ | probability transition matrix |

## 1.1 Markov Chains

For a countable *state space* $\mathcal{S}$, for instance $\mathcal{S} = \mathbb{N}_0 := \{0, 1, \dots\}$, we construct a sequence of evolving discrete R.V.s

$$(X_0, X_1, \dots, X_{n-1}, X_n) = (X_i)_{i \in \mathbb{N}_0} = (X_i)$$

where

$$\begin{aligned} X_0 &:= \text{ initial state at time } 0 \\ X_1 &:= \text{ state of process after time } 1 \\ &\vdots \\ X_n &:= \text{ state of process after time } n \end{aligned}$$

and call $(X_i)$ a *Markov chain* if it satisfies the *Markov property*.

**Markov Property**

A stochastic process is a *Markov chain* $(X_i)_{i \in \mathcal{S}}$ if it satisfies

$$\mathbb{P}\left( X_{n+1} = x_{n+1} \;\middle|\; \bigcap_{i=0}^{n} \{X_i = x_i\} \right) = \mathbb{P}(X_{n+1} = x_{n+1} \,|\, X_n = x_n) \qquad (1.1)$$

known as the *Markov property*. In words,

*Past and future are conditionally independent given the present.*

Given the present value $X_n = x_n$, the past values $X_0, \dots, X_{n-1}$ become irrelevant for predicting values of $X_{n+1}$.

## 1.2 Specifying Joint Probabilities

To specify a stochastic process, you must describe the joint distribution of its variables. For example, we know for R.V.s $X_0, X_1, X_2 \in \mathbb{N}_0$

$$\mathbb{P}(X_0 \leq 3, X_1 \leq 5, X_2 \leq 7) = \sum_{x=0}^{3} \sum_{y=0}^{5} \sum_{z=0}^{7} p(x, y, z)$$

So to specify the joint distribution of the three variables $X_0, X_1, X_2$ it is enough to specify their *joint probability function* $p(x, y, z)$. This must be some non-negative

function which sums to 1 over all triples $(x, y, z)$. Now for any sequence of three R.V.s,

$$\mathbb{P}(X_0 = x, X_1 = y, X_2 = z) = \\ \mathbb{P}(X_0 = x)\mathbb{P}(X_1 = y \,|\, X_0 = x)\mathbb{P}(X_2 = z \,|\, X_1 = y, X_0 = x)$$

For a Markov chain, this reduces to

$$\mathbb{P}(X_0 = x, X_1 = y, X_2 = z) = \\ \mathbb{P}(X_0 = x)\mathbb{P}(X_1 = y \,|\, X_0 = x)\mathbb{P}(X_2 = z \,|\, X_1 = y)$$

Now *transition matrices* $P_1$ and $P_2$ can be defined by

$$\mathbb{P}(X_1 = y \,|\, X_0 = x) = P_1(x, y)$$
$$\mathbb{P}(X_2 = z \,|\, X_1 = y) = P_2(y, z)$$

Then

$$\mathbb{P}(X_0 = x, X_1 = y, X_2 = z) = \mathbb{P}(X_0 = x)P_1(x, y)P_2(y, z)$$

Most commonly we assume that the chain has *homogeneous* transition probabilities. That means $P_1 = P_2 = P$ for a single transition matrix $P$. We deal with this idea formally in the next section.

## 1.3 Transition Mechanism

Suppose the state space $\mathcal{S}$ is finite. We can construct a *transition matrix*

$$P = P(x, y) \tag{1.2}$$

a 'set of rules' or 'mechanism' for moving between different states in $\mathcal{S}$. Rules of probability imply that (1.2) is a *stochastic matrix*.

**Stochastic Matrix**

A *stochastic matrix* is a non-negative matrix with all row sums equal to 1. With the usual convention of $x$ indexing rows and $y$ indexing columns, we say $P$ is *stochastic* if it satisfies

$$P(x, y) \geq 0 \quad \text{and} \quad \sum_y P(x, y) = 1$$

It should make intuitive sense that for a fixed row, summing its elements yields one. Given the present state $x$, you're bound to go somewhere.

For now, our Markov chains will possess *homogeneous transition probabilities*. All that means is we use the same matrix $P$ at each step in time. To make this more concrete, observe for a Markov chain

$$\mathbb{P}(X_0 = x_0, X_1 = x_1, X_2 = x_2, X_3 = x_3) \\ = \mathbb{P}(X_0 = x_0)P_1(x_0, x_1)P_2(x_1, x_2)P_3(x_2, x_3)$$

and by time homogeneity $P_1, P_2, P_3 = P$ and we have

$$\begin{aligned}\mathbb{P}(X_0 = x_0, X_1 = x_1, X_2 = x_2, X_3 = x_3) \\ = \mathbb{P}(X_0 = x_0)P(x_0, x_1)P(x_1, x_2)P(x_2, x_3)\end{aligned}$$

### 1.3.1 Absorbing States

When $P(x, x) = 1$, we say that $x$ is an *absorbing state*. If you start in an absorbing state, you are sure to be there next step, and there again after two steps, and so on. Put another way, once you arrive at an absorbing state, you never leave it.

## 1.4 Constructing the Joint Distribution

To get the chain going, we initialize the process with an assigned *initial distribution* $\lambda$ for $X_0$. That is

$$\mathbb{P}(X_0 = x_0) = \lambda(x_0)$$

Again, rules of probability require $\lambda$ is a probability distribution on the state space.

$$\lambda(x_0) \geq 0 \quad \text{and} \quad \sum_{x_0} \lambda(x_0) = 1$$

We finally have everything we need to completely specify the joint distribution of a Markov chain with homogeneous transition probabilities.

**Prescription for the Joint Distribution**

Let the Markov chain $(X_i)_{i \in \mathcal{S}}$ have finite state space $\mathcal{S}$, assigned initial distribution $\lambda$, and transition probability matrix $P$. Then for sequences of length $n + 1$, the joint distribution

$$(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n)$$

has the following prescription

$$\mathbb{P}\left(\bigcap_{i=0}^{n} \{X_i = x_i\}\right) = \lambda(x_0) \prod_{i=0}^{n-1} P(x_i, x_{i+1})$$

This construction is a proper assignment of a joint distribution, according to the rules of probability. One can check this by verifying 1. all the joint probabilities are non-negative (which is trivial) and 2. all the probabilities sum to one, that is

$$\sum_{x_0} \sum_{x_1} \cdots \sum_{x_{n-1}} \sum_{x_n} \lambda(x_0) \prod_{i=0}^{n-1} P(x_i, x_{i+1}) = 1$$

which one can show using mathematical induction on $n$. e.g. assuming it is true for $n - 1$ instead of $n$, in going from $n - 1$ to $n$ there is just one more summation,

which simplifies as it should using row sums of the transition matrix equal 1.

### Some Notation

- $\mathbb{P}_\lambda$ is used to show that $\mathbb{P}$ makes $X_0$ have distribution $\lambda$.
- $\mathbb{P}_x$ is used to show that $\mathbb{P}$ makes $X_0 = x$, i.e. $\lambda(x) = 1$ and $\lambda(x_0) = 0$ for $x_0 \neq x$. So under $\mathbb{P}_x$ the chain starts in state $x$.

## 1.5 Simulating a Markov Chain

We can simulate a Markov chain $(X_i)_{i\in\mathbb{N}_0}$ with a supply of uniform R.V.s

$$U_0, U_1, \ldots \sim \textbf{Uniform}(0,1)$$

For the initial state $X_0$ to have distribution $\lambda$, that is $X_0 \sim \lambda$, let

$$X_0 = \begin{cases} 0, & \text{if } 0 \leq U_0 < \lambda(0) \\ 1, & \text{if } \lambda(0) \leq U_0 < \lambda(0) + \lambda(1) \\ 2, & \text{if } \lambda(0) + \lambda(1) \leq U_0 < \lambda(0) + \lambda(1) + \lambda(2) \\ & \vdots \end{cases}$$

Now, if $X_0 = x_0$, define

$$X_1 = \begin{cases} 0, & \text{if } 0 \leq U_1 < P(x_0, 0) \\ 1, & \text{if } P(x_0,0) \leq U_1 < P(x_0,0) + P(x_0,1) \\ 2, & \text{if } P(x_0,0) + P(x_0,1) \leq U_1 < P(x_0,0) + P(x_0,1) + P(x_0,2) \\ & \vdots \end{cases}$$

and so on. Hence, given $X_0 = x_0$ and $X_1 = x_0$, we create intervals using the elements $P(x_1, \cdot)$. It is easy to implement this simulation using a language like `R`, `Python`, etc.

Often the row $P(x, \cdot)$ is a standard distribution, e.g. uniform or binomial or Poisson or geometric with parameters depending on $x$. Then there are built in packages for generating such variables which can be used instead of the crudest scheme indicated above.

## 1.6 Gambler's Ruin Chain

This is an exemplar Markov chain and we'll be referencing this problem extensively throughout the course. The setting is as follows:

> A gambler has an initial fortune of $\$a$, $(0 \leq a \leq N)$. At each play, the gambler wins a \$1 with probability $p$ and loses \$1 with probability $q = 1 - p$. The gambler plays until $X_n = N$ (quitting with a gain) or until $X_n = 0$ (ruined).

It is clear that

$$X_n := \text{ the gambler's capital at round } n$$

is a Markov chain on state space $\{0, 1, \ldots, N\}$ and has transition probabilities

$$P(0,0) = P(N,N) = 1,$$

$$P(x, x+1) = p, \text{ and}$$

$$P(x, x-1) = q = 1 - p$$

Observe states $\{0, N\}$ are *absorbing*. In words, once you win $\$N$, you leave the casino having won some money or you lost all of your money and you have no choice but to leave the casino. In either case, once you hit the boundary, say $X_m = N$, the chain remains there forever, that is $X_n = N$ for all $n > m$. Some natural questions that arise from this chain are

- What's the expected time until the $(X_n)$ hits $N$ or 0?
- Is it possible for the $(X_n)$ to get stuck on the interior of the state-space forever? That is, as $n \to \infty$, is $\mathbb{P}(X_n \in \{1, 2, \ldots, N-1\}) > 0$?

These questions will be addressed and studied in the coming weeks. A rendition of the setting is presented in the text. Reference. [1] Durrett's, *Essentials of Stochastic Processes* Section 1.1. A realization of the Gambler's Ruin is illustrated on the next page.

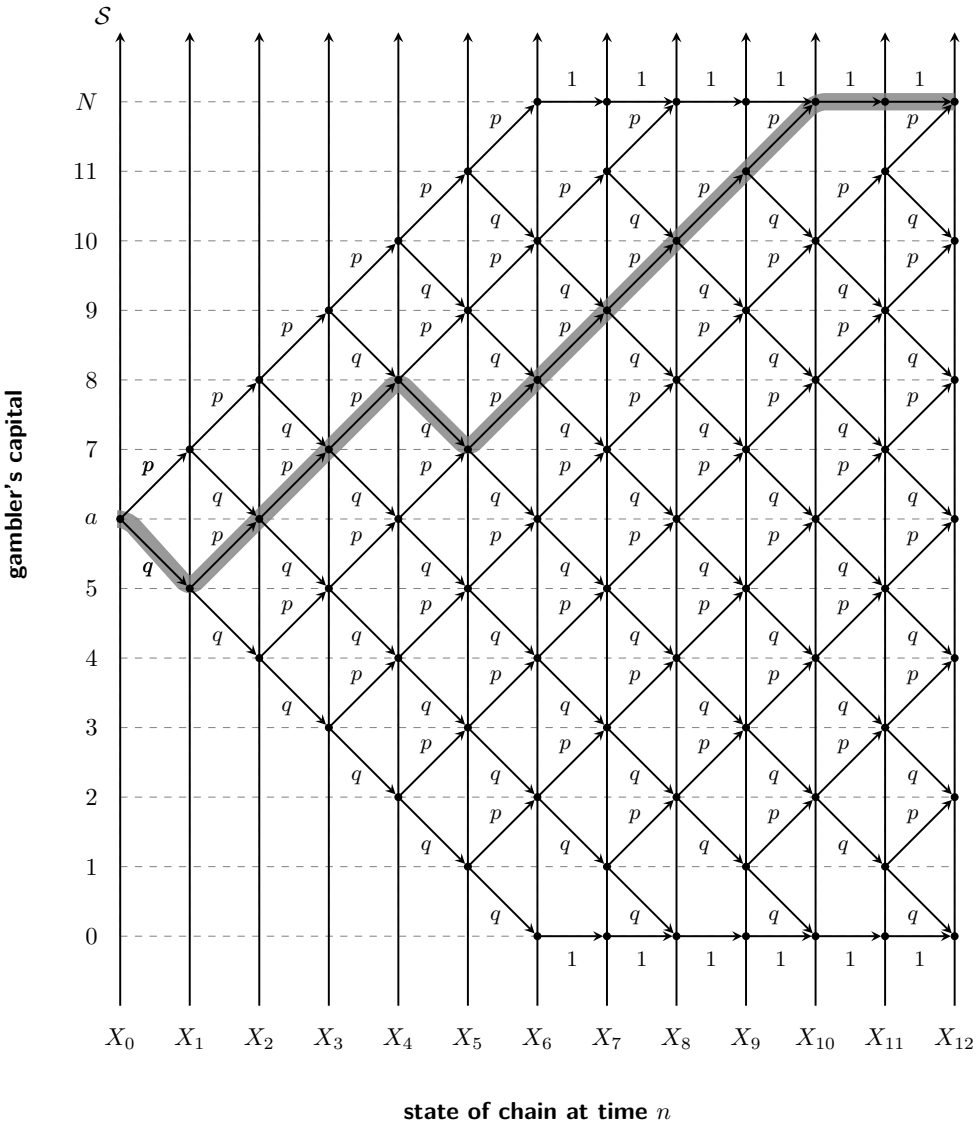



Figure 1.1: Realization of the gambler's ruin chain with $N = 12$, $a = 6$, and $X_{n=10} = N$. Each vertical axis represents the state of the chain $(X_n)$ at time $n$. At each step in time, we move $\uparrow$ with probability $p$ and $\downarrow$ with probability $q = 1 - p$. Observe at time $n = 10$, $(X_n)$ has hit a boundary state, that is $X_n \in \{0, N\}$, hence for all $n \geq 10$, $X_n = N$ with probability 1. The gray band traces the trajectory of the chain over time. One of the nice features of this realization is we need not worry about visualizing the chain in higher dimensions. Another realization of this chain will debut in lecture 3.

# LECTURE 2

# Transition Mechanism

Pitman reminds us that Wikipedia serves as a valuable resource for clarifying most basic definitions in this course.
Recall from Lecture 1 we worked with a *transition matrix* $P$ with rows $x$ and columns $y$. The $x$th row and $y$th column entry is $P(x, y)$. All entries are non-negative. All row sums are 1.
For the first step in the Markov chain, we have:

$$P(x, y) = \mathbb{P}(X_1 = y \,|\, X_0 = x).$$

With many steps, and homogeneous transition probabilities, also

$$P(x, y) = \mathbb{P}(X_{n+1} = y \,|\, X_n = x).$$

Pitman notes that the first problem on homework 1 is very instructional, which gets us to think about what exactly is the Markov property.

## 2.1 Action of a Transition Matrix on a Row Vector

Take an initial distribution $\lambda(x) = \mathbb{P}(X_0 = x)$. If we write $P(x, \cdot)$, we're taking the row of numbers in the matrix. With $N$ states we can simply consider sequences of length $N$ rather than $N$-dimensional space. To ensure we really know what's going on here, consider 2 steps (indexed 0 and 1). What is the distribution of $X_0$? Trivially, it's $\lambda$. Now what is the distribution of $X_1$? We need to do a little more. We don't know how we started, and we want to think of all the ways we could have ended up at our final state $X_1$.
To do this, we use the **law of total probability**, which gives:

$$\boxed{\mathbb{P}(X_1 = y) = \sum_{x \in S} \mathbb{P}(X_0 = x, X_1 = y).}$$

Now it just takes a little bit of calculation to go forward. Conditioning on $X_0$ (turning a joint probablity into a marginal for the first and a conditional given the

first) gives:

$$\begin{aligned}
\mathbb{P}(X_1 = y) &= \sum_{x \in S} \mathbb{P}(X_0 = x, X_1 = y) \\
&= \sum_{x \in S} \mathbb{P}(X_0 = x) \cdot \mathbb{P}(X_1 = y \,|\, X_0 = x) \\
\mathbb{P}(X_1 = y) &= \sum_{x \in S} \lambda(x) \underbrace{P(x, y)}_{\text{matrix entry}} \\
&= (\lambda P)(y) \text{ or equivalently, } = (\lambda P)_y
\end{aligned}$$

To have this fit with matrix multiplication, we must take $\lambda(x)$ to be a ROW VECTOR. Back in our picture going from one step to the next of a Markov chain, we use $x$ the ($n$th state) to index the row of the matrix and $y$ (the $n+1$th state) to index the column of the matrix $P(x, y)$.

### 2.1.1 Conclusion

There is a happy coincidence between the rules of probability and the rules of matrices, which implies that if a Markov Chain has $X_0 \sim \lambda$ (meaning random variable $X_0$ has distribution $\lambda$), then at the next step we have the following distribution

$$re\boxed{X_1 \sim \lambda P}$$

where argument $y$ is hidden. If you evaluate the row vector $\lambda P$ at entry $y$, you get $(\lambda P)_y = \mathbb{P}(X_1 = y)$. Although this may not be terribly exciting, Pitman notes this is fundamental and important to understand the connection between linear algebra and rules of matrices with probability. We will maintain and strengthen this connection throughout the course.

## 2.2 Action of a Transition Matrix on a Column Vector

Suppose $f$ is a function on $S$. Think of it as a **reward** in that if $X_1 = x$, then you get $\$f(x)$ (random monetary reward $f(X_1)$ where $X_1 \in S$ as an abstract object; these can be partitions or permutations or something very abstract, not necessarily numerical). Pitman notes some applications of Markov chains to Google's PageRank with a very big state space of web pages. Without being scared about the potential size of the **state space**, we open to some abstraction in our immediate example. Consider the Markov chain step from $X_0$ to $X_1$ and the conditional expectation:

$$\mathbb{E}\big[f(X_1) \,|\, X_0 = x\big] = \sum_y \underbrace{P(x, y)}_{\text{matrix}} \underbrace{f(y)}_{\text{col.vec.}}$$

where we could make some concrete financial definitions to apply our abstract problem if we wish.

Starting at state $x$, we move to the next state according to the row $P(x, \cdot)$. Recognize this as a matrix operation and we have, for the above:

$$\mathbb{E}\left(f(X_1) \,|\, X_0 = x\right) = (Pf)(x)$$

**Remark:** The function or column vector $f$ can be signed (there is no difficulty if we are losing money as opposed to gaining); it is more difficult to interpret the action on a signed row vector $\lambda$. But easy to interpret $\lambda P$ for a probability measure $\lambda$.

## 2.3 Two Steps

Now consider two steps in time

$$X_0 \xrightarrow{P} X_1 \xrightarrow{Q} X_2$$

Assume the Markov property. Now let's discuss the probability of $X_2$, knowing $X_0 = x$. That is,

$$\mathbb{P}(X_2 = z \,|\, X_0 = x)$$

where we have some mystery intermediate $X_1$. The row out of the matrix which we use for the intermediate is random.
We condition upon what we don't know in order to reach a solution. It should become instinctive to us soon to do such a thing: condition on $X_1$. This gives

$$\begin{aligned}\mathbb{P}(X_2 = z \,|\, X_0 = x) &= \sum_y \mathbb{P}\left(X_1 = y, X_2 = z \,|\, X_0 = x\right) \\ &= \sum_y P(x, y)Q(y, z)\end{aligned}$$

where in the homogeneous case, $P = Q$; however, here we prefer the more clear notation as above. Pitman jokes that generations of mathematicians developed a surprisingly compact form for this, which is matrix multiplication. If $P, Q$ are matrices, this is simply:

$$\sum_y P(x, y)Q(y, z) = PQ(x, z)$$

where we take the $x, z$th element of the resulting matrix $PQ$.

### 2.3.1 Review: Matrix Multiplication

Assuming $P, Q, R$ are $S \times S$ matrices, where $S$ is the label set of indices, then Pitman notes that indeed,

$$PQR := (PQ)R = P(QR)$$

via the associativity of matrix multiplication. This is true for all finite matrices. As a side comment, this is also true for infinite matrices, provided they are nonnegative $\geq 0$ (of course, if we have signed things, then summing infinite arrays in different

orders may cause issues). For our purposes, all our entries are nonnegative, so we have no issues.
Now, recall that typically, matrix multiplication is not commutative; that is,

$$PQ \neq QP$$

However, one easy (and highly relevant) case:
If our chain has homogeneous transition probabilities: $P, P, P, P$. If Pitman asks us what is the probability that $X_n = z$ if we knew $X_0 = x$, then we iterate what we found for 2 steps

$$\mathbb{P}(X_n = z \,|\, X_0 = x) = \underbrace{PPP\cdots P}_{n \text{ times}}(x, z) =: \boxed{P^n(x, z)}$$

Again, Pitman notes we have a very happy 'coincidink' (coincidence): If we take an $n$-step transition matrix (TM) of a markov chain (MC) with homogeneous probabilities $P$, this is equivalent to simply $P^n$, the $n$th power of matrix $P$. We can bash this out with computers, but Pitman notes there are techniques of diagonalizing and spectral theory to perform high powers of matrices. Realize that every technique here has an **immediate application** to Markov chains (with very many steps).
Note the *Chapman-Kolmogorov equations*

$$P^{m+n} = P^m P^n = P^n P^m$$

So powers of a single matrix do commute. These equations are easily justified either by algebra, or by probabilistic reasoning. See text Section 1.2 for details of the probabilistic reasoning.

## 2.4 Techniques for Finding $P^n$ for some $P$

Pitman wants to warn us that these ideas will be coming and eventually will be useful for this course. Especially, we consider matrices $P$ related to sums of independent random variables. The most basic example is a **Random Walk** on $\mathbb{N}_0 := \{0, 1, 2, \dots\}$.
In this problem one usually writes $S_n$ for the state instead of $X_n$. Our basic $X$ has $X_0, X_1, X_2, \dots$ i.i.d. according to some $P$. This is truly a trivial MC. All rows of $P$ are equal to some $p = (p_0, p_1, \dots)$ We consider:

$$S_n = X_0 + X_1 + \cdots + X_n = \text{ cumulated winnings in a gambling game}$$

(Ignore costs or losses for convenience, so natural state space of $S_n$ is $\mathbb{N}_0$ ).

## 2.5 First Example

Let $p \sim$ **Bernoulli**$(p)$ where values $0, 1$ have probabilities $q, p$, respectively. Then $S_n := X_0 + X_1 + \cdots + X_n$.

This admits the following (infinite) matrix:

$$\begin{bmatrix} * & 0 & 1 & 2 & 3 & 4 & 5 & \cdots \\ 0 & q & p & 0 & 0 & 0 & 0 & \cdots \\ 1 & 0 & q & p & 0 & 0 & 0 & \cdots \\ 2 & 0 & 0 & q & p & 0 & 0 & \cdots \\ 3 & 0 & 0 & 0 & q & p & 0 & \cdots \\ 4 & 0 & 0 & 0 & 0 & q & p & \cdots \\ 5 & 0 & 0 & 0 & 0 & 0 & q & \cdots \\ \vdots & & & & & & & \end{bmatrix}$$

Because we can only win \$1 at a time, we fill in the first row trivially.
Pitman asks us now to write down a formula for $P^n$. As a hint, he says to start with the top row.

$$P^n(0,k) = \mathbb{P}(\underbrace{X_1 + \cdots + X_n}_{n \text{ iid } \textbf{Bernoulli}(()p)} = k)$$

If this doesn't come quickly to us (the answer is trivial according to Pitman), then we should re-visit our 134 probability text (which for me happens to be by Pitman). To find $P^n$, we note $n = 1$ is known, so taking $n = 2$ for a state space of $X_0, X_1, X_2$ gives the probabilities:

$$\begin{aligned} P^2(0,0) &= q^2 \\ P^2(0,2) &= p^2 \\ P^2(0,1) &= 2pq, \end{aligned}$$

and this is the familiar **binomial distribution**. Our formula is:

$$\begin{aligned} P^n(0,k) &= \mathbb{P}(\underbrace{X_1 + \cdots + X_n}_{n \text{ iid } bern(p)} = k) \\ &= \boxed{\binom{n}{k} p^k q^{n-k}}. \end{aligned}$$

Now being at an initial fortune $i$, we have:

$$P^n(i,k) = \binom{n}{k-i} p^{k-i} q^{n-(k-i)}.$$

## 2.6 Second Example: More Challenging

Now consider the same problem, same setup, but now with $X_1, X_2, \dots$ are i.i.d. with the distribution $(p_0, p_1, p_2, \dots)$ (perhaps all strictly positive) instead of $(q, p, 0, 0, 0, \dots)$. We are interested in the distribution of our Markov Chain after $n$ steps. Taking the same method, it's enough to discuss the distribution of $S_n = X_1 + \cdots + X_n$, because we just shift by $i$ to $S_0 = i$.

Our matrix is now:

$$\begin{bmatrix} * & 0 & 1 & 2 & 3 & 4 & 5 & \cdots \\ 0 & p_0 & p_1 & p_2 & \cdots & \cdots & \cdots & \cdots \\ 1 & 0 & p_0 & p_1 & p_2 & \cdots & \cdots & \cdots \\ 2 & 0 & 0 & p_0 & p_1 & p_2 & \cdots & \cdots \\ 3 & 0 & 0 & 0 & p_0 & p_1 & p_2 & \cdots \\ 4 & 0 & 0 & 0 & 0 & p_0 & p_1 & \cdots \\ 5 & 0 & 0 & 0 & 0 & 0 & p_0 & \cdots \\ \vdots & & & & & & & \end{bmatrix}$$

Again, to get closer to induction, we take $n = 1$ to $n = 2$ steps (with $S_0 = 0$). In matrix notation, we have:

$$\mathbb{P}_0(S_2 = k) = \sum_{j=0}^{k} P(0, j)P(j, k),$$

where we stop at $k$ because we are only adding nonnegative variables. And in probability notation, where we start with $j$ and need to get to $k$ (so we move $k - j$) we have:

$$\mathbb{P}_0(S_2 = k) = \sum_{j=0}^{k} \mathbb{P}(X = j)\mathbb{P}(X = k - j),$$

and either way (of the above two), this ends up being equal to:

$$\mathbb{P}_0(S_2 = k) = \sum_{j=0}^{k} P_j P_{k-j}.$$

So in conclusion, we have found

$$P^2(0, k) = \sum_{j=0}^{k} P_j P_{k-j}$$

where we want to know the name of this operation: **discrete convolution** (so that we know what to look up!). This gets us from a distribution of random variables to the distribution of their sum. There is a "brilliant idea" (as given by Pitman) Consider the power series (of the generating function) $G(z) := \sum_{n=0}^{\infty} p_n z^n$, where taking

$$\left(p_0 + p_1 z + p_2 z^2 + \cdots\right)\left(p_0 + p_1 z + p_2 z^2 + \cdots\right)$$

yields that $\sum_{j=0}^{k} P_j P_{k-j}$ is simply the coefficient of a particular term. Pitman gives us a slick notation:

$$\begin{aligned} P^2(0, k) &= \sum_{j=0}^{k} P_j P_{k-j} \\ &= \left[z^k\right] \underbrace{\left(\sum_{n=0}^{\infty} p_n z^n\right)^2}, \end{aligned}$$

which is just the coefficient of $z^k$ in the under-braced expression.
Repeating this convolution, we move forward from $n = 2$, by induction on $n$ if you want to be careful:

$$P^n(0,k) = [z^k][G(z)]^n$$

**Example:** Pitman asks us to evaluate via Wolfram Alpha dice rolls $(p_0, p_1, \dots) = \left(\underbrace{\frac{1}{6}, \frac{1}{6}, \dots, \frac{1}{6}}_{6}, 0, \dots\right)$ and want to find: $P^4(0,5)$ for dice rolls $= \mathbb{P}(S_4 = 5)$.
We have

$$\begin{aligned}\left[\frac{1}{6}(z+z^2+z^3+z^4+z^5+z^6)\right]^4 = \frac{1}{6^4}(&z^{24}+4z^{23}+10z^{22}+20z^{21}\\ &+35z^{20}+56z^{19}+80z^{18}\\ &+104z^{17}+125z^{16}+140z^{15}\\ &+146z^{14}+140z^{13}+125z^{12}\\ &+104z^{11}+80z^{10}+56z^9+35z^8\\ &+20z^7+10z^6+\underbrace{4z^5}+z^4)\end{aligned}$$

which implies

$$P^4(0,5) = \frac{4}{6^4}$$

where we took the coefficient of the under-braced term (power of 5). Pitman credits the invento of this method, Laplace. This is unusually simple but demonstrates the general method. Of course, $P^4(0,5) = \frac{4}{6^4}$ is rather trivial because you can count the number of dice patterns on one hand. But the evaluations of $P^4(0,k)$ for all $4 \le k \le 24$ above are not so trivial. This method can be used to prove all the familiar properties of sums of independent discrete variables, e.g. sums of Poissons are Poisson. You should try it for that purpose.

# LECTURE 3

# Hitting Times, Strong Markov Property, and State Classification

## 3.1 Hitting Times

This discussion follows quite closely §1.3 of the text. See text for further developments and details. Consider a Markov Chain with fixed transition matrix $P$ and state space $\mathcal{S}$. Consider states $x, y \in \mathcal{S}$ [1]. We are interested in the *first hitting time* or *first passage time*

$$T_B := \min\{n \geq 1 : X_n \in B\}$$

for some target set of states $B$. In words, the first time at or after time 1 that the chain hits the set of states $B$. An immediate pedantic issue with is what if the chain never reaches $B$, that is $X_n \notin B \ \forall n$? In this case, we need to make the following (very useful) convention

$$\min\{\varnothing\} = \inf\{\varnothing\} := \infty$$

where $\infty$ is a conventional element assumed to be greater than every positive integer.

**Strong Markov Property (SMP)**

Start with $X_0, X_1, X_2, \dots$ which is a Markov chain with transition matrix $P$, and any initial distribution for $X_0$. Conditionally given $T_B = n < \infty$ and $X_n = y \in B$, the following process

$$(X_n, X_{n+1}, X_{n+1}, \dots)$$

is a copy of the original Markov Chain with transition matrix $P$ conditioned to start in state $y$.

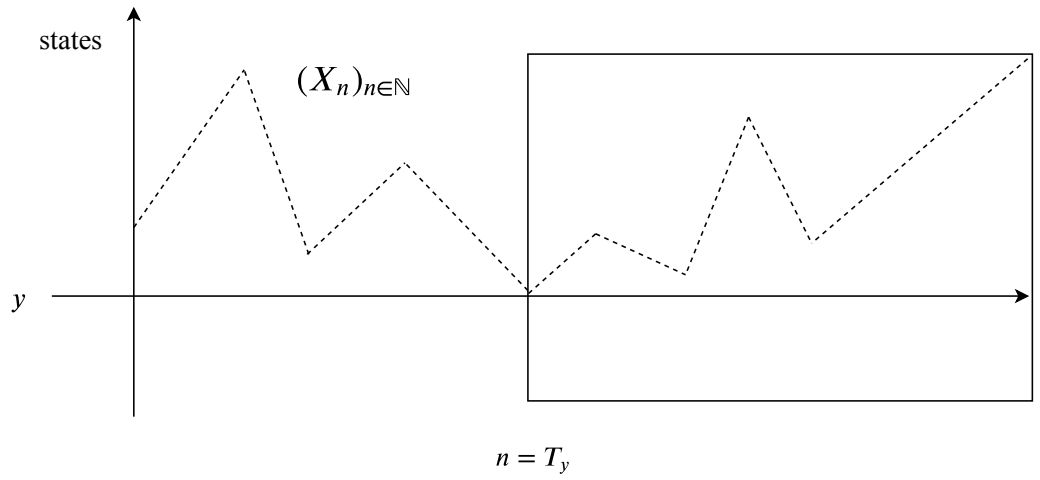
[1] sometimes $i, j \in \mathcal{S}$

In particular, the distribution of $X_{n+1} \,|\, X_n = y$ is $P(y, \cdot)$, that is, for any state $z$

$$\mathbb{P}(X_{n+1} = z \,|\, T_y = n, X_n = y) = P(y, z), \text{ also}$$
$$\mathbb{P}(X_{n+1} = z, X_{n+2} = w \,|\, T_y = n, X_n = y) = P(y, z)P(z, w)$$

and so on, an infinite list of equations.

*Proof.* See Durett page 14. □




**Remark:** In discrete time (even with a general state space), *all* Markov chains with a homogeneous transition mechanism have the Strong Markov Property. Now, we can use the SMP to discover and prove things about Markov chains.

## 3.2 Iterating

From $T_y^0 = 0$, for $k \geq 1$

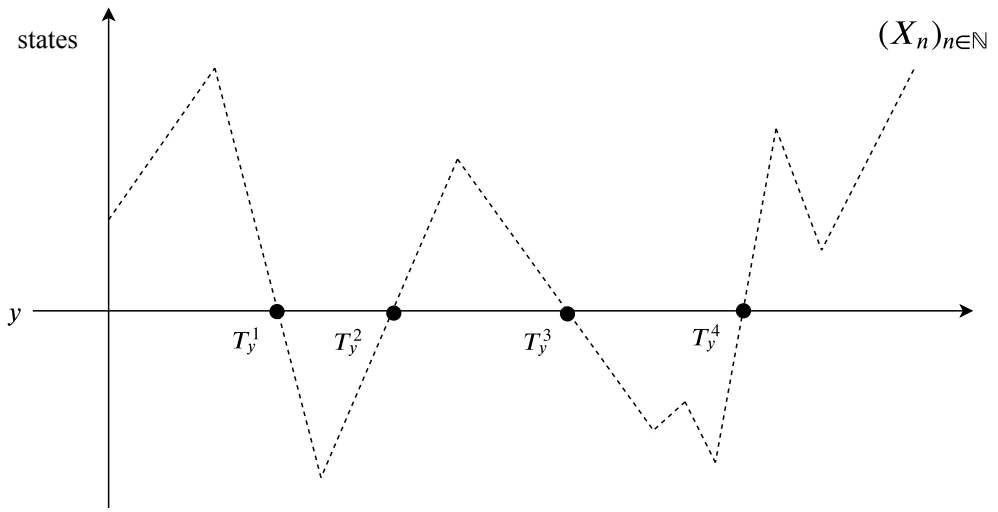
$$T_y^k := \min\{n > T_y^{k-1} : X_n = y\}.$$

Note $T_y^k$ is a $k^{\text{th}}$ iterate of the scheme for defining $T_y = T_y^1$, not the $k^{\text{th}}$ power of $T_y$. Suppose we have a path that hits the state $y$ a finite number of times $n \geq 1$, say exactly four times: $T_y, T_y^2, T_y^3, T_y^4$. Then by our convention, we say that $T_y^5 = \infty$. Consider the random variable which is the total number of hits of $y$, at any time $n \geq 1$, deliberately not counting a hit at time 0 if $X_0 = y$:

$$N_y := \sum_{n=1}^{\infty} \mathbf{1}(X_n = y).$$

The possible values of $N_y$ are $\{0, 1, 2, \ldots, \infty\}$, an infinite time horizon. By the logic of the definitions, there is the identity of events:

$$(N_y = 0) = (T_y = \infty).$$




As another example, consider $(N_y \geq 1)$, the complement of $(N_y = 0)$ because we include $\infty$ as a part of $(N_y \geq 1)$. Hence

$$(N_y \geq 1) = (T_y < \infty).$$

Recall $N_y := \sum_{n=1}^{\infty} \mathbf{1}(X_n = y)$ is simply counting the number of hits on $y$. Pitman asks the audience to find expressions in terms of $T_y^k$ for the left hand side of

$$\begin{aligned}
(N_y \geq 3) &= (T_y^3 < \infty) \\
(N_y = 3) &= (T_y^3 < \infty, T_y^4 = \infty) \\
(N_y \geq k) &= (T_y^k < \infty) \\
(N_y = k) &= (T_y^k < \infty, T_y^{k+1} = \infty).
\end{aligned}$$

Now let's discuss the probabilities. The SMP gives

$$\mathbb{P}_y(T_y^k < \infty) = \rho_y^k,$$

where $k$ on the RHS is a power, and $k$ on the LHS is an index. Now taking $k = 1$, we have the definition of $\rho_y$:

$$\mathbb{P}_y(T_y < \infty) = \rho_y$$

called the *first return probability* of state $y$. Now how to get from $\rho_y$ to $\rho_y^2$? Basically, this is by the SMP. Observe

$$(T_y^k < \infty) = (N_y \geq k).$$

which tells us that the probability of hitting $y$ at least $k \in \mathbb{N}_0$ times is

$$\mathbb{P}_y(N_y \geq k) = \rho_y^k$$

If we want to find the point probability that $N_y = k$, we take

$$\begin{aligned}\mathbb{P}_y(N_y = k) &= \mathbb{P}_y(N_y \geq k) - \mathbb{P}_y(N_y \geq k+1) \\ &= \rho_y^k - \rho_y^{k+1} \\ &= \boxed{\rho_y^k(1-\rho_y)}\end{aligned}$$

Now *either* $\rho_y = 1$ and this probability is 0 for all $k < \infty$, pushing all the probability to $\mathbb{P}_y(N_y = \infty) = 1$, *or* $\rho_y < 1$ in which case the probability distribution (starting at $y$) of $N_y := \sum_{n=1}^{\infty} \mathbf{1}(X_n = y)$ is geometric($p$) on $\{0, 1, 2, \ldots\}$ with parameter

$$p = 1 - \rho_y = \mathbb{P}_y(T_y = \infty) = \mathbb{P}_y(N_y = 0).$$

Notice, via the tail-sum formula for $\mathbb{E}$ of a non-negative integer valued random variable

$$\mathbb{E}_y N_y = \sum_{k=1}^{\infty} \mathbb{P}_y(N_y \geq k) = \sum_{k=1}^{\infty} \rho_y^k = \frac{\rho_y}{1-\rho_y} = \frac{q}{p}$$

for $q = \rho_y$ and $p = 1 - \rho_y$, in agreement with the standard formula for $\mathbb{E}$ of a geometric($p$) variable.

## 3.3 State Classification

There are two cases to consider.

(1) $y$ is *transient* : $0 \leq \rho_y < 1$. This implies that our expected number of visits is:

$$\mathbb{E}_y N_y = \frac{\rho_y}{1-\rho_y} < \infty$$

which implies

$$\mathbb{P}_y(N_y < \infty) = 1,$$

which says that if we have a transient state, then we only return to $y$ a finite number of times. In other words, after some point, the Markov chain never visits $y$ again.

(2) $y$ is *recurrent* : $\rho_y = 1$. In other words, $\mathbb{P}_y(N_y = \infty) = 1$ in that given any number of hits, we are sure to hit $y$ again.

### 3.3.1 Constructing $\rho_y$

Here is an explicit a formula:

$$\begin{aligned}\rho_y &= \mathbb{P}_y(T_y = 1) + \mathbb{P}_y(T_y = 2) + \mathbb{P}_y(T_y = 3) + \cdots \\ &= P(y,y) + \sum_{y_1 \neq y} P(y,y_1)P(y_1,y) + \sum_{y_1 \neq y} \sum_{y_2 \neq y} P(y,y_1)P(y_1,y_2)P(y_2,y) + \cdots\end{aligned}$$

But this is not so nice to work with.

**Exercise:** Show that for $n \geq 2$ the $n$th term $\mathbb{P}_y(T_y = n)$ can be expressed in matrix notation as $P(y, \cdot)K^{n-2}P(\cdot, y)$ for a suitable matrix $K$ to be determined. Note that $K$ is *sub-stochastic* with non-negative entries and row sums $\leq 1$.

## 3.4 Lemma 1.3

Reference. Durrett's, *Essentials of Stochastic Processes* Page 16. Take $B$ to be a set of states. Hypothesis: Suppose the probability starting at $x$ that $T_B \leq k$ is at least $\alpha > 0$ for some fixed $k$ and all $x$:

$$\mathbb{P}_x(T_B \leq k) \geq \alpha > 0 \text{ for all states } x$$

Then

$$\mathbb{P}_x(T_B > nk) \leq (1-\alpha)^n.$$

As an example where the hypothesis is obviously satisfied, consider the Gambler's Ruin chain with state 0 and $N$ as absorbing states. That is, $B = \{0, N\}$, with $P(i, i+1) = p$ and $P(i, i-1) = q$ for $0 < i < N$. Then this condition holds with $k = N$ and

$$\alpha = p^N + q^N > 0$$

because no matter where you start away from the boundary states, a sequence of either at most $N$ consecutive up steps or $N$ consecutive down steps will get you to the boundary.

*Proof.* The conclusion is obvious by taking complements if $n = 1$: $\mathbb{P}_x(T_B > k) \leq 1 - \alpha$ for all $x$. Now by induction on $n$. Observe that

$$\begin{aligned}
\mathbb{P}_x(T_B > (n+1)k) &= \mathbb{P}_x(T_B > nk \text{ and after time } nk \text{ before time } (n+1)k \text{ still don't hit } B) \\
&= \sum_{y \notin B} \mathbb{P}_x(T_B > nk, X_{nk} = y, \text{ do not hit } B \text{ before time } (n+1)k) \\
&= \sum_{y \notin B} \mathbb{P}_x(T_B > nk, X_{nk} = y)\mathbb{P}_y(T_B > k) \\
&\leq \sum_{y \notin B} \mathbb{P}_x(T_B > nk, X_{nk} = y)(1-\alpha) \\
&= \mathbb{P}_x(T_B > nk)(1-\alpha) \\
&\leq (1-\alpha)^n(1-\alpha) = (1-\alpha)^{n+1}.
\end{aligned}$$

□

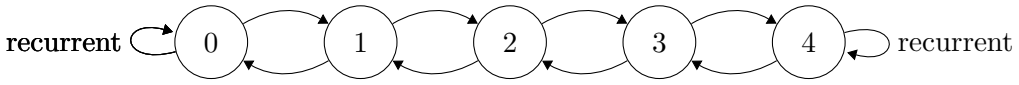
Pitman mentions Kai Lai Chung of Stanford who presents this Lemma in terms of a pedestrian repeatedly crossing the road. Depending on visibility and weather conditions, (current state), there is a varying chance of making it to the other side in $k$ steps. But suppose that no matter how favorable the visibility and weather conditions, there is always at least a small chance $\alpha$ that the pedestrian gets killed while crossing the road. Then, if the pedestrian repeatedly attempts to cross the road, eventually they will get killed, with a geometric bound as above on how long that takes. In the language of Markov chains, if there is always at least a some strictly postive chance of the chain reaching a boundary $B$ in the next $k$ steps, no matter where it starts, eventually the chain will hit such a boundary set. Observe that in standard real numbers,

$$0 \leq \mathbb{P}_x(T_B = \infty) \leq \mathbb{P}_x(T_B > nk) \leq (1-\alpha)^n, \ \forall_n$$

implies

$$\mathbb{P}_x(T_B = \infty) = 0$$

Back to Gambler's Ruin. Suppose $0 < p < 1$. In the language of transient and recurrent states, every state $x \notin \{0, N\}$ is transient! Moreover, $x \in \{0, N\}$ is recurrent. Here's a Gambler's Ruin chain for $N = 4$.




**Irreducible Matrix**

We say that a matrix $P$ is **irreducible** if

$$\forall \ x, y \in \mathcal{S}, \ \exists n : P^n(x, y) > 0$$

In words, for every pair of states $x, y$, it is possible to get from $x$ to $y$ in some number $n$ of steps. Here $n = n(x, y)$ is a function of $x, y$. If matrix $P$ is irreducible, then either

all states are recurrent **or** all states are transient

We then say the matrix $P$ is "recurrent" or "transient", meaning that it drives a chain all of whose states are recurrent or transient, as the case may be.

This and other properties of states of a chain with irreducible transition matrix $P$, which hold for one state iff they hold for all states, are called *solidarity properties*. Other examples are the conditions that $\mathbb{E}_x T_x < \infty$, and that state $x$ is *aperiodic* as discussed in the text, or that state $x$ has a particular period $d$.

Easy fact. (Pigeon hole principle: with a finite state space, and infinitely many steps, some state must be hit infinitely often): Suppose $\mathcal{S}$ is finite and $P$ is irreducible. Then $P$ is recurrent. Notice the Gambler's Ruin chain exhibits a matrix that is **not** irreducible, which can be seen via the definition above and the requirement that there exists some $n$ where $P^n(x, y) > 0$.

# LECTURE 4

# Exchangeability, Stationary Distributions, and Two State Markov Chains

## 4.1 Sampling without Replacement

Consider $(X_1, \ldots, X_N)$ an exhaustive random sample without replacement from a box of $N = A + B$ tickets, with $A$ labeled 1 (success) and $B$ labeled 0 (fail). Let $S_0 := 0$ and $S_n := X_1 + \cdots + X_n$ the number of 1s and $\bar{S}_n := n - S_n$ the number 0s in the first $n$ places of the sample. Then $((\bar{S}_n, S_n), 0 \le n \le N)$ is a Markov chain with transition matrix

$$P((f,s),(f+1,s)) = \frac{B-f}{A+B-f-s} \tag{4.1}$$

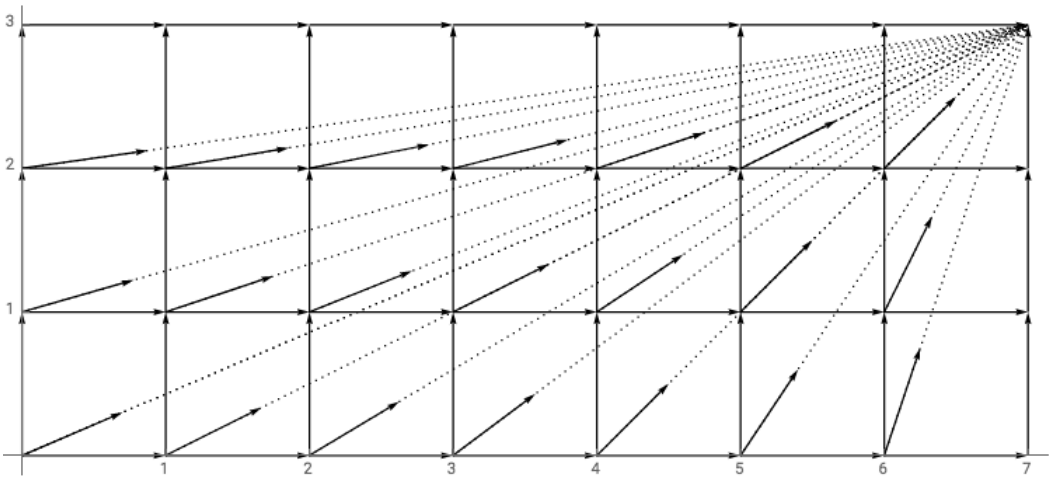
$$P((f,s),(f,s+1)) = \frac{A-s}{A+B-f-s} \tag{4.2}$$

and all other entries 0. In the following diagrams, with Cartesian coordinates $(f, s)$, the the horizontal scale counts the number of failures $f$, the vertical scale counts the number successes $s$, and the sum of coordinates is the number of draws $n = f + s$. The chain $W_n := (\bar{S}_n, S_n)$ starts at the origin $(\bar{S}_0, S_0) = (0, 0)$ at time $n = 0$, and terminates at $(B, A)$ at time $n = N = A + B$.

- each step right in this chain increments the first component $f$ of $(f, s)$ to $f+1$ for a failure (0) in the sequence $(X_1, \ldots, X_N)$.
- each step up in this chain increments of the second component $s$ of $(f, s)$ to $s + 1$ for a success (1) in the sequence $(X_1, \ldots, X_N)$.

Altogether there are $N = A + B$ steps, with $A$ steps up and $B$ steps right. All $\binom{A+B}{A}$ possible paths of the chain are equally likely.

Figure 4.1: Transition probabilities for sampling without replacement. Here $A = 3, B = 7$. The only possible transitions are one step up or one step right, following arrows on the grid of possible states $(f, s)$, with $0 \le f \le 7$ the number of failures (0) and $0 \le s \le 3$ the number of successes, after $f + s$ draws without replacement from 7 values 0 and 3 values 1. The pair of transition probabilities out of each state $(f, s)$ is represented by a vector with tail $(f, s)$ and head $(f + q, s + p)$ for $q = P((f, s), (f + 1, s)$ and $p = P((f, s), (f, s + 1)$ as above. The head of each vector is the conditional mean of the random vector $(\bar{S}_{n+1}, bS_{n+1})$ given $(\bar{S}_n = f, S_n = s)$ with $n = f + s$. All the transition vectors point towards the terminal state of the chain at $(B, A) = (7, 3)$ after $n = 10$ draws.




## 4.2 Exchangeability and Reversibility

It is an important general property of a sample without replacement $(X_1, \ldots, X_N)$ that these random variables are *exchangeable*, meaning that for every permutation $\sigma$ of $[N] := \{1, \ldots, N\}$

$$(X_{\sigma(1)}, \ldots, X_{\sigma(N)}) \stackrel{d}{=} (X_1, \ldots, X_N) \tag{4.3}$$

where $\stackrel{d}{=}$ denotes equality in distribution. [2] See Pitman *Probability* Section 3.6. In particular, this holds for the sample $(X_1, \ldots, X_N)$ of $A$ ones and $B$ zeros considered here. Except in degenerate cases, the sequence $(X_1, \ldots, X_N)$ is not Markov: given $X_1, \ldots, X_n$ the conditional probability that $X_{n+1} = 1$ is $(A - S_n)/(N - n)$ which is typically not just a function of $X_n$, but involves all of the previous values $X_1, \ldots, X_n$ through their sum $S_n$, the number of 1s in the first $n$ draws without replacement. However, in the model of sampling without replacement from $A$ values 1 and $B$ values 0, it is instructive to study the common joint distribution of every pair of draws

$$(X_{\sigma(1)}, X_{\sigma(2)}) \stackrel{d}{=} (X_1, X_2) \qquad (\sigma(1) \ne \sigma(2)). \tag{4.4}$$

The joint probability function of this pair of draws is obtained by assuming that all $(A + B)(A + B - 1)$ possible pairs of different tickets are equally likely to appear on

the first and second draws. By counting pairs of different tickets

$$\mathbb{P}(X_1 = 0, X_2 = 0) = \frac{B(B-1)}{(A+B)(A+B-1)} \tag{4.5}$$

$$\mathbb{P}(X_1 = 1, X_2 = 1) = \frac{A(A-1)}{(A+B)(A+B-1)} \tag{4.6}$$

$$\mathbb{P}(X_1 = 0, X_2 = 1) = \mathbb{P}(X_1 = 1, X_2 = 0) = \frac{AB}{(A+B)(A+B-1)} \tag{4.7}$$

The last equality of the two off-diagonal probabilities is important. In counting pairs of different tickets, this comes from $BA = AB$: the number of different ways to get 1 followed by 0 from the first two draws is equal to the number of different ways to get 0 followed by 1. In probabilistic terms, the equality (4.7) of off-diagonal probabilities gives the equality in distribution

$$(X_2, X_1) \stackrel{d}{=} (X_1, X_2) \tag{4.8}$$

Such a pair of random variables $(X_1, X_2)$ is called either *reversible* or *exchangeable*. In terms of a joint distribution table of numerical random variables $X_1$ and $X_2$, displayed in Cartesian coordinates with values of $X_1$ horizontal and values of $X_2$ vertical, such a distribution is symmetric with respect to reflection accross the set of diagonal values $(X_1 = X_2)$:

$$\mathbb{P}(X_2 = x, X_1 = y) = \mathbb{P}(X_1 = x, X_2 = y) \tag{4.9}$$

for all possible values $x$ and $y$. If $x = y$ this identity is trivial. If $X_1$ and $X_2$ have only two possible values 0 and 1, there are only two possible off-diagonal pairs $(0, 1)$ and $(1, 0)$. So for indicator variables, $(X_1, X_2)$ is reversible iff (4.9) holds for the single pair $(x, y) = (0, 1)$, as it does in (4.7).

In general, for $N \geq 2$, a sequence of random variables $(X_1, \ldots, X_N)$ is called *reversible* if (4.3) holds just for the single permutation $\sigma$ which reverses the order of indices, that is

$$(X_N, \ldots, X_1) \stackrel{d}{=} (X_1, \ldots, X_N). \tag{4.10}$$

For a random vector of length $N = 2$, reversible is the same as exchangeable, because there are only two permutations of $\{1, 2\}$, the identity permutation, for which there is nothing to check, and the permutation which switches 1 and 2. For a random vector of length $N \geq 3$ exchangeable implies reversible, but not conversely. For instance, if $N = 3$ there are $3! - 1 = 5$ permutations besides the identity, and reversibility only involves an identity in distribution for just one of these 5 permutations. See also further discussion below.

In sampling without replacement from $B$ values 0 and $A$ values 1, the first variable $X_1$ has distribution $\mathbb{P}(X_1 = i) = \pi_i$ given by

$$(\pi_0, \pi_1) = \frac{(B, A)}{A + B} \tag{4.11}$$

and the step from $X_1$ to $X_2$ is made according to the transition probability matrix

$$P = \begin{bmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{bmatrix} = \frac{1}{A+B-1} \begin{bmatrix} B-1 & A \\ B & A-1 \end{bmatrix} \tag{4.12}$$

This description of the joint distribution of $(X_1, X_2)$ is logically equivalent to the previous description of the joint probability function (4.5)-(4.7) by four applications of the product rule $\mathbb{P}(CD) = \mathbb{P}(C)\mathbb{P}(D \,|\, C)$. Either description implies $(X_1, X_2) \stackrel{d}{=} (X_2, X_1)$ and hence $X_2 \stackrel{d}{=} X_1$. More algebraically, the distribution of $X_2$ is determined by

$$\begin{aligned} \mathbb{P}(X_2 = 1) &= \mathbb{P}(X_1 = 0, X_2 = 1) + \mathbb{P}(X_1 = 1, X_2 = 1) \\ &= \mathbb{P}(X_1 = 0)\mathbb{P}(X_2 = 1 \,|\, X_1 = 0) + \mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 1 \,|\, X_1 = 1) \\ &= \frac{B}{(A+B)} \frac{A}{(A+B-1)} + \frac{A}{(A+B)} \frac{(A-1)}{(A+B-1)} \\ &= \frac{A(A+B-1)}{(A+B)(A+B-1)} = \frac{A}{A+B} \end{aligned}$$

The probability $\mathbb{P}(X_2 = 0)$ can be found similarly, or by

$$\mathbb{P}(X_2 = 0) = 1 - \mathbb{P}(X_1 = 1)$$

since the only possible values of $X_2$ are 0 and 1. The simple algebraic structure of this joint distribution of the pair of indicator variables $(X_1, X_2)$ derived from sampling without replacement from $A$ values 1 and $B$ values 0 is worth understanding thoroughly, especially the reversibility $(X_1, X_2) \stackrel{d}{=} (X_2, X_1)$ which implies $X_2 \stackrel{d}{=} X_1$. The algebraic structure of this joint distribution of dependent indicators $(X_1, X_2)$, with two parameters $A$ and $B$, in terms of which the algebra comes out very nicely, turns out to be shared by every pair of exchangeable indicator variables $X_1$ and $X_2$ which are not independent.

**Remark.** There is a simple construction of dependent indicator variables $(X_1, X_2)$ with various levels of dependence, which may be helpful. Draw a Venn diagram with two regions $V_1$ and $V_2$, each of area $p$, for some fixed $0 < p < 1$. Let $X_i$ be the indicator of $V_i$. Now move the regions around in the diagram to vary their overlap. There is no loss of generality in making each region a rectangle of height 1 over some base interval of length $p$. Now each $V_i$ may be identified with a subinterval of $[0, 1]$ of length $p$, and it is just a matter of moving around two subintervals of $[0, 1]$, each of length $p$, and considering the possible length of overlap of these two intervals $V_i$. You can treat each $X_i$ as a function $X_i(\omega)$ of $\omega \in [0, 1]$ with value $X_i(\omega) = 1$ if $\omega$ falls in some interval $V_i$ of length $p$, and value 0 if $\omega \notin V_i$. So $\mathbb{E}X_i = 1 \times p + 0 \times (1 - p) = p$. The most the two intervals can overlap is if they are identical, which makes $\mathbb{E}X_1X_2 = p$ and $\mathbb{P}(X_1 = X_2) = 1$, with correlation 1. By an elementary argument (Boole's inequality)

$$0 \leqq \mathbb{E}(1 - X_1)(1 - X_2) = 1 - 2p + \mathbb{E}(X_1X_2)$$

the least they can overlap is if

$$\mathbb{P}(V_1V_2) = \mathbb{E}(X_1X_2) = (2p-1)_+$$

which is 0 if $0 \leqq p \leqq 1/2$, and $2p-1$ if $1/2 < p \leqq 1$. This bound is achieved by $V_1 = [0,p]$ and $V_2 = [1-p,1]$. Any value of $\mathbb{P}(V_1V_2)$ in this range $[(2p-1)_+, p]$ determines a possible exchangeable joint distribution of $(X_1, X_2)$ with $\mathbb{E}X_1 = \mathbb{E}X_2 = p$, which is realized on $[0,1]$ by any two intervals of length $p$ with the assigned overlap. And for an allowed value of $\mathbb{E}X_1X_2$, Every exchangeable joint law of a pair of indicators $(X_1, X_2)$ is completely determined by the common value of $p := \mathbb{E}X_i$ and the value of $\mathbb{E}X_1X_2$ in $[(2p-1)_+, p]$. Always included in the range of possible values of $\mathbb{E}(X_1X_2)$ is the value $p^2$ for independent $X_i$. Thus

$$(2p-1)_+ < p^2 < p \text{ for } 0 < p < 1$$

as you should check by sketching graphs of all three functions of $p$ over $[0,1]$. Indicators $X_1$ and $X_2$ are called *positively dependent* or *negatively dependent* according to the sign of $\mathrm{Cov}(X_1, X_2) := \mathbb{E}X_1X_2 - \mathbb{E}X_1\mathbb{E}X_2$.

## 4.3 Stationary Distributions

For a pair of discrete random variables $(X_1, X_2)$, write either

$$X_1 \sim \pi \text{ and } (X_2 \,|\, X_1) \sim P(X_1, \cdot\,)$$

or

$$\mathbb{P}(X_1 \in \cdot) = \pi(\cdot) \text{ and } \mathbb{P}(X_2 \in \cdot \,|\, X_1) = P(X_1, \cdot)$$

to mean that $X_1$ has distribution $\pi$, and the conditional distribution of $X_2$ given $X_1 = x$ is given by the row $P(x, \cdot)$ of some transition probability matrix $P$, for every possible value $x$ of $X_1$. This prescription of a distribution $\pi$ for $X_1$ and the conditional distribution $P(X_1, \cdot)$ for $X_2$ given $X_1$ uniquely determines the joint distribution of $X_1$ and $X_2$, and is equivalent to the formula for the joint probability function of $(X_1, X_2)$

$$\mathbb{P}(X_1 = x, X_2 = y) = \pi(x)P(x,y)$$

as $x$ and $y$ range over all possible values of $X_1$ and $X_2$ respectively. The distribution of $X_2$ is then determined by the matrix operation $X_2 \sim \pi P(\,\cdot\,)$:

$$\begin{aligned}\mathbb{P}(X_2 = y) &= \sum_x \mathbb{P}(X = x)\mathbb{P}(Y = y \,|\, X = x) \\ &= \sum_x \lambda(x)P(x,y) = (\pi P)(y).\end{aligned}$$

In particular, for $X_1$ and $X_2$ with the same set of possible values,

$$X_1 \stackrel{d}{=} X_2 \iff \pi = \pi P \qquad \text{meaning} \tag{4.13}$$

$$\sum_x \pi(x)P(x,y) = \pi(y) \text{ for all states } y. \tag{4.14}$$

Then $\pi$ is called a *stationary* (or *invariant* or *equilibrium* or *steady state*) *distribution* for the transition matrix $P$. This condition $X_1 \stackrel{d}{=} X_2$ is implied by the stronger *reversibility condition*

$$(X_1, X_2) \stackrel{d}{=} (X_2, X_1) \iff \pi(x)P(x,y) = \pi(y)P(y,x) \text{ for all } x, y \tag{4.15}$$

when $\pi$ is called a *reversible equilibrium distribution* for the transition matrix $P$. The equations in (4.14) are called *balance equations* while those in (4.15) are called *detailed balance equations*. If there are $|S| = N$ states, there are $N$ different balance equations, and $\binom{N}{2}$ different detailed balance equations. For a prescribed transition matrix $P$, to solve either system of equations to obtain a stationary probability distribution $\pi$ you must add the constraint $\sum_x \pi(x) = 1$. Issues of existence and uniqueness of solutions of these balance equations are treated in the text and will be discussed further in following lectures. It is often easy to see directly that some distribution $\pi$ provides a reversible equilibrium for a particular transition matrix $P$. This just involves checking $\pi(x)P(x,y) = \pi(y)P(y,x)$ for $x \neq y$, which was already noticed above in the case of sampling without replacement, by counting outcomes. No summations were involved.

**Exercise:** The text on page 22 has a nice *sand metaphor* for the balance equations. Explain the meaning of detailed balance in terms of the sand metaphor.
Here are some easy consequences of these definitions, all of which you should be able to derive for yourself without consulting any text::

- If $(X_0, X_1, X_2, \dots, )$ is a Markov chain with homogeneous transition matrix $P$ and $X_0 \sim \pi$ with $\pi P = \pi$, then for all positive integers $n$ and $N$

  $$(X_0, \dots, X_N) \stackrel{d}{=} (X_n, \dots, X_{n+N}).$$

  A stochastic process $(X_0, X_1, X_2, \dots, )$ with this property is called *stationary*. In words: the finite dimensional distributions of a stationary process are invariant with respect to a shift in time.

- If a distribution $\pi$ solves the detailed balance equations for $P$ , then $\pi$ also solves the balance equations for $P$;

- If $X_1$ and $X_2$ are random variables, each with only two possible values, then $X_1 \stackrel{d}{=} X_2$ iff $(X_1, X_2) \stackrel{d}{=} (X_2, X_1)$;

- If $X_1$ and $X_2$ have three or more possible values, it is possible to have $X_1 \stackrel{d}{=} X_2$ without $(X_1, X_2) \stackrel{d}{=} (X_2, X_1)$. An example on three states is $(X_1, X_2)$ with transition matrix

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

  corresponding to deterministic rotation by one step around three states $0, 1, 2$ arranged in a circle. The unique equilibrium distribution is the uniform distribution $\pi = (1, 1, 1)/3$, and this equilibrium is not reversible. This is an

example of a *periodic chain* with *period* 3. In general for transition matrix $P$, the *period of a state* $x$ is the greatest common divisor $d(x)$ of the set of positive integers $n$ such that $P^n(x,x) > 0$. For an irreducible matrix $P$, Lemma 1.17 of the text shows that $d(x) \equiv d$ for some positive integer $d$, called the *period of* $P$.

- The above transition matrix $P$ on 3 states is *doubly stochastic*, meaning that all its row sums are 1 and all its column sums are 1. For an $S \times S$ matrix $P$ with $S$ finite, the uniform distribution $\pi$ on $S$ is $P$-invariant iff $P$ is doubly stochastic. (Text Theorem 1.14)

- If $\pi$ is $P$-invariant, then $\pi$ is $P^n$-invariant for every positive integer $n$. Here $P^n$ is the $n$th iterate of the transition matrix $P$, which is the $n$-step transition matrix for a Markov chain with homogeneous transition matrix $P$.

- In terms of a Markov chain $(X_0, X_1, \ldots)$ with $X_0 \sim \pi$ and homogeneous transition matrix $P$, an equilibrium $\pi$ for $P$ is reversible iff for every $N \geq 1$ there is the equality in distribution

$$(X_0, \ldots, X_N) \stackrel{d}{=} (X_N, \ldots, X_0).$$

See the text §1.4 and §1.5 for further discussion and many examples. Two less obvious but very important facts, treated in the text in §1.6, 1.7 and 1.8 are :

- if $P$ is irreducible with a finite number of states, then there is a unique stationary distribution $\pi$ for $P$, specifically

$$\pi_j = \frac{1}{\mathbb{E}_j T_j} \tag{4.16}$$

where $\mathbb{E}_j T_j$ is the mean return time of state $j$. This is also true more generally if $P$ is irreducible and *positive recurrent*, meaning that $\mathbb{E}_j T_j < \infty$ for some (hence all ) states $j$,

- If $P$ is irreducible and positive recurrent and *aperiodic*, meaning that some (and hence every) state $x$ has period 1, then

$$\lim_{n\to\infty} P^n(i,j) = \pi_j \tag{4.17}$$

as above for all states $j$.

- Consequently, for any Markov chain $(X_n)$ with countable state space $S$ and such a transition matrix $P$, no matter what the distribution of $X_0$, there is the convergence in distribution $X_n \stackrel{d}{\to} \pi$ as $n \to \infty$, meaning

$$\lim_{n\to\infty} \mathbb{P}(X_n = j) = \pi_j \qquad (j \in S).$$

**Exercise.** Explain exactly how each of these results can be deduced from specific theorems in the text.

## 4.4 Two State Transition Matrices

Suppose $A$ and $B$ are positive integers with $A+B = N \geq 3$, and consider $(X_1, \ldots, X_N)$ an exhaustive sample without replacement from $A$ values 1 and $B$ values 0. In the sample $(X_1, X_2, X_3)$ of size 3, every pair of variables has the reversible joint distribution of $(X_1, X_2)$ displayed in (4.5) - (4.7).

**Exercise.** Check, by calculations like (4.5) - (4.7) that this sequence $(X_1, X_2, X_3)$ of exchangeable indicators is not Markovian.
For $P$ the transition matrix of $(X_1, X_2)$ displayed in (4.12), with parameters $(B, A)$, the iterates $P^n$ of $P$ have no obvious meaning in terms of the exhaustive sample $(X_k, 1 \leq k \leq N)$. In particular, the conditional distribution of $X_3$ given $X_1$ is provided by $P$, not by $P^2$, as you can see from the formula for $P^2$ for a two state Markov matrix $P$ (Homework 2). Rather, this example of $(X_1, X_2, X_3)$ derived from sampling without replacement is non-Markovian and exchangeable. The single transition matrix $P$ provides the conditional distribution of $X_i$ given $X_j$ for every $i \neq j$.
Observe that the matrix $P$ defined by (4.12), with two parameters $A$ and $B$, has row sums 1 not only for all positive integers $A$ and $B$, but also for any choice of real parameters $A$ and $B$ with $A + B - 1 \neq 0$. In fact, this construction generates every $2 \times 2$ transition matrix $P$ except for the relatively uninteresting *Bernoulli*($p$) *matrices*

$$\begin{bmatrix} q & p \\ q & p \end{bmatrix} \qquad (0 \leq p \leq 1, p + q = 1).$$

For $p = A/(A+B)$ the matrix above is associated with sampling without replacement from a population of $A$ ones and $B$ zeros. For general $0 \leq p \leq 1$, the Bernouli($p$) matrix corresponds to an unlimited sequence of independent Bernoulli($p$) trials. You can easily check the following proposition:

**Proposition 0.1.** *Let $P$ be a $2 \times 2$ transition matrix with $P_{01} \neq P_{11}$. Then $P$ is of the algebraic form* (4.12)*, that is*

$$\begin{bmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{bmatrix} = \frac{1}{A + B - 1} \begin{bmatrix} B - 1 & A \\ B & A - 1 \end{bmatrix} \tag{4.18}$$

*for a unique pair of real parameters $(B, A)$:*

$$\begin{bmatrix} B - 1 & A \\ B & A - 1 \end{bmatrix} = \frac{1}{P_{01} - P_{11}} \begin{bmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{bmatrix} \tag{4.19}$$

*Assume further that $P_{01} + P_{10} > 0$, to exclude the trivial case $P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ corresponding to $(B, A) = (0, 0)$. Then the unique stationary distribution for $P$ is $\pi = (\pi_0, \pi_1)$ defined by*

$$(\pi_0, \pi_1) = \frac{(P_{10}, P_{01})}{P_{10} + P_{01}} = \frac{(B, A)}{A + B}; \tag{4.20}$$

*This $\pi$ is a reversible equilibrium for $P$: if $X_1 \sim \pi$ then the joint distribution of $X_1$ and $X_2$ is given by the formulas (4.5)–(4.7) for sampling without replacement, without the requirement that $A$ and $B$ are positive integers. This makes*

$$Cov\,(X_1, X_2) := \mathbb{E}X_1X_2 - (\mathbb{E}X_1)(\mathbb{E}X_2) = \frac{-AB}{(A+B)^2(A+B-1)}. \tag{4.21}$$

*The range of parameters $(A, B)$ in this construction has two connected components:*

- *$A \geq 1$ and $B \geq 1$, when $X_1$ and $X_2$ are negatively dependent;*
- *$A = -a \leq 0$ and $B = -b \leq 0$, with $a + b > 0$, when $X_1$ and $X_2$ are positively dependent; then in terms of $a = -A \geq 0$ and $b = -B \geq 0$*

$$\begin{bmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{bmatrix} = \frac{1}{a+b+1} \begin{bmatrix} b+1 & a \\ b & a+1 \end{bmatrix}$$

*The transition matrix (4.4) of independent Bernoulli($p$) trials is recovered as the limit case when either $A$ and $B$ both tend to $+\infty$, or $A$ and $B$ both tend to $-\infty$, with $A/(A+B) \to p$.*

**Exercise.** Check that the formula (4.20) for the stationary distribution $\pi$ of a two state chain agrees with the general formula $\pi_i = 1/\mathbb{E}_iT_i$ in (4.16), by directly evaluating $\mathbb{E}_iT_i = \sum_{n=1}^{\infty} \mathbb{P}_i(T_i \geq n)$ for the two state chain.

## 4.5 Two State Transition Diagrams

To fully understand the dependence between the variables $X_1$ and $X_2$ in a two-state Markov chain, and how this affects the distribution of $X_1 + X_2$, regard each $X_i$ as the indicator of success on trial $i$ in a pair of dependent trials. Let $S_2 := X_1 + X_2$ be the number of successes in the two trials, and

$$\bar{S}_2 := (1 - X_1) + (1 - X_2) = 2 - S_2$$

the number of failures in the two trials. With $(S_0, \bar{S}_0) := (0, 0)$ and $(S_1, \bar{S}_1) := (X_1, 1 - X_1)$, the pair of dependent indicators $(X_1, X_2)$ is now encoded in the sequence

$$W_0 := (\bar{S}_0, S_0); \qquad W_1 := (\bar{S}_1, S_1); \qquad W_2 := (\bar{S}_2, S_2).$$

So $(W_0, W_1, W_2)$ is a Markov chain with 6 states:

- $(0, 0)$ is the initial state of $W_0 = (\bar{S}_0, S_0)$,
- $(0, 1)$ and $(1, 0)$ are the two possible states of $W_1 = (1-X_1, X_1)$, corresponding to $(X_1 = 1)$ and $(X_1 = 0)$ respectively
- $(0, 2)$ and $(1, 1)$ and $(2, 0)$ are the three possible states of $W_2 = (\bar{S}_2, S_2)$, corresponding to the events

$$\begin{aligned} (W_2 = (0, 2)) &= (S_2 = 2) = (X_1 = 1, X_2 = 1) \\ (W_2 = (1, 1)) &= (S_2 = 1) = (X_1 = 0, X_2 = 1) \cup (X_1 = 1, X_2 = 0) \\ (W_2 = (2, 0)) &= (S_2 = 0) = (X_1 = 0, X_2 = 0). \end{aligned}$$

There are two motivations for this proliferation of states:

- the distribution of $S_2 = X_1 + X_2$, the number of successes in the two dependent trials, is naturally of interest; this is encoded in the distribution of $W_2$.
- each of the $2 \times 2 = 4$ possible values of $(X_1, X_2)$ corresponds to two consecutive transitions of the chain $(W_0, W_1, W_2)$; vectors representing probabilities of these transitions are easily displayed graphically, as in Figure 4.1 for the cumulative counts in sampling without replacement.

For the transition matrix $P$ as in (4.12) derived from $(X_1, X_2)$ a sample of size 2 without replacement from $A$ values 1 and $B$ values 0, the transition diagram of $(\bar{S}_n, S_n)$ for $0 \le n \le 2$ is just the bottom left corner of the larger diagram already displayed in Figure 4.1 for $A = 3$ and $B = 7$, involving just the first two steps away from $(0, 0)$. See Figure 4.2. The special feature of this transition diagram, that lines through the various probability vectors all pass through the point $(B, A)$, is essentially an algebraic property of the transition rules for sampling without replacement. Remarkably, this algebraic property extends to the the setting of the above proposition, as follows:

**Corollary 0.1.** *Let a $2 \times 2$ transition probability matrix $P$ with $P_{01} \neq P_{11}$ be represented in the form (4.18) for a pair of real parameters $(B, A)$. Consider a Cartesian plane of pairs of real numbers $w = (f, s)$, with the six pairs indexed by non-negative integers $f$ and $s$ with $f + s \le 2$ representing possible states of the chain $W_i := (\bar{S}_i, S_i)$ for $i \in \{0, 1, 2\}$, derived as above from a pair of indicator variables $(X_1, X_2)$ with transition matrix (4.18). For each of the states $w = (0, 0)$ or $(0, 1)$ or $(1, 0)$ represent the two transition probabilities of the $W$-chain out of state $w$ by a vector pointing from $w$ to $w + v(w)$, where $v(w)$ is the following probability vector:*

$$v(0,0) = \lambda(\cdot) = (\lambda_0, \lambda_1) \text{ is the distribution of } X_1 \tag{4.22}$$

$$v(0,1) = P(1,\cdot) = (P_{10}, P_{11}) \text{ is the distribution of } X_2 \text{ given } X_1 = 1 \tag{4.23}$$

$$v(1,0) = P(0,\cdot) = (P_{00}, P_{01}) \text{ is the distribution of } X_2 \text{ given } X_1 = 0. \tag{4.24}$$

*Regard these three probability vectors, together with the probability vector $\lambda P$ representing the unconditional distribution of $X_2$, and the vector with components $(B, A)$, as five points in the $(f, s)$-plane. Then:*

*(i) $(B, A)$ is the unique point of intersection of the lines through $w$ and $w + v(w)$ for $w = (0, 1)$ and $w = (1, 0)$.*

*(ii) For each initial distribution $\lambda$ of $X_1$, the line through $\lambda$ in direction $\lambda P$ passes through $(B, S)$.*

*(iii) The point*

$$\lambda + P\lambda = \mathbb{E}(\bar{S}_2, S_2) \tag{4.25}$$

*is the point of intersection of the upsloping line through $\lambda$ and $(B, A)$ and the downsloping line $\{(f, s) : f + s = 2\}$.*

*(iv) For $(B, A) \neq (0, 0)$, the unique stationary distribution $\pi$ for $P$ is the point $(\pi_0, \pi_1) = (B, A)/(A + B)$ where the line from $(0, 0)$ to $(B, S)$ intersects the line $\{(f, s) : f + s = 1\}$.*

*Proof.* Part ((i)) is implied by the cases $\lambda = (0, 1)$ and $\lambda = (1, 0)$ of part ((ii)), and parts ((iii)) and ((iv)) also follow easily from part ((ii)). So it suffices to check part ((ii)). By the assumption that $\lambda$ is a probability vector, $\lambda_0 + \lambda_1 = 1$. So the representation (4.18) of $P$ in terms of $A$ and $B$ makes

$$(\lambda P)_1 = \frac{\lambda_0 A + \lambda_1 (A - 1)}{A + B - 1} = \frac{A - \lambda_1}{A + B - 1} \tag{4.26}$$

$$(\lambda P)_0 = \frac{\lambda_0 (B - 1) + \lambda_1 B}{A + B - 1} = \frac{B - \lambda_0}{A + B - 1} \tag{4.27}$$

In Cartesian coordinates $(f, s)$ with horizontal coordinate $f$ counting failures, that is values $X_i = 0$, and vertical coordinate $s$ counting successes, that is values $X_i = 1$, the slope of the probability vector representing the distribution $\lambda P$ of $X_2$ is therefore

$$\frac{(\lambda P)_1}{(\lambda P)_0} = \frac{A - \lambda_1}{B - \lambda_0}$$

which is the slope of the line through the points $\lambda = (\lambda_0, \lambda_1)$ and $(B, A)$. □

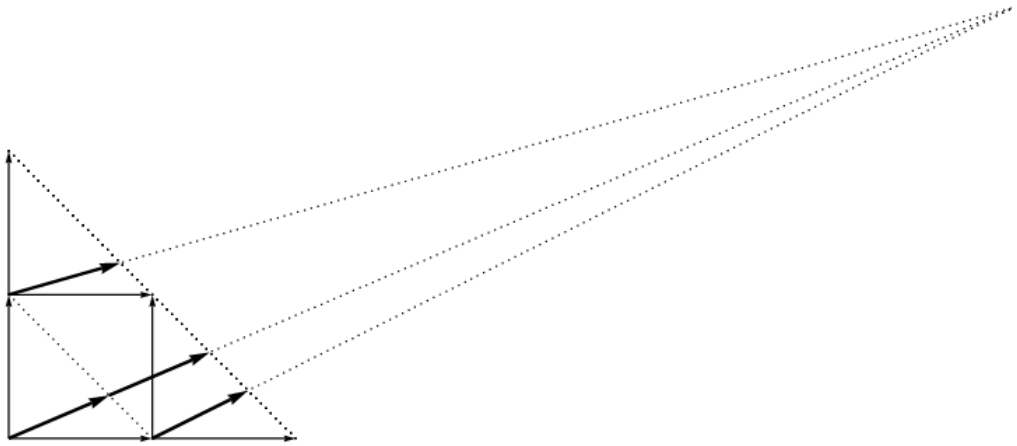
**Remarks.** Parts ((i)) and ((iv)) of the above corollary were pointed out by Bruno de Finetti in his study of exchangeable sequences of random variables. I do not know a reference for parts ((ii)) and ((iii)). It is a curious feature of this geometric construction of the vector $\lambda P(\,\cdot\,)$ from vectors representing $P(0, \cdot)$ and $P(1, \cdot)$, that the basic decomposition $\lambda P(\,\cdot\,) = \lambda_0 P(0, \,\cdot\,) + \lambda_1 P(1, \,\cdot\,)$ is not very apparent from the geometry. This makes it hard to give a comparably simple construction of the two inverse probability vectors $\mathbb{P}(X_1 \in \cdot \,|\, X_2 = j)$ for $j = 0, 1$ which are given by Bayes' rule:

$$\mathbb{P}(X_1 = i \,|\, X_2 = j) = \frac{\lambda_i P(i, j)}{(\lambda P)_j}. \tag{4.28}$$

**Exercise.** Show that if the probability vectors $P(0, \,\cdot\,)$ and $P(1, \cdot)$ are both drawn emanating from $(0, 0)$ (rather than from $(1, 0)$ and $(0, 1)$ as in Figure 4.3) , so the tips of both $P(0, \,\cdot\,)$ and $P(1, \,\cdot\,)$ fall on the downsloping line $\{(f, s) : f + s = 1\}$, then $\lambda P(\,\cdot\,)$ is the vector emanating from $(0, 0)$ whose tip is on the same downsloping line, a fraction $\lambda_1$ of the way along the directed line segment from $P(0, \,\cdot\,)$ to $P(1, \,\cdot\,)$. Embellish this diagram by making $\lambda P(\,\cdot\,)$ the top right corner of a parallelogram with two sides which are initial segments of the vectors $P(0, \,\cdot\,)$ and $P(1, \,\cdot\,)$. Each of the four terms $\lambda_i P(i, j)$ should now be apparent as a length on or other of the two axes.

**Problem.** How best to visualize Bayes' rule geometrically?

Figure 4.2: Transition vector diagram for a sample of size 2 without replacement.



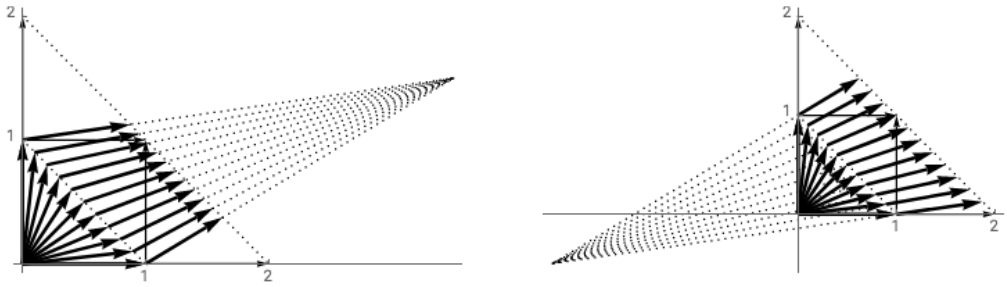
This diagram just amplifies the bottom left corner of the transition diagram of Figure 4.1 for sampling without replacement. The chain $(W_0, W_1, W_2)$ makes 2 steps through 6 states $(f, s)$ for non-negative counts $f$ and $s$ of failures and successes with $f + s \leq 2$, driven by $(X_1, X_2)$ a sample of size 2 without replacement from 3 ones (successes) and 7 zeros (failures). Starting from $W_0 = (0, 0)$ the first transition is to $W_1 = (0, 1)$ (up) or $W_1 = (1, 0)$ (right) according to whether $X_1 = 1$ or 0. The next step to $W_2 = (X_1 + X_2, 1 - X_1 + 1 - X_2)$ is up or right according to whether $X_2 = 1$ or 0. After these two steps, $W_2$ is in one the states $(0, 2), (1, 1)$ or $(2, 0)$.

The transition probability vector

- out of $(0, 0)$ gives the stationary distribution $(7/10, 3/10)$ for $X_1$.
- out of $(0, 1)$ gives the distribution $(P_{10}, P_{11}) = (7/9, 2/9)$ of $X_2$ given $X_1 = 1$.
- out of $(1, 0)$ gives the distribution $(P_{00}, P_{01}) = (6/9, 3/9)$ of $X_2$ given $X_1 = 0$.

The distribution of $X_2$, which is identical to the distribution of $X_1$, is represented by copy of the vector for the stationary distribution of $X_1$, added to the tip of that vector. Observe that all the probability vectors point towards the point $(B, A) = (7, 3)$, representing the total numbers of failures and successes if the process of sampling without replacement is continued to an exhaustive sample of size 10.

Figure 4.3: Transition vector diagrams for two-state Markov chains.




The left hand diagram shows the $(f, s)$-Cartesian plane for an indicator chain with $P_{01} = 3/8$ and $P_{11} = 1/8$ corresponding to $(B, A) = (7, 3)/2$. The geometric structure is very similar to that of Figure 4.2 for $(X_1, X_2)$ a sample of size 2 without replacement from a population of 7 zeros and 3 ones. Now $B$ and $A$ are no longer integers, but the algebraic prescription of transition probabilities (4.18) still defines a $2 \times 2$ transition probability matrix. Here

- the transition vector out of $(1, 0) \longleftrightarrow (X_1 = 0)$ adds $P(0, \cdot) = (5, 3)/8$
- the transition vector out of $(0, 1) \longleftrightarrow (X_1 = 1)$ adds $P(1, \cdot) = (7, 1)/8$.

In accordance with Corollary 0.1, these transition vectors point to $(B, A) = (7, 3)/2$. The diagram shows the 11 initial probability vectors $\lambda = (i, 10-i)/10$ for $0 \leq i \leq 10$, emanating from the origin. Added to the tip of each of these vectors $\lambda$ is the corresponding probability vector $\lambda P$, which always points from $\lambda$ to $(B, A)$. The stationary probability vector is $\pi = (7, 3)/10$, the unique vector such that both $\lambda$ and $\lambda P$ point directly to $(B, A) = (7, 3)/2$. The right hand diagram is the corresponding geometric description of the two state indicator chain with $P_{01} = 1/8$ and $P_{11} = 3/8$ corresponding to $(B, A) = (-5, -1)/2$. This diagram for positively dependent $(X_1, X_2)$ is similar to the left hand diagram for negatively dependent $(X_1, X_2)$, except that each vector $\lambda P$ added to $\lambda$ points away from $(B, A)$ instead of towards $(B, A)$. Now the stationary probability vector is $\pi = (B, A)/(A + B) = (5, 1)/6$, which does not equal any of the displayed initial probability vectors $\lambda = (\lambda_0, \lambda_1)$, with $\lambda_1$ ranging over a multiples of $1/10$ as in the left hand diagram.

# LECTURE 5

# Recurrence Classes, $x$-Blocks, and Limit Theorem

## 5.1 Key Points for Homework

Pitman gives a few key pointers (which are from the textbook) that may help with finishing the homework due tonight.

- Recall the definition of an *irreducible* chain. That is,

$$\forall x, y \in \mathcal{S},\ \exists n\ : P^n(x, y) > 0$$

  This forbids a random walk on a graph with 2 or more components (closed classes). Most of the chains we commonly deal with (and in our homework) are irreducible.

- Fact: (See Theorem 1.7 in Durett). If $P$ is irreducible and if there is a stationary probability vector $\pi$ for $P$ (that is, we can solve $\pi P = \pi$ where $\sum_x \pi(x) = 1, \pi(x) \geq 0$), then all the states are positive recurrent, i.e. the chain is positive recurrent.

## 5.2 Positive and Null Recurrence

**Positive Recurrence**

We say that an irreducible chain (or transition matrix) is *positive recurrent* when, for some or for all $x$

$$\mathbb{E}_x T_x < \infty$$

Note that

$$\mathbb{E}_x T_x = \sum_{n=1}^{\infty} \mathbb{P}_x(T_x \geq n).$$

You should check that if $\mathbb{E}_x T_x < \infty$ for some $x$ and $P$ is irreducible, then

$$\mathbb{E}_x T_x < \infty, \quad \forall x \in \mathcal{S}$$

This is closely related to the formula $\pi(x) = 1/\mathbb{E}_x T_x$

**Null Recurrence**

If a state is recurrent, but not positive recurrent (i.e. $\mathbb{P}_x(T_x < \infty) = 1$, but $\mathbb{E}_x T_x = \infty$), then we say that $x$ is *null recurrent*.

## 5.3 Review: Mean Return Time

Pitman reminds us that there is a formula relating the mean return time and the stationary probability (Theorem 1.21 Durett):

$$\pi(x) = \frac{1}{\mathbb{E}_x T_x}$$

As a simple corollary, this formula directly implies that $\pi$ is unique. There is no doubt about this for a stationary measure in terms of the mean recurrence time. If we discuss a system of countably infinite space, our traditional linear algebra may fail. This result provides an interpretation beyond a system of finitely many equations and unknowns.

Conversely, if $P$ is irreducible and positive recurrent, then there exists this $\pi$. This is almost trivial, but of course we have to check that $\pi$ is a stationary probability.

## 5.4 Example: Symmetric Random Walk

Consider a simple (symmetric) random walk with equal probability of going either direction on $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$. We take the usual notation $S_n$ for the walk. Start at $x = 0$, so that

$$S_n := \Delta_1 + \Delta_2 + \cdots + \Delta_n$$

where $\Delta_k$ is $+1$ or $-1$, each with probability $\frac{1}{2}$. This gives

$$P^n(0,0) = \begin{cases} 0 & \text{, if } n \text{ is odd} \\ \binom{2m}{m} \left(\frac{1}{2}\right)^{2m} & \text{, if } n = 2m \text{ is even} \end{cases}$$

Now Pitman notes we can tell recurrence or transience by looking at the fact that the total number of visits to 0 follows a geometric distribution with parameter $(1 - \rho_0)$

$$\mathbb{E}_0(\text{total \# visits to 0}) = \sum_{n=1}^{\infty} P^n(0,0)$$

But we know that $\binom{2m}{m}(\frac{1}{2})^{2m}$ is the same as the probability of $m$ heads and $m$ tails in $2m$ tosses. Increasing tosses gives a very "flat" normal curve because the mean of $\mathbb{E}_0 S_{2m} = 0$ and the variance tends to infinity, because the variance of each summed term is 1, the mean square is

$$\mathbb{E}_0 S_{2m}^2 = \underbrace{1 + 1 + \cdots + 1}_{2m} = 2m$$

We call this *diffusion*, in that on average the center of our distribution goes no where, but the distribution spreads out and flattens. Using Stirling's formula[1]

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

and applying this to our earlier expression to show that

$$P^{2m}(0,0) \sim \frac{C}{\sqrt{m}}$$

where $C$ is some constant and $\sim$ means the ratio tends to 1 as $n \to \infty$.

### 5.4.1 Recurrence versus Transience

To see recurrence versus transience, we look at (from earlier)

$$\sum_{n=1}^{\infty} P^n(0,0) = \sum_{m=1}^{\infty} P^{2m}(0,0) \sim \sum_{n=1}^{\infty} \frac{C}{\sqrt{m}} = \infty$$

(A rather paradoxical fact) This implies that the expected return time to 0 is infinite

$$\mathbb{E}_0 T_0 = \infty$$

although we are sure to eventually return with probability 1. Recall the definition of recurrent gives

$$\mathbb{P}_x(T_x < \infty) = 1 \iff \mathbb{P}_x(T_x \geq n) \to 0, \text{ as } n \to \infty.$$

Also, we should know that positive recurrence implies recurrence, but the converse is not necessarily true. Pitman summarizes that on our homework, we can quote the result: If we have a stationary probability measure, then the chain is *positive recurrent*.

## 5.5 Notion of $x$-Blocks of a Markov chain

Start at $x$ (for simplicity) or wait until we hit $x$. Then look at the successive return times $T_x^{(i)}$ which is the $i^{\text{th}}$ copy of $T_x$. Now recall this has the Strong Markov Property, which gives us two things:

[1] or Normal Approximation

(i) Every $T_x^{(i)}$ has the same distribution as $T_x$.

(ii) Further, they are independent copies. That is, $T_x^{(1)}, T_x^{(2)}, \ldots$ are independent.

Now Pitman mentions a variation on this theme of $x$-blocks, which explains many things.

### 5.5.1 Example: $x$-Blocks

Let $N_{xy}^{(i)} :=$ the # of visits to $y$ in the $i^{\text{th}}$ $x$-block of length $T_x$. In our previous in-class example, this gives a sequence

$$2, 0, 6, 0, 4, 2, \ldots$$

Now for some book keeping, consider what happens if we sum over all states $y$. Of course, this just gives the length of $T_x^{(i)}$ by "Accounting 101."

$$\sum_{y \in \mathcal{S}} N_{xy}^{(i)} = T_x^{(i)}$$

Note we must agree that $N_{xx}^{(i)} = 1$ for this to work. Now this implies that there is a formula involving expectations. Taking expectation starting at $x$

$$\sum_{y \in \mathcal{S}} \mathbb{E}_x N_{xy}^{(i)} = \mathbb{E}_x T_x^{(i)}$$

where this is really the same equation for all $i$ by the Strong Markov Property. Fix $x, y$ and look at $N_{xy}^{(1)}, N_{xy}^{(2)}, \ldots$, each of which

(i) $N_{xy}^{(i)}$ has the same distribution as $N_{xy} := N_{xy}^{(1)}$.

(ii) Further, the $N_{xy}^{(i)}$ are independent and identically distributed.

Pitman reminds us that as we return to $x$, via the Strong Markov Property, nothing of the past changes our expectations or distributions going forward.

## 5.6 Positive Recurrent Chains ($P$ irreducible)

Notice that if $\mathbb{E}_x T_x < \infty$, and we define $N_{xy}$ as we have earlier, then we can let

$$\begin{aligned} \mu(x, y) &:= \mathbb{E}_x(N_{xy}) \\ \mu(x) &:= \mathbb{E}_x T_x = \text{ mean length of } x\text{-block} \end{aligned}$$

Correspondingly to our Accounting 101, we write

$$\sum_{y \in \mathcal{S}} \mu(x, y) = \mu(x) < \infty$$

Further, we can show (see text for details) that if we sum

$$\sum_y \mu(x,y)P(y,z) = \mu(x,z)$$

or in other words, $\mu(x,\cdot)$ is a stationary measure, **not** a stationary probability, as it is an unnormalized measure). That is

$$\mu(x,\cdot)P = \mu(x,\cdot)$$

This is important because it gives us a simple explicit construction of a stationary measure $\mu(x,\cdot)$ for every state $x$ in state space $\mathcal{S}$ of a positive recurrent (PR) irreducible chain with matrix $P$. Notice that this is not just any measure. By convention, we say that the number of times we visit $x$ in the duration of $T_x$ is 1 (this is necessary to satisfy our constructions today). That is, we must not count a visit twice, and we must set

$$\mu(x,x) := 1$$

in order to get

$$\sum_{y\in\mathcal{S}} \mu(x,y) = \mu(x) < \infty$$

Now to get a stationary probability measure, we take

$$\pi(y) = \frac{\mu(x,y)}{\sum_z \mu(x,z)} = \frac{\mu(x,y)}{\mu(x)}$$

and this does **not** depend on $x$. We can take any reference state and we get the same thing when we look at these ratios.

### 5.6.1 Explanation of the Key Formula

We may ask why we have

$$\sum_y \mu(x,y)P(y,z) = \mu(x,z)$$

Recall that $\mu(x,y)$ is the expected number of hits on $y$ before $T_x$. That is,

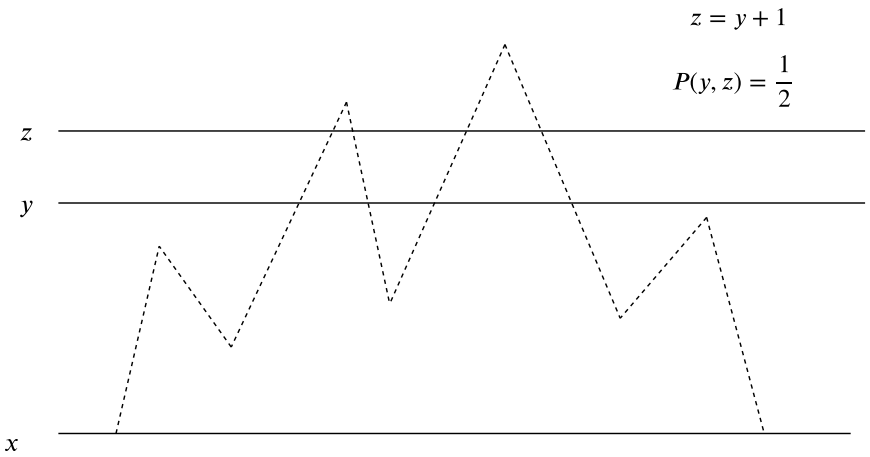
$$\mu(x,y) = \mathbb{E}_x(\#\text{ of hits on } y \text{ before } T_x)$$

Now, every time we hit $y$, then $P(y,z)$ is the probability that the next step is to state $z$. Therefore, at least intuitively, $\mu(x,y)P(y,z)$ has a particular meaning. That is

$$\mu(x,y)P(y,z) = \mathbb{E}_x(\ \#\text{ of transitions } y\to z \text{ before } (\leq) T_x\ )$$

The distribution of a single $x$-block gives the following formulas for the invariant probability measure $\pi$

$$\pi(x) = \frac{1}{\mathbb{E}_x T_x}, \quad \frac{\pi(y)}{\pi(x)} = \mu(x,y)$$




## 5.7 Limit Theorem

If we let $N_n(y) := \sum_{k=1}^n \mathbf{1}(X_k = y) = \#$ of hits on $y$ in first $n$ steps, then

$$\begin{aligned}\mathbb{E}_x \frac{N_n(y)}{n} &= \text{ mean } \# \text{ hits on } y \text{ per unit time up to } n \\ &= \frac{1}{n}\sum_{k=1}^{n} P^k(x,y) \to \pi(y)\end{aligned}$$

We have this Cesáro mean convergence always for irreducible positive recurrent chains, these themselves do not necessarily converge, but their average converges. Now if we additionally impose *aperiodicity*, we have

$$P^n(x,y) \to \pi(y)$$

always for irreducible and positive recurrent and aperiodic transition matrix $P$. See §1.8 in the text for the proof.

## 5.8 Review and Audience Questions

A null recurrent chain has a stationary measure with reference state $x$ assigned as measure 1. If we do this on a simple symmetric random walk, we find that the expected time spent in every state on an $x$-block is 1, which explains why we expect to spend so much time to return back to $x$. If we have a stationary probability measure

$$\mathbb{E}_\pi \frac{1}{n}\sum_{k=1}^{n} P^k(x,y) = \pi(y)$$

we can argue that the stationary measure must be approached in the limit (at least in the Césaro sense) and is therefore unique as the limit must be unique.

# LECTURE 6

## First Step Analysis and Harmonic Equations

Pitman opens to questions regarding irreducible, aperiodic, recurrent (both positive and null), or transient. For a nice transition probability matrix, there exists a stationary probability $\pi$ so that

$$\lim_{n\to\infty} P^n = \begin{pmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{pmatrix} \tag{6.1}$$

Pitman asks us to recall the single most important formula regarding recurrent chain and its expected time for returning to $x$ starting from $x$, $\mathbb{E}_x T_x$. For a nice (irreducible, positive recurrent),

$$\mathbb{E}_x T_x = \frac{1}{\pi(x)} \tag{6.2}$$

where $\pi(x)$ is the long run average (fraction of) time spent in state $x$. Recall that we defined that we hit $x$ exactly once per $x$-cycle on average, which is equal to once per $\mathbb{E}$ cycle. This makes sense intuitively, where expecting to take a long time before returning to state $x$ corresponds to not being in state $x$ as often.

### 6.1 Hitting Places

Recall that we used the notation

$$T_A := \min\{n \geq 1 : X_n \in A\}$$
$$T_x := \min\{n \geq 1 : X_n = x\}$$

This is not trivial for $X$ with $X_0 = x$. For analysis of hitting places (and time), it's often easier to have our discrete-time sequence start at 0. Hence we define

$$V_A := \min\{n \geq 0 : X_n \in A\} \tag{6.3}$$

Pitman notes that this is not a universal notation and we might see $T, V, \tau$ used for this definition, but for this text and course, we will use $V_A$ for this purpose.

**Thoerem 1.28 (Durett p. 55)**

Consider a Markov chain with state space $S$. Take two non empty, (necessarily disjoint) $A, B \subseteq S$. Let $C := S - (A \cup B)$ and assume $C$ is finite

**Assumptions** Suppose we have $h : S \to \mathbb{R}$ such that

$$h(a) = 1, \ \forall\, a \in A \tag{6.4}$$

$$h(b) = 0, \ \forall\, b \in B \tag{6.5}$$

$$h(x) = \sum_y P(x, y)h(y), \ \forall\, x \in C \tag{6.6}$$

Suppose also that

$$\mathbb{P}_x(V_{A\cup B} < \infty) > 0, \ \forall\, x \in C$$

Then

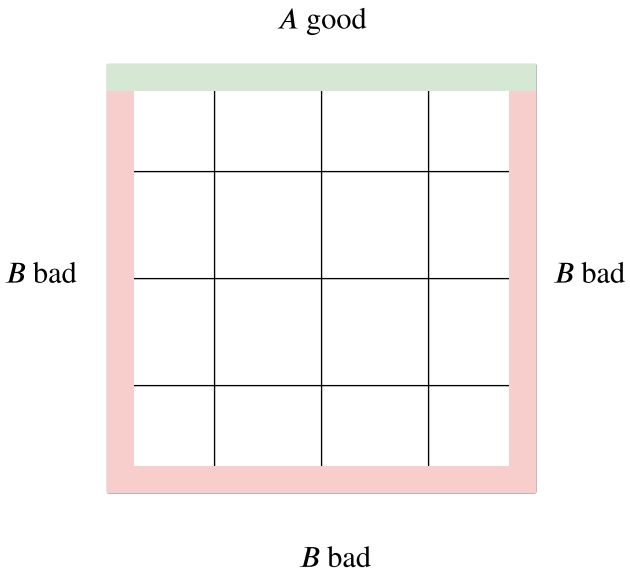
$$h(x) = \mathbb{P}_x(V_A < V_B) \tag{6.7}$$

The point of the theorem is that for a typical Markov chain $X$, and disjoint sets of states $A$ and $B$, the chance that $X$ hits $A$ before $B$ can be found, as a function of the initial state $x$, by solving the system of linear equations (6.6) subject to the obvious boundary conditions (6.4) and (6.5). Very commonly, we'll write the equation (6.6) in matrix notation, where $h$ is a *column vector* as $h(x) = (Ph)(x)$ for $x \in C$. It is very convenient to assume (with no loss of generality) that both $A$ and $B$ are absorbing sets of states. Then (6.6) reduces to simply $h = Ph$, an equality of row vectors indexed by $x \in S$. This is because $h(x) = Ph(x)$ holds trivially for any absorbing state $x$ (meaning $P(x, x) = 1$). Note that because $C$ is assumed finite, a variant of Durrett's Lemma 1.3 shows that (6.1) is equivalent to

$$\mathbb{P}_x(V_{A\cup B} < \infty) = 1, \ \forall\, x \in C \tag{6.8}$$

For a chain with infinite state space $S$ and $C$ infinite, this condition is adequate as a replacement for (6.1), provided it is assumed that $h$ is a bounded or non-negative function, so there are no problems with the definition of $Ph$.

Intuitively, regard $A$ and $B$ as sets of boundary states. Graphically, it is convenient to place $A$, the set of target states, at the top of a 2 (or higher dimensional) lattice, and place $B$ as all three remaining boundaries of the lattice (left, right, bottom edges).




## 6.2 Method of Solution (Durret p.54)

Pitman notes that the method is more important than the solution here. From the text, "Let $h(x)$ be the probability of hitting $A$ before $B$, starting from $X$." We call this technique **first step analysis**. "By considering what happens at the first step." That is, we assert that we start at $X_0 := x$, and we condition on time $X_1$, the value of the chain at time 1. Generalizing, let $Y$ be any nonnegative (for simplicity) random variable, which is a function of $X_0, X_1, X_2, \ldots$ some Markov chain with transition matrix $P$. Consider $\mathbb{E}_x Y$ as a function of $x$, and notice we can write, by summing out all states $z \in S$, using $\sum_{z \in S} \mathbb{1}(X_1 = z) = 1$

$$\begin{aligned}
\mathbb{E}_x Y &= \mathbb{E}_x \sum_{z \in S} \mathbb{1}(X_1 = z) Y \\
&= \sum_z \mathbb{E}_x \big[ \mathbb{1}(X_1 = z) Y \big] \\
&= \sum_z \mathbb{P}_x(X_1 = z) \mathbb{E}_x(Y \mid X_1 = z) \\
&= \sum_z P(x, z) \mathbb{E}_x(Y \mid X_1 = z)
\end{aligned}$$

which is simply computing the $\mathbb{P}_x$ expectation of $Y$ by conditioning on $X_1$ Commonly, $Y$ can be written as a function of $X_1, X_2, \ldots$ and this can be further simplified. We can do this for instance when $Y$ is the indicato $Y = \mathbb{1}(V_A < V_B)$ in the setting of the above theorem. Then

$$\boxed{\mathbb{E}_x Y = \mathbb{P}_x(V_A < V_B)} \tag{6.9}$$

Something else is true as well via first step analysis. Take $x \notin A \cup B$. Look at the probability that $V_A$ happens before $V_B$, provided that we know $X_1 = z$. Now if $z$ is one of the boundary cases, this is trivial. So we treat in cases, using the Markov property

$$\mathbb{P}_x(V_A < V_B \,|\, X_1 = z) = \begin{cases} 1 & , z \in A \\ 0 & , z \in B \\ \mathbb{P}_z(V_A < V_B) & , \text{else} \end{cases}$$

as you should convince yourself.

> Does this probability of hitting $A$ before $B$ have anything to do with $P(c, \cdot)$ for $c \in A \cup B$?

We agree on the edge cases, for starting in $A$ or $B$. Now we make this key observation, which is not mentioned in the text. Because of our definitions, namely the possibility of being there at time zero, the answer is NO!

With this in mind, we modify the problem at hand to make the entire set of states $A \cup B$ absorbing. That is, $P(c, c) := 1, \forall\, c \in A \cup B$. That is to say when we arrive, we stick there, and we solve the problem under these circumstances.
Notice that we agreed by conditioning on $X_1$ that

$$h(x) := \mathbb{P}_x(V_A < V_B), \text{ for} x \notin A \cup B$$

solves the *Harmonic equation*

$$\boxed{h(x) = \sum_y P(x, y) h(y)} \tag{6.10}$$

Notice that if we make $A \cup B$ absorbing, then this harmonic equation above is true for ALL $x \in A \cup B$. Now we arrive at a reformulation of the theorem.

**Pitman's Version of Durett's Theorem**

Assume that

- $P$ has $A \cup B$ as absorbing states and
- $\mathbb{P}_x(\text{hit } A \cup B \text{ eventually}) = 1, \forall\, x \in S$

then

$$h(x) := \mathbb{P}_x(\text{hit } A \text{ before } B)$$

is the *unique* bounded or non-negative solution of $h = Ph$, subject to the *boundary condition* that $h = \mathbb{1}_A$ (the indicator of $A$) on $A \cup B$.

This is fundamentally the same as Durett's theorem, but with some tinkering, we have a more elegant statement as here. Notice that $h = Ph$ is a very special equation,

whose as solutions solve various problems. In order to understand this equation, it is important to understand what is $Pf$ for a function (column vector) $f$ (assume either nonnegative or bounded so that we can make sense of the summations). Then the action of the transition matrix $P$ on a column vector $f$ gives us:

$$\big(Pf\big)(x) = \sum_{y \in S} P(x,y) f(y),$$

summing over all $y$ in the state space. $P(x,y)$ gives the probability distribution over values $y$, depending on the initial state $x$ and $f(y)$ simply gives the return from state $y$. Hence directly by our notation, we have:

$$\big(Pf\big)(x) = \mathbb{E}_x f(X_1).$$

Hence

$$\big(Ph\big)(x) = \mathbb{E}_x h(X_1)$$

as the meaning of $(Ph)(x)$. Another way to say this is by looking at the conditional expectation (knowing $X_0$)

$$\mathbb{E}\big[h(X_1) \,|\, X_0\big] = \big(Ph\big)(X_0)$$

Pitman makes the following claim: If $h = Ph$ (that is, $h$ solves the harmonic equation), then the expectation (starting at $x$) of $h$ of any variable $(X_n)$ is

$$\mathbb{E}_x\big[h(X_n)\big] = h(x)$$

which is true by $n = 1$ by $(Ph)(x) = \mathbb{E}_x h(X_1)$ from above (that is, $h = Ph$). Now, this is true for $n = 1, 2, 3, \ldots$ by induction and the Markov property. If we trust this for now (we may revisit this later), we may want to assume that

$$h = \begin{cases} 1, & \text{on } A \\ 0, & \text{on } B, \end{cases}$$

then we can write

$$h(x) = \mathbb{E}_x h(X_n) = \sum_{y \in S} P^n(x,y) h(y),$$

as our familiar notation for a Markov chain. Then we can equivalently write this as a summation over the three state cases

$$h(x) = \sum_{y \in A} P^n(x,y) h(y) + \sum_{y \in B} P^n(x,y) h(y) + \sum_{y \in S - A - B} P^n(x,y) h(y)$$

Recall that we've set $A \cup B$ to be absorbing, so the first two terms are simply

$$\sum_{y \in A} P^n(x,y) h(y) = \mathbb{P}(V_A \leq n)$$

$$\sum_{y \in B} P^n(x,y) h(y) = 0$$

Hence

$$h(x) = \mathbb{P}_x(V_A \leq n) + 0 + \sum_{y \in S-A-B} P^n(x,y)h(y)$$

Now if we take $n \to \infty$, then

$$\lim_{n\to\infty} h(x) = \lim_{n\to\infty} \mathbb{E}_x h(X_n) = P_x(V_A < \infty) + \underbrace{\lim_{n\to\infty} \sum_{y \in S-A-B} P^n(x,y)h(y)}_{=0}$$

$$= \mathbb{P}_x(V_A < \infty)$$

because $\mathbb{P}_x(\text{hit } A \cup B \text{ eventually}) = 1$ via our assumption, and the sum which tends to 0 is bounded above by the maximum absolute value of $h(y)$ over $y \in S - A - B$ times $\mathbb{P}_x(V_{A\cup B} > n)$ which tends to 0 by the assumption that $\mathbb{P}_x(V_{A\cup B} < \infty) = 1$.

## 6.3 Canonical Example: Gambler's Ruin for a Fair Coin

The state space is $S := \{0, 1, 2, \ldots, N\}$, and the goal state is $A = \{N\}$, and the bad state is $B = \{0\}$. The transition matrix is then

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

Now let $(X_n)$ be the simple random walk with absorbing states $\{0, N\}$. Then

$$\mathbb{P}_x(\text{hit } N \text{ before } 0) = h(x)$$

is desired. The equation $h = Ph$ becomes

$$h(x) = \frac{1}{2}h(x+1) + \frac{1}{2}h(x-1) \text{ for } 0 < x < N$$

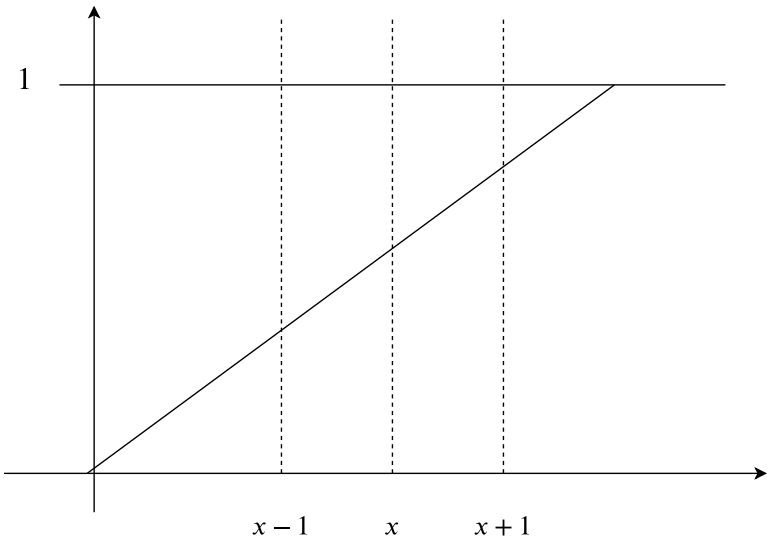
and we set the boundary conditions

$$h(N) := 1 \qquad h(0) := 0$$

Now the harmonic equation $h(x) = \frac{1}{2}h(x+1) + \frac{1}{2}h(x-1)$ says that the graph of $h(x)$ is a straight line, over integer values $x$, passing through 0 and 1. Hence $h(x) = \frac{x}{N}$ is the unique solution to this system of equations. Here it is easy that we are certain to eventually hit the boundary states. Hence for the simple symmetric random walk started at $0 \leq x \leq N$

$$\mathbb{P}_x(\text{hit } N \text{ before } 0) = \frac{x}{N}$$

which is a famous result as due to Abraham de Moivre around 1730.




### 6.3.1 Gambler's Ruin with a Biased Coin

What are the harmonic equations? We reason that this results in the same equations, with slight modifications

$$h(x) = ph(x+1) + qh(x-1) \quad (0 < x < N)$$

which we may solve via algebra as done in Durett (p. 58). Pitman shows us a more clever way, related to the idea of a *martingale*. There is a discussion of this problem in the context of martingales at the end of the text, as aspects of a hitting-time problem (we will revisit this at the end of the course). Observe that $h(x) = x$ is no longer harmonic when $q \neq p$ (biased coin). Now by some good guesswork you can discover that

$$h(x) = \left(\frac{q}{p}\right)^x$$

is a harmonic function for the $p - q$ walk. We check this

$$\begin{aligned} Ph(x) &= p\left(\frac{q}{p}\right)^{x+1} + q\left(\frac{q}{p}\right)^{x-1} \\ &= \left(\frac{q}{p}\right)^x = h(x) \end{aligned}$$

This is a bit clever, but it is not a bad idea to try a solution of the form $h(x) = r^x$ of the harmonic equations, and if you do that you will get a quadratic which forces $r = 1$ (boring) or $r = q/p$ (very useful) as above. As soon as we have found this

$h(x)$, we can argue as before: from $h = Ph$ get $h = P^n h$ and so for each $n \geq 0$

$$h(x) = \mathbb{E}_x \left(\frac{q}{p}\right)^{X_n}$$

$$= \left(\frac{q}{p}\right)^N \mathbb{P}_x(\text{hit } N \text{ before } n) + \left(\frac{q}{p}\right)^0 \mathbb{P}_x(\text{hit } 0 \text{ before } n) + \sum_{y \notin \{0,N\}} \cdots$$

Now taking $n \to \infty$, this final term goes to zero. Hence in the limit
and additionally

$$\mathbb{P}_x(\text{hit } N) + \mathbb{P}_x(\text{hit } 0) = 1$$

Now we have two equations and two unknowns. Solve these, and you get the solution found by Durrett on p.58.

# LECTURE 7

# First Step Analysis Continued

## 7.1 First Step Analysis: Continued

The simple idea here is to derive equations by conditioning on step 1. We can find all sorts of things about Markov chains by doing exactly this. Pitman notes that the text keeps doing this technique without explicitly pointing it out. Recall that first step analysis for a Markov chain $(X_0, X_1, X_2, \dots)$, we consider some random variable

$$Y = Y(X_0, X_1, X_2, \dots)$$

If we know $\mathbb{E}_x Y$ for all states $x$ and we want to compute the expectation of $Y$ for a chain with $X_0$ assigned a probability distribution $\lambda = \lambda(x) \;\; x \in S$, denoted $\mathbb{E}_\lambda Y$, we would take

$$\mathbb{E}_\lambda Y = \sum_{x \in S} \lambda(x) \mathbb{E}_x Y$$

Put simply, the expectation of a random variable $Y$ is the expectation of the expectation of $Y$ conditioned on $X_0$. That is,

$$\mathbb{E}(Y) = \mathbb{E}\big[\mathbb{E}(Y \mid X_0)\big].$$

We may want to condition on $X_1$ as well, which is how we derived the harmonic equations from the previous lecture. Let's look at an example where we can do this again.

### 7.1.1 Example: Mean Hitting Times

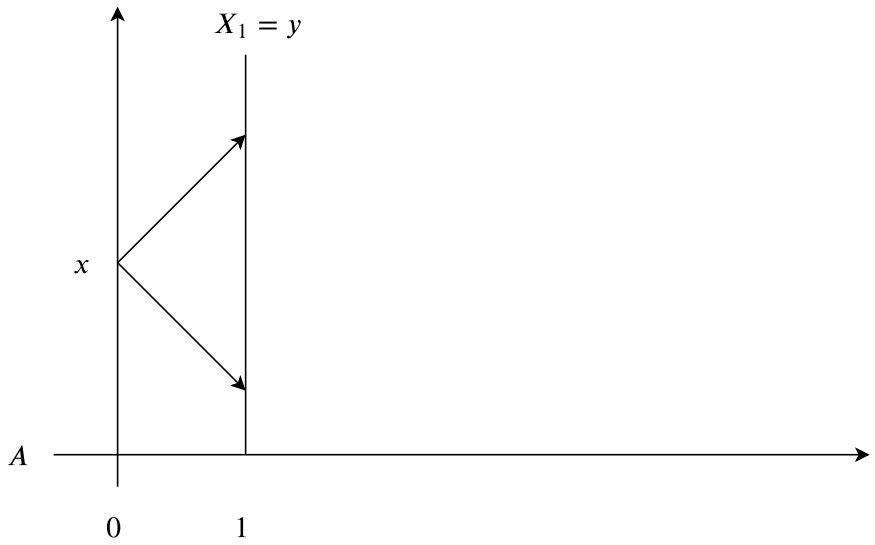
Suppose we have a set of states $A$ (we can make them absorbing as a matter of technique, it makes no difference to the answer), and consider

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

We want to find $\mathbb{E}_x V_A$ for any initial state $x$. If $x \in A$, then we trivially have $\mathbb{P}_x(V_A = 0) = 1$ and hence $\mathbb{E}_x V_A = 0$. Now, for any state $x$ define a function for the mean

$$m(x) = m_A(x) := \mathbb{E}_x V_A$$

where we drop the subscript $A$ as it is understood from context. We want equations for $m(x)$.




From $x$, we hit $X_1 = y$ with probability $P(x, y)$. Now given $X_0 = x$, $X_1 = y$, for $x \notin A$ we have

$$\mathbb{E}(V_A \,|\, X_0 = x, X_1 = y) = 1 + \mathbb{E}_y(V_A)$$

Notice that this is correct if $y \in A$. If we happen to hit $A$ =at time 1, then $V_A = 1$ and the second term $\mathbb{E}_y(V_A)$ is zero. Additionally, this is correct if $y \notin A$, that is, because $x \notin A$ we are certain to take at least 1 step, with $\mathbb{E}_x V_A \geq 1$. This means that we can write down a system of equations, relating to the mean times

$$m(x) = 1 + \sum_{y \in S} P(x, y) m(y) \qquad (x \notin A)$$

This system should be solved together with the *boundary condition*

$$m(x) = 0 \qquad (x \in A)$$

If we have only a finite number of non-absorbing states, then we have a finite number of linear equations and this number of unknowns.

In the text, Theorem 1.29 on page 62 states that as long as we can reach the boundary from the any state in the interior (in some number of steps) with positive probability, provided there are only a finite number of interior states, this system of equations will have a unique solution. In practice, in examples, you just write down the system of linear equations and solve them by standard methods or software.

### 7.1.2 Application: Duration of a Fair Game

The usual Gambler's Ruin for a fair coin. Text Example 1.52 on page 66, We start with $\$x$ and play for $\pm\$1$ gains with equal probability until we hit either \$0 or some

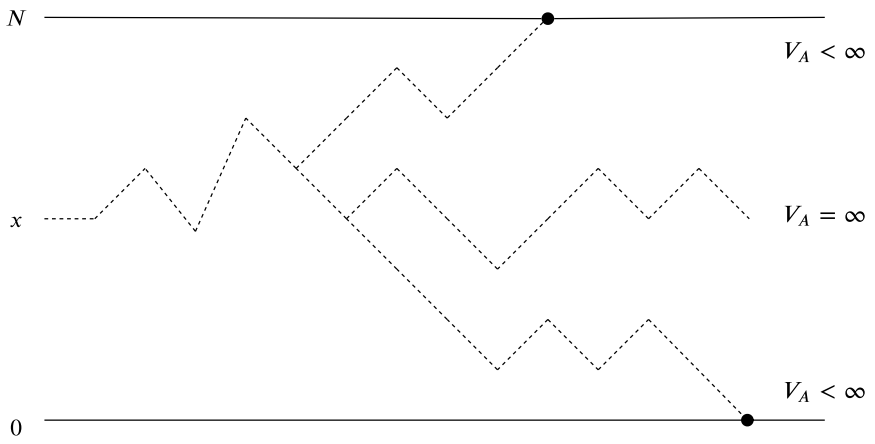
$\$N$. Last lecture, we showed

$$\mathbb{P}_x(\text{reach } N \text{ before } 0) = \frac{x}{N}$$

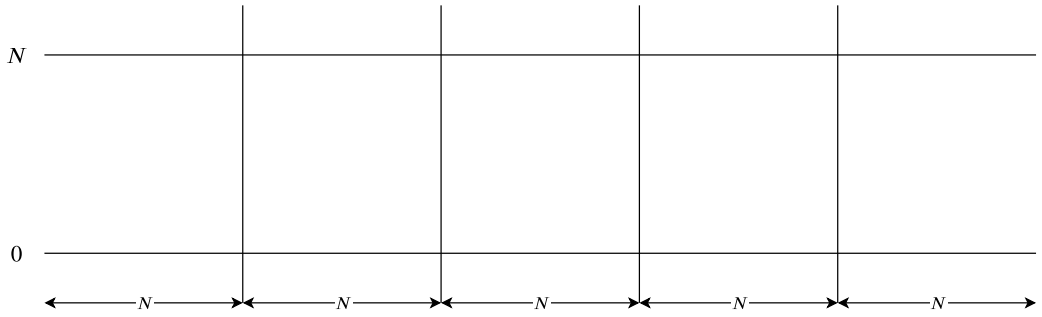
Set $A := \{0, N\}$ as our absorbing states and $V_A$ the duration of the game, where

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

Recall that there remains the scenario of never hitting the boundary $A$, but we have already found before that the probability assigned to this uncountable infinite number of never-ending paths is zero.




To see this, notice that for any 'block' of $N$ steps, there is a strictly positive probability that we hit a boundary state. That is $\mathbb{E}_x V_A < \infty$ with probability 1 and for all $x \in S$.




We use this argument to form the geometric bound as we have before in a previous lecture. To find $m(x) := \mathbb{E}_x V_A$ we first write out the boundary conditions. That is,

$$m(0) = m(N) = 0$$

Now the nontrivial cases, we again break into two parts

$$m(x) = 1 + \frac{1}{2}m(x+1) + \frac{1}{2}m(x-1) \quad 0 < x < N$$

Then we solve for this system of equations. Recall that previously we considered the simpler system $h(x) = \frac{1}{2}h(x+1) + \frac{1}{2}h(x-1)$, which implies that $h(x)$ is linear, more pedantically *affine*. For constants $a, b$, we have

$$h(x) = ax + b$$

Now consider the addition of a quadratic term:

$$m(x) = cx^2 + ax + b$$

Then we observe

$$\frac{1}{2}c(x+1)^2 + \frac{1}{2}c(x-1)^2 = cx^2 + \underbrace{\frac{1}{2}c(2x) + \frac{1}{2}c(-2x)}_{=0} + c$$

Now from this sort of consideration, $m(x)$ as above solves the equation

$$\frac{1}{2}m(x+1) + \frac{1}{2}m(x-1) = c + m(x)$$

Hence, we conclude that our system of equations is solved by a quadratic function of the form $m(x) = cx^2 + ax + b$ , where $c = -1$.

### 7.1.3 Summarizing our Findings

Observe

$$g_1(x) = ax + b \implies \frac{1}{2}g_1(x+1) + \frac{1}{2}g_1(x-1) = g_1(x)$$

$$g_2(x) = cx^2 \implies \frac{1}{2}g_2(x+1) + \frac{1}{2}g_2(x-1) = g_2(x) + c$$

These together imply

$$g(x) = cx^2 + ax + b = (g_1 + g_2)(x) \implies \frac{1}{2}g(x+1) + \frac{1}{2}g(x-1) = g(x) + c$$

Hence we have that

$$m(x) := cx^2 + bx + a$$

solves our equations from earlier if and only if $c = -1$. Then plugging this in, we have

$$m(x) = -x^2 + bx + a$$

and additionally recall that $m(0) = m(N) = 0$. There's only one quadratic that satisfies these, namely

$$m(x) = -x(x - N) = \boxed{x(N - x)}$$

In summary, with the idea to try a quadratic (which Pitman notes is not too different from noticing before that a harmonic function for the fair gambler's ruin chain must be linear), finding the exact solution is not hard. See text for solution of the mean duration of an unfair game, and many further examples.

## 7.2 Conditioning on other variables

Commonly in the analysis of Markov chains it is effective to condition on $X_0$ or on $X_1$. Also common to condition on $X_n$ (example in homework). Now we would like to consider that these may not be the only variables on which we would like to condition. There may be more clever techniques, where we employ our imagination to find a more apt conditioning variable, often a suitable random time. Also, exploiting the addition rule for expectation, after breaking a random variable into a sum of two variables, should be kept in mind.

### 7.2.1 Runs in Independent Bernoulli($p$) Trials

We want to find the mean time until we see $N$ successes in a row. Let $\tau_N$ be the random number of trials required. e.g. for $N = 3$ if the outcome of the trials is

$$(X_1, X_2, \ldots) = (0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, \ldots)$$

then $\tau_N = 11$. Note that in treating Bernoulli trials, and more generally i.i.d. sequences, it is customary to start indexing by 1, whereas for general discussion of Markov chains it is customary to start indexing by 0. Python programmers should like the Markovian convention.

The exact distribution of $\tau_N$ is tricky. You can find its generating function if you like, but we only want the expectation here, which is relatively simple. Of course, because $\mathbb{P}(\tau_N \geq N) = 1$

$$\begin{aligned}\mathbb{E}\tau_N &= \sum_{k=N}^{\infty} k\ \mathbb{P}(\tau_N = k) \\ &= \sum_{k=N}^{\infty} \mathbb{P}(\tau_N \geq k)\end{aligned}$$

but neither the point probabilities in the first equality nor the tail probabilities needed for the second tail sum formula sum have a simple formula. Hence we ask, "What should we condition on?" As a student suggests, try $\tau_{N-1}$. That is, having a string of $N-1$ ones in a row.

$$\tau_N = \tau_{N-1} + \Delta_N, \quad \text{where } \Delta_N = \begin{cases} 1 \text{ with probability } p \\ 1 + \text{a copy of } \tau_N \text{ otherwise} \end{cases}$$

If you are eager to see the run of $N$ ones, you are disappointed if the trial following trial $\tau_{N-1}$ is a 0, as this means you must start over again. However, this *regeneration* of the problem is exactly what is needed to help find the sequence of means $\mu_N$

Define $\mu_N := \mathbb{E}\tau_N$, then the above observation gives

$$\mu_N = \mu_{N-1} + 1 + q\mu_N$$

where rearranging gives

$$\mu_N = \frac{\mu_{N-1} + 1}{p}$$

We test this

$$\mu_1 = \frac{1}{p}$$

by the mean of geometric. Similarly,

$$\mu_2 = \frac{\left(\frac{1}{p} + 1\right)}{p} = \frac{1+p}{p^2}$$

and repeating this gives

$$\mu_N = \frac{1 + p + p^2 + \cdots + p^{N-1}}{p^N}$$

In summary, we solved this problem by noticing that to get to $N$ in a row, we needed to first get to $N-1$ in a row, and then condition on the next trial. Here is another approach:

### 7.2.2 Conditioning on the First Zero

Define $G_0$ as the first $n \geq 1$ such that $X_n = 0$ (that is, wait for the first 0). In other words, $G_0$ is one plus the length of the first run of 1s. Then $G_0 \sim \textbf{Geometric}(q)$, where $q$ is the failure probability. It seems reasonable to try to find $\mathbb{E}\tau_N$ by conditioning on $G_0$, as $G_0$ is closely related to $\tau_N$, and we know the distribution of $G_0$. If $G_0 > N$, then $\tau_N = N$. On the other hand, if $G_0 = g \leq N$, then the problem starts over: there is the equality in distribution

$$(\hat{\tau}_N - g \,|\, G_0 = g) \stackrel{d}{=} \tau_N \qquad (0 < g < N)$$

meaning that conditional given $G_0 = g$) the remaining time $\tau_N - g$ has the same distribution as $\tau_N$. This is by a rather obvious form of the Strong Markov Property for Bernoulli trials.

Therefore, by conditioning on $G_0$, we have

$$\mathbb{E}\tau_N = \left[\sum_{g=1}^{N} \mathbb{P}(G_0 = g)(g + \mathbb{E}\tau_N)\right] + \mathbb{P}(G_0 > N)N$$

Now let $\mu_N := \mathbb{E}\tau_N$, so that the earlier equation gives

$$\mu_N = \sum_{g=1}^{N} p^{g-1} q(g + \mu_N) + p^N N$$

We look at a simple $N = 2$ case. Here in this solution, we have:

$$\mu_2 = p^0 q(1 + \mu_2) + pq(2 + \mu_2) + p^2 \cdot 2$$
$$\mu_2(1 - q - pq) = q + 2pq + 2p^2$$

hence easily $\mu_2 = (1 - p)/p^2$ as before. You can easily check this method gives the same conclusion as before for general $N$.

# LECTURE 8

# Infinite State Spaces and Probability Generating Functions

## 8.1 Infinite State Spaces

§1.11 is starred in the text, but is not optional for our course. We'll discuss techniques for both finite and infinite state spaces, in particular

- probability generating functions
- potential kernel (AKA) Green matrix

Pitman gives a list of additional resources with nice problems worth trying. See Bibliography for further details.


[3] Grimmett, Geoffrey R. and Stirzaker, David R. *Probability and Random Processes*

[4] Asmussen, S{}o ren *Applied Probability and Queues*

[5] Norris, J. R. *Cambridge Series in Statistical and Probabilistic Mathematics*

[6] Feller, William. *An Introduction to Probability Theory and its Applications*


## 8.2 Review of Mathematics: Power Series

Know the following by heart, because it'll be on the midterm.

### 8.2.1 Binomial Theorem

The most important case of the binomial expansion

$$(1+x)^n = \sum_{k=0}^{n} \binom{n}{k} x^k$$

where we should observe

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1}$$

and it is an important insight that the numerator is a polynomial in $n$. Pitman comments that no one realized why this is important until about 1670. The reason is that this form can be extended to other powers, namely $n := -1, \frac{1}{2}, \frac{-1}{2}$, or any real number $n \to r \in \mathbb{R}$. Now look at

$$(1+x)^r = \sum_{k=0}^{\infty} \binom{r}{k} x^k \qquad (|x| < 1)$$

which is valid for all real $r$ and all real or complex $x$ with $|x| < 1$. Notice that the combinatorial meaning of $r$ 'choose' $k$ makes sense only for $r = n$ a positive integer and $k$ a non-negative integer. But the meaning of $\binom{r}{k}$ is extended to all real numbers $r$ and all non-negative integers $k$ by treating $\binom{n}{k}$ as a polynomial of degree $k$ in $n$, then substituting $r$ in place of $n$ in this polynomial.
This is the instance with $f(x) = x^r$ of the *Taylor expansion* of a function $f$ about the point 1

$$f(1+x) = f(1) + f'(1)x + \frac{f''(1)}{2!}x^2 + \cdots$$

which for suitable $f$ is valid for $|x| < R$, where $R$ is the radius of convergence. Usually for our purposes, $R \geq 1$. As another Taylor expansion (around 0 instead of 1)

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

We get exponentials arising as limit of binomial probabilities (e.g. the Poisson distribution). Also, recall that the geometric distribution converges to the exponential distribution with suitable scaling.

## 8.3 Probability Generating Functions

Suppose we have a non-negative integer-valued random variable $X$, which for simplicity will have non-negative integer values $X \in \{0, 1, 2, \dots\}$.

**Probability Generating Function (PGF)**

The *probability generating function* for a discrete $X \in \{0, 1, 2, \dots\}$ is

$$\phi_X(z) := \mathbb{E}z^X$$

We usually take $0 \leq z \leq 1$. When discussing PGFs, we may push $z$ to $|z| \leq 1$, but in this course we will work entirely with PGFs defined as a function of an argument $z \in [0, 1]$.

Then $\phi_X(z) \in [0,1]$ too, and there are many contexts in which $\phi_X(z)$ acquires meaning as the probability of something. Now we can write the above as a power series. Recall that

$$\mathbb{E}g(X) = \sum_{n=0}^{\infty} \mathbb{P}(X=n)g(n) \tag{8.1}$$

so

$$\phi_X(z) := \mathbb{E}Z^X = \sum_{n=0}^{\infty} \mathbb{P}(X=n)z^n = \sum_{n=0}^{\infty} P_n z^n$$

where $p_n := \mathbb{P}(X=n)$. We worked with PGFs very briefly in a previous lecture, for dice probabilities, namely taking $X$ uniform on $\{1,2,3,4,5,6\}$, and we looked at

$$\phi_X(z) = \frac{1}{6}(z + \ldots + z^6)$$

Recall this is where Pitman asked us to look at powers of this expansion in Wolfram Alpha. Notice that by convention, $0^0 = 1$, so $\phi_X(0) = \mathbb{P}(X=0)$. Now for any PGF, we have

$$\begin{aligned}
\frac{d}{dx}\phi_X(z) &= \frac{d}{dz}\sum_n \mathbb{P}(X=n)z^n \\
&= \sum_n \mathbb{P}(X=n)\frac{d}{dz}z^n \\
&= \sum_n \mathbb{P}(X=n)nz^{n-1}
\end{aligned}$$

and so we see that

$$\mathbb{E}X = \frac{d}{dz}\phi_X(z)\Big|_{z=1^-}$$

where we must approach $z=1$ from the left if the radius of convergence $R$ is exactly $R=1$, but typically $R>1$ and you can just evaluate the derivative at $z=1$.
Perhaps we'd like to compute the variance. We ask, what happens if we differentiate twice?

$$\left(\frac{d}{dz}\right)^2 \phi_X(z) = \sum_{n=0}^{\infty} \mathbb{P}(X=n)n(n-1)z^{n-2}$$

Again we'd like the $z$ factor to go away, so we set $z := 1$ and we have

$$\begin{aligned}
\mathbb{E}\left[X(X-1)\right] &= \sum_{n=0}^{\infty} \mathbb{P}(X=n)n(n-1) \\
&= \left(\frac{d}{dz}\right)^2 \phi_X(z)\Big|_{z=1^-}
\end{aligned}$$

Recall that $X_\lambda \sim$ **Poisson** $(\lambda)$ if and only if:

$$\mathbb{P}(X_\lambda = n) = \frac{e^{-\lambda}\lambda^n}{n!},$$

which via the generating function implies

$$\phi_{X_\lambda}(z) = \sum_{n=0}^{\infty} \frac{e^{-\lambda}\lambda^n z^n}{n!} = e^{-\lambda}e^{\lambda z} = e^{\lambda(z-1)}$$

Easily from the above analysis by $d/dz$, or otherwise,

$$\begin{aligned} \mathbb{E}X_\lambda &= \lambda \\ \mathbb{V}\mathrm{ar}(X_\lambda) &= \lambda \end{aligned}$$

A (good) question arises whether $\phi_X(z)$ is a probability. The answer is yes, because after all the range of values is between 0 and 1, and any such function can be interpreted as a probability. Notably, we have

$$\phi_X(z) = \mathbb{P}(X \leq G_{1-z})$$

where $G_p$ for $0 \leq p \leq 1$ denotes a random variable independent of $X$ with the geometric $(p)$ distribution on $\{0, 1, \ldots\}$: for $n \geq 0$
Then

$$\mathbb{P}(G_p = n) = (1-p)^n p, \text{ and } \mathbb{P}(G_p \geq n) = (1-p)^n.$$

In summary, we can think of a probability generating function as a probability, and we only need that $G_{1-z}$ is independent of $X$.
Now if $X, Y$ are independent, then

$$\begin{aligned} \mathbb{E}z^{X+Y} &= \mathbb{E}\left[z^X z^Y\right] \\ &= \left[\mathbb{E}z^X\right]\left[\mathbb{E}z^Y\right] \\ &= \phi_X(z)\,\phi_Y(z) \end{aligned}$$

Hence the PGF of a sum of independent variables is the product of their PGFs.

**Example:** Let $G_p \sim$ **Geometric**$(p)$ on $\{0, 1, 2, \ldots\}$. Then

$$\mathbb{P}(G_p = n) = (1-p)^n p, \text{ for } n = 0, 1, 2, \ldots$$

Now if we want to look at the probability generating function, we have

$$\mathbb{E}(z^{G_p}) = \sum_{n=0}^{\infty} q^n p z^n = \frac{p}{1-qz}$$

for $p + q = 1$ and $|z| < 1$. Now we look at

$$T_r := G_1 + G_2 + \cdots + G_r$$

where $r = 1, 2, 3, \ldots,$ and $G_i$ are all independent geometrically distributed with the same parameter $p$. The interpretation is to see $G_p$ as the number of failures before the first success. That is, the number of 0s before the first 1 in independent **Bernoulli**()) 0,1 trials. Then similarly,

$$T_r = T_{r,p} = \text{ number of 0s before } r^{\text{th}} \text{ 1 in indep. } \textbf{Bernoulli}(p) \text{ 0,1 trials}$$

Looking at i.i.d. copies of $G_p$ we use generating functions

$$\begin{aligned}
\mathbb{E}z^{T_r} &= \left(\frac{p}{1-qz}\right)^r = p^r(1-qz)^{-r} \\
&= p^r\left(1+(-qz)\right)^{-r} \\
&= \sum_{n=0}^{\infty}\binom{-r}{n}(-qz)^n, \\
&= p^r\sum_{n=0}^{\infty}\frac{(r)_{n\uparrow}}{n!}q^n z^n
\end{aligned}$$

where we simply plug into Newton's binomial formula. Notice this is

$$\mathbb{E}z^{T_r} = p^r\sum_{n=0}^{\infty}\frac{(r)_{n\uparrow}}{n!}q^n z^n$$

where

$$(r)_{n\uparrow} := r(r+1)\cdots(r+n-1)$$

$$\frac{(r)_{n!}}{n!} = \binom{r+n-1}{n}$$

From 134, we know this to be the negative binomial distribution. The above formula can be derived directly by counting: $\binom{r+n-1}{n}$ is the number of ways to place the $n$ failures in the first $r+n-1$ trials, and the last $(n+r)^{\text{th}}$ trial must be a 1. But the generating function technique used above is instructive, and can be applied to more difficult problems.

## 8.4 Probability Generating Functions and Random Sums

Suppose we have $Y_1, Y_2, \ldots$ i.i.d. non-negative integer random variables, with probability generating function

$$\phi_Y(z) = \mathbb{E}z^{Y_k} = \sum_{n=0}^{\infty}\mathbb{P}(Y_k = n)z^n$$

the same generating function for all $Y_k$. Now consider another random variable, $X \geq 0$, integer valued, assumed independent of the sequence of $Y's$, and look at:

$S_X = Y_1 + Y_2 + \ldots + Y_X$, the sum of $X$ independent copies of $Y$. Then

$$S_n = Y_1 + \cdots + Y_n$$
$$S_X = Y_1 + \cdots + Y_X$$

Now if $X = 0$ with 0 copies of $Y$, then our convention is to set the empty sum to give 0. We wish to find the PGF of $S_X$. The random index $X$ is annoying, so try conditioning on it

$$\begin{aligned}
\mathbb{E}z^{S_x} &= \sum_{n=0}^{\infty} \mathbb{P}(X = n)\mathbb{E}\left(z^{S_n}\right) \\
&= \sum_{n=0}^{\infty} \mathbb{P}(X = n)\left[\phi_Y(z)\right]^n \\
&= \phi_X\left[\phi_Y(z)\right]
\end{aligned}$$

which is a composition of generating functions. In the middle line, recognize this is a generating function, just evaluated at a different location. Notice that for this to hold, we needed to assume that $X$ is independent of $Y_1, Y_2, \ldots$. Also, there are some easy consequences for moments which you can derive directly or by generating functions, especially $\mathbb{E}S_X = (\mathbb{E}Y)\mathbb{E}X$ and you can get a formula for $\mathbb{E}S_X^2$ and hence the variance of $S_X$.

## 8.5 Application: Galton-Watson Branching Process

Assume that we're given some probability distribution (offspring distribution)

$$p_0, p_1, p_2, \ldots$$

Start with some fixed number $k$ of individuals in generation 0, where each of these $k$ individuals has offspring with distribution according to $X$. Our common notation is

$$Z_n := \#\text{ of individuals in generation } n$$

and so we have the following equality in distribution

$$(Z_1 \mid Z_0 = k) \stackrel{d}{=} X_1 + X_2 + \cdots + X_k$$

where the $X_i$ are i.i.d. $\sim p$. Continuing the problem, given $Z_0, Z_1, \ldots, Z_n$ with $Z_n = k$, then $Z_{n+1} \sim X_1 + \cdots + X_k$. It's intuitive to draw this as a tree, where individuals of generation 0 have some number of offspring and some have none. We create a branching tree from one stage to the next. Clearly,$(Z_n)$ is a Markov chain on $\{0, 1, 2, \ldots\}$. Note that state $k = 0$ is absorbing, which fits with the convention of empty sums i.e. summing 0 copies of the offspring variable gives 0.

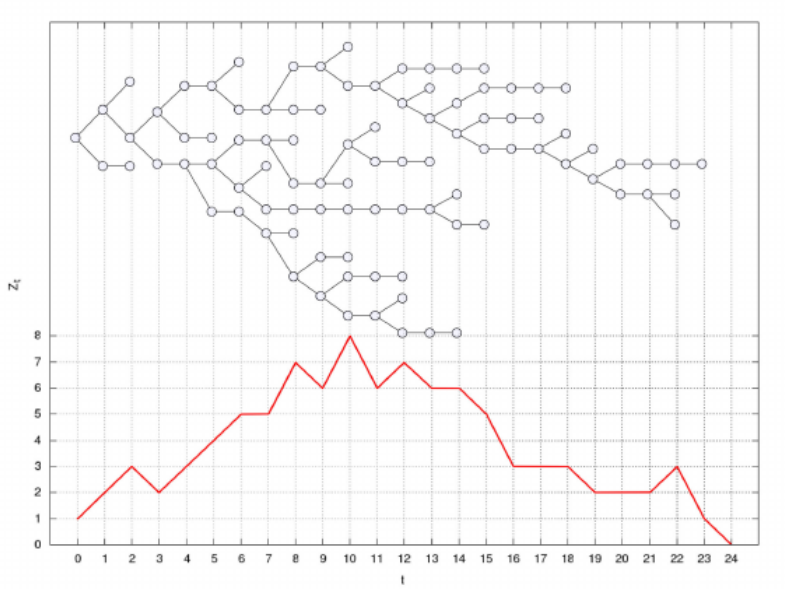



Figure 8.1: See [7] A realization of the Galton-Watson process. At the top, the tree associated to the process is shown, starting from the left ($Z_0 = 1$). At the bottom, the evolution of the number of elements originated in each generation $t$ are displayed.

Here's a visualization of the branching process.

Now, we should expect that generating functions should be helpful, as we are iterating random sums. We'll iterate the composition of generating functions. For simplicity, start with $z_0 = 1$. Let $\phi_n(s) = \mathbb{E}\left(s^{Z_n}\right)$ for $0 \leq s \leq 1$. We see that

$$Z_{n+1} = \text{ sum of } Z_n \text{ copies of } X$$

Hence

$$\phi_1(s) = \sum_{n=0}^{\infty} p_n s^n = \mathbb{E}s^X$$

which we define as the **offspring generating function**. To find $\phi_2$, we look at $\phi_1(\phi_1(s))$. That is,

$$\begin{aligned} \phi_2(s) &= \text{ PGF of sum of } Z_1 \text{ copies of } X \\ &= \phi_1\left[\phi_1(s)\right] \end{aligned}$$

Continuing, we similarly have

$$\begin{aligned} \phi_3(s) &= \text{ PGF of sum of } Z_2 \text{ copies of } X \\ &= \phi_1(\phi_1(\phi_1(s))) \end{aligned}$$

and so on. Now Pitman presents the famous problem of finding the probability of

extinction

$$\mathbb{P}_1(\text{extinction}) = \mathbb{P}_1(Z_n = 0 \text{ for large } n)$$

$$= \lim_{n\to\infty} \mathbb{P}_1(Z_n = 0)$$

Now we ask, how do we find $Z_n = 0$? We basically have a formula for this. What is the probability that $Z_1 = 0$? This is simply

$$\mathbb{P}_1(Z_1 = 0) = p_0$$

Then

$$\mathbb{P}_1(Z_2 = 0) = \phi(\phi(0)) = \phi(p_0)$$

and similarly,

$$\mathbb{P}_1(Z_3 = 0) = \phi(\phi(\phi(0))) = \phi(\phi(p_0))$$

and so on. See figure 8.2. This gives the exact formula in general that

$$\mathbb{P}_1(Z_n = 0) = \phi_{n-1}(0)$$

where $\phi_{n-1}$ is the $n$ th iterate of the offspring generating function $\phi$. Because $\phi$ is continuous, it follows that the extinction probability

$$s_0 := \lim_{n\to\infty} \phi_n(s)$$

is a root of the equation

$$s = \phi(s)$$

In general $s_0$ is the least root $s$ of this equation with $0 \leq s \leq 1$. Note that $s = 1$ is always a root. Even if you aren't a fan of generating functions, you should note that they are inescapable in the solution to the branching extinction problem.
By analysis of the graph of $\phi$, which is convex with $\phi(0) = p_0$ and derivative $\phi'(1-) = \mu$, there are three cases:

- *supercritical* ($\mu > 1$): then there is a unique root $s_0$ with $0 \leq s_0 < 1$ and $\phi(s_0) = s_0$. This is the extinction probability.
- *subcritical* ($\mu < 1$): then the unique root is $s_0 = 1$: extinction is certain;
- *critical and non-degenerate* ($\mu = 1$) and $p_0 > 0$: then $s_0 = 1$. See figure 8.3. See figure 8.3. Because the generating function $\phi$ is convex, the only root returned from fixed point iteration is precisely at 1. This implies that if $\mu = 1$ and $p_0 > 0$ the probability of extinction $\mathbb{P}_1(\text{extinction}) = 1$. The fluctuations of $Z_n$ in this case lead with probability one to extinction.

There is a very annoying case for branching processes that we should not forget, which is the *degenerate case* where $p_1 := \mathbb{P}(X = 1) = 1$, which just makes the population stay at 1, $\mathbb{P}_1(Z_n = 1) = 1$ for all $n$, and the extinction probability is 0. There is no random fluctuation in it. The book presents this conclusion in different ways.

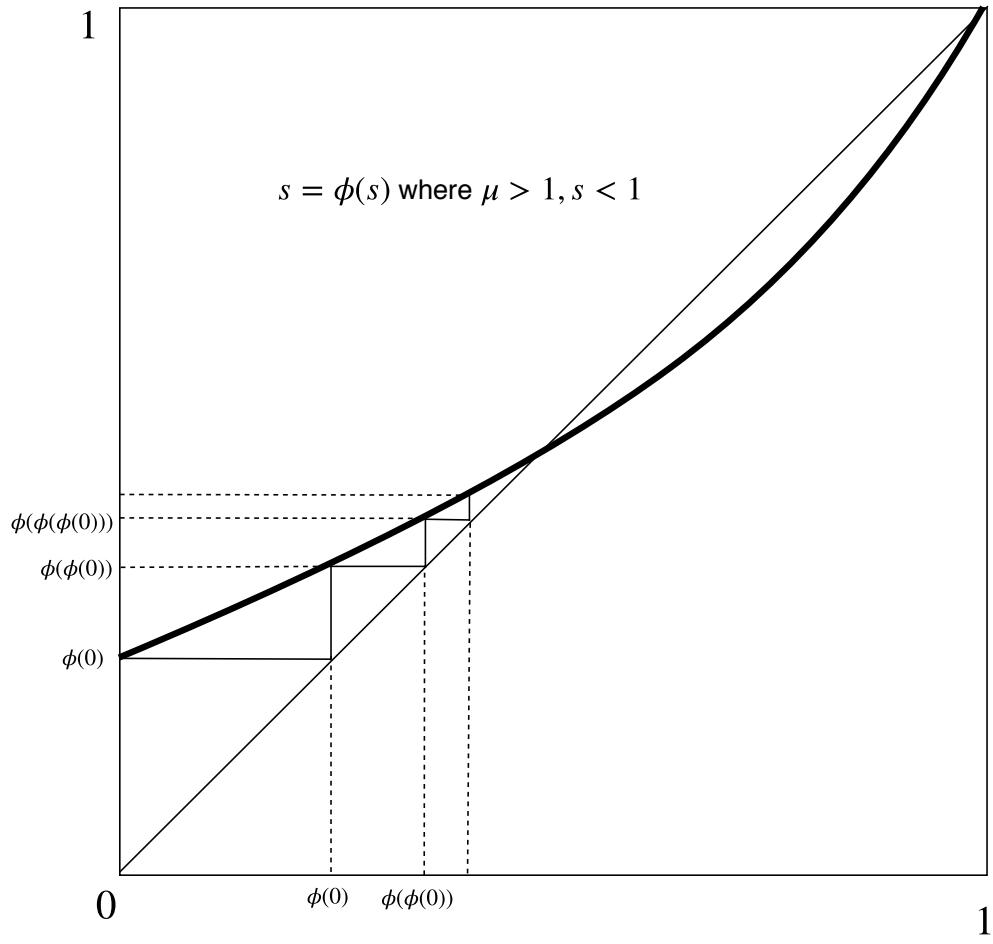



Figure 8.2: Supercritical. As an example, we sketch ($\phi(s)$ with respect to $s$ for the generating function of **Poisson** $(3/2)$. This gives a fixed point iteration returning the unique root $s$ of $s = \phi(s)$ with $s < 1$. Here the mean is larger than 1 ($\phi'(1) = \mu > 1$).

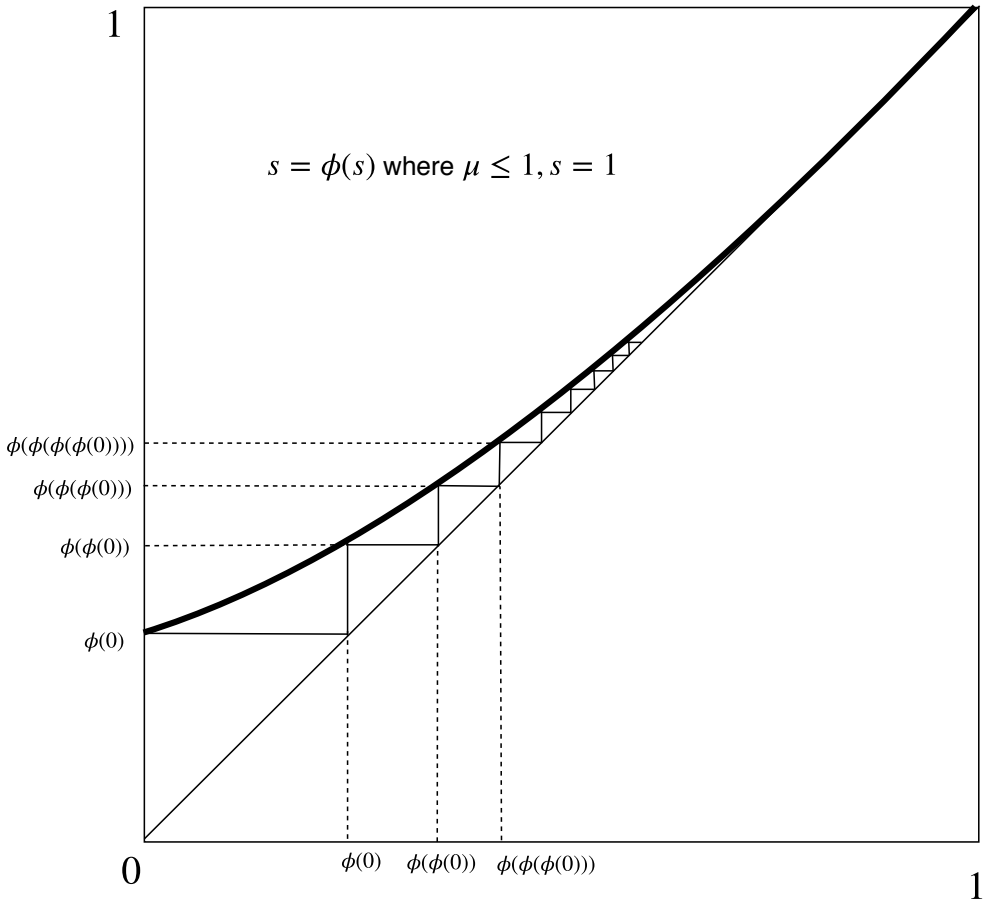



Figure 8.3: Subcritical or critical non-degenerate. We see that if $p_0 > 0$ and $\mu := \sum_n np_n \leq 1$, then the probability generating function is a convex curve with slope $\leq 1$ at 1 and value $p_0 > 0$ at 0. So the curve cannot ever cross the diagonal $s$. There is a no root of $\phi(s) = s$ with $s < 1$, and hence $\phi(\phi(\phi(\cdots(0)))) \to 1$ as $n \to \infty$

# LECTURE 9

# Potential Theory (Green Matrices)

## 9.1 Potential Theory (Green Matrices)

Green was an English mathematician who mainly worked on differential equations. This is a concept borrowed from differential equations, applied to our present context. Let $P$ be a transition matrix on a countable state space $S$. Define

$$\begin{aligned} G(x,y) &:= \sum_{n=0}^{\infty} P^n(x,y) \\ &= \mathbb{E}_x \sum_{n=0}^{\infty} \mathbb{1}(X_n = y) \\ &= \mathbb{E}_x N_y \end{aligned}$$

As a bit of book-keeping, we define

$$N_y := \text{ total \# visits to } y \text{ including a visit at time } n = 0$$

We can tell if states are transient or recurrent by looking at the Green matrix. Recall that we saw before that

$$G(x,x) < \infty \iff x \text{ is transient}$$

$$G(x,x) = \infty \iff x \text{ is recurrent}$$

**Remark:** If $S$ is finite, then it is obvious that $G(x,x) = \infty$ for some $x$. Particularly, some state must be recurrent. Let us start with the case where we have *infinite state space* and a *transient chain*. Recall that we defined

$$\begin{aligned} T_y &:= \min\{n \geq 1 : X_1 = y\} \\ V_y &:= \min\{n \geq 0 : X_n = y\} \end{aligned}$$

The following is generally true with no assumptions. There is a relation between $G(x, y)$ and $G(y, y)$. Always, $G(x, y) \leq G(y, y)$, and the ratio is a hitting probability

**Key Fact**

For all $x$ and $y$ (including $x = y$),

$$G(x, y) = \mathbb{P}_x(V_y < \infty)G(y, y)$$

Why is this formula true? Simply, $\mathbb{E}_x N_y$ is computed by conditioning on the event $(N_y > 0)$ which is identical to the event $(V_y < \infty)$. On this event, once we get to $y$, the chain starts over as if from $y$. To be more formal, we would cite the Strong Markov Property. Now let's look at a key example.

### 9.1.1 Example

Consider a simple random walk on the integers $\mathbb{Z} := \{\dots, -1, 0, 1, \dots\}$, where from any integer, we go left one with probability $q$ and right one with probability $p$. Pitman notes that we have something similar to this on homework 4, where we are moving on a circle. However, we could very easily unwrap the circle into the integer line. Let's compute the potential kernel. First of all, we ask if we really need the first parameter $x$, where $x, y \in \mathbb{Z}$. The definition of subtraction (translation invariance of the transition matrix) gives us:

$$G(x, y) = G(0, y - x)$$

So it's enough to discuss $G(0, y)$. Because of the key fact above, most of the action is looking at $G(y, y) = G(0, 0)$, and hence

$$G(x, y) = \underbrace{G(0, 0)}\,\underbrace{\mathbb{P}_x(\text{event of hitting } y)}$$

These two are our key ingredients. Let's first compute $G(0, 0)$. Notice that this random walk can only come back on even $n$ (this is a periodic walk).

$$\begin{aligned} G(0, 0) &:= \sum_{n=0}^{\infty} P^n(0, 0) \\ &= \sum_{m=0}^{\infty} P^{2m}(0, 0) \\ &= \sum_{m=0}^{\infty} \binom{2m}{m} p^m q^m \end{aligned}$$

and comparing this against the case where $p = q = \frac{1}{2}$ and adjusting, we have:

$$G(0, 0) = \sum_{m=0}^{\infty} \binom{2m}{m} 2^{-2m} (4pq)^m$$

Now Pitman states the following fact and requires that we perform this tedious computation once in our life

$$\binom{2m}{m}2^{-2m} = \frac{\left(\frac{1}{2}\right) m\uparrow}{m!} = \frac{\left(\frac{1}{2}\right)\left(\frac{1}{2}+1\right)\cdots\left(\frac{1}{2}+m-1\right)}{m(m-1)\cdots 1}$$

and we know (from the previous lecture) that for $|x| < 1$

$$(1+x)^r = \sum_{m=0}^{\infty}\binom{r}{m}x^r$$

$$\implies (1-x)^{-r} = \sum_{m=0}^{\infty}\binom{-r}{m}(-x)^r$$

$$= \boxed{\sum_{m=0}^{\infty}\frac{(r)_{m\uparrow}}{m!}x^m}$$

where we call this the negative binomial expansion. Bringing this back to the problem at hand (recognizing the negative binomial coefficient), we have

$$G(0,0) = \sum_{m=0}^{\infty}\frac{\left(\frac{1}{2}\right)_{m\uparrow}}{m!}(4pq)^m$$

by negative binomial expansion with $r := \frac{1}{2}$ and $x := 4pq$. Then this gives

$$G(0,0) = (1-4pq)^{-\frac{1}{2}}$$

Pitman reminds that in using this expansion, we should always be cautious for convergence in ensuring $|x| < 1$. Hence in this problem, provided $4pq < 1$ (equivalent to $p \neq \frac{1}{2}$)
Notice that if $p = \frac{1}{2}$, then in our formula we have $(1-1) = 0$ to a negative power, which gives us $\infty$. We can easily check

$$\binom{2m}{m}\left(\frac{1}{2}\right)^{2m} \sim \frac{c}{\sqrt{m}} \qquad (m \to \infty)$$

This precisely gives $G(0,0) = \infty$ in the case $p = \frac{1}{2}$. Hence

$$G(0,0) = \begin{cases} \infty, & p = \frac{1}{2} \\ (1-4pq)^{-\frac{1}{2}}, & p \neq \frac{1}{2} \end{cases}$$

Pitman says we can be a bit cuter about this. Notice

$$1 - 4pq = 1 - 4p + 4p^2 = (2p-1)^2$$

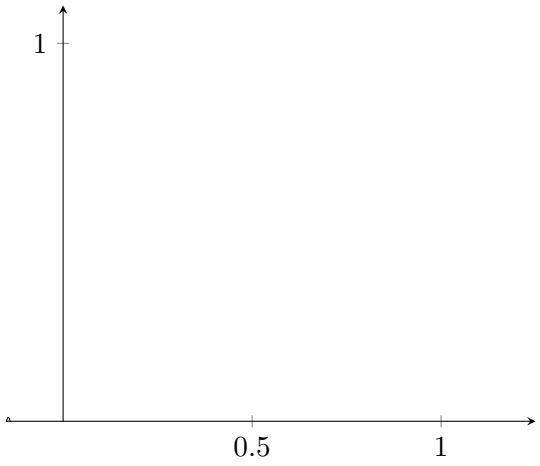



Figure 9.1: Graph of $4p(1-p)$

Therefore,

$$\begin{aligned} G(0,0) &= \left[(2p-1)^2\right]^{-\frac{1}{2}} \\ &= \frac{1}{|2p-1|} \\ &= \frac{1}{2\left|p-\frac{1}{2}\right|} \end{aligned}$$

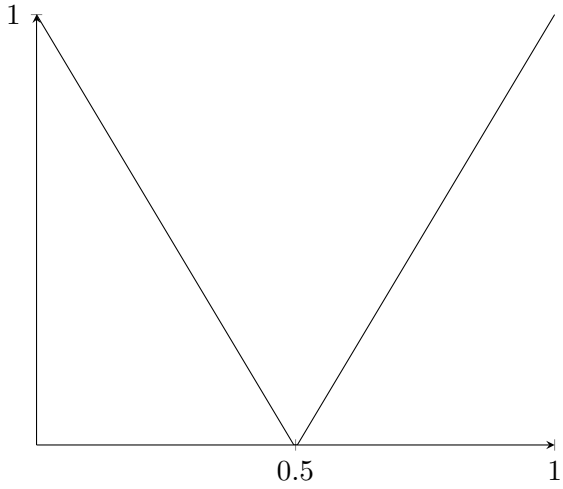



Figure 9.2: Graph of $|2x-1|$. Similar graph present in homework, with the exception of a smoother valley towards $1/2$, as compared to the sharp corner present here.

## 9.2 Escape Probability

We'll continue the $p, q$ walk. Consider the probability, starting at 0, that we never come back, which we write as $\mathbb{P}_0(T_0 = \infty)$, and set this equal to $w$.
Now we ask, what is $w$ in terms of $G(0,0)$? Recall that $G(0,0) = \mathbb{E}_0 N_0$, the expected number of hits on 0. Again, our convention is to count the starting position at time 0. Then

$$\mathbb{P}_0(N_0 = 1) = \mathbb{P}_0(T_0 = \infty) = w$$

which is familiar from a past discussion. That is, under $P_0$, starting at 0, $N_0$ has a very friendly distribution. *Brief Recap: Geometric Distribution* Recall that for $N \sim \textbf{Geometric}(p)$ on $\{1, 2, 3, \dots\}$ with probability of success is $p$. Then

$$\mathbb{P}(N = n) = (1-p)^{n-1} p$$

Hence for our problem

$$\mathbb{P}_0(N_0 = n) = (1-w)^{n-1} w$$

so

$$N_0 \sim \textbf{Geometric}(w)$$

Hence we can do away with our placeholder $w$, so that $w = \frac{1}{G(0,0)}$, so that

$$\mathbb{P}_0(T_0 = \infty) = \frac{1}{G(0,0)} = 2|p - 1/2|$$

The homework problem graph is very similar to this, but with a curve instead of a sharp 'valley' at $1/2$

## 9.3 More Formulas for Simple Random Walks (SRW)

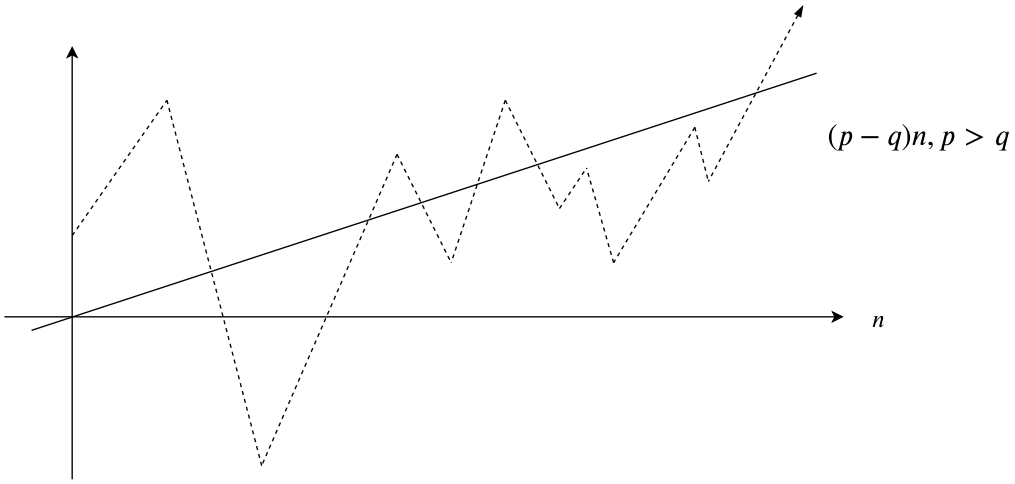
We consider the same context as our example. Let $S_n = \Delta_1 + \cdots + \Delta_n$, where each $\Delta_i$ is either $+1$ or $-1$ with probability $p$ or $q$, respectively. What is the probability that $S_n$ tends to $+\infty$? This depends on $p$. In the recurrent case, this probability will be zero. Now if $p > q$, then we have a "drift" for our expectation that carries us off to $\infty$. That tells us that

$$\mathbb{P}(S_n \to +\infty) = \begin{cases} 0, & p < q \\ 1, & p > 1 \end{cases}$$

or more neatly, using indicator notation

$$\mathbb{P}(S_n \to +\infty) = \mathbb{1}\left(p > \frac{1}{2}\right)$$

$$\mathbb{P}(S_n \to -\infty) = \mathbb{1}\left(p < \frac{1}{2}\right)$$




Now assume $p > q$. Then for $x \geq 1$, we have

$$\mathbb{P}_x(T_0 < \infty) = \left(\frac{q}{p}\right)^x$$

by Gambler's ruin probability (in the limit) from a previous lecture. Now $\frac{q}{p} < 1$, so

$$\mathbb{P}_x(T_0 = \infty) = 1 - \left(\frac{q}{p}\right)^x$$

$$\mathbb{P}_{-x}(T_0 < \infty) = 1$$

because the drift up ($\uparrow$) takes us to $+\infty$ with probability 1 and must hit 0 along the way.

## 9.4 Green's Matrix for Finite State Space $S$

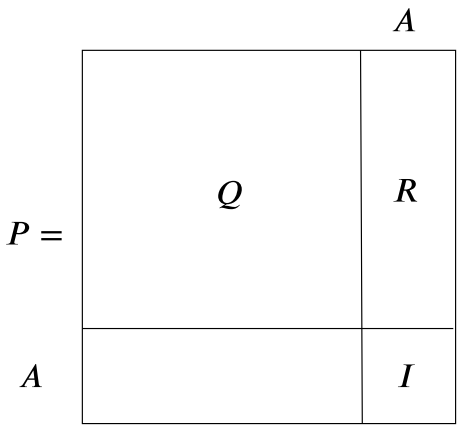
The entries $G(x, y)$ are only interesting if $y$ is a transient state. Take the example where the set $A$ is an absorbing set of states, and let $S - A$ (or $S \setminus A$) be interior states, and

$$\mathbb{P}_x(T_A < \infty) = 1 \qquad \forall x \in S$$

In the Gambler's Ruin example, take equal probability ($\frac{1}{2}\uparrow, \frac{1}{2}\downarrow$). Let $S := \{0, 1, \ldots, N\}$ and $A = \{0, N\}$. In the matrix (see next page), $Q$ is a $(S-A)\times(S-A)$ matrix. We claim that for $x, y \in S - A$

$$G(x, y) = \sum_{n=0}^{\infty} P^n(x, y) \quad \text{(from earlier)}$$

$$= \sum_{n=0}^{\infty} Q^n(x, y)$$




because $P^n(x,y)$ is a sum of products along paths through interior states only, and $P(w,z) = Q(w,z))$ for all transitions $(w,z)$ contributing to such products. Notice that $Q$ is **not** stochastic; in fact, we say that $Q$ is "sub-stochastic". We see

$$(Q\mathbb{1})(x) = \mathbb{P}_x(X_1 \notin A)$$

which will sometimes be less than 1. In any case, it's $\leq 1$. We want to focus only on the non-degenerate (interesting) part of $G$. So, assuming $\mathbb{P}_x(V_a < \infty) > 0$ (there is some positive probability), then $G(x,a) = \infty$ for every $a$. This follows from

$$G(x,a) = \mathbb{P}_x(V_a < \infty) \underbrace{G(a,a)}_{=\sum_{n=0}^{\infty} 1 = \infty}$$

**Remark** Now we'll throw away all the absorbing states for our discussion, so that all our matrices are indexed by $S - A$. We're shrinking our matrix to focus on the interesting portion of our potential kernel. Hence

$$G = \sum_{n=0}^{\infty} Q^n, \text{ on } S - A$$

$$= I + Q + Q^2 + Q^2 + \cdots$$

and compare against

$$QG = Q + Q^2 + Q^3 + \cdots$$

and subtracting these gives

$$G - QG = G - GQ = G(I - Q) = I$$

which implies

$$\boxed{G = (I - Q)^{-1}} \qquad (\star\star\star)$$

Computationally, this boils down to simply using our computers to crunch the inverse. Of course, for large matrix powers, we may run into underflow, overflow, or computationally singular matrices. However, there are ways to treat this issue within numerical linear algebra.

### 9.4.1 Return to Gambler's Ruin

We'd like to find $G(x, \cdot)$, which is the row $x$ of the Green matrix for Gambler's Ruin. Take

$$G - GQ = I \implies G = I + GQ$$

Now the row $G(x, \cdot)$ is determined in general by

$$G(x, y) = \mathbb{1}(x = y) + \sum_z G(x, z)Q(z, y)$$

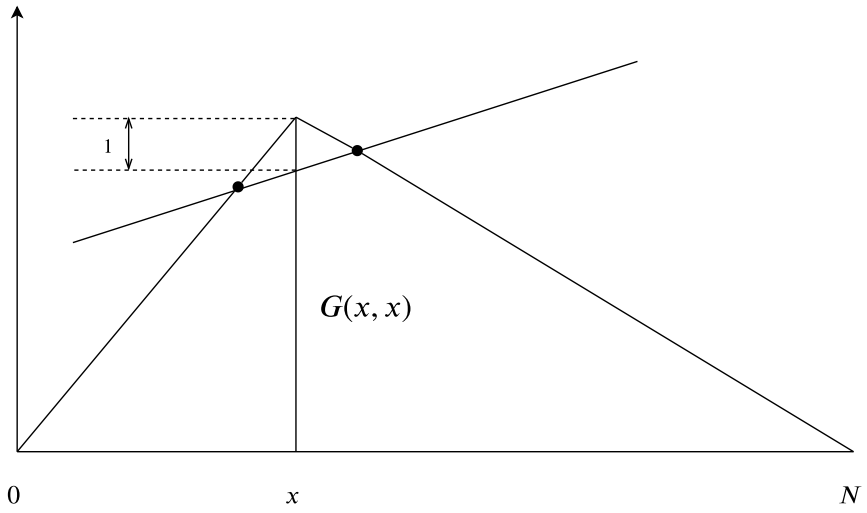
which for the Gambler's Ruin chain becomes

$$G(x, y) = \mathbb{1}(x = y) + \frac{1}{2}G(x, y - 1) + \frac{1}{2}G(x, y + 1)$$

with the boundary conditions that

- if $y = 1$ then $G(x, 0) = 0$
- if $y = N - 1$ then $G(x, N) = 0$

Now we'd like to graph $y \mapsto G(x, y)$.




If we ignore the indicator term, the right two terms gives a straight line from where we start at $x$, to the boundary, where the function must vanish. So the graph is a tent with its peak at $x$, and the equation for $x$ says this peak value $G(x, x)$ is 1 greater than the average of values to its right and left. That indicates how high the peak is, so the equations determine $G(x, \cdot)$ completely. From this consideration

$$G(x, y) = \begin{cases} \dfrac{y}{x}G(x, x), & 0 \le y \le x \\ \dfrac{N - y}{N - x}G(x, x), & x \le y \le N \end{cases}$$

Also

$$G(x,x) - 1 = \frac{1}{2}\left(\frac{x-1}{x} + \frac{N-(x+1)}{N-x}\right) G(x,x)$$

which is easily solved to give

$$G(x,x) = \frac{2x(N-x)}{N}$$

We have two important checks on this calculation. We know

$$\mathbb{P}_x(\text{ hit 0 before } N) = 1 - \frac{x}{N} = \frac{N-x}{N}$$

But we can compute this by conditioning on $T_0$: for $0 < x < N$

$$\begin{aligned}
\mathbb{P}_x(\text{ hit 0 before } N) &= \sum_{n=0}^{\infty} \mathbb{P}_x(T_0 = n+1) \\
&= \sum_{n=0}^{\infty} \mathbb{P}_x(X_n = 1, X_{n+1} = 0) \\
&= \sum_{n=0}^{\infty} \mathbb{P}_x(X_n = 1)\,\frac{1}{2} \\
&= G(x,1)\,\frac{1}{2} = \frac{1}{x}\frac{2x(N-x)}{N}\frac{1}{2} = \frac{N-x}{N}
\end{aligned}$$

Similarly, by the same method, get

$$\mathbb{P}_x(\text{hit } N \text{ before } 0) = \frac{x}{N}$$

as before. Also from before, we found $\mathbb{E}_x V_{0N} = x(N-x)$ which we can now check using the Green matrix

$$\mathbb{E}_x V_{0N} = \mathbb{E}_x \sum_{y=1}^{N-1} N_y \tag{9.1}$$

$$= \sum_{y=1}^{N-1} G(x,y) \tag{9.2}$$

$$= \sum_{y=1}^{x-1} G(x,y) + G(x,x) + \sum_{y=x+1}^{N-1} G(x,y) \tag{9.3}$$

$$= \cdots = x(N-x) \tag{9.4}$$

where you can easily fill in the "$\cdots$."

### 9.4.2 Conclusion

Whenever it is possible to evaluate the Green matrix $G$ you have immediate access to both hitting probabilities and mean hitting times as in the example above. See the text around (1.26) on page 62 for other applications of the same method.

# LECTURE 10

# The Fundamental Matrix of a Positive Recurrent Chain

## 10.1 Comments on Homework 5

This week's homework is posted as a worksheet on bCourses, and there is one correction as posted on Piazza. The first two problems are quite easy. The first is on branching processes.

For the second problem, we'll be frustrated if we don't know about Poisson thinning. This is Poisson thinning in disguise within a Markov chain.
In words, this means that if we have a Poisson ($\mu$) number of independent Bernoulli($p$) trials, the count of successes is Poisson ($\lambda p$). We should use the 'obvious' generating function to show this.

Now, a hint for the Kac identities, which are about a stationary process of 0s and 1s. With an informal notation, the claim is that in a stationary sequence of 0s and 1s $\mathbb{P}(1\underbrace{000}_{n}) = \mathbb{P}(\underbrace{000}_{n}1)$. This is much weaker than assuming reversibility (this is only for one such pattern). As a hint:

$$\mathbb{P}(1000) = \mathbb{P}(*000) - \mathbb{P}(0000),$$

where $*$ acts as a wild-card and can be a 0 or 1. Once you have worked with this idea, it takes about 4 lines to solve the problem.

The exercise on tail generating functions is just routine for us to get to get practice with generating functions.

The renewal generating function problem is relatively easy from the result of #4. We'll discuss a bit of renewal theory today in-class.

## 10.2 Renewal Generating Functions

For our discussions today, we assume $P$ is irreducible.

We look at the summation

$$\sum_{n=0}^{\infty}\left(u_n - \frac{1}{\mu}\right) = \sum_{n=0}^{\infty}\left[P^n(0,0) - \pi(0)\right],$$

where there is an underlying Markov chain, $P$ is our transition matrix with stationary probability $\pi$, and we assume irreducible and aperiodic.
We take the state 0 to be a special state. Then in the language of Renewal theory (by definition, returning to our initial state),

$$u_n = P^n(0,0) = \mathbb{P}_0(\text{return to 0 at time } n) = \mathbb{P}(\text{renewal at } n).$$

and the mean inter-renewal time is $\mu = 1/\pi(0)$. This formula arises from looking at a natural generalization of the Potential kernel (Green matrix):

$$G := \sum_{n=0}^{\infty} P^n = (I - P)^{-1},$$

for a Markov matrix $P$. This matrix $G$ is very useful for transient chains, when all entries of $G$ are finite. We may ask, what is the Green matrix if $P$ is recurrent? In general

$$G(x,y) = \mathbb{E}_x \sum_{n=0}^{\infty} \mathbf{1}(X_n = y) = \sum_{n=0}^{\infty} P^n(x,y).$$

In words, this is $\mathbb{E}_x(\#$ of hits on $y$ with infinite time horizon). We know

$$G(x,y) = \infty, \forall_{x,y\in S} \iff P \text{ is recurrent}$$

Finite state irreducible $P$ are very interesting. They have a stationary distribution $\pi$, with $\pi P = \pi$. However, $G(x,y) = \infty$ for all $x$ and $y$, which is of no interest. However,, there is another matrix associated with a positive recurrent chain which is nearly as informative as $G = (1-P)^{-1}$ for a transient chain.
Assume for simplicity that $S$ is finite, and that $P$ is aperiodic. Then we know a lot about $P^n$. We know

$$P^n(x,y) \to \pi(y) \text{ as } n \to \infty.$$

Informally, the process loses track of its starting state $x$, and no matter what the value $x$ of $X_0$ the distribution of $X_n$ given $X_0 = x$ approaches the limit distribution $\pi$ as $n \to \infty$.
As an aside, we invent the notation $\mathbf{1}$ for a column vector of all 1s: so $\mathbb{1}(x) = 1$, for all $x \in S$. So $\pi P = \pi$, and $\pi\mathbb{1} = 1$.

A bit 'cuter': we use matrix notation to write simply:

$$P^n \to \Pi := \mathbb{1}\pi \text{ as } n \to \infty$$

where the limit matrix $\Pi$ is the matrix with all rows equal to $\pi$. Notice that $P^n \to \Pi$ rapidly as $n \to \infty$. If we're careful about this,

$$\left|P^n(x,y) - \pi(y)\right| \le c\rho^n \text{ for some } c < \infty \text{ and } 0 < \rho < 1.$$

This implies that

$$\sum_{n=0}^{\infty} \left|P^n(x,y) - \pi(y)\right| < \infty,$$

which then implies that

$$\sum_{n=0}^{\infty} (P^n - \Pi)$$

exists, entrywise as a limit matrix.
Look what happens when we square $(P - \Pi)$. Recall that matrix multiplication is not commutative; however, we can still perform the expansion:

$$\begin{aligned}(P-\Pi)^2 &= (P-\Pi)(P-\Pi)\\ &= P^2 - P\Pi - \Pi P + \Pi^2\\ &= P^2 - \Pi - \Pi + \Pi\\ &= P^2 - \Pi\end{aligned}$$

as you can easily show. In general, the product of any number of factors $P$ and $\Pi$ is $\Pi$, so long as there is at least one $\Pi$. Hence,

$$(P-\Pi)^n = P^n - \Pi \text{ for } n = 1, 2, \ldots$$

where we need only use the binomial theorem to evaluate the coefficient of $\Pi$. Beware that

$$(P-\Pi)^0 = I \neq I - \Pi = P^0 - \Pi$$

which means that the $n = 0$ term must be treated separately in calculations such as the following:

$$\begin{aligned}\sum_{n=0}^{\infty} (P^n - \Pi) &= I - \Pi + \sum_{n=1}^{\infty} (P^n - \Pi)\\ &= I - \Pi + \sum_{n=1}^{\infty} (P-\Pi)^n\\ &= \sum_{n=0}^{\infty} (P-\Pi)^n - \Pi\end{aligned}$$

Recall that

$$\sum_{n=0}^{\infty} K^n = (I-K)^{-1},$$

for suitable $K$ (like a sub-stochastic matrix). Now this is the case for $K := P - \Pi$, and we can define:

$$\boxed{Z := \sum_{n=0}^{\infty} (P-\Pi)^n = (I - P + \Pi)^{-1}}$$

which is called the *fundamental matrix* of the irreducible, recurrent Markov chain. Our homework is to look at

$$\sum_{n=0}^{\infty} \left( u_n - \frac{1}{\mu} \right),$$

for a renewal sequence $u_n$. We can always write this, for a suitable Markov matrix $P$ with invariant probability $\pi$ , $\Pi = \mathbb{1}\pi$ and fundamental matrix $Z$ as above, (see "renewal chain" early in the text), as

$$\sum_{n=0}^{\infty} \left( u_n - \frac{1}{\mu} \right) = \sum_{n=0}^{\infty} (P^n(0,0) - \pi(0)) = Z(0,0) - \pi(0) \tag{10.1}$$

To summarize this discussion:

- for an aperiodic recurrent $P$ with finite state space $S$ and $\pi P = \pi$, let $\Pi := \mathbb{1}\pi$. Then the matrix $I - P + \Pi$ has an inverse $Z$ as above.

This result is true also for aperiodic $P$, except that the usual partial sums of the series diverge, and so the series must be evaluated using an Abel sum:

$$Z := (I - P + \Pi)^{-1} = \lim_{s \uparrow 1} \sum_{n=0}^{\infty} (1-P)^n s^n$$

as discussed further in the homework for the diagonal entries of $Z$ corresponding to (10.1).

**Example:** Consider the period 2 transition probability matrix

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \implies \Pi = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix} \implies I - P + \Pi = \begin{bmatrix} 3/2 & -1/2 \\ -1/2 & 3/2 \end{bmatrix}$$

which is easily inverted to give

$$Z = (I - P + \Pi)^{-1} = \begin{bmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{bmatrix} \tag{10.2}$$

In this example $P^n(0,0)$ for $n = 0, 1, 2, \dots$ gives the sequence $(1, 0, 1, 0, 1, 0, 1, 0, \dots)$, with $\pi(0) = \frac{1}{2}$ and $\mu = 1/\pi(0) = 2$. Notice that

$$(P^n(0,0) - \pi(0), n = 0, 1, 2, \dots) = \left(\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}, \cdots\right),$$

which gives an oscillating sum if left untreated. Using an Abel sum, we can check that

$$\lim_{s\uparrow 1} \left(P^n(0,0) - \pi(0)\right) s^n = \lim_{s\uparrow 1} \frac{1}{2} \sum_{n=0}^{\infty} (-s)^n = \lim_{s\uparrow 1} \frac{1}{2} \frac{1}{(1+s)} = \frac{1}{4} \tag{10.3}$$

Whereas according to (10.1) and (10.2) the same limit is evaluated as

$$Z(0,0) - \pi(0) = \frac{3}{4} - \frac{1}{2} = \frac{1}{4}. \tag{10.4}$$

As detailed in later sections there are lots of formulas for features of a recurrent Markov chain in terms of entries of the fundamental matrix $Z$. For transient matrices, we express things in terms of entries of the Green matrix. For recurrent chains, we express things in terms of entries of $Z$.

## 10.3 Variance of Sums Over a Markov chain

A first indication of the importance of the matrix $Z$ comes from computation of variances in the Central Limit Theorem (CLT) for Markov chains.
A key special case arises when $P = \Pi$. This means that under $\mathbb{P}_\lambda$ with $X_0 \sim \lambda$ all the following variables are iid with $X_1 \sim \pi$, $X_2 \sim \pi$, and so on.
In this (iid) case with $P = \Pi$, and more generally, we look at:

$$S_n(f) := \sum_{k=1}^{n} f(X_k),$$

which we may regard as the reward from $n$ steps of the chain if we are paid $\$f(x)$ for each value $x$. In the iid case, and more generally under $\mathbb{P} = \mathbb{P}_\pi$ with $X_0 \sim \pi$, so the chain is stationary,

$$\begin{aligned} \mathbb{E}S_n(f) &= n\mathbb{E}f(X_1) \\ &= n\pi f, \end{aligned}$$

where $\pi f = \sum_x \pi(x) f(x)$ is a real number. Notice that the states $X_n$ of the chain can be abstract, but we assume $f$ to take on numerical values, so that we may discuss the expectation and variance of sums like $S_n(f)$. Continuing in the iid case $P = \Pi$, we have

$$\mathbb{V}\mathrm{ar}(S_n(f)) = n\mathbb{V}\mathrm{ar}(f(X_1)),$$

and $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$ gives, in matrix notation:

$$\sigma^2(f) := \mathbb{V}\mathrm{ar}(X_1(f)) = \pi f^2 - (\pi f)^2 .$$

Then by the Central Limit Theorem for sums of iid random variables, provided $\pi f^2 < \infty$,

$$\mathbb{P}\left(\frac{S_n(f) - n\pi f}{\sigma(f)\sqrt{n}} \le z\right) \to \Phi(z),$$

the standard normal CDF. So what about for a Markov chain?

### 10.3.1 The Mean

Assuming now that $(X_n)$ is an irreducible finite state Markov chain with $X_0 \sim \lambda$

$$\begin{aligned}
\mathbb{E}_\lambda \sum_{k=1}^{n} f(X_k) &= \lambda \left(\sum_{k=1}^{n} P^k\right) f \\
&= n\lambda \left(\frac{1}{n}\sum_{k=1}^{n} P^k\right) f \\
&\sim n \underbrace{\lambda \Pi}_{=\pi} f = n\pi f, \text{ as } n \to \infty
\end{aligned}$$

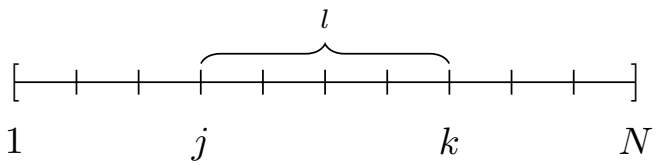
Notice that the mean per step $\pi f$ is the same as in the iid case. If $\lambda = \pi$ the formula for the mean of $S_n(f)$ is still exactly $n\pi f$, but for $\lambda \neq \pi$ this formula only holds asympotically in the limit as $n \to \infty$ instead of exactly.

### 10.3.2 The Variance

At least in the aperiodic case, because of the convergence in distribution of $X_n$ to $\pi$, it is intuitively obvious that the behavior of $\mathbb{V}\mathrm{ar}S_n(f)$ for large $n$ can't depend much on the initial distribution $\lambda$. We certainly see this for the mean, and it holds here for the variance as well, even for periodic chains. So the most important case for variance computations is the stationary case with $\lambda := \pi$. Then we can compute

$$\begin{aligned}
\mathbb{V}\mathrm{ar}_\pi \left(S_n(f)\right) &= \mathbb{V}\mathrm{ar}_\pi \left(f(X_1) + f(X_2) + \cdots + f(X_n)\right) \\
&= \sum_{k=1}^{n} \mathbb{V}\mathrm{ar}_\pi f(X_k) + 2 \sum_{1 \le j < k \le n} \mathbb{C}\mathrm{ov}\left[f(X_j), f(X_k)\right] \\
&= n\, \mathbb{V}\mathrm{ar}_\pi f(X_1) + 2\sum_{l=1}^{n-1} (n-l)\mathbb{C}\mathrm{ov}\left[f(X_0), f(X_l)\right]
\end{aligned}$$

because stationarity of the process implies $\mathbb{C}\mathrm{ov}\left[f(X_j), f(X_k)\right] = \mathbb{C}\mathrm{ov}\left[f(X_0), f(X_{k-j})\right]$ and for $1 \le l \le n-1$ there are $n-l$ pairs $1 \le j < k \le n$ with $k - j = l$.




Now we are interested in what happens for large $n$. We see that

$$\frac{\mathbb{V}\mathrm{ar}_\pi(S_n(f))}{n} = \mathbb{V}\mathrm{ar}_\pi f(X_1) + 2\sum_{l=1}^{n-1}\left(1-\frac{l}{n}\right)\mathbb{C}\mathrm{ov}\left[f(X_0), f(X_l)\right]$$

To simplify, assume that $\pi f = 0$ which is the same as $\mathbb{E}_\pi f(X_k) = 0$ because $X_k \sim \pi$ under $\mathbb{P}_\pi$, the probability with the stationary measure. Then

$$\frac{\mathbb{V}\mathrm{ar}_\pi(S_n(f))}{n} = \pi f^2 + 2\sum_{l=1}^{n-1}\left(1-\frac{l}{n}\right)\mathbb{E}_\pi f(X_0)f(X_l),$$

where we get rid of the subtraction terms by our assumption. Now we ask how to compute $\mathbb{E}_\pi\left[f(X_0)f(X_l)\right]$, where we must respect the joint distribution between $X_0$ and $X_l$, which involves $P^l$. Use Markov chain properties and condition on $X_0$:

$$\begin{aligned}
\mathbb{E}_\pi\left[f(X_0)f(X_l)\right] &= \mathbb{E}_\pi\left[f(X_0)(\overbrace{\mathbb{E}_\pi f(X_l)|X_0}^{=(P^l f)(X_0)})\right] \\
&= \mathbb{E}_\pi f(X_0)(P^l f)(X_0) \\
&= \mathbb{E}_\pi\left(f\cdot P^l f\right)(X_0) \\
&= \pi(f\cdot P^l f),
\end{aligned}$$

where $(f\cdot g)(x) = f(x)g(x)$; in other words, not matrix multiplication of column vector times column vector, which does not make sense.

### 10.3.3 The Central Limit Theorem

For $f$ with $\pi f = 0$ we have derived the following exact formula

$$\frac{\mathbb{V}\mathrm{ar}_\pi\left[S_n(f)\right]}{n} = \pi f^2 + 2\sum_{l=1}^{n-1}\left(1-\frac{l}{n}\right)\pi\left(f\cdot P^l f\right).$$

Now we would like to see what happens when $n$ is large (as in the Central Limit Theorem). Let $n\to\infty$. Assume $P$ is aperiodic, so that $P^l \to \Pi$, and use $1-\frac{l}{n}\uparrow 1$. Then

$$\frac{\mathbb{V}\mathrm{ar}_\pi\left[S_n(f)\right]}{n} \xrightarrow{n\to\infty} \sigma^2(f) \tag{10.5}$$

where the asymptotic variance $\sigma^2(f)$ per step of the chain can be evaluated from the limit of the exact formula above as

$$\begin{aligned}
\sigma^2(f) &= \pi f^2 + 2\sum_{l=1}^{\infty} \pi\left(f \cdot P^l f\right) \\
&= \pi f^2 + 2\pi f\left(\sum_{l=1}^{\infty} P^l f\right) \\
&= \pi f^2 + 2\pi f \sum_{l=1}^{\infty}\left(P^l - \Pi\right) f \text{ because } \pi f = 0 \\
&= \pi f^2 + 2\pi f\,(Z - I)\, f \text{ where } Z := \sum_{l=0}^{\infty}(P-\Pi)^l \\
&= \boxed{2\pi f \cdot Zf - \pi f^2}.
\end{aligned}$$

To summarize, the fundamental matrix $Z$ arises naturally in the formula for the limiting variance per unit time in the sum $S_n(f)$ of values of a function $f$ of a stationary Markov chain.

**Central Limit Theorem for Markov Chains**

The CLT works for any irreducible and positive recurrent Markov chain and any $f$ with $\pi f^2 < \infty$ with this evaluation of the asymptotic mean and variance.

See Asmussen's text APQ [4] for a more careful statement and proof.

## 10.4 Further Applications of the Fundamental Matrix

The following material was not covered in lecture, but serves to review many of the basic properties of Markov chains, and provides further applications of the fundamental matrix of a recurrent Markov chain.

### 10.4.1 Stopping times

First recall from the text, page 13, the definition of a stopping time $T$ for a discrete time stochastic process $(W_n) = (W_n, n = 0, 1, 2\ldots)$ with countable state space. That is, $T$ is a random variable with values in $\{0, 1, 2, \ldots, \infty\}$, such that for each $n = 0, 1, 2, \ldots$ the event $(T = n)$ is determined by the values of $W_0, \ldots, W_n$. Formally, the indicator function $\mathbf{1}(T = n)$ is a function of $(W_0, \ldots, W_n)$:

$$\mathbf{1}(T = n) = f_n(W_0, \ldots, W_n) \qquad (n = 0, 1, 2, \ldots) \tag{10.6}$$

for some $\{0, 1\}$-valued function $f_n$ of $n+1$ variables $W_k \in S$ for $0 \le k \le n$. Because

$$1 - \mathbf{1}(T > n) = \mathbf{1}(T \le n) = \sum_{k=0}^{n} \mathbf{1}(T = k) \qquad (n = 0, 1, 2, \ldots)$$

equivalent condition are

$$\mathbf{1}(T \le n) = g_n(W_0, \ldots, W_n) \qquad (n = 0, 1, 2, \ldots) \tag{10.7}$$

for some $\{0, 1\}$-valued function $g_n$ of the same variables, and

$$\mathbf{1}(T > n) = h_n(W_0, \ldots, W_n) \qquad (n = 0, 1, 2, \ldots) \tag{10.8}$$

for some $\{0, 1\}$-valued function $h_n$ of the same variables. Intuitively, for each $n$ you can tell whether or not any of the events $(T = n)$, $(T > n)$ or $(T \le n)$ has occurred just by looking at the variables $W_0, \ldots, W_n$. It is only ever necessary to check one of the above conditions (10.6) (10.7) (10.8) for each $n = 0, 1, 2, \ldots$, because each of these conditions implies both the others by simple manipulation of indicator variables. Examples of stopping times $T$ are the now familiar first hitting times $T = V_A$, allowing a hit at time 0, with

$$\mathbf{1}(V_A > n) = \prod_{k=0}^{n} \mathbf{1}(W_k \notin A) \qquad (n \ge 0) \tag{10.9}$$

$$\mathbf{1}(V_A = n) = \mathbf{1}(W_n \in A) \prod_{k=1}^{n} \mathbf{1}(W_k \notin A) \qquad (n \ge 0). \tag{10.10}$$

with the convention for $n = 0$ in the product $\prod_{k=1}^{n}$ that the empty product $\prod_{k=1}^{0} = 1$ (usual convention that an empty product equals 1, which corresponds by taking logs to the convention that an empty sum equals 0). Similarly, the first passage times $T_A := \min\{n \ge 1 : W_n \in A\}$ are stopping times, with

$$\mathbf{1}(T_A > 0) = 1 \tag{10.11}$$

$$\mathbf{1}(T_A > n) = \prod_{k=1}^{n} \mathbf{1}(W_k \notin A) \qquad (n \ge 1) \tag{10.12}$$

$$\mathbf{1}(T_A = 0) = 0 \tag{10.13}$$

$$\mathbf{1}(T_A = n) = \mathbf{1}(W_n \in A) \prod_{k=1}^{n-1} \mathbf{1}(W_k \notin A) \qquad (n \ge 1). \tag{10.14}$$

Note that

- for both $T = V_A$ and $T = T_A$ the logical description of $\mathbf{1}(T > n)$ as a product of indicators, corresponding to an intersection of events, is slightly simpler that the corresponding description of $\mathbf{1}(T = n) = \mathbf{1}(T > n - 1) - \mathbf{1}(T > n)$.
- the definition of a stopping time $T$ relative to $(W_n)$ does not involve any assumptions about the distribution of the process $(W_n)$.

Now let $(W_n)$ be some process defined on a probability space with underlying probability measure $\mathbb{P}$, and let $(X_n)$ be another process derived as $X_n = x_n(W_0, \ldots, W_n)$ for some function $x_n$ of $W_0, \ldots, W_n$. Common examples are

- sums and products, $X_n = W_0 + \cdots + W_n$ and $X_n = W_0 \cdots W_n$.
- expansions of the state space to allow extra randomization: $W_n = (X_n, Y_n)$ for some process $(Y_n)$.

Let $P$ be a transition probability matrix on the countable state space of $X$. Say that $(X_0, X_1, \ldots)$ is *Markov with transition matrix $P$ relative to the history of* $(W_n)$, abbreviated *Markov* $(P)$ *relative to* $(W_n)$, if

- $X_n = x_n(W_0, \ldots, W_n)$ for some function $x_n$ of $W_0, \ldots, W_n$
- for every $n = 0, 1, 2, \ldots$, every choice of states $x$ and $y$ of the process $X$, and every event $A_n$ with $\mathbf{1}_{A_n} = f_n(W_0, \ldots, W_n)$ for some function $f_n$ of $W_0, \ldots, W_n$,

$$\mathbb{P}(A_n \text{ and } X_n = x \text{ and } X_{n+1} = y) = \mathbb{P}(A_n \text{ and } X_n = x)P(x, y) \qquad (10.15)$$

which is equivalent by $\mathbb{P}(B \,|\, A) = \mathbb{P}(AB)/\mathbb{P}(A)$ to

$$\mathbb{P}(X_{n+1} = y \,|\, X_n = x \text{ and } A_n) = P(x, y) \qquad (10.16)$$

for all choices of $x$ and $A_n$ with $\mathbb{P}(X_n = x \text{ and } A_n) > 0$.

Typically, (10.15) is used for computations. Taking $\mathbf{1}_{A_n}$ to be a function of $X_0, \ldots, X_n$ shows that if $(X_n)$ is Markov $(P)$ relative to the history of $(W_n)$, then $(X_n)$ is Markov $(P)$ relative to its own history. This is just the usual time-homogeneous Markov property of $(X_n)$. If $(X_n)$ is Markov relative to a richer history $(W_n)$ than just its own history, it means that given $X_n = x$ any additional information in $(W_0, \ldots, W_n)$, beyond what is already encoded in $(X_0, \ldots, X_n)$, is of no use in predicting the next value $X_{n+1}$: the distribution of this variable given $(X_n = x)$ is $P(x, \cdot)$, no matter what is known about $(W_0, \ldots, W_n)$ besides that the event $(X_n = x)$ has occurred.

Suppose now that $T$ is a stopping time relative to the history $(W_n)$ and that the process $X$ is Markov $(P)$ relative to $(W_n)$, under a probability measure $\mathbb{P} = \mathbb{P}_\lambda$ which assigns $X_0$ some arbitrary initial distribution $\lambda$. Then (10.15) for $A_n = (T = n)$ reads

$$\mathbb{P}_\lambda(T = n \text{ and } X_n = x \text{ and } X_{n+1} = y) = \mathbb{P}_\lambda(T = n \text{ and } X_n = x)P(x, y) \qquad (10.17)$$

so summing over $n = 0, 1, 2, \ldots$ gives

$$\mathbb{P}_\lambda(T < \infty \text{ and } X_T = x \text{ and } X_{T+1} = y) = \mathbb{P}_\lambda(T < \infty \text{ and } X_T = x)P(x, y) \qquad (10.18)$$

That is, for any initial distribution $\lambda$ of $X_0$

- under $\mathbb{P}_\lambda$ given $T < \infty$ and $X_T = x$ the distribution of $X_{T+1}$ is $P(x, \cdot)$ .

This iterates easily to give the *Strong Markov Property*:

- under $\mathbb{P}_\lambda$ given $T < \infty$ and $X_T = x$ the process $(X_T, X_{T+1}, \cdots)$ has the same distribution as the original chain $(X_0, X_1, \ldots)$ under $\mathbb{P}_x(\cdot) := \mathbb{P}_\lambda(\cdot \,|\, X_0 = x)$.

This formulation of the strong Markov property is exactly as in Durrett's Theorem 1.2 on page 13, except that Durrett derives the result only for stopping times $T$ of $(X_n)$ itself. The above argument shows that the Strong Markov Property holds also for stopping times $T$ of any process $(W_n)$ such that $(X_n)$ is Markov $(P)$ relative to the history of $(W_n)$. Such stopping times can involve additional randomization beyond the chain $(X_n)$, and are sometimes called *randomized stopping times of* $(X_n)$. Examples of such stopping times involving extra randomization are

- $T$ that is independent of $(X_n)$, taking $W_n = (T, X_n)$;
- $T := \min\{n \geq 0 : X_n = Y\}$ for a state $Y$ chosen independently of $(X_n)$, taking $W_n = (X_n, Y)$;
- $T := \min\{n \geq 0 : p(X_n) \leq U_n\}$ for some function $p : S \to [0,1]$ and $(U_n)$ i.i.d. uniform $[0,1]$ variables independent of $(X_n)$. Here $W_n := (X_n, U_n)$. So given you have not stopped before time $n$, after observing values of both $X_k$ and $U_k$ for $0 \leq k < n$, and $X_n = x$ you stop at time $n$ with probability $p(x)$, according to whether or not the current uniform variable $U_n \leq x$.

So the Strong Markov Property holds in examples such as these involving extra randomization.

### 10.4.2 Occupation Measures for Markov chains

Similarly to (10.19), the general formula (10.15) for $A_n = (T > n)$ reads

$$\mathbb{P}_\lambda(T > n \text{ and } X_n = y \text{ and } X_{n+1} = z) = \mathbb{P}_\lambda(T > n \text{ and } X_n = y)P(y, z). \quad (10.19)$$

Summing this over $n = 0, 1, 2, \ldots$ and all states $x$ gives

$$\mathbb{E}_\lambda \sum_{n=0}^{\infty} \mathbf{1}(T > n, X_{n+1} = z) = \sum_y \mathbb{E}_\lambda \left( \sum_{n=0}^{\infty} \mathbf{1}(T > n, X_n = y) \right) P(y, z). \quad (10.20)$$

For any initial probability distribution $\lambda$, and any stopping time $T$ of some history $(W_n)$ relative to which $(X_n)$ is Markov with transition matrix $P$, define measures $\lambda G_T$ and $\lambda P_T$ on the state space of the chain as follows: for $y \in S$

$$\lambda G_T(y) := \mathbb{E}_\lambda \sum_{n=0}^{\infty} \mathbf{1}(T > n, X_n = y) = \mathbb{E}_\lambda \sum_{n=0}^{T-1} 1(X_n = y) \quad (10.21)$$

$$\lambda P_T(y) := \mathbb{E}_\lambda \sum_{n=0}^{\infty} \mathbf{1}(T = n, X_n = y) = \mathbb{P}_\lambda(T < \infty, X_n = y). \quad (10.22)$$

Observe that

- the *pre-$T$ occupation measure* $\lambda G_T(\cdot)$ describes the expected numbers of hits of various states $y$ counting only times $n$ with $0 \le n < T$.

- $\lambda P_T(\cdot)$ is the distribution of $X_T$ on the event $T < \infty$; so $\lambda P_T(\cdot)$ is a sub-probability measure with total mass $\lambda P_T \mathbf{1} = \mathbb{P}_\lambda(T < \infty) \in [0, 1]$.

For purposes of matrix operations, each of these measures on $S$ should be treated as a *row vector*.

- for any non-negative function $f$ with $\lambda G_T |f| < \infty$

$$\lambda G_T f = \sum_{y \in S} \lambda G_T(y) f(y) = \mathbb{E}_\lambda \sum_{n=0}^{T-1} f(X_n) \in [0, \infty] \tag{10.23}$$

- In particular, for $f = \mathbf{1}$, the function with constant value 1, the total mass of $\lambda G_T(\cdot)$ is

$$\lambda G_T \mathbf{1} = \sum_{y \in S} \lambda G_T(y) = \mathbb{E}_\lambda T \in [0, \infty] \tag{10.24}$$

- If this expectation $\mathbb{E}_\lambda T < \infty$, then (10.23) holds with a finite expectation $\lambda G_T f$ for every bounded $f$.

Let $\delta_x$ be the row vector $\delta_x(y) = \mathbf{1}(x = y)$ with mass 1 at $x$ and mass 0 elsewhere. Let $G_T(x, \cdot) := \delta_x G_T(\cdot)$ be the pre-$T$ occupation measure and $P_T(x, \cdot) := \delta_x P_T(x, \cdot)$ the distribution of $X_T$ on $(T < \infty)$ for a chain started in state $x$. Let $G_T$ and $P_T$ denote the $S \times S$ matrices with these rows. As the notation suggests, and is justified by conditioning on $X_0$,

$$\lambda G_T(\cdot) = \sum_x \lambda(x) G_T(x, \cdot) \text{ and } \lambda P_T(\cdot) = \sum_x \lambda(x) P_T(x, \cdot). \tag{10.25}$$

Then (10.20) gives the following general *occupation measure identity* for any stopping time $T$ of a Markov chain $(X_n)$, including randomized stopping times, as discussed above:

$$\lambda G_{T+1} = \lambda G_T + \lambda P_T = \lambda + \lambda G_T P. \tag{10.26}$$

This identity is just two different ways of evaluating the pre-$(T + 1)$ occupation measure $\lambda G_{T+1}$:

- firstly by peeling off the contribution of the last term in (10.21)for $n = T$ on the event $(T < \infty)$, and

- secondly by peeling off the first term for $n = 0$, and evaluating the remaining terms by (10.20).

For every stopping time $T$ of a Markov chain with transition probability matrix $P$, this identity relates the initial distribution $\lambda$ of $X_0$ and the pre-$T$ occupation

measure $\lambda G_T$ to the distribution $\lambda P_T$ of $X_T$. The occupation measure identity can also be written more compactly as an identity of matrices:

$$G_{T+1} = G_T + P_T = I + G_T P \tag{10.27}$$

where for each $x \in S$ the identity of row $x$ in (10.27) is the identity (10.26) for $\lambda = \delta_x$. As a simple example, if $T = N$ has constant value $N \geq 0$, then (10.27) reduces to

$$\sum_{n=0}^{N} P^n = \left(\sum_{n=0}^{N-1} P^n\right) + P^N = I + \left(\sum_{n=0}^{N-1} P^n\right) P \tag{10.28}$$

which is obviously true for any matrix $P$ with well defined powers, not just for transition matrices.

Call a measure $\mu$ on $S$ *locally finite* if $\mu(z)$ is finite for every $z \in S$. Provided the initial distribution $\lambda$ and the stopping time $T$ are such that the occupation measure $\lambda G_T$ is locally finite, which is typically obvious by geometric bounds, the occupation measure identity (10.26) can be rearranged as

$$\lambda G_T(I - P) = \lambda - \lambda P_T. \tag{10.29}$$

You easily can check that this rearrangement of the occupation measure identity is justified in each of the following cases. These cases include several instances of (10.29) which have been discussed in previous lectures and in the text.

(a) The state space $S$ is infinite, every state $z \in S$ is transient, without any restriction on the stopping time $T \in [0, \infty]$. In particular $T = \infty$ is allowed for such a chain. Then $P_T = P_\infty$ is the zero matrix, and $G_T = G_\infty$ with

$$G_\infty = \sum_{n=0}^{\infty} P^n = (I - P)^{-1}$$

the Green matrix associated with $P$, as discussed in Lecture 9.

(b) $T = V_A$, the first hitting time of $A$, counting a hit at time $n = 0$, for any subset $A$ of $S$ such that $\mathbb{P}_x(V_A < \infty) > 0$ for all $x \notin A$. Let $Q_A(x, y) := P(x, y)\mathbf{1}(x \notin A, y \notin A)$ be the restriction of $P$ to $(S - A) \times (S - A)$, and write simply $I$ for the similarly restricted identity matrix. Then, as discussed in Lecture 9,

$$G_{V_A}(x, y) = \left(\sum_{n=0}^{\infty} Q_A^n\right)(x, y) = (I - Q_A)^{-1}(x, y) \text{ for } x \notin A \text{ and } y \notin A \tag{10.30}$$

and $G_{V_A}(x, y) = 0$ else. The evaluation of rows of the pre-$V_A$ occupation matrix $G_{V_A}$ was detailed in Lecture 9 for the fair Gambler's Ruin chain on $S = \{0, 1 \ldots, N\}$ with $A = \{0, N\}$, using a the occupation measure identity (10.29) to argue that in this case the graph of the function $y \mapsto G_{V_A}(x, y)$ is linear on each of the intervals $[0, x]$ and $[x, N]$, vanishing at 0 and $N$, with peak value $G_{V_A}(x, x)$ which is 1 greater than the average of neighboring values $\frac{1}{2}G_{V_A}(x, x - 1) + \frac{1}{2}G_{V_A}(x, x + 1)$.

(c) $\lambda$ and $T$ are such that $\lambda G_T$ is locally finite, with $\mathbb{P}_\lambda(T < \infty) = 1$ and $X_T \stackrel{d}{=} X_0$. Then (10.29) becomes $\lambda G_T(I - P) = 0$, so $\lambda G_T$ is a $P$-invariant measure.

(d) In particular, if $P$ is irreducible and recurrent, $\lambda = \delta_x$ and $T = T_x$ for any fixed state $x$, then and $X_0 \stackrel{d}{=} X_{T_x}$, so $\mu_x(\cdot) := G_{T_x}(x, \cdot)$, the occupation measure of a single $x$-block of the chain, gives a strictly positive, locally finite $P$-invariant measure:

$$\mu_x(\cdot) = \mu_x(\cdot)P \text{ with } 0 < \mu_x(y) < \infty \qquad (y \in S). \tag{10.31}$$

This is Durrett Theorem 1.24 on page 48. The above derivation of this result follows essentially the same steps as Durrett's proof.

### 10.4.3 Positive recurrent chains: the ergodic theorem

Focusing now on the case when $P$ is irreducible and positive recurrent, formula (10.31) gives a different invariant measure $\mu_x(\cdot)$ for each $x \in S$. Normalizing $\mu_x(\cdot)$ by its total mass

$$m_{xx} := \mu_x(\cdot)\mathbf{1} = \mathbb{E}_x T_x$$

gives a stationary probability measure $\pi$ for $P$. It appears at first as if this stationary probability measure $\pi_x(\cdot) = \mu_x(\cdot)/m_{xx}$ for $P$ might depend on the choice of reference state $x$. But it does not, for it can be shown in many ways that there can be at most one invariant probabiity measure for an irreducible transition matrix $P$. Hence the basic formula

$$\pi(x) = \frac{1}{m_{xx}} \qquad (x \in S) \tag{10.32}$$

for the unique stationary probability measure $\pi$ for an irreducible and positive recurrent chain. In particular, this uniqueness of $\pi$ and (10.32) are consequences of the following *ergodic theorem* for an irreducible chain (Durrett's Theorem 1.22): If $\pi$ is any stationary probability measure for an irreducible $P$, then for every initial distribution $\lambda$ of $X_0$, and every function $f$ with $\pi|f| < \infty$, there is the convergence

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \to \pi f := \sum_y \pi(y)f(y) \text{ as } n \to \infty. \tag{10.33}$$

In particular, for $f(y)$ the indicator function $f_x(y) = \mathbf{1}(y = x)$, this states that the long run limiting relative frequency of times the chain hits state $x$ is $\pi f_x = \pi(x)$. Technically, the convergence (10.33) holds *almost surely*, meaning with $\mathbb{P}_\lambda$ probability one, no matter what the intial distribution $\lambda$. A complete formulation and proof of this result is beyond the scope of this course. But this notion of *almost sure convergence* implies what is called *convergence in probability*: for every initial distribution $\lambda$

$$\mathbb{P}_\lambda\left(\left|\frac{1}{n}\sum_{k=1}^{n} f(X_k) - \pi f\right| > \epsilon\right) \to 0 \qquad (\forall \epsilon > 0). \tag{10.34}$$

If $S$ is finite, and more generally if $\pi f^2 < \infty$, this can be proved, as in the case of i.i.d. $X_i$, when $P = \Pi := \mathbf{1}\pi$ is the transition matrix with all rows equal to $\pi$, by

bounding the variance, as discussed earlier, and using Chebychev's inequality, From either kind of convergence of averages of $f(X_k)$ over the path of the Markov chain, granted that the limiting average value is a constant $\pi f$, this constant is obviously unique. Hence the uniqueness of $\pi$. To summarize: for an irreducible transition matrix $P$ with countable state space $S$, the following conditions are equivalent:

- $m_{xx} < \infty$ for some (hence all ) $x \in S$ ;
- $P$ has an invariant measure $\mu$ with $\mu(x) \geq 0$ for all $x \in S$ and $0 < \mu\mathbf{1} < \infty$ ;
- $P$ has unique invariant probability measure $\pi(x) = 1/m_{xx}$ .

Then the transition matrix $P$, or the associated Markov chain, is called *positive recurrent*, (10.33) and (10.34) imply

$$\frac{1}{n}\sum_{k=1}^{n} P^k \to \Pi := \mathbf{1}\pi \text{ as } n \to \infty. \tag{10.35}$$

If $P$ is aperiodic, then (10.35) can be strengthened (Durrett Theorem 1.23 on page 52) to

$$P^n \to \Pi \text{ as } n \to \infty \tag{10.36}$$

In both (10.35) and (10.36), and other formulas involving limits of matrices in this course, the meaning of convergence is that each entry of the matrix on the left converges to the corresponding entry of the matrix in the right. For transition matrices on both sides, as in (10.35) and (10.36), the rows on both sides are probability measures, meaning the entries are non-negative with row sums 1. Convergence of entries of a sequence of transition matrices then implies convergence of each row of probability measures in the metric of *total variation distance* between probability measures. For instance, in (10.36), for each initial state $x$,

$$\sum_y |P^n(x,y) - \pi(y)| \to 0 \text{ as } n \to \infty \tag{10.37}$$

as in the last estimate in Durrett's proof of (10.36) on page 53. This implies things like

$$\mathbb{E}_x f(X_n) = (P^n f)(x) = \sum_y P^n(x,y) f(y) \to \pi f \text{ as } n \to \infty \tag{10.38}$$

for every bounded function $f$, and that this convergence holds uniformly over all $f$ with $|f(x)| \leq b$ for any finite bound $b$.

### 10.4.4 Occupation measures for recurrent chains

Suppose now that the transition matrix $P$ is irreducible and recurrent. For any non-empty set of states $A \subseteq S$ in the state space $S$ let $T_A := \min\{n \geq 1 : X_n \in A\}$ be the first passage time to $A$. Recurrence of the chain implies $\mathbb{P}_x(T_A < \infty) = 1$ for all $x \in S$. Let $G_A(x, \cdot)$ denote the pre-$T_A$ occupation measure starting from state $x$. A fairly complete analysis of the process $(X_n, 0 \leq n \leq T_A)$, with initial fixed value $x \notin A$ and final random value $X_{T_A} \in A$ is obtained by consideration of the pre-$T_A$ occupation matrix $G_A$ with rows $G_A(x, \cdot)$. Observe that

- if $x \notin A$ then $G_A(x,y) = 0$ for $y \in A$, because the pre-$T_A$ occupation measure only counts hits on $A$ strictly before time $T_A$.
- these rows $G_A(x,\cdot)$ for $x \notin A$ are therefore completely determined by the restriction of the matrix $G_A$ to $(S-A)\times(S-A)$, as displayed in (10.30), which is just the inverse of the restriction $I-Q_A$ of $I-P$ to $(S-A)\times(S-A)$.
- if $a \in A$ then $G_A(a,y) = \mathbf{1}(a=y)$ for $y \in A$, since the only possible hit of $y$ strictly before $T_A$ is a hit at time 0, and conditioning on $X_1$ gives

$$G_A(a,y) = \mathbf{1}(a=y) + \mathbf{1}(y \notin A)\sum_{x \notin A} P(a,x)G_A(x,y) \qquad (a \in A) \tag{10.39}$$

  So the rows $G_A(a,\cdot)$ for $a \in A$ are easily determined from the rows $G_A(x,\cdot)$ for $x \notin A$.

Consequently, to compute the entire pre-$T_A$ occupation matrix $G_A$ for any non-empty subset $A$ of states of a recurrent Markov chain, the main task is to compute the restriction of $G_A$ to $(S-A)\times(S-A)$ by inverting the restriction of $I-P$ to $(S-A)\times(S-A)$ matrix $(I-Q_A)$.
Observe that if

$$V_A := \min\{n \ge 0 : X_n \in A\} = T_A\mathbf{1}(X_0 \notin A)$$

then

$$\mathbb{P}_x(T_A = V_A) = 1 \qquad (x \notin A).$$

So for initial states $x \notin A$ the the pre-$T$ occupation measures $G_T(x,\cdot)$ are identical for $T = T_A$ and $T = V_A$:

$$G_A(x,\cdot) := G_{T_A}(x,\cdot) = G_{V_A}(x,\cdot) \qquad (x \notin A).$$

However, for $x = a \in A$, instead of (10.39) there is the trivial evaluation $G_{V_A}(a,\cdot) = 0$, which is sometimes convenient as a boundary condition. For instance, this condition fits well with the equations satisfied by the function

$$m_A(x) := \mathbb{E}_x V_A = \begin{cases} 0, & x \in A \\ \mathbb{E}_x T_A = \sum_{y\in S} G_A(x,y) = G_A(x,\cdot)\mathbf{1}, & x \notin A. \end{cases}$$

If $P$ is positive recurrent, then $m_A(x) < \infty$ for all $x$. By conditioning on $X_1$, the function $m(x) = m_A(x)$, regarded as a column vector indexed by states $x \in S$, solves the system of equations

$$m(x) = 1 + \sum_{y\in S} P(x,y)m(y) \text{ with } m(a) = 0 \text{ for } a \in A. \tag{10.40}$$

According to Theorem 1.29 on page 62 of the text, if $S-A$ is finite, this system of equations has unique solution $m(x) = m_A(x)$. With the present assumption that $P$ is irreducible and positive recurrent, this uniqueness can also be shown if $S-A$ is

infinite. Once this mean function $m_A(x)$ is found for starting states $x \notin A$, for a starting state $a \in A$ the mean first return time to $A$ is found, either by summing (10.39) over $y \in S$, or by conditioning on $X_1$:

$$\mathbb{E}_a T_A = 1 + \sum_{x \notin A} P(a,x) m_A(x) \qquad (a \in A). \tag{10.41}$$

Starting from any state $X_0 = x$, as well as the mean first passage times $\mathbb{E}_x T_A$, the distributions and moments of many other functions of the path $(X_n, 0 \le n \le T_A)$ are easily expressed in terms of the pre-$T_A$ occupation matrix $G_A$, using consequences of the Strong Markov Property. Consider for instance, the random count of hits on $y$ before $T_A$

$$N_{yA} := \sum_{n=0}^{T_A - 1} \mathbf{1}(X_n = y)$$

whose mean given $X_0 = x$ is $\mathbb{E}_x N_{yA} = G_A(x,y)$. If $y \in A$ this count is just the constant random variable $\mathbf{1}(x = y)$, either 1 or 0 depending on the starting state $x$. For $y \notin A$ let

$$p_{xyA} := \mathbb{P}_x(N_{yA} > 0) = \frac{G_A(x,y)}{G_A(y,y)}$$

be the probability starting from $x$ of reaching $y$ before $T_A$, where the second equality is due to the Strong Markov Property. Then $N_{yA}$ has the modified geometric distribution

$$\mathbb{P}_x(N_{yA} = 0) = 1 - p_{xyA} \tag{10.42}$$

$$\mathbb{P}_x(N_{yA} = n) = p_{xyA}(1 - e_{yA})^{n-1} e_{yA} \qquad (n = 1, 2, \ldots). \tag{10.43}$$

where the escape probability

$$e_{yA} = \mathbb{P}_y(N_{yA} = 1) = \mathbb{P}_y(T_A < T_y) = \frac{1}{G_A(y,y)}$$

is determined by the mean $G_A(y,y)$ of the $\mathbb{P}_y$ distribution of $N_{yA}$, which is geometric $(e_{yA})$ on $\{1, 2, \ldots\}$. Thus for $y \notin A$, due to the Strong Markov Property

- for each starting state $x$ with $p_{xyA} > 0$ the $\mathbb{P}_x$ distribution of $N_{yA}$ given $(N_{yA} > 0)$ does not depend on $x$, and is identical to the $\mathbb{P}_y$ distribution of $N_{yA}$ which is geometric $(e_{yA})$ on $\{1, 2, \ldots\}$.

These formulas for the distribution of $N_{yA}$ starting from varous states $x$ are just variations of the results presented in Durrett's text for $T = \infty$ instead of $T = T_A$ in formula (1.5) on page 18 and Lemma 1.11 on page 19. These formulas show that the path of a recurrent Markov chain stopped when it first hits some set of states $A$ is full of counting variables $N_{yA}$ with modified geometric distributions, whose parameters can be read from the pre-$T_A$ occupation matrix $G_A$.
This analysis is of particular interest when $A = \{a\}$ for a single state $a$. Then the pre-$T_a$ occupation matrix

$$G_a(x,y) = \mathbb{E}_x N_{ya} = \mathbb{E}_x \sum_{k=0}^{T_a - 1} \mathbf{1}(X_n = y) \tag{10.44}$$

gives the expected number of hits on $y$ before $T_a$, starting from any state $x$. However, if there are for instance a finite number $N$ of states, for each target state $a$ a different $(N-1)\times(N-1)$ submatrix of $I-P$ must be inverted to obtain the matrix $G_a$. Or if it is desired to obtain the full matrix of mean first passage times $(m_{xa}, x, a \in S)$, a corresponding system of $N-1$ linear equations (10.40) must be solved to obtain the $m_{xa}$ for $x \neq a$, and thence $m_{aa} = 1 + \sum_{x\neq a} P(a,x)m_{xa}$. But to do this for all states $a \in S$ involves repeating this computational process $N$ times over.

### 10.4.5 The fundamental matrix of a positive recurrent chain

There is a much more efficient way to evaluate all the Green matrices $G_a(x,y)$ and first passage moments $m_{xa} = G_a(x,\cdot)\mathbf{1}$ of a positive recurrent Markov chain with transition matrix $P$. This involves the following three step process:

- Step 1. Find the invariant probability vector $\pi$. This can be done more or less quickly, depending on the structure of $P$. Before getting into linear algebra, it is best to first try the shortcuts for this:
  - check the detailed balance equations to see if there is a reversible equilibrium, as if there is it will be easy to find.
  - check if $P$ is doubly stochastic, in which case the uniform distribution is invariant.
  - check if there is some family of initial distributions $\lambda$ for which it is particularly easy to find $\lambda P$, and look for $\pi$ amongst these initial distributions (as in Problem 2 of Worksheet 5).

  Worst case, if none of these shortcuts works,
  - either solve the system of equations $\pi P = \pi$ and $\pi\mathbf{1} = 1$ by linear algebra,
  - or compute the pre-$T_a$ occupation kernel $G_a$ for some fixed state $a$ by linear algebra, and evaluate $\pi(x) = G_a(a,x)/G_a(a,\cdot)\mathbf{1}$.
- Step 2. Set $\Pi := \mathbf{1}\pi$, and invert $I - P + \Pi$ to obtain the *fundamental matrix*

$$Z := (I - P + \Pi)^{-1} \tag{10.45}$$

- Step 3. Read the Green matrices and first passage moments from the simple formulas for all these quantities in terms of entries of $Z$, as detailed below.

The method of Steps 2 amd 3 will now be motivated by studying the relation between of the matrix $I - P + \Pi$ and the Green matrix $G_a$ for some arbitrary fixed target state $a$, assuming for simplicity that the state space $S$ is finite. But it is known that all the formulas obtained in this case are valid for any irreducible and positive recurrent chain, just with finite sums replaced by infinite sums, with some care to ensure the infinite sums are convergent. Further motivation for the study of the fundamental matrix $Z$ derived from an irreducible positive recurrent Markov matrix $P$ is provided by the fact that $Z$ is involved in evaluating the variance of the sum

$\sum_{k=1}^{n} f(X_k)$ appearing in the ergodic theorem (10.33), which is needed to provide normal approximations to the distribution of this sum for large $n$. Here is a more formal statement of Step 2.

**Theorem 1** (Existence of the fundamental matrix $Z$)**.** *Let $P$ be an irreducible transition matrix with finite state space $S$. Let $I$ be the $S \times S$ identity matrix, let $\pi$ with*

$$\pi P = \pi \text{ and } \pi \mathbf{1} = 1$$

*be the unique invariant probability measure for $P$, and let $\Pi := \mathbf{1}\pi$ be the $S \times S$ transition matrix with all rows equal to $\Pi$. Then*

1. *The matrix $I - P + \Pi$ is invertible, with inverse $Z$ that is uniquely determined by either one of the two matrix equations*

$$(I - P + \Pi)Z = I = Z(I - P + \Pi) \tag{10.46}$$

2. *This matrix $Z$ shares with $P$ the basic properties*

$$\pi Z = \pi, \qquad Z\mathbf{1} = \mathbf{1}, \qquad \text{hence} \qquad \Pi Z = Z\Pi = \Pi \tag{10.47}$$

3. *There is the identity $(I - P)^n = P^n - I$ for $n \geq 1$, but not for $n = 0$. If $P$ is aperiodic, then*

$$Z = \sum_{n=0}^{\infty} (P - I)^n = I + \sum_{n=1}^{\infty} (P^n - \Pi) \tag{10.48}$$

   *meaning that if $Z_N := \sum_{n=0}^{N} (I - P)^n$ is the matrix of partial sums up to $N$, then $Z(x, y) = \lim_{N \to \infty} Z_N(x, y)$ for all $x, y \in S$.*

4. *If $P$ is periodic, the parital sums $Z_N$ of the series* (10.48) *series are not convergent. But whatever the period of $P$, the sum can be evaluated by Abel's method:*

$$Z = \lim_{s \uparrow 1} \sum_{n=0}^{\infty} (I - P)^n s^n = I + \lim_{s \uparrow 1} \sum_{n=1}^{\infty} (P^n - \Pi) s^n \tag{10.49}$$

*Proof.* By linear algebra, a square matrix $M$ is invertible iff there is no non-trivial linear relation among its column vectors: that is, the equation $Mg = 0$ for an unknown column vector $g$ has unique solution $g = 0$. So to establish the existence of $Z$ it is enough to show that the equation

$$(I - P + \Pi)g = 0$$

for an unknown column vector $g$ implies $g = 0$. Pre-multiply this equation by the row vector $\pi$ to see it implies $\pi g = 0$, hence $\Pi g = 0$, and $(I - P)g = 0$. So $g$ is a $P$-harmonic column vector with

$$g = Pg = P^n g \qquad (n = 1, 2, \ldots)$$

But for an irreducible $P$ with finite state space, every $P$-harmonic $g$ is constant: $g = m\mathbf{1}$ for some scalar multiplier $m$, and $\pi g = 0$ forces $m = 0$, hence $g = 0$. To confirm that $g = m\mathbf{1}$ for some $m$, let $m := \min_{y \in S} g(y)$ and let $x$ be a state with $g(x) = m$. Then

$$0 = g(x) - m = (P^n[g - m\mathbf{1}])(x) = \sum_y P^n(x,y)[g(y) - m] \geq 0.$$

Hence $P^n(x,y)[g(y) - m] = 0$ for every state $y$. By irreducibility of $P$, there is an $n$ such that $P^n(x,y) > 0$, hence $g(y) = m$. So $g = m\mathbf{1}$. This proves the first assertion that $I - P + \Pi$ is invertible. The remaining assertions are then easily checked, using $\Pi P = P\Pi = \Pi$ to show that any finite product of factors $P$ and $\Pi$ containing at least one $\Pi$ reduces to $\Pi$, and hence $(P - I)^n = P^n - \Pi$ for $n \geq 1$ by using the binomial expansion of $0 = (1 - 1)^n$ to identify the coefficient of $\Pi$ in the expansion of $(P - I)^n$. In the aperiodic case, the difference $|(P^n - \Pi)(x,y)|$ is bounded by some constant time $\rho^n$ with $0 < \rho < 1$ by the estimate in Durrett's proof of (10.36) on page 53. This implies the series in (10.48) is absolutely convergent, hence that $Z$ defined by (10.48) satisfies (10.46). The discussion of the Abel sum (10.49) in the periodic case just requires a bit more analysis. □

**Corollary 1.1** (Solution of Poisson equations)**.** *For a finite state irreducible transition probability matrix $P$, with $\Pi = \pi\mathbf{1}$ and $Z = (I - P + \Pi)^{-1}$ as above,*

*(a)* [Poisson equation for a column vector] *For each given column vector $f$*

$$\text{there exists a solution } g \text{ of } (I - P)g = f \qquad \Longleftrightarrow \qquad \pi f = 0 \tag{10.50}$$

*in which case, for each scalar $m$*

$$\text{the unique solution } g \text{ of } (I - P)g = f \text{ and } \pi g = m \text{ is } g = Zf + m\mathbf{1}. \tag{10.51}$$

*(b)* [Poisson equation for a row vector] *For each given row vector $\delta$*

$$\text{there exists a solution } \mu \text{ of } \mu(I - P) = \delta \qquad \Longleftrightarrow \qquad \delta\mathbf{1} = 0 \tag{10.52}$$

*in which case, for each scalar $m$*

$$\text{the unique solution } \mu \text{ of } \mu(I - P) = \delta \text{ and } \mu\mathbf{1} = m \text{ is } \mu = \delta Z + m\pi. \tag{10.53}$$

*Proof.* Dealing with the case of row vectors, suppose that $\mu$ is a solution of the Poisson equation $\mu(I - P) = \delta$ for some given vector $\delta$. Post-multiply the Poisson equation by $\mathbf{1}$ to see that $\delta\mathbf{1} = 0$ is necessary for existence of any solution $\mu$. Continuing to assume that $\mu(I - P) = \delta$, let $m := \mu\mathbf{1}$, so $(\mu - m\pi)\Pi = 0$ and the Poisson equation gives $(\mu - m\pi)(I - P) = \delta$. Add these two equations to get

$$(\mu - m\pi)(I - P + \Pi) = \delta \qquad \text{hence} \qquad \mu - m\pi = \delta Z$$

by post-multiplication by $(I-P+\Pi)^{-1} = Z$. Finally, these steps are easily reversed to show for every $\delta$ with $\delta\mathbf{1} = 0$ and every scalar $m$ that $\mu := \delta Z + m\pi$ solves $\mu(I-P) = \delta$ and $\mu\mathbf{1} = m$. The corresponding result for column vectors is obtained by a dual argument, using pre-multiplication of the Poisson equation $(I-P)g = f$ by $\pi$ to obtain the necessary condition $\pi f = 0$. □

Note the striking duality exposed by the above analysis of the two Poisson equations

$$(1-P)g = f \qquad \text{and } \mu(I-P) = \delta$$

according to whether the transition matrix $P$ is regarded as an operator on the left of column vectors $g$ or on the right of row vectors $\mu$. For each action of $P$, the Poisson equation with 0 on the right side is solved by any $P$-invariant vector:

- the only $P$-invariant column vectors are the constant function $g = m\mathbf{1}$ for some scalar $m$;
- the only $P$-invariant row vectors are scalar multiples $\mu = m\pi$ of the invariant probability $\pi$.

In each case, the solution of the Poisson equation is clearly unique only up to addition of $P$-invariant vector, that is some multiple of $\mathbf{1}$ or of $\pi$ as the case may be. The key points are that for each given column vector $f$ or row vector $\delta$ on the right side of the Poisson equation, subject to the obvious necessary condition for existence of a solution (that is $\pi f = 0$ or $\delta\mathbf{1} = 0$),

- a specific solution of the Poisson equation is obtained as either $Zf$ or $\delta Z$;
- the most general solution of the Poisson equation is then either $Zf + m\mathbf{1}$ or $\delta Z + m\pi$, as the case may be.

**Corollary 1.2** (Occupation measures derived from the fundamental matrix)**.** *For an irreducible Markov chain with transition matrix $P$ with invariant probability $\pi$ and fundamental matrix $Z = (I-P+\Pi)^{-1}$, let $T$ be any randomized stopping time of the chain, and for a given initial distribution $\lambda$ let $\lambda G_T$ be the pre-$T$ occupation measure*

$$\lambda G_T(z) := \mathbb{E}_\lambda \sum_{n=0}^{T-1} 1(X_n = z) \qquad (z \in S) \tag{10.54}$$

*regarded as a row vector, with total mass $\lambda G_T\mathbf{1} = \mathbb{E}_\lambda T$. Assuming $\mathbb{E}_\lambda T < \infty$, the occupation measure $\lambda G_T$ is related to the initial distribution $\lambda$ and the final distribution $\lambda_T$ of $X_T$ by the Poisson equation*

$$\lambda G_T(I-P) = \lambda - \lambda_T \tag{10.55}$$

*whose solution is*

$$\lambda G_T = (\lambda - \lambda_T)Z + (\mathbb{E}_\lambda T)\pi \tag{10.56}$$

*Proof.* The Poisson equation (10.55) for $\lambda G_T$ is read from the general occupation measure identity (10.26), using $\mathbb{E}_\lambda T < \infty$ to justify finiteness of the measure $\lambda G_T$. Then (10.56) is read from the general form (10.53) of the solution of the Poisson equation. □

For the stopping time $T = T_y := \min\{n \geq 1 : X_n = y\}$ this formula (10.56) establishes a close connection between the fundamental matrix $Z$ and the pre-$T_y$ occupation matrix $G_y$ with $(x, z)$ entry

$$G_y(x, z) := \mathbb{E}_x N_{zy} \text{ where } N_{zy} := \sum_{n=0}^{T_y - 1} \mathbf{1}(X_n = z)$$

is the number of hits on $z$ strictly before $T_y$, counting a hit at time $n = 0$ if $x = z$. Abbreviate

$$m_{xy} := \mathbb{E}_x T_y = \sum_z G_y(x, z) = G_y(x, \cdot)\mathbf{1}$$

for the mean first passage time from $x$ to $y$, and

$$m_{\lambda y} := \mathbb{E}_\lambda T_y = \sum_x \lambda(x) m_{xy} = \sum_z (\lambda G_y)(z) = \lambda G_y \mathbf{1}$$

for the mean first passage time to state $y$ starting from $X_0 \sim \lambda$. Formula (10.56) in this case reads

$$\lambda G_y(\cdot) = \lambda Z(\cdot) - Z(y, \cdot) + m_{\lambda y}\pi(\cdot) \tag{10.57}$$

The choice of the stationary initial distribution $\lambda = \pi$ is of special interest. The feature of $Z$ noted in (10.47) that $\pi Z = \pi$ makes (10.57) for $\lambda = \pi$ simplify to

$$\pi G_y(\cdot) = -Z(y, \cdot) + (1 + m_{\pi y})\pi(\cdot) \tag{10.58}$$

which rearranges as

$$Z(y, \cdot) = (1 + \pi G_y \mathbf{1})\pi(\cdot) - \pi G_y(\cdot). \tag{10.59}$$

Thus the fundamental matrix $Z = (I - P + \Pi)^{-1}$ can be described in probabilistic terms as follows:

- row $y$ of $Z$ is a scalar multiple $m\pi$ of the stationary probability $\pi$, minus the pre-$T_y$ occupation measure for the chain started with $X_0 \sim \pi$, where $m = 1 + \mathbb{E}_\pi T_y = 1 + \pi G_y \mathbf{1}$ is determined by the row sum $Z(y, \cdot)\mathbf{1} = 1$.

In particular, evaluating the row vector (10.59) at $y$ gives the diagonal entries of the fundamental matrix $Z$ in terms of $\pi$ and mean first moments:

$$Z(y, y) = (\mathbb{E}_\pi T_y)\pi(y) = \sum_x \pi(x) m_{xy} \pi(y) = \frac{\mathbb{E}_y T_y^2 + m_{yy}}{2m_{yy}^2} \tag{10.60}$$

where $m_{yy} = \mathbb{E}_y T_y$ and the last equality is due to Kac's identity (Problem 3 of Worksheet 5)

$$\mathbb{P}_\pi(T_y = n) = \mathbb{P}_y(T_y \geq n)/m_{yy} \qquad (n = 1, 2, \ldots) \tag{10.61}$$

which shows how the distribution of $T_y$ for $X_0 \sim \pi$ determines the distribution of $T_y$ given $X_0 = y$ and vice versa. As a checks on these formulas: summing (10.61) over $n$ gives 1 on the left side by the assumed recurrence of the chain, and 1 on the right side by the tail sum formula for $\mathbb{E}_y T_y$. Also, (10.60) can be checked by renewal theory, using the infinite sum (10.48) and probability generating functions, as indicated in Problem 5 of Worksheet 5.
Similarly, taking $\lambda = \delta_x$ in (10.57) gives

$$Z(x,y) - Z(y,y) = \mathbf{1}(x = y) - m_{xy}\pi(y). \tag{10.62}$$

Hence, by adding (10.60) and (10.63), there is a general formula for entries of $Z$ in terms of mean first passage times:

$$Z(x,y) = \mathbf{1}(x = y) + (m_{\pi y} - m_{xy})\pi(y). \tag{10.63}$$

The case $x = y$ of (10.63) reduces to (10.60) by $\pi(x) = 1/m_{xx}$, which allows the diagonal mean first passage times $m_{xx}$ to be determined from $\pi$. Subtract (10.63) from (10.60) to see that similarly, all the off diagonal mean first passage times can be read from the fundamental matrix $Z$ and $\pi$:

$$m_{xy} = \frac{Z(y,y) - Z(x,y)}{\pi(y)} \qquad (x \neq y) \tag{10.64}$$

As a check on these formulas, the simple special case that $P$ governs a sequence of i.i.d. random variables with distribution $\pi$ corresponds to

$$P = \Pi \iff Z = I \iff m_{xy} = m_{yy} \text{ for all } x. \tag{10.65}$$

### 10.4.6 Exercises.

1. Deduce that

$$G_y(x,z) = (m_{xy} - m_{xz}1(x \neq z) + m_{yz}1(y \neq z))\pi(z). \tag{10.66}$$

This can be rearranged as

$$m_{xy} + m_{yz}1(y \neq z) = G_y(x,z)m_{zz} + m_{xz}1(x \neq z) \tag{10.67}$$

which is just two different ways of calculating $\mathbb{E}_x T_{yz}$ where $T_{yz}$ is the first visit to $z$ at or after time $T_y$. See Chapter 2 of the Aldous-Fill book `https://www.stat.berkeley.edu/users/aldous/RWG/book.html` for further discussion.

2. Check directly as follows, without discussion of the Poisson equation, that $Z$ defined by (10.63) is the inverse of $I - P + \Pi$. Recall that by a first step analysis, for all pairs of states $x$ and $z$, including $x = z$,

$$m_{xz} = 1 + \sum_{y \neq z} P(x,y)m_{yz} \tag{10.68}$$

For $Z$ defined by (10.63), use (10.68) to evaluate the $(PZ)(x,z)$, and show that $PZ = Z - I + \Pi$. Check also that $\Pi Z = \Pi$, and hence that $Z(I - P + \Pi) = I$.

3. Evaluate all entries of $Z$ explicitly for a two state chain.

4. Evaluate one row of $Z$ explicitly for a random walk on three states in a circle, with probabilities $p$ and $q$ for clockwise and anticlockwise steps, with $p+q=1$. How can the other rows of $Z$ be derived from this row with no further calculation?

5. (Asmussen Proposition 7.1 [4]) Show that for each $f$ with $\pi f = 0$ the unique solution $g$ of the Poisson equation $(I-P)g = f$ with $g(y) = 0$ is

$$g(x) = \delta_x G_y f = \mathbb{E}_x \sum_{n=0}^{T_y - 1} f(X_n)$$

6. Is there a corresponding description of the solution of $\mu(I-P) = \delta$ with $\delta \mathbf{1} = 1$ and $\delta(y) = 0$?

## 10.5 References

The material on pre-$T$ occupation measures is based on my *Occupation measures for Markov chains* Adv. Appl. Probab. 9, 69-86 (1977). See Asmussen's *Applied probability and queues* (§I.7[4]) and Chapter 2 of the Aldous-Fill book `https://www.stat.berkeley.edu/users/aldous/RWG/book.html` for further discussion of the fundamental matrix $Z$. Beware that Aldous-Fill use the notation $Z$ for $Z - \Pi$, and $T_x^+$ for our $T_x$ and $T_x$ for our $V_x$.

# LECTURE 11

# Poisson Processes, Part 1

## 11.1 Introduction: Poisson Processes

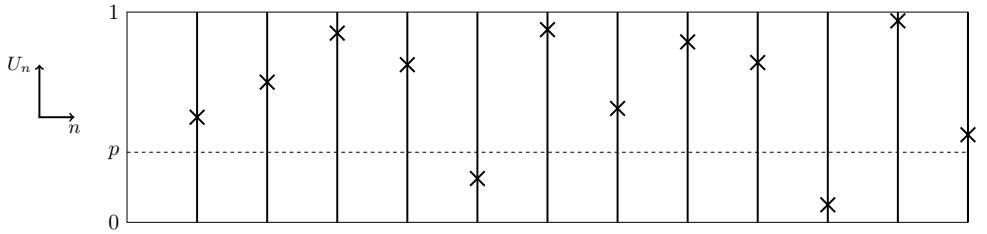
We'll do a quick review of basic properties of the Poisson distribution. Recall that we have the **Binomial** $(n, p)$distribution of

$$S_n(p) := X_1(p) + X_2(p) + \cdots + X_n(p),$$

where the $X_i(p)$ are independent **Bernoulli**$(p)$ variables. We have a nice construction of all these variables at once. Take $U_1, U_2, \ldots$ iid over interval $[0, 1]$ and let

$$X_n(p) := \mathbf{1}\{U_n \leq p\}$$

Observe




| $n$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_n(p)$ | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |

Now we look at when $n \to \infty$, with $p \downarrow 0$, so that $np \equiv \mu$ is fixed. In terms of the above construction, we are simply lowering the dotted line at level $p$, and pushing $n = \mu/p$ to infinity (through integer values, or round $\mu/p$ to the nearest integer, it makes no difference in the limit), to keep the expected number of points $\times$ below the bar in $n$ trials to equal $\mu$, either exactly for simplicity as in the following calculations,

or approximately enough so that $np \to \mu$ as both $n \to \infty$ and $p \downarrow 0$. Working with $np \equiv \mu$ exactly

$$\mathbb{E}S_n(p) = np \equiv \mu$$

so that

$$\begin{aligned}\mathbb{P}(S_n(p) = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{(n)_{k\downarrow}}{k!} \left(\frac{\mu}{n}\right)^k \left(1 - \frac{\mu}{n}\right)^{n-k} \\ &\to e^{-\mu} \frac{\mu^k}{k!} \text{ as } n \to \infty \text{ and } p \downarrow 0\end{aligned}$$

This is how the Poisson distribution arises as the limit of binomial distributions with large $n$ and small $p$. A more careful argument (see text Theorem 2.9 on page 104) shows that for any collection of independent indicator variables $X_1, \ldots, X_n$ with $p_k = \mathbb{P}(X_k = 1)$ with sum $S_n$ so $\mathbb{E}S_n = p_1 + \cdots + p_n\mu$, the **Poisson** $(\mu)$ distribution will be a good approximation to the distribution of $S_n$ whenever $\mu(\max_k p_k)$ is small. In fact the total variation distance between this distribution of $S_n$ (depending on $p_1, \ldots, p_n$) and **Poisson** $(\mu)$ is at most $\mu(\max_k p_k)$, which for any fixed $\mu$ is small provided all the $p_k$ are sufficiently small. Now when $\mu \geq 0$, let $N_\mu$ or $N(\mu)$ denote a random variable with this **Poisson** $(\mu)$ limit law

$$\mathbb{P}(N_\mu = k) = e^{-\mu} \frac{\mu^k}{k!} \quad k = 0, 1, 2, \ldots$$

Some basics

$$\mathbb{E}N_\mu = \mu \quad \text{and} \quad \mathbb{V}\text{ar}(N_\mu) = \mu$$

Notice that in the **Binomial** $(n, p)$ setup

$$\mathbb{V}\text{ar}(S_n(p)) = np(1-p) = \mu(1-p) \to \mu$$

as $p \downarrow 0$. So the variance for Poisson is the limit of binomial variances, as should be expected. You should check these moment formulas by summation. Additionally, by probability generating functions, we have

$$G_{N(\mu)}(z) := \mathbb{E}z^{N(\mu)} = \sum_{n=0}^{\infty} z^n e^{-\mu} \frac{\mu^n}{n!} = e^{-\mu} e^{\mu z} = e^{-\mu(1-z)}$$

Take the derivatives $\frac{d}{dz}, \frac{d^2}{dz^2}$ at $z = 1$. Gives us $\mathbb{E}N_\mu$ and $\mathbb{E}N_\mu(N_\mu - 1)$, hence the formula for variance.

**Exercise** There is a pretty formula for $\mathbb{E}\binom{N_\mu}{k}$. Find it and prove it.

## 11.2 Sum of Independent Poissons

Take $N_1, N_2, \ldots, N_m$ independent Poissons with means $\mu_1, \mu_2, \ldots, \mu_m$. Then

$$N_1 + \cdots + N_m \sim \textbf{Poisson}\,(\mu_1 + \cdots + \mu_m)$$

This is intuitively clear as a limit of the corresponding result for sums of independent binomials with parameters $(n_i, p)$ with $n_i p \equiv \mu_i$ and $p \downarrow 0$. Alternatively, we can easily derive it with probability generating functions. Or by induction on $m$, from the case $m = 2$, using the convolution formula and the binomial theorem to evaluate

$$\mathbb{P}(N_1 + N_2 = n) = \sum_{k=0}^{n} \mathbb{P}(N_1 = k)\mathbb{P}(N_2 = n - k)$$

## 11.3 Poissonization of the Multinomial

What is the distribution of the following random vector?

$$\Big((N_1, N_2, \ldots, N_m) \,|\, N_1 + N_2 + \cdots + N_m = n\Big)$$

for independent Poisson$(\mu_i)$ variables $N_i$? For $m = 2$, the last computation suggested above shows as a byproduct that

$$(N_1 \mid N_2 = n) \stackrel{d}{=} \textbf{Binomial}\,(n, p) \text{ for } p = \mu_1/(\mu_1 + \mu_2).$$

For general $m$, we can compute the above conditional distribution using Bayes rule. We should perform this computation once in our lives, as this is the basis of the entire theory of Poisson processes. It is only of interest to consider $n_1, \ldots, n_k$ with $n_1 + \cdots + n_m = n$, the given value for the sum. So consider

$$\begin{aligned}
&\mathbb{P}(N_1 = n_1, N_2 = n_2, \ldots, N_m = n_m \mid N_1 + N_2 + \cdots + N_m = n) \\
&\qquad = \frac{\mathbb{P}(N_1 = n_1, N_2 = n_2, \ldots, N_m = n_m)}{\mathbb{P}(N_1 + \cdots + N_m = n)} \\
&\qquad = \frac{\dfrac{e^{-\mu_1}\mu_1^{n_1}}{n_1!} \cdots \dfrac{e^{-\mu_m}\mu_m^{n_m}}{n_m!}}{\dfrac{e^{-(\mu_1+\cdots+\mu_m)}(\mu_1 + \cdots + \mu_m)^{n_1+\cdots+n_m}}{(n_1 + \cdots + n_m)!}} \\
&\qquad = \frac{n!}{n_1! \cdots n_m!}\, p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m}
\end{aligned}$$

Notice how the exponentials cancel across the numerator and denominator, with $\mu = \mu_1 + \cdots + \mu_m$ and $p_k = \frac{\mu_k}{\mu}$. We recognize this as the familiar *multinomial* distribution. Hence for $N = N_1 + \cdots + N_m$ as before, we have

$$(N_1, \ldots, N_m \,|\, N = n) \stackrel{d}{=} \textbf{Multinomial}(n, p_1, p_2, \ldots, p_m)$$

In words, any vector of independent Poisson counts $N_1, \ldots, N_m$ conditioned on the total count equal to $n$ is distributed like the counts of values with probabilities $(p_1, \ldots, p_m)$ in $n$ Multinomial trials. To state the conclusion more formally:

**Poissonization of the Multinomial**

The following two conditions on a random vector of counts $N_1, \ldots, N_m$ are equivalent:

(1) the $N_1, \ldots, N_m$ are independent Poissons with means $\mu_1, \ldots, \mu_m$.

(2) $N_1 + \cdots + N_m$ is **Poisson** $(\mu_1 + \cdots + \mu_m)$ and given $N_1 + \cdots + N_m = n$, the vector $(N_1, \ldots, N_m)$ is multinomial$(n, p_1, \ldots, p_m)$ with $p_k = \frac{\mu_k}{\mu_1 + \cdots + \mu_m}$.

This statements is both important and easy to check once formulated. It is not obvious at first, but it becomes very familiar and quite intuitive as you work with Poisson processes.

**Corollary**

Suppose we have $Y_1, Y_2, \ldots$ iid with probability distribution

$$\mathbb{P}(Y_i = k) = p_k,$$

for some probability distribution $(p_1, \ldots, p_m)$ on $\{1, \ldots, m\}$. Assume that $N$ is independent of $Y_1, Y_2, \ldots$ and $N \sim$ **Poisson** $(\mu)$. Define

$$N_k := \sum_{i=1}^{N} \mathbb{1}(Y_i = k),$$

which is the number of $Y$ values equal to $k$ in the $N$ trials (hence equal to 0 if $N = 0$).
Then $N_1, N_2, \ldots, N_m$ are independent **Poisson** $(\mu_1, \ldots, \mu_m)$.

*Proof.* By constuction, $N_1 + N_2 + \cdots + N_m = N$, and given $N = n$, the $(N_1, \ldots, N_m)$ are multinomial$(n, p_1, \ldots, p_k)$. Now plug into the theorem with $\mu + k = p_k \mu$ so that $\mu_1 + \cdots \mu_m = \mu$. □

**Remark** It is not really necessary to have an infinite iid sequence $Y_i$ to construct counts $N_k$ as above. All that is needed is a Poisson variable $N$, and for each $n \geq 1$, conditionally given $N = n$ a sequence $Y_1, \ldots, Y_n$ of iid variables with the required distribution $(p_1, \ldots, p_m)$. Then the same conclusion holds, for the same reasons.
To summarize, if in multinomial trials, we randomize the number of trials $N$ with a Poisson distribution, we make the counts independent. This is highly non obvious at first, but this is a key fact which will be exploited heavily for Poisson Processes.

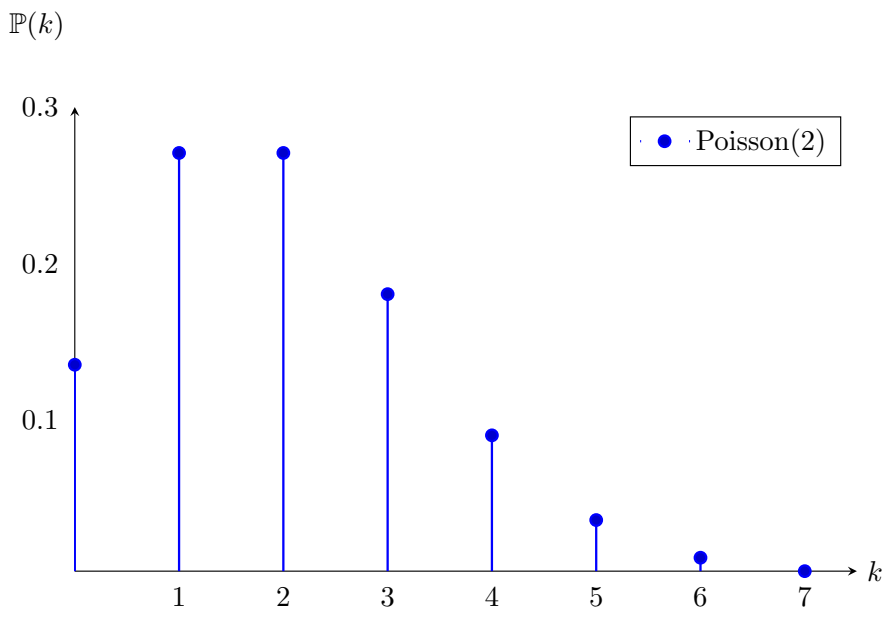
See the text Theorem 2.1 on page 108, Theorem 2.12 on page 110 and Theorem 2.15 on page 115 for several slight variations on this theme, expressed in terms of Poisson processes. The underlying distributional relation that makes all these results work is the Poissonization of the multinomial.

## 11.4 Poisson Point Processes (PPP)

We'll presume that we've seen the Poisson processes on a line. It is more interesting and intuitive to start looking at a PPP in a strip of the plane. The setting is as follows

- Let $\Delta_i \sim \textbf{Poisson}\,(\lambda)$ be the Poisson count for box $i$ with intensity $\lambda > 0$, some fixed rate per unit area.
- If $N_t$ is the number of points in $[0, t] \times [0, 1]$, then it follows
- $N_t = \Delta_1 + \Delta_2 + \ldots + \Delta_t \sim \textbf{Poisson}\,(\lambda t)$ for $t \in \mathbb{Z}^+$.

We can obtain insight for the likely number of counts for $\Delta_i \sim \textbf{Poisson}\,(2)$, by looking at its histogram. Take note of the the fact that for a random variable with distribution $\textbf{Poisson}\,(m)$, the values with highest probability occur at $m$ and $m-1$.




With this in mind, given $\Delta_i = n$, we throw down $n$ iid points with uniform probability in the $i^{\text{th}}$ square $[i, i+1] \times [0, 1]$. Study the following diagrams.

### 11.4.1 PPP Strips

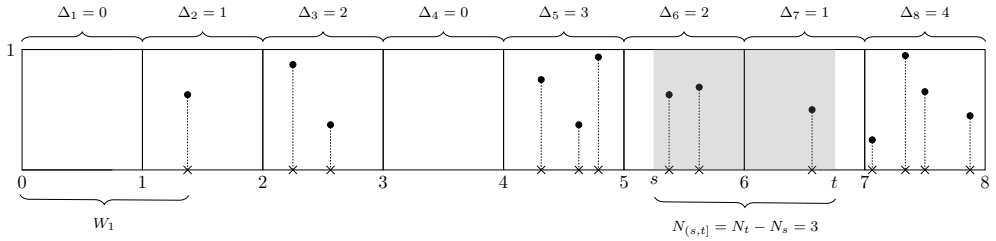



Figure 11.1: We project each "dot" onto the horizontal axis. This exploits two things: (1) the time between arrivals, the first shown below the strip, denoted $W_1$ and (2) the counts of arrivals between an interval say $(s, t]$. Recall that $N_t$ denoted the counts before and including time $t$. Hence for $N_{(s,t]}$ we simply take the difference between $N_t$ and $N_s$.

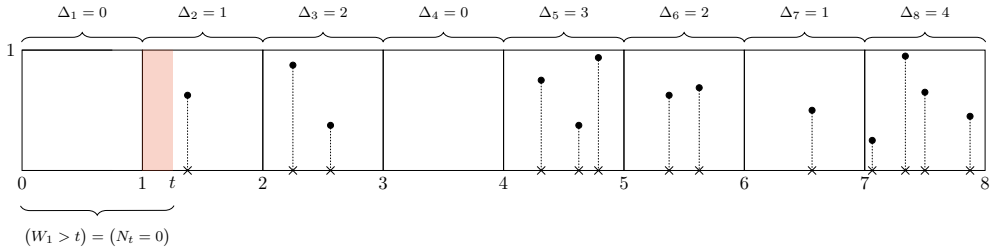



Figure 11.2: Notion of duality. In particular, the first hit arriving after time $t$, the event $W_t > 0$ is the same as $N_t = 0$.

For example, take independent uniform $[0, 1]$ variables $X_{i,j}$ and $Y_{i,j}$, independent of all the $N_i$, and for each $i = 1, 2, \ldots$ place points at

$$(i - 1 + X_{i,j}, Y_{i,j}) \text{ for } 1 \leq i \leq N_{i,j}$$

with of course no points at all in the $i$th square if $N_i = 0$
Now Pitman asks what happens if we project this strip down onto a line. Notice that the probability of projecting two or more points onto the same point on the line is 0. Or in other words, there are no multiple points (repeated $X$-values) in the strip. Let $W_1, W_2, W_3, \ldots$ be the spacings between points along the $X$-axis.

**Case:** $0 < t < 1$

Notice $\mathbb{P}(W_1 > t)$ is simply the probability that there there are no points to the left of $t$.

$$\mathbb{P}(W_1 > t) = \mathbb{P}(N_{\lambda t} = 0) = e^{-\lambda t},$$

by design and the Poissonization of the binomial.

**Case:** $1 < t < 2$

Here, we use independence to have:

$$\begin{aligned}\mathbb{P}(W_i > t) &= \mathbb{P}(N_1 = 0 \text{ and count in } [1,t] \times [0,1] = 0) \\ &= \mathbb{P}(N_1 = 0)\mathbb{P}(\text{count in } [1,t] \times [0,1] = 0) \\ &= e^{-\lambda} e^{-\lambda(t-1)} \\ &= e^{-\lambda t}.\end{aligned}$$

This result makes us very happy, and we claim this can be continued inductively for arbitrary $t > 0$.

### Repeating this discussion for counts

Let $N_t$ be the number of points to the left or equal to $t$. Then we simply have:

$$N_t \sim \textbf{Poisson}\,(\lambda t).$$

This is obvious for integer $t = 1, 2, \ldots$, and true also for non-integer $t > 0$ by design and Poissonization of the binomial. Now consider $0 \le s \le t$. Observe that $N_t - N_s$ is the number of points in $(s, t]$. If $s, t$ are integers, it is obvious that $N_t - N_s$ is **Poisson** $(\lambda(t-s))$ independent of $N_s$. Now if they are not integers, this still works via the Poissonization of the binomials involved in their fractional parts. That is, $N_t - N_s$ is the number of points in $(s, t]$. We have that

$$N_t - N_s \sim \textbf{Poisson}\,(\lambda(t-s)),$$

and notice that this count is **independent** of the $N_s$. Continuing this, we recover the usual definition of the Poisson point distribution on the half-line from 0 to $\infty$. (Text, page 101). Thus we see that the construction of a PPP in the strip, just indicated, projects to a PPP with constant rate $\lambda$ on $[0, \infty)$, just be ignoring the $Y$-values. Of course, it was not necessary even to involve the $Y$-values, but the pictures are nicer with both $X$ and $Y$ values, and this construction also serves to explain many further properties of Poisson processes (text Section 2.4).
Let $0 \le t_1 < t_2 < t_3 \le \cdots \le t_n$. Then

$$N(t_1), N(t_2) - N(t_1), \ldots, N(t_n) - N(t_{n-1})$$

are independent Poisson with parameters

$$\lambda t_1, \lambda(t_2 - t_1), \ldots, \lambda(t_n - t_{n-1}).$$

**Theorem**

The Poisson finite dimensional distributions indicated above for a counting process $(N(t), t \ge 0)$ are equivalent to spacings $W_1, W_2, \ldots$ iid **Exponential** $(\lambda)$, where $W_1, W_1 + W_2, \ldots$ are the arrival times on the $X$-axis.

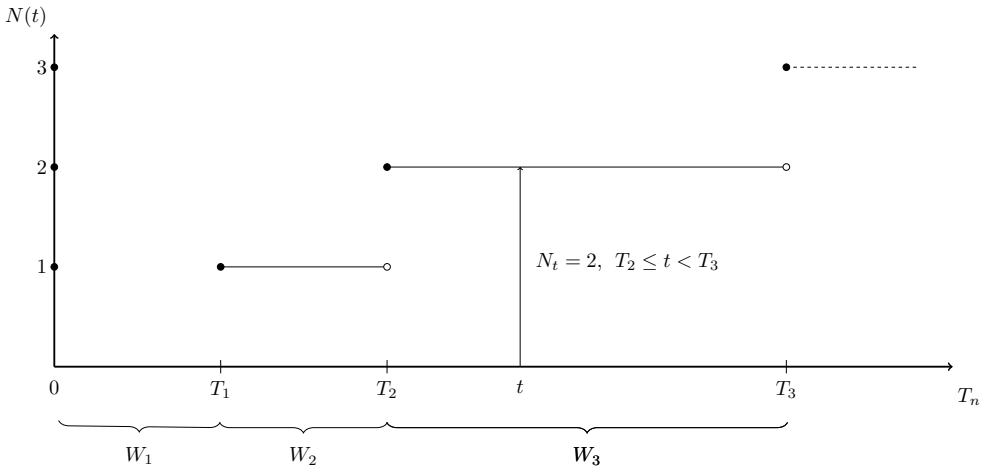



Figure 11.3: Poisson Arrival Process with counting variable $N(t) = N_t$, inter-arrival times $0 < W_1, W_2, \ldots$, and arrivals $T_1 < T_2 < \ldots$

This picture has a continuous time axis and a discrete count (vertical) axis. There are formulas that correspond to the picture. Because $N_t$ is a count, we write it as a sum of indicators:

$$N_t = \sum_{n=1}^{\infty} \mathbb{1}(T_n \le t),$$

which is simply counting the $T_i$ less than or equal to $t$. Then for the inverse direction, we have the attained min:

$$T_n = \min\{t : N_t = n\} = \min\{t : N_t \ge n\},$$

where the second expression is generally true even if we jump over a value. Pitman reminds that these are very important formulas. Now we want to notice a key duality implied by the above definitions. Note that this duality is valid for any counting process $(N_t)$ and $(T_n)$ as above, without making any distributional assumptions. It applies for instance also to counting processes, where $T_n = W_1 + \cdots + W_n$ for $W_n$ that might not be independent or not identically distributed.

**Key Duality**

For counting process $N_t$ and arrival times $T_n$

$$(T_n > t) = (N_t < n)$$

Equivalently,

$$(T_n \le t) = (N_t \ge n).$$

To check this, we logically check implication in both ways. In fact, $t \to N_t$ by definition is increasing ($\uparrow$) and right-continuous, and no missing value (that is to say no repeated points).

## 11.5 Applications

$$\begin{aligned}\mathbb{P}(T_n \le t) &= \mathbb{P}(N_t \ge n) \\ &= 1 - \mathbb{P}(N_t < n) \\ &= 1 - \sum_{k=0}^{n-1} e^{-\lambda t}\frac{(\lambda t)^k}{k!}.\end{aligned}$$

Pitman asks us to find the density, given the CDF. To do so, we differentiate. This gives

$$\begin{aligned}f_{T_n}(t) &= \frac{d}{dt}\mathbb{P}(T_n \le t) \\ &= \frac{d}{dt}\left(1 - \sum_{k=0}^{n-1} e^{-\lambda t}\frac{(\lambda t)^k}{k!}\right),\end{aligned}$$

which gives a telescoping sum. After evaluating, we get a gamma distribution, $\textsf{Gamma}(n, \lambda)$.

## 11.6 Secret Method

Pitman gives a secret method so that we are sure to get this right (no book will tell us this). This is much better than performing the telescoping sum.
Look at an infinitesimal interval $[t, t+dt]$, so that

$$\begin{aligned}\mathbb{P}(t \le T_n \le t+dt) &= \mathbb{P}(n-1 \text{ points in } [0,t] \text{ and } \ge 1 \text{ points in } [t, t+dt] + o(dt) \\ &= \left[\frac{e^{-\lambda t}(\lambda t)^{n-1}}{(n-1)!}\right]\left[\frac{\cancel{e^{-\lambda dt}}(\lambda dt)^1}{1!} + o(dt)\right] + o(dt) \\ &= e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}\lambda\, dt + o(dt)\end{aligned}$$

where $o(dt)$ indicates a term which is negligible compared to $dt$ in calculus limit as $dt \to 0$ ( for the possibility of more than 1 point in the interval of length $dt$). So this gives the answer that $T_n$ has probability density

$$f_{T_n}(t) = \frac{e^{-\lambda t}\lambda^n t^{n-1}}{(n-1)!} \qquad (t > 0)$$

which is the $\textsf{Gamma}(n, \lambda)$ density. See also text Theorem 2.2 for the more usual derivation of the same density from the important representation

$$T_n = W_1 + \cdots + W_n$$

for independent exponential $\lambda$ spacings $W_n$, denoted $\tau_n$ by Durrett.

# LECTURE 12

# Poisson Processes, Part 2

## 12.1 Theorem 2.10

*Some notation:* Pitman will be employing $X$'s instead of Durret's $Y$'s.

**Theorem 2.10 (Durrret)**

Let $X_1, X_2$ be a sequence of i.i.d. random variables, each with the same distribution as $X$, and define

$$S_N = X_1 + X_2 + \ldots + X_N \tag{12.1}$$

where $N \geq 0$ is a random index independent of the sequence $X_1, X_2, \ldots$, and summing zero terms yields 0, formally $S_N = 0$ if $N = 0$. Then

(*i*) $\mathbb{E}(S_N) = \mathbb{E}(N)\mathbb{E}(X)$ provided $\mathbb{E}N < \infty$ and $\mathbb{E}|X| < \infty$

(*ii*) $\mathbb{V}\mathrm{ar}(S_N) = \mathbb{E}(N)\mathbb{V}\mathrm{ar}(X) + \mathbb{V}\mathrm{ar}(N)(\mathbb{E}X)^2$
provided $\mathbb{E}N < \infty$ and $\mathbb{E}X^2 < \infty$

(*iii*) If $N \sim$ **Poisson** $(\mu)$, then $\mathbb{E}(N) = \mathbb{V}\mathrm{ar}(N) = \mu$. Implying

$$\mathbb{V}\mathrm{ar}(S_N) = \mu\mathbb{V}\mathrm{ar}(X) + \mu(\mathbb{E}X)^2 = \mu\ \mathbb{E}(X^2)$$

*Proof.* Durret's proof is fine. But here is a shorter derivation. Recall that for any numerical random variable $Y$ with $\mathbb{E}|Y| < \infty$, and any random variable $X$

$$\mathbb{E}(Y) = \mathbb{E}\Big[\mathbb{E}(Y \,|\, X)\Big] \tag{12.2}$$

In words,

> "The expectation is the expectation of the conditional expectation"

And for any random variable $Y$ with $\mathbb{E}Y^2 < \infty$ and any random variable $X$

$$\mathbb{V}\text{ar}(Y) = \mathbb{E}\Big[\mathbb{V}\text{ar}(Y \mid X)\Big] + \mathbb{V}\text{ar}\Big[\mathbb{E}(Y \mid X)\Big] \tag{12.3}$$

> "The variance of is the expectation of the conditional variance plus the variance of the conditional expectation"[1]

These formulas apply in an obvious way to our set-up with $Y = S_N$ and $X = N$. Now to the proof.

*For conclusion* $(i)$

$$\mathbb{E}(S_N) = \mathbb{E}\Big[\mathbb{E}(S_N \mid N)\Big] = \mathbb{E}\big[N\mathbb{E}(X)\big] = \mathbb{E}(N)\mathbb{E}(X)$$

*For conclusion* $(ii)$

$$\begin{aligned}\mathbb{V}\text{ar}(Y) &= \mathbb{E}\Big[\mathbb{V}\text{ar}(S_N \mid N)\Big] + \mathbb{V}\text{ar}\Big[\mathbb{E}(S_N \mid N)\Big] \\ &= \mathbb{E}\Big[N\mathbb{V}\text{ar}(X)\Big] + \mathbb{V}\text{ar}\Big[N\mathbb{E}(X)\Big] \\ &= \mathbb{E}(N)\mathbb{V}\text{ar}(X) + \mathbb{V}\text{ar}(N)(\mathbb{E}X)^2\end{aligned}$$

*For conclusion* $(iii)$ just plug in the fact that $N \sim$ **Poisson** $(\mu)$ has mean and variance both equal to $\mu$. □

## 12.2 Generalization to a Stopping Time $N$

In the case where $N$ is not necessarily independent of $X_1, X_2, \ldots$, we study `Wald's Identities`.

### 12.2.1 Wald's Identities

We have this notion of $N$ being a stopping time for the sequence $X_1, X_2, \ldots$. For example, $N = \min\{n \geq 1 : X_n \in B\}$ or $N = \min\{n \geq 1 : X_n \in S_n\}$. We realize,

- $(N = n)$ is *determined* by $(X_1, X_2, \ldots X_n)$
- $(N \leq n)$ and $(N > n)$ are *determined* by $(X_1, X_2, \ldots X_n)$
- $(N < n)$ and $(N \geq n)$ are *determined* by $(X_1, X_2, \ldots X_{n-1})$

[1] Pitman suggests reciting this 9 times before one sleeps (if you haven't yet internalized it already).

**Claim** If both $\mathbb{E}N$ and $\mathbb{E}|X| < \infty$, and $N$ is a stopping time of the sequence $X_1, X_2, \ldots$, then $\mathbb{E}(S_N) = \mathbb{E}(N)\mathbb{E}(X)$ is still true. i.e. It is as if $N$ and $X$ are independent. This is not intuitive at first, we sketch the proof and show the claim is true first for $X \geq 0$.

$$S_N = \sum_{k=1}^{N} X_k = \sum_{k=1}^{\infty} X_k \ \mathbf{1}\{k \leq N\}$$

Notice $(k \leq N) = (N > k-1) = (N \leq k-1)^c$ is determined by $X_1, \ldots, X_{k-1}$, hence independent of $X_k$. Applying expectation

$$\mathbb{E}(S_N) = \mathbb{E}\left(\sum_{k=1}^{\infty} X_k \ \mathbf{1}\{k \leq N\}\right) = \sum_{k=1}^{\infty} (\mathbb{E}X_k)\mathbb{P}(N \geq k)$$

Since $\mathbb{E}X_k = \mathbb{E}X$ for all $x$, and exploiting the tail-sum formula, this is

$$(\mathbb{E}X) \underbrace{\sum_{k=1}^{\infty} \mathbb{P}(N \geq k)}_{\text{tail-sum}} = (\mathbb{E}X)(\mathbb{E}N)$$

Here the swap of $\mathbb{E}$ and $\Sigma$ is justified by non-negativity of all variables. If the $X_i$ are signed, first split each $X_i = X_i^+ - X_i^-$ into the difference of two positive variables, work on the positive and negative parts separately, then argue that both pieces are finite so OK to take their difference.

Note. What really makes this argument work is that for each $k$ the event $N \geq k$ is independent of $X_k$. One way to achieve this is to assume $N$ is independent of each $X_k$. Another way is to assume $N$ is a stopping time, so $(N \geq k)$ is determined by $X_1, \ldots, X_{k-1}$. A slight generalization that includes both cases is to assume there is some richer sequence of random variables $(W_0, W_1, \ldots)$ such that

- a) $N$ with values in $\{0, 1, 2, \ldots\}$ is a stopping time relative to $(W_0, W_1, \ldots)$, meaning $(N \leq n)$ is determined by $(W_0, \ldots, W_n)$.
- b) each variable $X_n$ for $n \geq 1$ has the same distribution as $X$, and $X_n$ is independent of $(W_0, \ldots, W_{n-1})$.

The proof is exactly the same. There is also a companion Wald identity for variances. Under the same assumptions a) and b) above,

- if $\mathbb{E}N < \infty$ and $\mathbb{E}X^2 < \infty$ and $\mathbb{E}X = 0$ then $\mathbb{E}S_N = 0$ and $\mathbb{E}S_N^2 = \mathbb{E}N(EX^2)$

To see this, assume first that $N$ is bounded above by say $m$, and compute

$$\mathbb{E}(S_N^2) = \mathbb{E}\left(\sum_{j=1}^{m} X_j 1(j \leq N)\right)\left(\sum_{k=1}^{m} X_k 1(k \leq N)\right)$$

be expanding the product. The diagonal terms easily give $\mathbb{E}N(EX^2)$, and the off-diagonal terms are all 0 if $\mathbb{E}X = 0$. Some care is required to pass to the limit of unbounded stopping times $N$, using convergence in mean square of $S_{N \wedge m}$ to $S_N$ as $m \to \infty$.

## 12.3 Poisson Thinning

We turn back to looking at the Poisson Point Process (PPP) on the strip $[0, \infty) \times [0, 1]$. Recall that points are distributed with intensity $\lambda$ per unit area: each square $[n, n+1] \times [0, 1]$ contains a Poisson($\lambda$) number of points, independently from one square to the next, and given $m$ points in one of these squares, these points are distributed i.i.d. according to area in the square.
The general formula for $\mathbb{V}\mathrm{ar}(S_N)$ for a stopping time $N$ when $\mathbb{E}X \neq 0$ is not so simple. Exercise: find the formula, and note it contains an $\mathbb{E}NS_N$ which is easy to evaluate if $N$ is independent of $X_1, X_2, \ldots$, but not always so easy to evaluate.

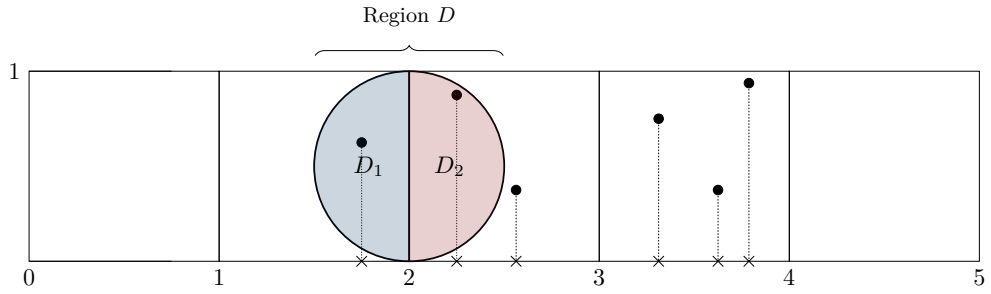



Figure 12.1: PPP on $[0, \infty) \times [0, 1]$ with region $D$.

Consider the region $D$, a circle with diameter one shown above. Let $N(D)$ be the number of points in $D$. So $N(D) = 2$ for the particular realization of the Poisson scatter illustrated. We seek to compute the probability of this event

$$\mathbb{P}\Big(N(D) = 2\Big)$$

We observe

- $N(D) = N(D_1) + N(D_2)$, where $D_1$ is the left half-disk and $D_2$ the right half-disk.
- $\text{Area}(D) = \frac{\pi}{4}$, and
- by symmetry, $N(D_1)$ and $N(D_2)$ are i.i.d.

So to re-frame the question,

> Given that a point falls in $(1, 2] \times [0, 1]$, what is the probability that it falls into $D$?

It is clear that $\text{Area}(D_1) = \frac{1}{2}\left(\frac{\pi}{4}\right)$. We can then see

$$N(D_1) \stackrel{d}{=} N(D_2) \sim \textbf{Poisson}\left(\lambda \times \frac{\pi}{8}\right)$$

That is, we derive the distribution of $N(D_1)$ from the known Poisson $(\lambda)$ distribution of $N((1, 2] \times [0, 1]$ by Poisson *thinning*. Hence

$$N(D) \sim \textbf{Poisson}\left(\lambda \times \frac{\pi}{8} + \lambda \times \frac{\pi}{8}\right) = \textbf{Poisson}\left(\lambda \times \frac{\pi}{4}\right) = \textbf{Poisson}\left(\lambda \times \text{Area}(D)\right)$$

The conclusion holds when we translate the disk to the left by $1/4$. Or any other amoount. Check it!

### 12.3.1 Poisson Thinning for a General Region

Let $D$ be any sub-region of the strip with finite area as depicted below.

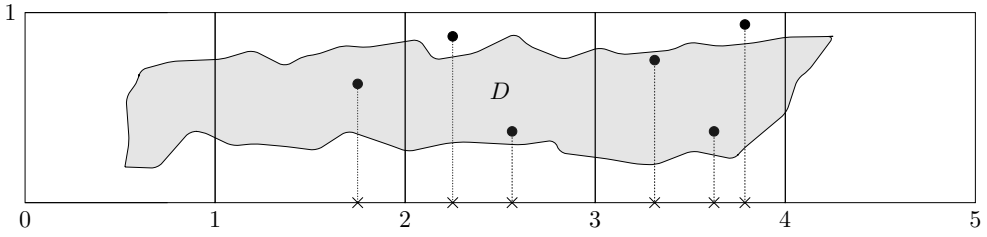



Figure 12.2: PPP on $[0, \infty) \times [0, 1]$ with general region $D$.

For simplicity, let $m < \infty$. We partition $D$

$$D = D_1 \cup D_2 \cup \cdots \cup D_m$$

where

$$D_i = D \cap \{(i, i + 1] \times [0, 1]\}$$

and it follows

$$N(D) = \underbrace{N(D_1) + N(D_1) + \ldots + N(D_m)}_{\text{independent Poisson counts}}$$

where by construction and Poisson thining

$$N(D_i) \sim \textbf{Poisson}\,(\lambda \times \text{Area}(D_i))$$

Hence by additing independent Poisson variables, *no matter what the shape of the region $D$, provided its area is well defined (technically, $D$ is a measurable subset of the strip), the number of Poisson points that fall in $D$ has* **Poisson** $(()\lambda Area(D))$ *distribution*

### 12.3.2 Poisson Thinning for Two General Regions

Suppose $D$ and $F$ are disjoint regions as depicted below. By Poisson thinning

$$N(D) \sim \textbf{Poisson}\,(\lambda \times \text{Area}(D)) \quad \text{and} \quad N(F) \sim \textbf{Poisson}\,(\lambda \times \text{Area}(F))$$

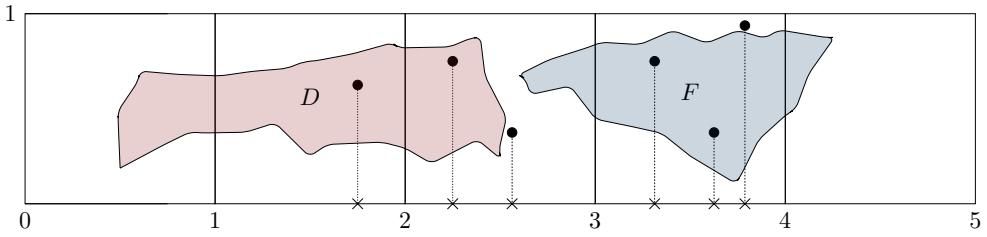
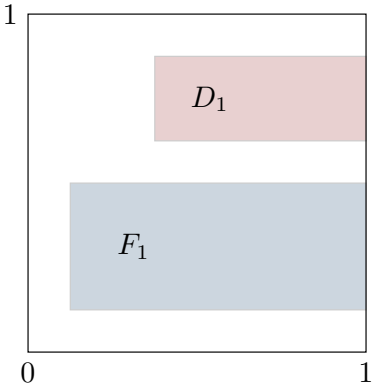



Figure 12.3: PPP on $[0,\infty) \times [0,1]$ with two disjoint regions $D$ and $F$.

Observe that for each block on the strip we have two independent counts. In total, $m$ independent pairs

$$N(D) = \boxed{N(D_1)} + \boxed{N(D_2)} + \ldots + \boxed{N(D_m)}$$
$$N(F) = \boxed{N(F_1)} + \boxed{N(F_2)} + \ldots + \boxed{N(F_m)}$$




Within each square $i$ there is a count for $D_i$, a count for $F_i$ and a count for the rest of the square. Hence, by Poissonizaton of the trinomial, the counts $D_i$ and $F_i$ in each square are independent, and they are independent between squares by construction, so we conclude:

- If $D$ and $F$ are disjoint sub-regions of the strip, then the counts $N(D)$ and $N(F)$ are independent Poisson variables with means $\lambda$Area$(D)$ and $\lambda$Area$(F)$.

The extension of this argument to 3 or more disjoint regions is obvious. The counts in any number of disjoint regions are independent Poisson variables.

## 12.4 General Measures

A *measure* $\mu$ on a space $\mathcal{S}$ is a function of subsets of $\mathcal{S}$. e.g. length, area, volume, probability, subject countable additivity

$$\mu(A) = \mu(A_1) + \mu(A_2) + \dots$$

Define

$$(\mathcal{S}, \mathfrak{S}, \mu) = \text{ measure space}$$

where $\mathfrak{S}$ (curly S) is the domain of measurable sets $A$, and $\mu(A)$ is a measure. Say a stochastic process

$$\Big(N(A), A \in \mathfrak{S}\Big) \text{ is a PPP with intensity } \mu \text{ iff}$$

- $N(A) \sim$ **Poisson** $(\mu(A))$
- If $A_1, A_2, \dots, A_m$ are disjoint measurable sets, then the $N(A_i)$ are independent.

In the case $\mathcal{S} = [0, \infty) \times [0, 1]$ and $\mu = \lambda \times$ Area we constructed a PPP($\mu$) by throwing down Poisson numbers of points i.i.d. with probability proportional to area in each of a sequence of unit squares. It was not important in this construction that the squares were unit squares, or even that they were squares. Any way of covering the strip with rectangles would work the same: give each rectangle $R$ a Poisson $\mu(R)$ number of points, and given there are $m$ points in the rectangle, let each of them be uniformly distributed. You could do the same with triangles and get the same Poisson process. More generally, the same argument shows that if $\mu$ is any $\sigma$-finite measure, meaning

$$\mathcal{S} = \mathcal{S}_1 + \mathcal{S}_2 + \dots \tag{12.4}$$

is the disjoint union of sets $\mathcal{S}_i$ with $\mu(\mathcal{S}_i) < \infty$ for all $i = 1, 2, \dots$, then
then we can construct a PPP($\mu$), by throwing down independent **Poisson** $(\mu(\mathcal{S}_i))$ numbers of points in $\mathcal{S}_i$, and given $n$ points in $\mathcal{S}_i$, assign them be i.i.d. locations according to $\mu(\,\cdot \,|\, \mathcal{S}_i)$, meaning for $\cdot = A$,

$$\mathbb{P}(\text{point in } A) = \mu(A \,|\, \mathcal{S}_i) := \frac{\mu(A \cap \mathcal{S}_i)}{\mu(\mathcal{S}_i)}.$$

The result will be a PPP($\mu$), no matter what the choice of partition (12.4) used in this construction. This is a very powerful and general way of thinking about Poisson processes, which is more useful in many respects than the logically equivalent description in the case $\mathcal{S} = [0, \infty)$ and $\mu$ is $\lambda$ times length measure, that the spacings between the points are i.i.d. exponential ($\lambda$) variables.

# LECTURE 13

# Midterm Preparation

## 13.1 Midterm Announcements

**coverage** text and readings to date (Markov Chains and Poisson Processes)

**format** 5 questions each 3-4 parts each, similar in style to past midterms

## 13.2 Old Midterm Exam Problems

See aggregate file of exam problems on bCourses.

### Problem 18

Read text and lecture notes

### Problem 19

- random walk on $\{0, 1, \dots\}$ with state 0 absorbing, i.e. $P(0, 0) = 1$
- this chain is driven by a probability distribution $p_0, p_1, \dots$
- from $i > 0$ this walk moves to $i + j - 1$ with probability $p_j$
- say $X$ has distribution $p_0, p_1, \dots$, walk hits increments by $X - 1$ until it hits 0
- $f_{ij} = \mathbb{P}_i(\text{ever hits } j \text{ at time } n \geq 1)$

Say you know $f_{10}$, then $f_{i,i-1} = f_{10}$ for all $i$. What can we say about $f_{20}$? Notice that to get from 2 to 1 the walk must pass through 1 along the way! How is this useful? Observe, by the *strong Markov property*

$$f_{20} = f_{21} f_{10}$$

which implies

$$f_{i0} = f_{10}^i$$

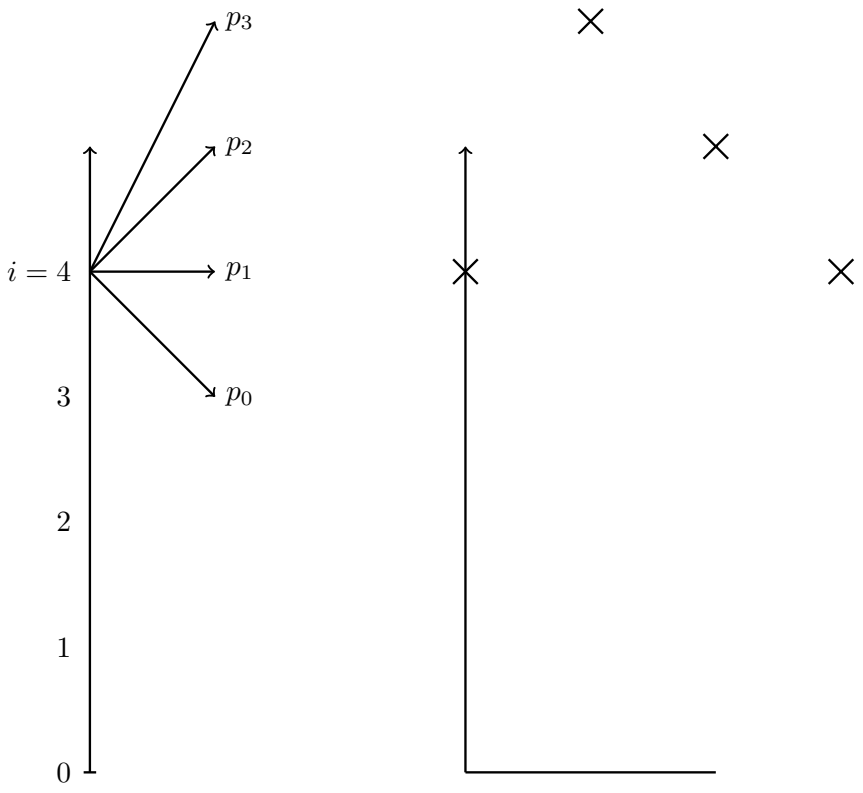



Figure 13.1: Illustration of problem 19's random walk, starting at 4.

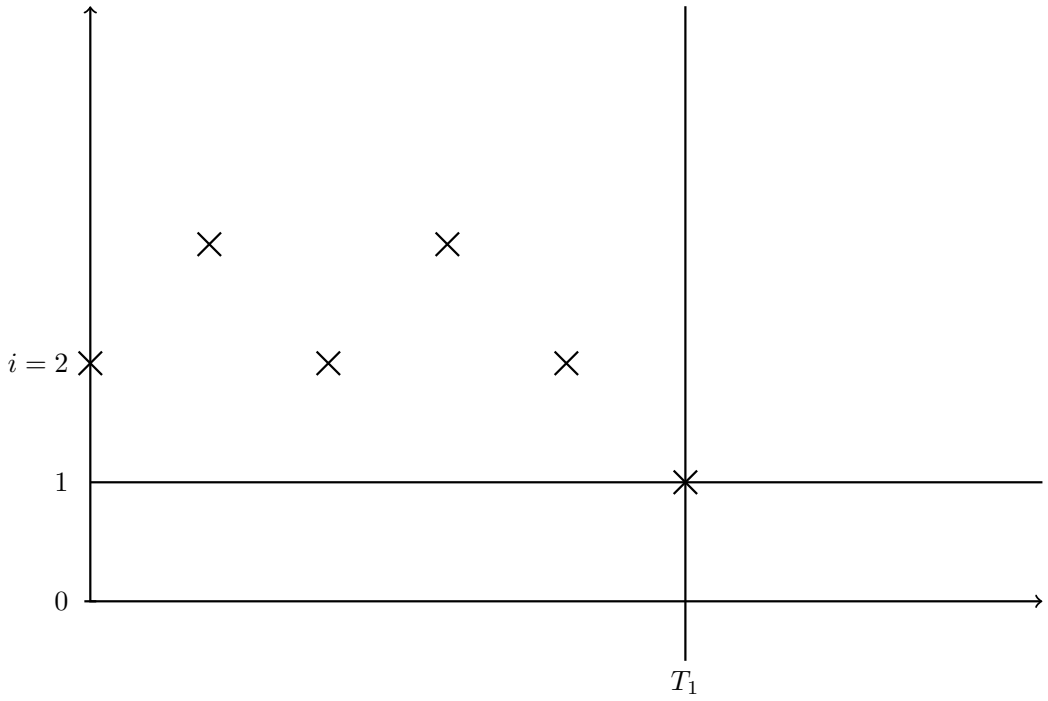



Figure 13.2: Illustration of problem 19's random walk, starting at 2 and necessarily passing through 1 to get to 0.

So let's find $f_{10}$ be a first step analysis, which combined with the above observation gives the harmonic equation

$$f_{10} = \sum_{k=0}^{\infty} p_k f_{10}^k = \phi(f_{10}) \qquad (*)$$

An old friend has returned, notice $(*)$ is the equation for the extinction probability for a branching process, starting with one individual and offspring generating function $p_0, p_1, \ldots$. Recall that the problem has given you

$$\mu = \sum_k k p_k$$

The mean of the $p$ distribution is driving this random walk. One technical remark, we must assume that $p_1 < 1$ for the statements of the problem to be true. For if $p_1 = 1$, the walk stays fixed where it starts (all states are absorbing). In the branching process, the population stays the same forever, because each person has one child. If you think a technical condition is missing on the exam, add it in. If there's an ambiguity, exploit it to your advantage.

In this problem $\mu \leq 1$. Let's sketch the PGF for $\mu \leq 1$.

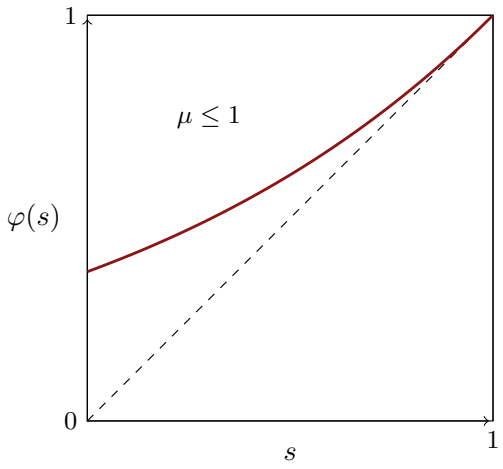



Figure 13.3: Notice that the unique root of $s = \varphi(s)$ is at $s = 1$ if $\mu \leq 1$ and we forbid $p_1 = 1$.

We can then see

$$f_{10} = f_{i0} = 1 \implies f_{ij} = 1 \qquad (0 \leq j \leq i)$$

For the rest of the problem, solve using first step analysis. For $j > 1$

$$f_{ij} = p_0 f_{i-1,j} + p_1 f_{i,j} + p_2 f_{i+1,j} + \ldots$$

$$= \sum_{k=0}^{\infty} p_k f_{i-1+k,j}$$

Notice if $i - 1 + k > j$, then $f_{i-1+k,j} = 1$, which implies $\sum$ from some point on $=$ tail probability of $p_\ell + p_{\ell+1} + \ldots = \mathbb{P}(X \geq \ell)$.

## 13.3 Intuition for Constructing Graphs of PGFs

This is done by example. Take a look.

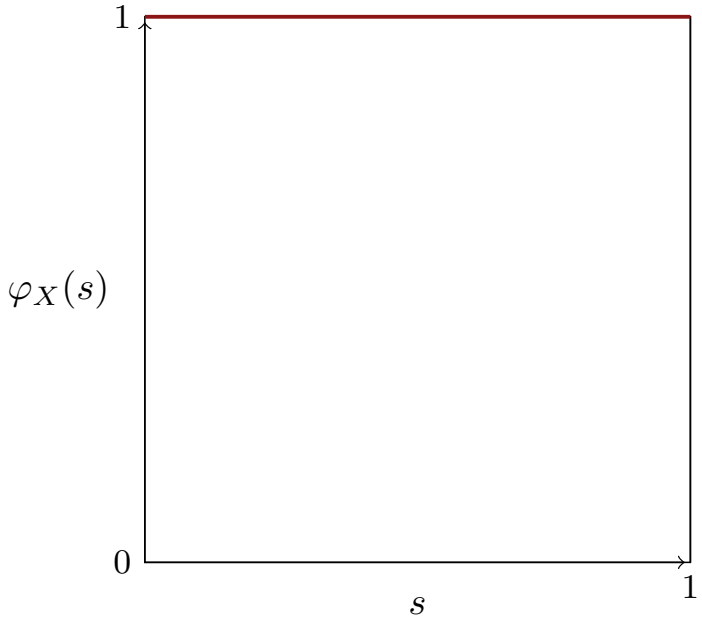



Figure 13.4: Plot of probability generating function for $X = 0$

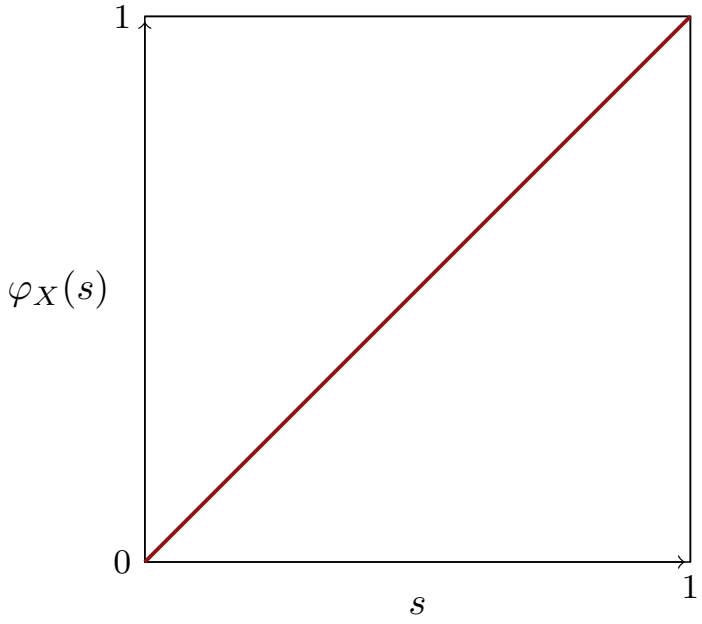



Figure 13.5: Plot of probability generating function for $X = 1$

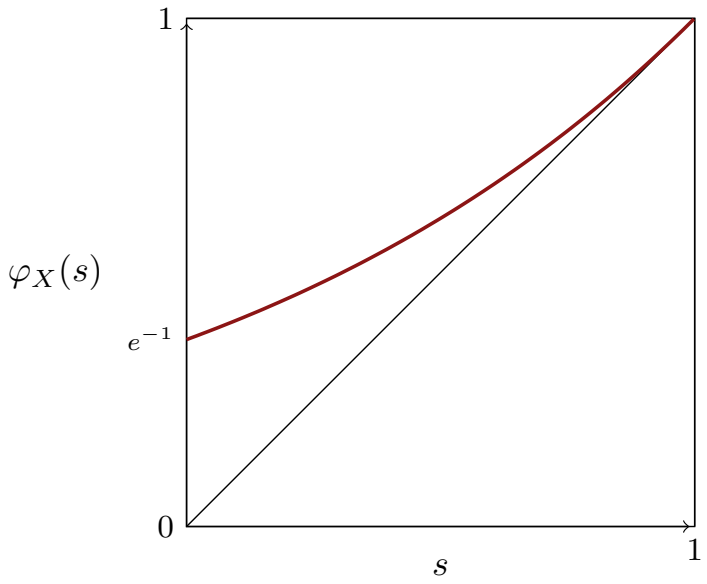



Figure 13.6: Plot of probability generating function for $X \sim$ **Poisson(1)** and diagonal. Observe the $\varphi_X(s) = e^{-(1-s)}$ is convex above the diagonal with a slope of 1 at $(1, 1)$.

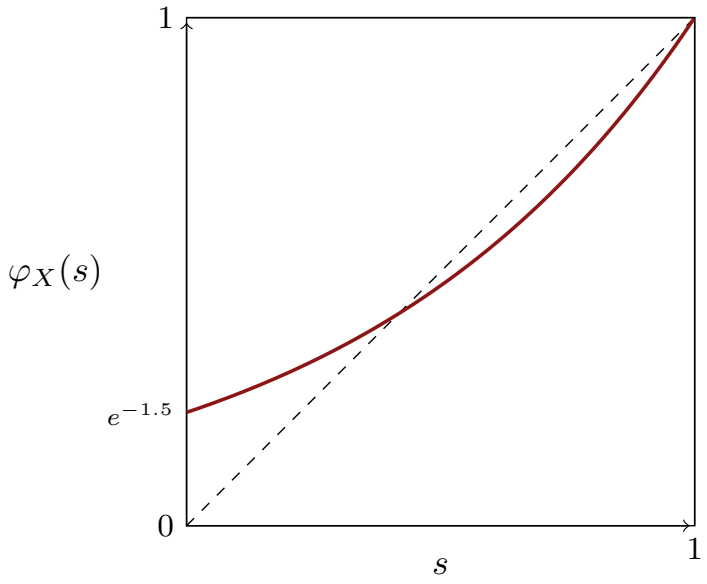



Figure 13.7: Plot of probability generating function for $X \sim$ **Poisson** $(3/2)$ and diagonal. Observe the $\varphi_X(s) = e^{-\frac{3}{2}(1-s)}$ is convex below the diagonal of $3/2$ at $(1, 1)$.

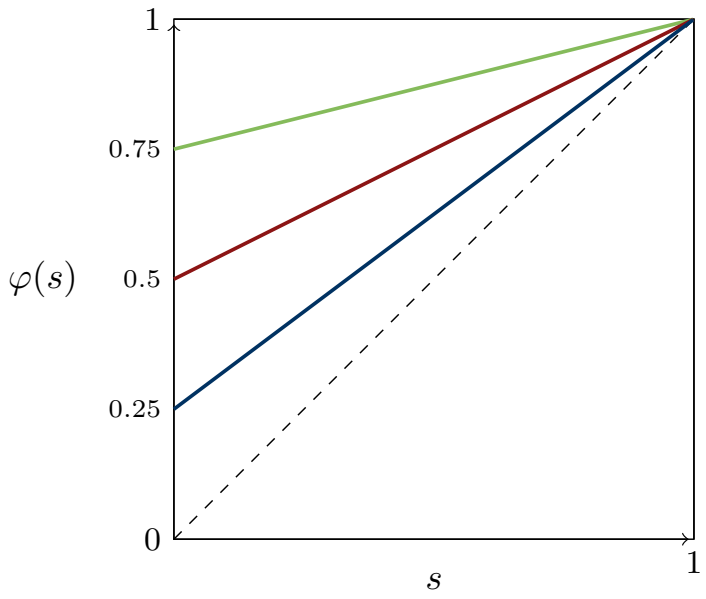



Figure 13.8: Plot of probability generating function for **Bernoulli**$(p)$ with parameters $p = \frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$ in green, red, and blue respectively. In general, the probability generating function for a **Bernoulli**$(p)$ random variable is $\varphi(s) = 1 - p + ps$

## 13.4 Discussion: Geometric-Negative Binomial PGFs

For a **Geometric**$(p)$ random variable defined on $\{0, 1, \dots\}$

$$p_n = p(1-p)^n$$

and its PGF is

$$\varphi(s) = \sum_{n=0}^{\infty} p_n s^n = \sum_{n=0}^{\infty} p(1-p)s^n = \frac{p}{1-qs}$$

Now let $X_1, X_2, \dots$ be i.i.d. **Geometric**$(p)$ random variables defined on $\{0, 1, \dots\}$. We seek to find the distribution of

$$F_n = X_1 + X_2 + \ldots + X_n$$

Clearly,

$$\mathbb{E}s^{F_n} = \mathbb{E}s^{X_1} \cdots \mathbb{E}s^{X_n} = \left(\frac{p}{1-qs}\right)^n$$

As an instance of this structure, $F_n$ may be constructed as the number of failures before the $n^{\text{th}}$ success in **Bernoulli**$(p)$ trials.

$$\underbrace{0\ 0\ 0}_{X_1}1\ \ \underbrace{0\ 0\ 0}_{X_2}1\ \ \underbrace{0\ 0}_{X_3}1$$

So how to get the probabilities in the distribution of $F_n$ out of the generating function? Observe

$$\left(\frac{p}{1-qs}\right)^n = p^n(1-qs)^{-n} = p^n \sum_{k=0}^{\infty} \binom{-n}{k}(-qs)^k = \sum_{k=0}^{\infty} \mathbb{P}(F_n = k)s^k$$

which is good for $|qs| < 1$. As a reminder, recall

$$(1-x)^{-1} = 1 + x + x^2 + \ldots \qquad |x| < 1$$

Back to the question of point probabilities for $F_n$, we simply inspect the formula conclude

$$\mathbb{P}(F_n = k) = p^n \binom{-n}{k}(-q)^k$$

In general, to recover the probabilities in a distribution on $\{0, 1, 2, \ldots\}$ from its generating function $\phi(s)$, use standard series formulas to expand $\phi(s)$ in powers of $s$. The coeeficients of these powers are the probabilities. Hence the uniqueness theorem: Two random variables with the same pgf have the same distribution.

### Problem 21

We have a Markov chain on $\{0, 1, \ldots, N\}$ and from the problem statement

$$X_{n+1} \,|\, X_n \sim \textbf{Binomial}\left(N, \frac{X_n}{N}\right)$$

We are tasked to show

$$\lim_{n\to\infty} \mathbb{P}_k(X_n = n) = \frac{k}{N} \tag{$\star$}$$

Given a distribution, the mean gives the most basic information about it. For **Binomial** $(n, p)$, its mean is $np$. Hence

$$\mathbb{E}(X_{n+1} \,|\, X_n) = N \cdot \frac{X_n}{N} = X_n \implies \mathbb{E}X_{n+1} = \mathbb{E}X_n$$

and this is true for $X_0 = k$, by the Markov property. Observe

$$\mathbb{E}_k X_n = \mathbb{E}_k X_0 = k$$

and hence

$$k = \sum_{x=0}^{N} x\mathbb{P}_k(X_n = k)$$

splitting the sum

$$\underbrace{\sum_{x=0}^{N-1} x\mathbb{P}_k(X_n = k)}_{\text{terms vanish to 0, as } n\to\infty} + N\mathbb{P}_k(X_n = N)$$

We then conclude ($\star$). We've actually seen this argument before with respect to he vanishing terms. The chain has $\{0, N\}$ absorbing and the chain wobbles between these states until it hits one of those boundaries. The chance that it does not eventually hit a boundary is zero, by the geometric bounds argument; no matter where you start, there's some path to hitting the boundary.

### Problem 27

We have a Markov chain $(X_n)$ with finite state space $S$, probability transition matrix $P$, absorbing set $B$, and $T = \min\{n \geq 1 : X_n \in B\}$. We also assume $\mathbb{P}_i(T < \infty) = 1$. For part a, we are tasked with deriving a formula for

$$\mathbb{P}_i(X_{T-1} = j, X_T = k) \tag{$*$}$$

where $i, j \notin B$ and $k \in B$. In terms of[1]

$$W = (G \text{ before}) = (I - Q)^{-1}$$

Observe,

$$Q(x, y) = P(x, y) \qquad x, y \in B$$

and

$$W = I + Q + Q^2 + \dots$$

Looking at

$$\big(Q\mathbb{1}\big)(x) = \sum_{y \in B^c} P^n(x, y) = \mathbb{P}_x(T > n) \downarrow 0$$

Going back to $(*)$

$$\begin{aligned}
\mathbb{P}_i(X_{T-1} = j, X_T = k) &= \sum_{n=1}^{\infty} \mathbb{P}_i(X_{n-1} = j, X_n = k, T = n) \\
&= \sum_{m=0}^{\infty} \mathbb{P}_i(X_m = j, T > m, X_{m+1} = k) \\
&= \sum_{m=0}^{\infty} \mathbb{P}_i(X_m = j, T > m) P(j, k) \\
&= \sum_{m=0}^{\infty} Q^m(i, j) P(j, k) \\
&= \big(I - Q\big)^{-1}(i, j) P(j, k)
\end{aligned}$$

□

[1] $I$ and $Q$ are $B^c \times B^c$ matrices.

# LECTURE 14

# Renewal Theory, Part 1

### Class Announcements

- The midterm had quite a large spread, and it did require that we have some manipulative skill to perform actions on the material (not simply true/false checks).
- To prepare for the exams, we should practice examples to really get familiar with the material. The purpose of this course is to acquire concepts and skills and be able to translate them into applications.
- This week's homework is an overhang of Poisson process material that we've covered in lecture and text.

## 14.1 Introduction to Renewal Theory

Renewal Theory serves as an introduction into continuous-time parameter models. Renewal theory provides an important technique for analysis of more complicated stochastic processes such as queuing systems. We start from a basic counting process in continuous time, e.g. a Poisson Process with $\lambda$ constant rate. We can also make $\lambda(t)$ vary, as in this week's homework. In a typical counting process model, we have that[1]

$$N_t = \sum_{n=1}^{\infty} \mathbf{1}(T_n \leq t) \tag{14.1}$$

where $(T_n, n \geq 1)$ is an increasing sequence of arrival times, usually with $0 < T_1 < T_2 < \cdots$ indicating no multiple points (simultaneous arrivals). We've seen an illustration of this process before (see next page), where time is on the horizontal axis and on the vertical, we have a discrete count. This illustration follows logically

[1] $N(t) = N_t$ when convenient.

from (14.1). Logically and by understanding what the variables mean, we have the fundamental *inversion* or *duality* relation

$$(N_t \leq n) \iff (T_n \geq t)$$

relating $(T_n, n = 1, 2, \dots)$ with typically continuous values at discrete time points $n$ and $(N_t, t \geq 0)$ which has discrete values over continuous time $t$. Notice that if we have a model for the distribution of $(T_n, n \geq 1)$, it determines the distribution of the process $(N_t, t \geq 0)$ and vice versa. We have seen this in a first model

$$\text{PPP}(\lambda) : T_n = \text{sum of } n \text{ iid } \textbf{Exponential}\,(\lambda)$$

which is logically equivalent to $(N_t, t \geq 0)$ with stationary independent increments with Poisson distributions

$$N_t \sim \textbf{Poisson}\,(\lambda t)$$

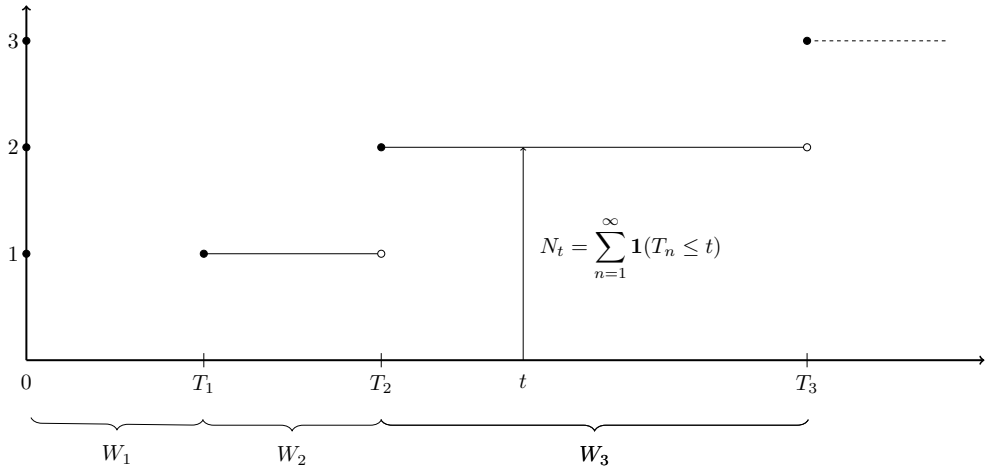



Figure 14.1: Poisson Process with arrival time $0 < T_1 < T_2 \dots$, inter-arrival times $0 < W_1, W_2, \dots$, and counting process $(N_t, t \geq 0)$

## 14.2 Renewal Theory

Now, we consider a model for Renewal Theory. We define[2]

$$T_n = X_1 + X_2 + \cdots + X_n$$

where the $X_i > 0$ are iid with the same non-negative distribution. Note that the duality relation (14.1) *holds* in renewal theory.

[2]Note that Durret uses $t_i$ as the inter-arrival times and $\tau_i$ as exponentially distributed inter-arrival times. Previously we used $W_i$ for iid exponentials, but $W$ has other meanings in queuing theroy. $X_i$ was selected after an in-class vote.

We say that a renewal occurs at time $T_n$ and say that $(N_t, t \geq 0)$ is the renewal counting process. Often, such a sequence of random times $T_n$ and such a renewal process are found embedded in (meaning they are functions of) some more complicated stochastic processes. We have seen this before in a familiar discrete time setup, i.e. Strong Markov Property.

Take a Markov chain with some reference state 0 (start at state 0). Let $T_n$ be the time of the $n$th return to 0. Then $X_1 = T_1, X_2, X_3, \ldots$ are iid copies of $X_1 = T_1$ by the Strong Markov Property. This is essentially discrete renewal theory. We've worked with this in a past homework with a generating function problem with tail probabilities. To consider the more interesting (continuous) scenario, in general, we may imagine $T_n$ as an $n$th "regeneration time" in some stochastic process.

### 14.2.1 Example: Queueing Model

Consider a queueing model with the following diagram of a set of 'busy cycles.' Let $Q(t)$ denote the number of individuals (customers, packets, items .. ) in the queue at time $t$. We start with an empty queue at $Q(t) = 0$, and the value goes up as more items arrive and are being attended to, and with more time, some of those are 'finished' (depart from the queue) and replaced with new items in the queue. We'll see in the text that several models of queues are analyzed in terms of their busy cycles via renewal theory. We notice that if we just do our renewal counting process, there is a lot less going on than otherwise. This renewal process is a component of a much more complex stochastic process. This is motivation as to how we can think in terms of renewal theory before looking closer into queueing models.

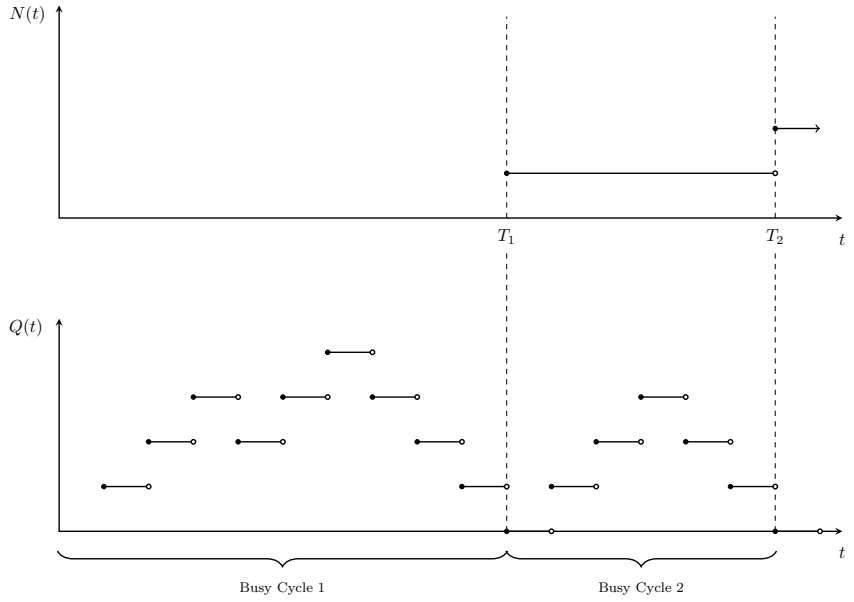



Figure 14.2: Renewal process from busy cycles of an underlying queuing process.

### 14.2.2 Example: A Janitor Replacing Lightbulbs

Take $X_1, X_2, \ldots$ as the lifetimes of lightbulbs, which are iid (e.g. from the same factory with the same level of quality control). Suppose that a new bulb is installed at time 0, and as each bulb burns out, it is replaced by a new one.

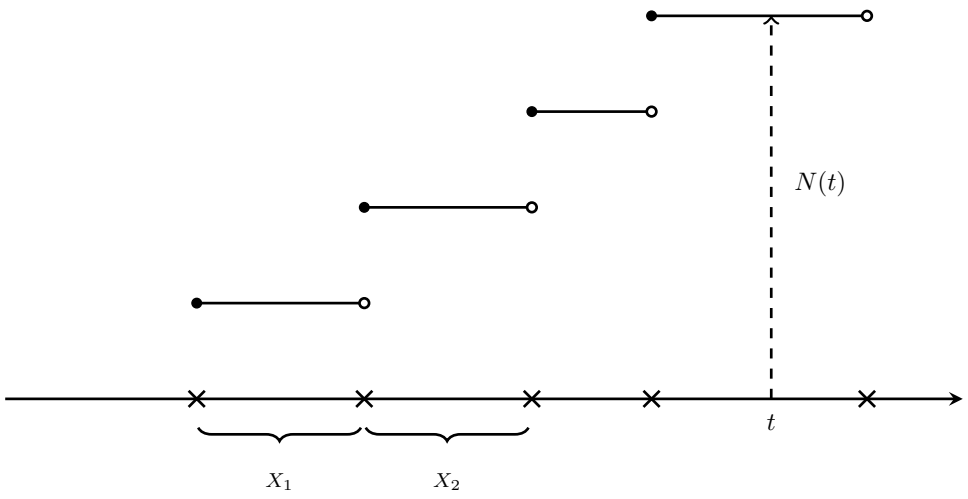



Figure 14.3: Janitor light-bulb replacement illustration.

We want the Poisson process to be a special case, so we take the convention of defining

$$N(t) := \#\text{ of replacements by time } t = \sum_{n=1}^{\infty} \mathbf{1}(T_n \leq t)$$

where $T_n := X_1 + \cdots + X_n$ is the time of the $n$th replacement. Of course, this can be more complicated than a Poisson process, because there the lifetimes of the light bulbs might not follow an exponential distribution. We want the abstraction and the additional flexibility to accommodate other lifetime distributions and things like busy cycles of queues.

Now we ask, what can we say about the *renewal counting process* $(N(t), t \geq 0)$? If the $X_i$ are iid **Exponential** $(\lambda)$, then $N(t), t \geq 0$ has the very special property of independent increments. In general, this is in fact characteristic of the exponential distribution, at least in contiuous time. (If we did this in discrete time, we have a similar special role for the geometric distribution, which corresponds to binomial counts with independent increments). Here, we take spacings to be generic and counts no longer have independent increments. However, for $N_t = N(t)$, we still have the duality

$$(T_n > t) \iff (N_t < n).$$

If we can find the CDF of $T_n$ for all $n$, then we get the CDF of $N_t$ for all $t$. The difficulty is that there are very few models for which we have an explicit formula for the above probability for general $n$ and $t$. We can use transforms (Laplace, MGF's), but then we have to invert, which can be a difficult task. The main parts of renewal theory get away from explicit formulas and deal with much more general things which hold even when we cannot get an explicit handle on the details.

The main idea is that we know a lot about the sums of random variables $T_n = X_1 + \ldots + X_n$, especially the Law of Large numbers

$$\frac{T_n}{n} \to \mathbb{E}X =: \mu = \frac{1}{\lambda} \text{ as } n \to \infty$$

Here $X := X_1 = T_1$ is a generic representative of the iid sequence whose partial sums are the $T_n$. Here $\mu$ is the mean time between renewals, and $\lambda := 1/\mu$ is the mean rate of renewals per unit time. If $X$ is exponential($\lambda$), then $\mu = 1/\lambda$ is a familiar formula. In general, for a renewal process replacing a Poisson process, the analog of the rate $\lambda$ of the PPP is the long run rate per unit time of renewals, which is $\lambda = 1/\mu$ for $\mu$ the mean inter-renewal time. Beware that in the text an the literature you never know without careful reading whether $\lambda$ or $\mu$ represents a rate or a mean or an inverse mean.

## 14.3 Weak Sense of Convergence

First of all, we'll define a weak sense of convergence. Assume $\mathbb{E}X < \infty$. Fix $x > 0$ and look at the probability that $T_n/n$ differs from its expected value $\mathbb{E}X$ by some small value $\epsilon$ (which we may set to $10^{-6}$ for instance):

$$\underbrace{\mathbb{P}\left(\left|\frac{T_n}{n} - \mathbb{E}X\right| > \epsilon\right)}_{\leq \frac{\sigma^2}{n\epsilon^2}} \to 0.$$

We have an honest proof to say that this probability in the case of a finite variance is bounded by $\frac{\sigma^2}{n}$, by Chebyshev's inequality, where $\sigma^2$ is the variance of $X$. The conclusion is also valid just assuming $\mathbb{E}X < \infty$, as shown in more advanced treatments of probability theory. The convergence promised by this weak sense of convergence can be very slow. Typically, if we assume more moments of $X$ are finite, we may have a faster rate of convergence, e.g. geometrically fast if $\mathbb{E}e^{\theta X} < \infty$ for some $\theta > 0$, by Chernoff bounds `https://en.wikipedia.org/wiki/Chernoff_bound`.

### 14.3.1 Strong Law of Large Numbers

This is implicitly used by Durrett in the ergodic theory of Markov chains and is used again in his treatment of renewal theory. The stronger statement is:

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{T_n}{n} = \mathbb{E}X\right) = 1$$

We think of this as the probability of an event. The difficulty is this is an infinite time horizon event. We cannot say if the event has happened at any given (finite) point in time, but this does not stop us thinking about it. As a quick explanation[3], we consider some axioms of probability. Essentially, we define $\mathbb{P}(F)$ for a collection of events $F$ that is closed under finite and *countable* set operations.

[3]For a full explanation, take Math 104, Math 202, and Stat 205.

That means that we can look at things like

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} F_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(F_n) \tag{14.2}$$

for disjoint $F_n$. We call this the *infinite sum rule* that we use for mathematical convenience. We'll assume that $0 \leq \mathbb{P}(F) \leq 1$ and $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ for $A \cap B = \varnothing$, and $\mathbb{P}(\Omega) = 1 \implies \mathbb{P}(\varnothing) = 0$. It can be shown assuming these basic rules of $\mathbb{P}$ for finite set operations that the infinite sum rule (14.2) is equivalent to each of the following conditions:

- $F_1 \supseteq F_2 \supseteq \ldots$ then $\mathbb{P}\left(\bigcap_n F_n\right) = \lim_{n\to\infty} \mathbb{P}(F_n)$
- $F_1 \subseteq F_2 \subseteq \ldots$ then $\mathbb{P}\left(\bigcup_n F_n\right) = \lim_{n\to\infty} \mathbb{P}(F_n)$

So the infinite sum rule allows the probabilities of limiting intersections or unions of events to be computed as either decreasing or increasing limits, as the case may be. In each case, the existence of the limit is obvious, assuming basic properties of real numbers. What is not obvious is that if probabilities of infinite intersections and unions are defined like this, then repeated use of these limit rules does not ever lead to two different probabilities for the same event, after representing the event in different ways. However, it is shown by *measure theory* that for all the probability models discussed in this course, this is not a problem. i.e. that these models can be defined in a way that is consistent with the infinite sum rule.
Now look at the event that $\left(\frac{T_n}{n} \to \mu\right)$

$$\left(\frac{T_n}{n} \to \mu\right) = \left(\forall\, \epsilon > 0,\ \exists\, N : \forall\, n \geq N,\ \left|\frac{T_n}{n} - \mu\right| \leq \epsilon\right)$$

$$= \bigcap_{\epsilon>0} \bigcup_{N>0} \bigcap_{n\geq N} \left(\left|\frac{T_n}{n} - \mu\right| \leq \epsilon\right)$$

which is just a succession of infinite intersections/unions/intersections, to which the above limit tules can be applied, to evaluate the probability of $(T_n/n \to \mu)$. Mathematicians discovered that doing this gives the event probability 1 (the limit exists and equals $\mu$ with probability 1). In this way, we combine the axioms of probability and limits to prove the Strong Law of Large Numbers. We will not rigorously go over the SLLN in this course. Pitman reminds that we take this detour simply to elaborate upon the convergence that Durett simply assumes, where it is stated that there is convergence of random variables without indicating the precise meaning of this convergence.

Consider a Poisson Process with rate $\lambda = \frac{1}{2}$ with the graph of $t \to N_t$. See depiction on next page. If we look only at the time points $t = 1\times10^6, 2\times10^6, \ldots$ at each of these points, the weak law of large numbers only gives us that with very overwhelming probability, we will be in the bands. The fact that we have these bands tells us

little about the behavior in between. The only obvious thing is to use Boole's inequality to sum a million inequalities to get a union bound, between one multiple of a million and the next. But there are better way to estimate the probabilities of such deviations, leading to the strong law of large numbers, which tells us that we can essentially get a uniform band around a linear approximation. Hence, the Strong Law of Large Numbers gives this as a bound that as we follow along time, the Poisson process will always stay within the bounds, with overwhelming probabilty. The Poisson process is simpler in this respect than a renewal process, because both $(N_t, t \geq 0)$ and $(T_n, n \geq 0)$ have stationalry independent increments, to which the strong law can be applied directly. For a renewal process, with have only that the $T_n$ increments satisfy the Strong Law. That is,

$$\frac{T_n}{n} \to \mathbb{E}X \text{ with probability 1.}$$

However, this can be turned into:

$$\frac{N_t}{t} \to \frac{1}{\mathbb{E}X} \quad \text{(with probability 1, by duality or inversion)}$$

For a more careful treatment and discussion, see Durett. Intuitively, if either one of the functions $\mapsto T_n$ or $t \mapsto N_t$ is close enough to a straight line for large $n$ or $t$, as the case may be, then also the inverse function must be close to a straight line, with the inverse slope. Hence all the back and forth between $\mu$ and $1/\mu$ in renewal theory.

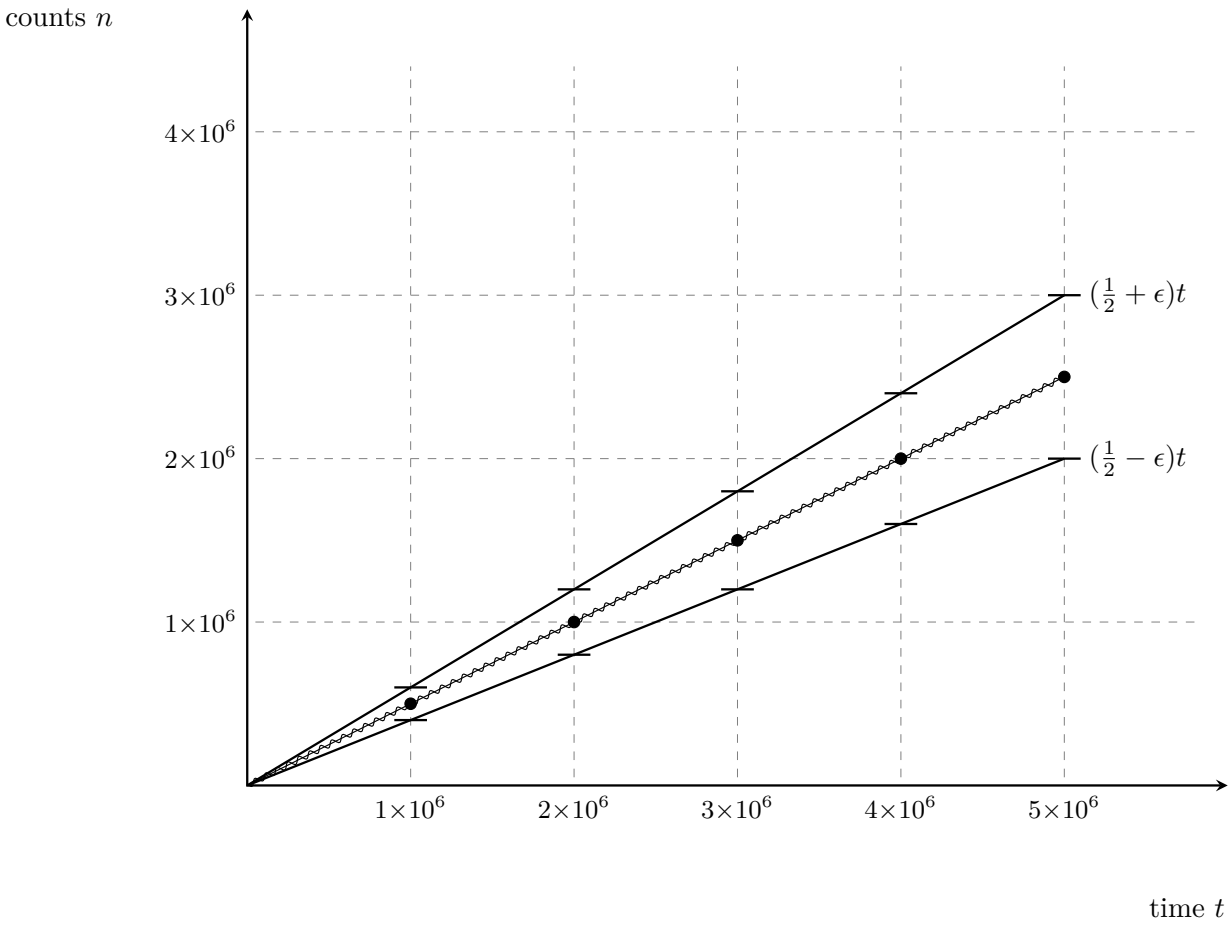



Figure 14.4: Poisson Process with radical rescaling.

**Strong Law of Large Numbers (SLLN) in Renewal Theory (RT)**

Strong Law of Large Numbers (SLLN) in Renewal Theory (RT): If $\mathbb{E}X < \infty$ and $N(t) = \sum_{n=1}^{\infty} \mathbb{1}(T_n \leq t)$ gives a renewal counting process, then

$$\frac{N(t)}{t} \to \frac{1}{\mathbb{E}X}$$

## 14.4 Renewal Reward Theorem

This is a simple theorem which is a variation on the Strong Law of Large Numbers. Suppose we have a renewal process, and in each cycle of length $X_i$, we get a reward

$R_i$. Let $R(t) :=$ cumulative reward up to time $t$. Then this is

$$R(t) = \sum_{i=1}^{N(t)} R_i + \underbrace{\text{something from the current cycle}}.$$

where the last term respects the idea that the reward up to $t$ may include only a part of the reward in the complet cycle covering time $t$. One of the main ideas is that whatever this last term is, it will be negligible compared the left term under reasonable assumptions when $t$ is large.
Although we have a random number of cycles, this is almost deterministic as

$$\frac{R(t)}{t} \to \frac{\mathbb{E}(R)}{\mathbb{E}(X)}$$

We arrive at this intuitively, by saying that the long run average reward per cucle is $\mathbb{E}R$, and the long run average number of cycles per unit time is $1/\mathbb{E}X$. But this is a mathematical fact that can be proven as a limit with probability one, like the law of large numbers. This holds in the sense of the Strong Law of Large numbers, with the minimal assumptions to make sense of the limit, that $\mathbb{E}|R| < \infty$ and $\mathbb{E}X < \infty$. See text for further details.

# LECTURE 15

# Renewal Theory, Part 2

### Class Announcements

Today we'll be finishing up our discussion on renewal theory. There are many examples in the text, which you should study for motivation. Homework next week will encompass renewal theory.

## 15.1 Age and Residual Life

Reference: Section 3.3 of text. Also, we have already covered aspects of discrete time theory (Section 3.3.1) in homework involving the Kac identity and tail sum formulas in renewal theory.

Recall that we've looked at points on a line according to a Poisson process. We allow a more interesting structure with $X_1, X_2, \ldots$ iid with $\mathbb{P}(X_i > 0) = 1$. Then construct

$$T_n = X_1 + \cdots + X_n$$

the time for the $n$th renewal. Then via indicators, we have the number of renewals up to and including time $t$ is

$$N(t) := \sum_{n=1}^{\infty} \mathbb{1}(T_n \leq t)$$

with now familiar duality relation between the distributions of $T_n$ for $n = 1, 2, \ldots$ and the distribution of $N(t)$ for $t \geq 0$.

We look at various associated stochastic processes, for example the *age process*, where we define

$$A_t = \text{ the age of component in use at time } t = t - T_{N(t)}$$

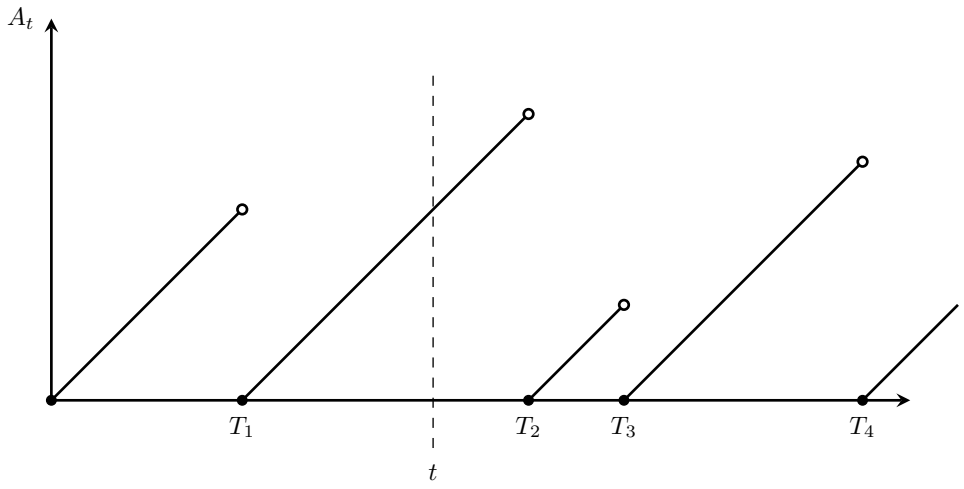



Figure 15.1: Component age diagram. Observe that in this realization, for the particular $t$ shown you should see $N(t) = 1$ and $A_t = t - T_1$. The vertical distance from $(t, 0)$ to $(t, A_t)$ equals the horizontal distance between $(T_1, 0)$ and $(t, 0)$.

This yields us a sort of 'sawtooth' process. We look for a closed-form formula for this process. Set the convention $T_0 := 0$ and notice we have the formula

$$A_t = t - T_{N(t)}$$

In some sense, we have a renewal at time 0, but by convention, we do not count it in $N(t)$. For today's discussion (for convenience and to not enter the study of null-recurrent processes), we assume $\mathbb{E}X_1 < \infty$. Here the $X_i$ are the inter-renewal times, which are the lifetimes in the light bulb metaphor, and can have other meanings, e.g. length of a busy period in a queuing model, or return time of some state in a suitably Markovian process.

The exact distribution of $A_t$ for a fixed time $t$ is easy in the Poisson case, but it's a lot more difficult in general. In the **Poisson** $(\lambda)$ case, when $N(t)$ is Poisson$(\lambda t)$, consider the probability $\mathbb{P}(A_t > a)$, by looking back $a$ units of time from time $t$. Then in terms of renewals, we have no Poisson points in the interval $a$, which gives (provided $a < t$)

$$\begin{aligned}\mathbb{P}(A_t > a) &= \mathbb{P}(\text{no renewals in } (t-a, t)) \\ &= \mathbb{P}(N(t-a, t) = 0) \\ &= e^{-\lambda a}\end{aligned}$$

Also in the Poisson case

$$\mathbb{P}(A_t = t) = e^{-\lambda t}$$

and in general $\mathbb{P}(0 \leq A_t \leq t) = 1$ so for an umbounded liftime distribution the distribution of $A_t$ will always have an atom of size $\mathbb{P}(X_1 > t)$ at the value $t$. Notice via right-tails that in the Poisson case

$$\mathbb{P}(A_t > a) \xrightarrow{t\to\infty} e^{-\lambda a}$$

$$\mathbb{P}(A_t \leq a) \xrightarrow{t\to\infty} 1 - e^{-\lambda a}$$

via the obvious convergence of CDFs. That is, the limit distribution of $A_t$ as $t \to \infty$ is exponential with rate $\lambda$ and mean $\frac{1}{\lambda}$.

### 15.1.1 General Lifetime Distribution of $X$

We may ask, what happens for a general lifetime distribution of $X$? Previously, we took $X \sim$ **Exponential** $(\lambda)$, which turned out to be the same as the limit distribution of $A_t$ as $t \to \infty$. But this is very special, in fact even a characterization of the Poisson process among all renewal processes.

To look at this, we approach this problem indirectly. Let's suppose that the limit distribution of $A_t$ can be understood in terms of long run averages. That is, look at $A_\infty$, a variable with the limit distribution. Then interpret, in the context of renewal theory

$$\mathbb{P}(A_\infty > a) = \text{long run fraction of time } t \text{ that } A_t > a.$$

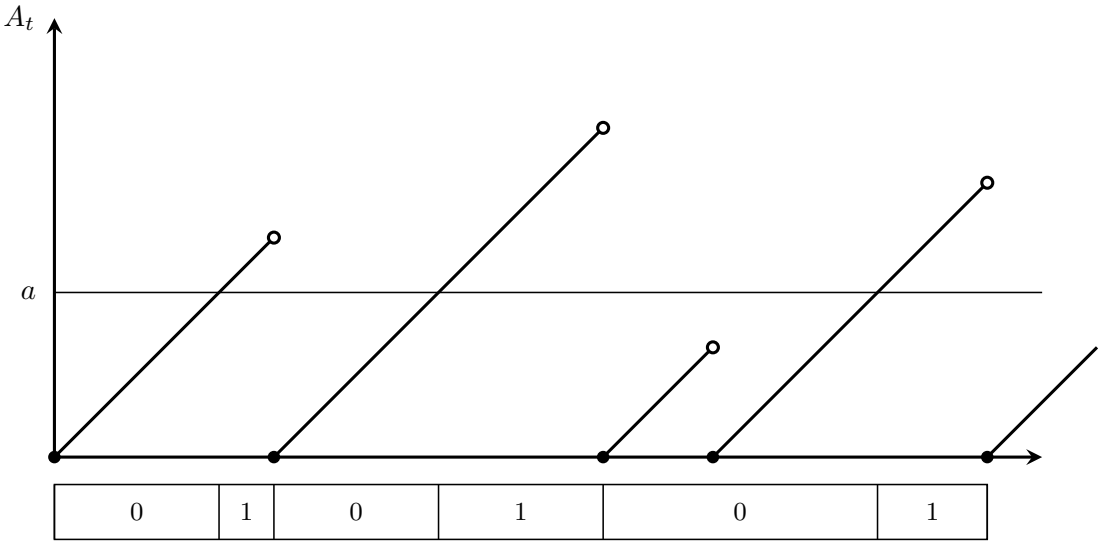



Figure 15.2: Underlying zero-one process, where $X := \mathbf{1}\{A_t > t\}$

Sometimes, the age will be above $a$, and sometimes below $a$. This yields an indicator process, where we have 0 when the age is $\leq a$ and 1 when age is $> a$. This is the

indicator process

$$(\mathbb{1}(A_t > a), t \geq 0)$$

Then we can apply the Renewal Reward Theorem from last lecture. Suppose that we get a reward $R_n$ from each renewal interval $[T_{n-1}, T_n]$ of length $X_n$. Then define

$$R(t) := \text{accumulated reward up to time } t$$

$$= \sum_{n=1}^{N(t)} R_n + (\text{things in last cycle})$$

We discuss the reward per unit time and have via the Strong Law of Large numbers

$$\frac{R(t)}{t} \to \frac{\mathbb{E}(R)}{\mathbb{E}(X)}$$

with probability 1 of convergence.
Here in the present problem, define

$$R_n = (X_n - a)\mathbb{1}(X_n > a) = (X_n - a)_+$$

where $x_+ := \max(x, 0)$ Then we see that $R(t)$ is the total length of times $s \leq t$ : $A_s > a$. This implies

$$\frac{R(t)}{t} \to \frac{\mathbb{E}R_1}{\mathbb{E}X_1}$$

by the Renewal Reward Theorem. So we argue this is the limit probability

$$\mathbb{P}(A_\infty > a) = \lim_{t\to\infty} \mathbb{P}(A_t > a)$$

This limit exists via an argument of Ergodic theory. Then we have the formula

$$\mathbb{P}(A_\infty > a) = \frac{\mathbb{E}(X_1 - a)_+}{\mathbb{E}X_1}$$

However, we can do a little better than this. We claim that $A_\infty$ has probability density

$$\mathbb{P}(A_\infty \in da)/da = \frac{\mathbb{P}(X_1 > a)}{\mathbb{E}X_1} \qquad (a > 0)$$

This is basically the tail-integral formula for $\mathbb{E}X_1$ and $\mathbb{E}(X_1 - a)_+$. Pitman draws us a picture to convince us of this fact.

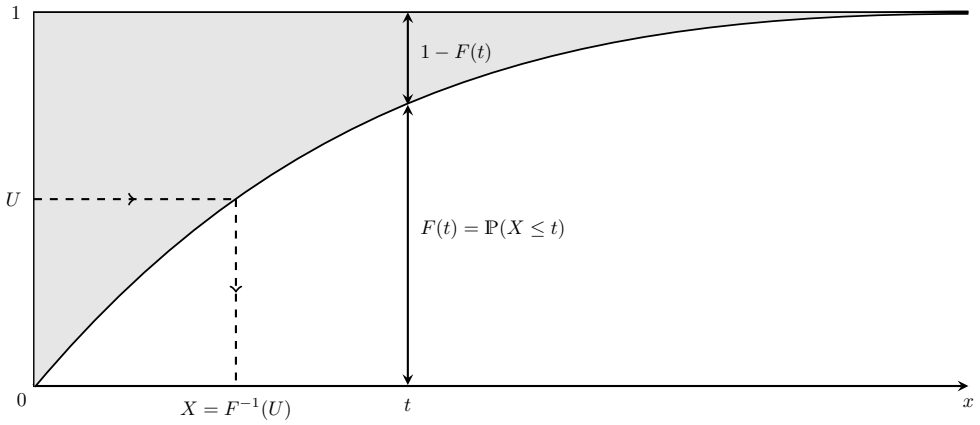



Figure 15.3: For a non-negative $X$, with CDF $F(x) = \mathbb{P}(X \leq x)$, the expectation $\mathbb{E}X = \int_0^\infty (1 - F(x))dx$ is the shaded area between the graph of $F(x)$ and level 1 as shown above. For a uniform random variable $U$ along the vertical, we can create $X$ as the inverse CDF of $U$. (Usual construction for simulation)

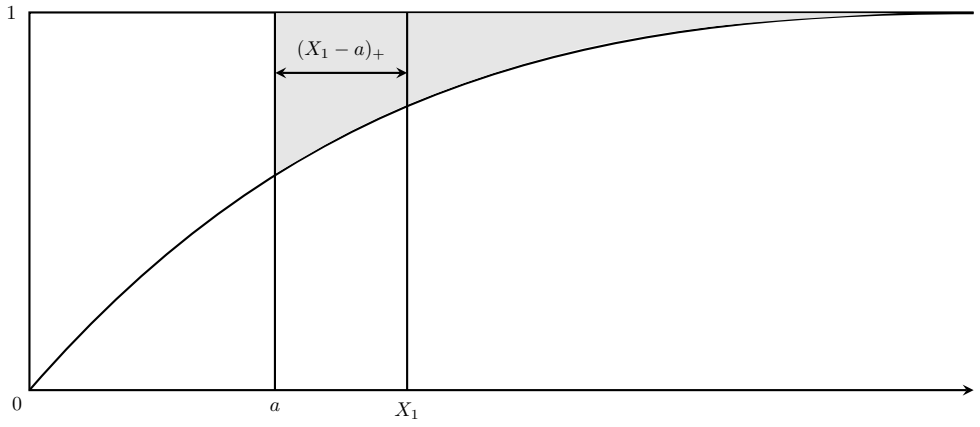



Figure 15.4: Realization of $\mathbb{E}(X_1 - a)_+$

To see where the mean is, Pitman reminds us of the tail integral formula (or integration by parts) for the mean

$$\mathbb{E}(X_1) = \int_0^\infty (1 - F(t))\ dt$$

$$= \int_0^\infty f(t)t\ dt.$$

This is the continuous analog of the tail sum formula in the discrete case.

Now we introduce a level $a$ and we ask what is $\mathbb{E}(X_1 - a)_+$. Convince yourself that by another application of the tail integral formula,

$$\begin{aligned}\mathbb{E}(X_1 - a)_+ &= \text{area to right of } a \text{ and above the CDF} \\ &\leq \mathbb{E}X_1.\end{aligned}$$

Bringing this into our previous formula, this gives

$$\mathbb{P}(A_\infty > a) = \frac{\mathbb{E}(X_1 - a)_+}{\mathbb{E}X_1} = \frac{\int_a^\infty (1 - F(t))\ dt}{\mathbb{E}X_1}$$

Now it is easy to take $-\frac{d}{da}$ to obtain the density

$$\frac{\mathbb{P}(A_\infty \in da)}{da} = -\frac{d}{da}\mathbb{P}(A_\infty > a) = \frac{1 - F(a)}{\mathbb{E}X_1} = \frac{\mathbb{P}(X_1 > a)}{\mathbb{E}X_1}$$

Notice that conclusion requires application of 'tools of the trade' in terms of thinking about tail integrals to represent expectations in a different way.

### 15.1.2 A Quick Check

Take $X_1 \sim$ **Exponential** $(\lambda)$. Then $1 - F(t) = e^{-\lambda t}$ and $\mathbb{E}X_1 = \frac{1}{\lambda}$. Then our formula above gives

$$\frac{\mathbb{P}(A_\infty \in da)}{da} = \frac{e^{-\lambda a}}{1/\lambda} = \lambda e^{-\lambda a}$$

which is the familiar density of **Exponential** $(\lambda)$, agreeing with what we found more easily in this case using the Poisson counting variables. So the general formula specializes well in the Poisson case when we already know the result.

## 15.2 Exercise

Let $Z_t$ be the residual life at $t$, or equivalently the remaining lifetime of component in use at $t$.

(1) Find a formula for $Z_t$

(2) Describe the limit distribution of $Z_t$ as $t \to \infty$

**Solution.** To arrive at the formula, we draw a picture of the process (see next page)

Then we should expect

$$Z_t = T_{N(t)+1} - t$$

and in the long-run, we know that the limiting probability should be represented by the fraction of time that the process is above a level.

Over a long stretch of time, the difference between the time $A$ is above a level and the time $R$ is above a level comes only comes from the current cycle. So we try

$$\frac{\mathbb{P}(Z_\infty \in dz)}{dz} = \frac{\mathbb{P}(X_1 > z)}{\mathbb{E}X_1}$$

and argue that limit distributions are the same. This is nontrivial but is true by application of the Renewal Reward Theorem. Pitman notes that we should also see Durett 3.3, which looks into the joint distribution:

$$\lim_{t\to\infty} \mathbb{P}(A_t > a, Z_t > b) = \frac{\mathbb{P}(X_1 > a + b)}{\mathbb{E}X_1}$$

which gives a generalization for two variables by the same token of the Renewal Reward Theorem. Note that setting $a = 0$ or $b = 0$ the two variable formula specializes as it should.

□

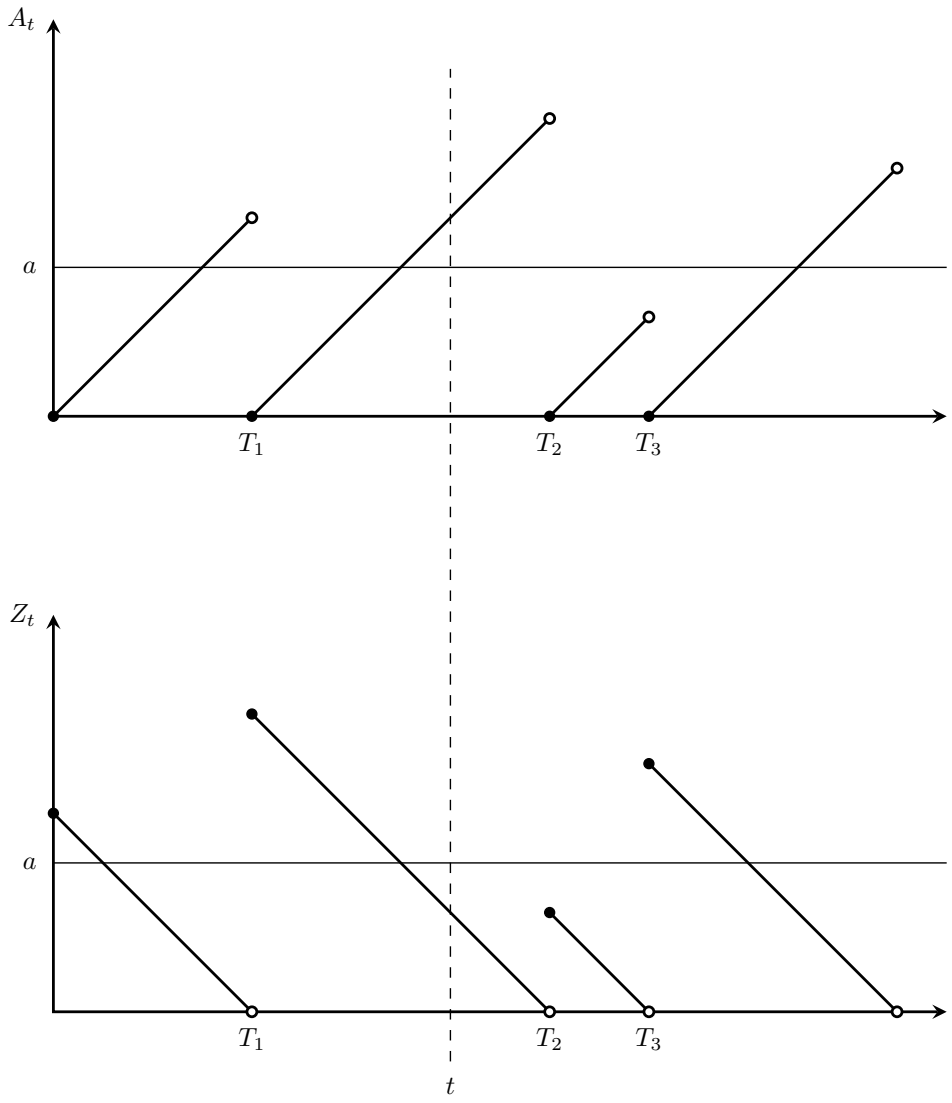



Figure 15.5: We superimpose graphs of the age of a component at time time $t$, $A_t$ and its corresponding residual life at time $t$, $Z_t$.

## 15.3 Queueing Theory: Terminology

This is David Kendall's notation from 1953. Our formalism is the notation of the form

$$A/B/C$$

where $A$ describes the input, $B$ the service time distribution, and $C$ is the number of servers. That is,

$$A \in \{M, G\}$$

where $M$ is Markov meaning **Poisson** $(\lambda)$ for some $\lambda > 0$ and $G$ or $GI$ is a General Input or Renewal Process Input, (the $I$ is often regarded as indicating independence) and $\lambda$ is the arrival rate, or equivalently the inverse mean $\lambda = 1/\mathbb{E}X$ where $X$ is the generic inter-arrival time.

$$B \in \{M, G\}$$

where $M$ is Markov or **Exponential** $(\mu)$ service and $G$ is the general service time distribution. Note that $\mu$ is the rate of service.

$$C \in \{1, 2, 3, \ldots, \infty\}$$

is the number of servers. See text and `https://en.wikipedia.org/wiki/Kendall%27s_notation` for further discussion.

### 15.3.1 Main Examples

(1) We've seen before in the satellite problem "$M/G/\infty$". The satellites are put up at **Poisson** $(\lambda)$ times, where each satellite lives some lifetime $(G)$, and there is no limit to the number of satellites in orbit. In the text, there is a simple analysis of this model by Poisson tricks in Theorem 2.13.

(2) We can also consider $M/G/1$ Queues, with **Poisson** $(\lambda)$ input, 1 server, any service time distribution, and assume that service times are iid, independent of arrival times. See text Section 3.2.3.

Let $Q_t$ be the queue length process, where $S_n$ is the $n$th service time and $T_n$ is the $n$th arrival time.

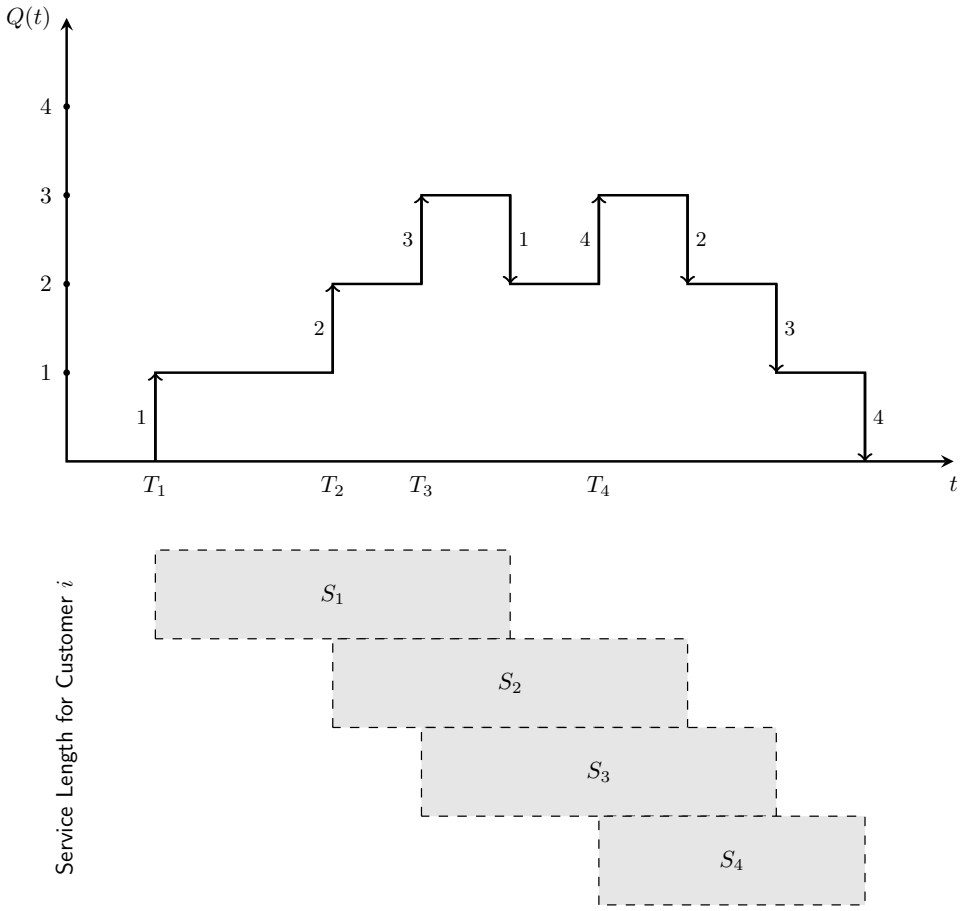



Figure 15.6: Depiction of a $M/G/1$ FIFO queue process with it's underlying service lengths.

There's a notion of 'Queue-Discipline'. We consider FIFO = First In First Out, and LIFO = Last In First Out. Considering queues, we are typically interested in stability of the queue, average wait times, or questions that do not depend on queue discipline, so we will not pursue this in detail. All the 1 server queues are stable if and only if the arrival rate $\lambda$ is less than the service rate $\mu$ (that is, $\lambda < \mu$). This gives positive recurrence. The case of equality $\lambda = \mu$ gives null recurrence. This means that the queue clears out eventually, but the expected time for this event to return is infinite.

## 15.4 Little's Formula

This is a famous formula, relating mean rates. Take $\lambda$ to be the arrival rate in a stable queue ($i.e. \lambda < \mu$), and let $L$ be the long-run average queue length, and let $W$ be the stationary state mean waiting time (not a random variable, in queue + service) per customer.

$$L = \lambda W$$

The text presents a long-run cost argument (Durett p. 131), and we give a similar treatment.

In this picture, we have

$$L = \lim_{t \to \infty} \frac{1}{t} \int_0^t Q_s \, ds$$

where $Q_s$ is the number of customers in the system at time $s$. This limit exists by a law of large numbers argument.

Our original picture is the composition of this sum of bars. The integral is the sum of lengths of these bars (size 1 for each customer) over time. Then these bars are being initiated at rate $\lambda$.

# LECTURE 16

# Continuous Time Markov Chains, Part 1

### Class Announcements

We're moving onto chapter 4, Continuous Time Markov Chains. There are elements from discrete-time Markov chains, some elements of Poisson processes, and some elements of renewal theory. Beyond that we'll cover Brownian motion and Gaussian Processes.

## 16.1 Continuous Time Markov Chains

The idea here is that we want to have a model for a continuous time process

$$(X_t,\ t \geq 0)$$

where $t$ is a now a continuous nonnegative variable and the range of $X_t$ is some countable state space. We say this process has the Markov property with a *transition semi-group* $(P_t, t \geq 0)$ if

$$\mathbb{P}\Big(X_{s+t} = y \mid X_s = x,\ \text{any event determined by } (X_n, 0 \leq u \leq s)\Big) = P_t(x, y)$$

for a future target $y$, given that at time $s$, we arrive at state $x$. See depiction on next page.

To make sense of this, we need some things. Let us start with a finite $S$ (or countably infinite) on $\{0, 1, 2, \dots\}$. Then for $x, y \in S$, for each fixed $t \geq 0$, $P_t(x, y)$ should be a transition probability matrix. There is a consistency issue as we vary $s$ and $t$ in this specification. Recall that in discrete time, we have that $P_1 = P$ is one transition matrix, then we would make $P_n = P^n$, which happens if and only if $P_n P_m = P_{n+m}$.

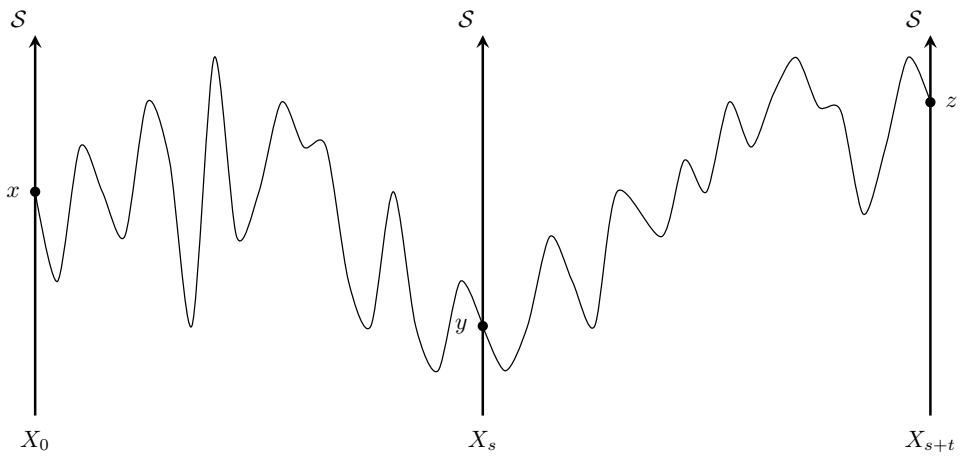



Figure 16.1: Depiction of a continuous time Markov Chain moving from state $x$ to $z$ through an intermediate state $y$.

This must be so by the addition rule of probability. Now by this same discussion in continuous time, let's use $\mathbb{P}_x$ as usual to specify $\mathbb{P}_x(X_0 = x) = 1$. This simply forces the initial state to be $x$. More or less by definition, or assumption, we have

$$\begin{aligned} P_{s+t}(x,z) &= \mathbb{P}_x(X_{s+t} = y) \\ &= \sum_{y \in S} \mathbb{P}_x\left(X_s = y, X_{s+t} = z\right) \quad \text{(conditioning on } X_s\text{)} \\ &\vdots \\ &= \sum_{y \in S} P_s(x,y) P_t(y,z) \end{aligned}$$

From this, we see that analogously to discrete time Markov chains, we must have the identity

$$P_{s+t} = P_s P_t = P_t P_s$$

which we call the semigroup property of the family of transition matrices driving a continuous time Markov chain. The second equality follows from the first equality, under the fact that $s + t = t + s$.

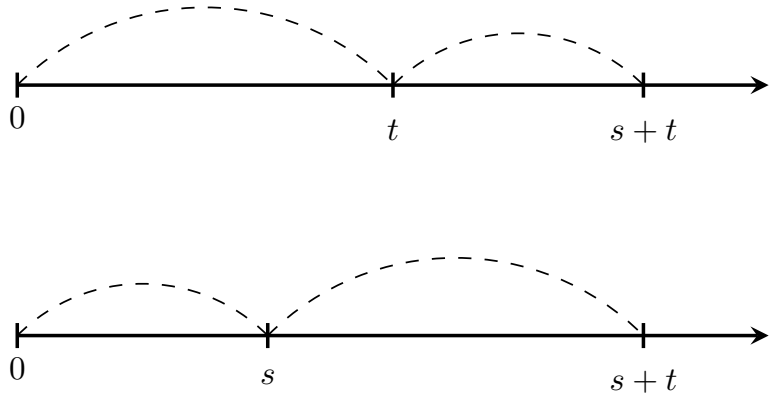



Figure 16.2: Realization of $P_{s+t} = P_s P_t = P_t P_s$.

We may ask, how do we create such a transition mechanism? There are two answers which end up more or less the same but from different perspectives.

### 16.1.1 Derivation 1

First, we can view this via analysis and see that the only candidate for a mechanism like this is to take

$$P_t = e^{Qt},$$

for a suitable matrix $Q$. Now the natural question is to ask how we compute this.

$$e^{Qt} = \sum_{n=0}^{\infty} \frac{(Qt)^n}{n!} = \sum_{n=0}^{\infty} \frac{t^n}{n!} Q^n$$

via the Taylor series expansion. Now this is well defined whenever $Q$ is a finite matrix or an infinite matrix with bounded elements. That is, we have some finite bound $\lambda$ where $|Q(i,j)| \leq \lambda < \infty$ for all $i, j$. Additionally, we can approach this the compound-interest way

$$e^{Qt} = \lim_{n \to \infty} \left( I + \frac{Qt}{n} \right)^n$$

which we can check is true easily by binomial expansion. Moreover, notice that (by manipulation of power series, just as for exponentials of real or complex numbers)

$$e^{Q(s+t)} = e^{Qs+Qt} = e^{Qs} \cdot e^{Qt}$$

and so if we set $P_t := e^{Qt}$, then we see that $P_{s+t} = P_s P_t$ as desired. It remains to discuss what sort of $Q$ give transition matrices $P_t$.

### 16.1.2 Derivation 2

Alternatively, we can view this via an explicit construction of the process $(X_t, t \geq 0)$ from suitable building blocks. A good start to do this to consider some transition matrix $U$ (from Durett example 4.1) for a discrete time Markov chain, say $Y_0, Y_1, Y_2, \ldots$ with transition matrix $U$. Consider a Poisson process $(N(t), t \geq 0)$ of rate 1, with iid **Exponential** (1) spacings. Additionally, we assume that the Poisson point process (the times of jumps) is independent of the discrete chain (where the jumps are going). The Poisson process says when the continuous time Markov chain should (try) to move. Now we write down and consider the process $X_t$ which is by definition

$$X_t := Y_{N(\lambda t)}$$

where $\lambda > 0$ is a rate parameter, so $(N(\lambda t), t \geq 0)$ is a Poisson process with rate $\lambda$. We can see explicitly $N(\lambda t) \sim$ **Poisson** $(\lambda t)$. We claim that we get a continuous time Markov chain like this. In fact, we can make every chain with finite state space like this, modulo some technical assumptions (e.g. that the paths of the chain are step functions). Each of the spacings between jumps of the PPP with rate $\lambda$ is **Exponential** $(\lambda)$. Now there is one minor complication in that it can happen that the continuous time chain does not move at one (the case where the discrete chain has a self-transition). But otherwise the continuous time chain jumps at the time of each point of the PPP($\lambda$). See depiction on next page.

### 16.1.3 Aside: Sum of Exponentials with Random Geometric Index

Let $\mathcal{E}_1, \mathcal{E}_2, \ldots$ are iid **Exponential** $(\lambda)$ with mean $\frac{1}{\lambda}$ and $G(p) \sim$ **Geometric**$(p)$ on $\{1, 2, \ldots\}$ with mean $\frac{1}{p}$, then $\mathcal{E}_1 + \mathcal{E}_2 + \cdots + \mathcal{E}_{G(p)}$ is **Exponential** $(p\lambda)$. That the distribution is exponential can be seen by a Poisson thinning argument. Or by conditioning on $G(p) = n$ to compute the density, and summing out $n$. To see the rate $p\lambda$ is correct, condition on $G(p)$ and notice

$$\mathbb{E}\left(\mathcal{E}_1 + \mathcal{E}_2 + \cdots + \mathcal{E}_{G(p)}\right) = (\mathbb{E}\mathcal{E}_1)\left(\mathbb{E}G(p)\right) = \frac{1}{\lambda}\frac{1}{p} = \frac{1}{p\lambda}$$

This fact about geometric sums of exponential variables explains why the Poisson construction ensures that the holding time in any particulat state, of the continuous time chaim made by timing the steps of a discrete chain with a Poisson process, must be exponentially distributed. This can also be seen as a general property of continuous time chains, by application of the Markov property to establish that the holding time of a state must have the memoryless property, hence be exponential.

Returning to our previous argument, let us find $Q$ for the Poisson construction

$$\begin{aligned} P_t(x,y) &= \mathbb{P}_x\Big(Y_{N(\lambda t)} = y\Big) \\ &= \sum_{n=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^n}{n!} U^n(x,y) \\ &= e^{-\lambda t} \sum_{n=0}^{\infty} \frac{(\lambda t U)^n}{n!}(x,y) \\ &= \left[e^{-\lambda t} \cdot e^{\lambda t U}\right](x,y) \\ &= e^{Qt} \end{aligned}$$

where

$$Q := \lambda(U - I)$$

and $I = P^0$ is the identity matrix. Here, $\mathbb{P}_x$ has $Y_0 = x$ and $N$ in a Poisson point process is independent of $Y$, and the $(x, y)$ at the right is the evaluation of a matrix at row $x$ and column $y$.

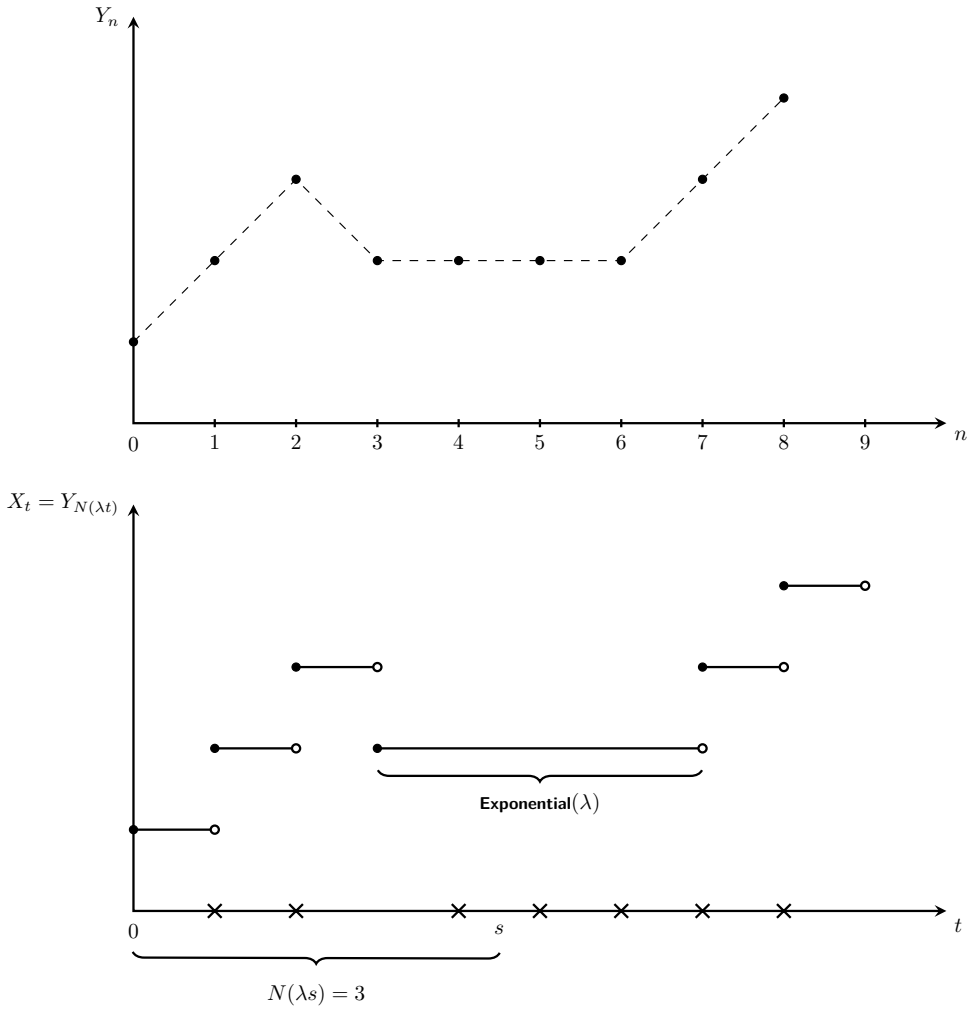



Figure 16.3: Realization of $(X_t, t \geq 0)$ where $X_t := T_{N(\lambda t)}$ and its embedded discrete time Markov Chain $Y_n$. Observe the *holding times* are exponentially distributed, even for successive transitions i.e. moving from $x$ to $x$. This construction is presented in Durret Example 4.1.

## 16.2 Simple Example 1: Boring Discrete Markov Chain

Take $Y_n := n$ for $n = 0, 1, 2, \ldots$, which is the (very boring and simple) deterministic Markov chain that increases by 1 at each time. Then the discrete time one step transition matrix is

$$U(x, y) = \mathbf{1}(y = x + 1)$$

This gives

$$X_t = Y_{N(\lambda t)} = N(\lambda t)$$

which is simply a Poisson process with rate $\lambda$.

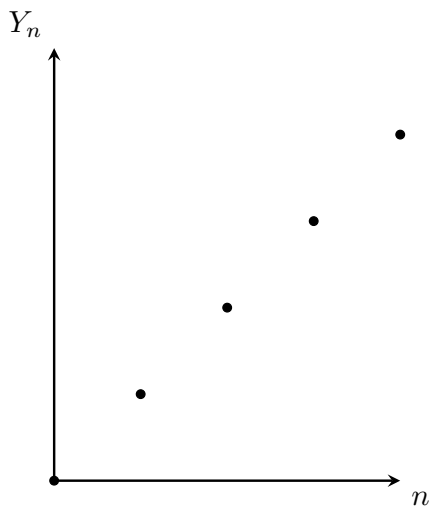



Figure 16.4: Realization of $Y_n := n$ for $n = 0, 1, 2, \ldots$.

## 16.3 Simple Example 2: Compound Poisson Process

Consider $Y_n = \Delta_1 + \Delta_2 + \cdots + \Delta_n$ for some iid $\Delta_i$. Then add a Poisson number of these terms:

$$X_t = \Delta_1 + \Delta_2 + \cdots + \Delta_{N(\lambda t)}$$

which is called a compound Poisson process. One instance if the Negative Binomial process as per Homework 7. Note that probability generating functions are very handy for this example, as we have seen in our homework.

## 16.4 Interpretation of Rate Matrix $Q$

We agree that we are creating our semigroup by the matrix formalism in that

$$P_t = e^{Qt}$$

Now this implies that something nice happens when we differentiate

$$\implies \frac{d}{dt}P_t = Qe^{Qt} = e^{Qt}Q = QP_t = P_tQ$$

where via Taylor's expansion,

$$\frac{d}{dt}\left(I + Q + \frac{Q^2t^2}{2!} + \frac{Q^3t^3}{3!} + \cdots\right) = 0 + Q + \frac{Q^2}{1!}t + \frac{Q^3t^2}{2!} + \cdots$$

$$= Q\left(I + \frac{Qt}{1!} + \frac{Q^2t^2}{2!} + \cdots\right)$$

$$= Qe^{Qt}$$

Notice that $Q$ commutes with every power of itself, so we get (1) above, which are the Kolmogorov, backwards, and forwards equations respectively.

Formally, $(P_t, t \geq 0)$ is defined by a matrix system of differential equations, given by the set of Kolmogorov's differential equations $P_0 = I$ and $\frac{d}{dt}P_t = QP_t = P_tQ$. This is fairly abstract, and we will focus on the interpretation of the $Q$ matrix. From the formalism and equations introduced above, we have

$$Q = \frac{d}{dt}P_t\Big|_{t=0^+} = \lim_{t \downarrow 0}\left(\frac{P_t - I}{t}\right)$$

which we can perform for each entry one at a time. Now, notice that for a column vector $\mathbb{1}$ of all entries 1,

$$P_t\mathbb{1} = \mathbb{1}$$

Hence,

$$\frac{d}{dt}P_t\mathbb{1} = \frac{d}{dt}\mathbb{1} = 0$$

the zero vector. This implies that

$$Q\mathbb{1} = \lim t \downarrow 0 \frac{d}{dt}\left(P_t\mathbb{1}\right) = 0$$

**Remark:** Now this tells us that all row sums of a rate matrix $Q$ must be identically zero. Additionally, the off-diagonal entries (for $i \neq j$) are

$$Q(i,j) = \lim_{t \to 0}\frac{P_t(i,j)}{t} \geq 0$$

because this is the limit of non-negative things. Hence the main diagonal entries $Q(i,i)$ MUST all be negative to compensate so that the row sums are all 0. Now let us define

$$\lambda_i := -Q(i,i) \geq 0$$

which has the interpretation that the holding time of the chain in state $i$ is **Exponential** $(\lambda_i)$. Then for $i \neq j$, the

$$\frac{Q(i,j)}{\lambda_i}$$

are transition probabilities when the chain jumps.

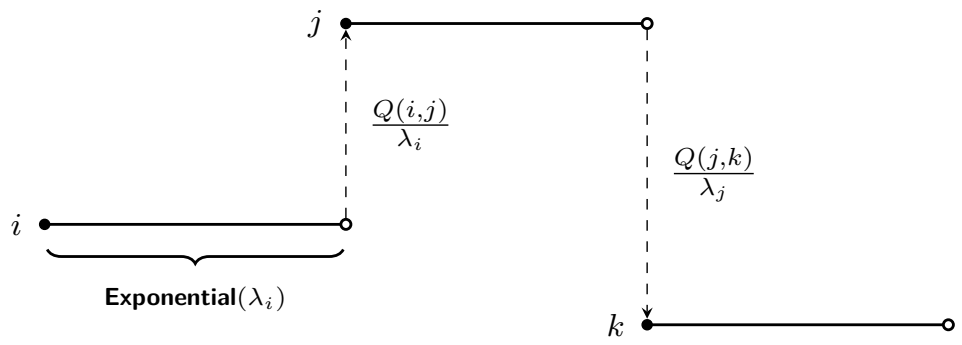



Figure 16.5: Depiction of a continuous Markov chain jumping between states through the $Q$ mechanism.

We keep moving around in the state space, and every time we hold, we hold for an exponential time. This can be read from the rate matrix $Q$. This is a high level view of the continuous time Markov chain theory. See text for many examples.

# LECTURE 17

# Continuous Time Markov Chains, Part 2

This lecture is fairly informal. We continue our study of the building blocks for Continuous Time Markov Chains, how they are assembled, and their properties. A homework problem illustrate this motivation. Durrett problem 3.12.

## 17.1 Repairman Problem

From this week's homework, recall the cute problem regarding a repairman who completes tasks according to an **Exponential** $(\mu)$ distribution. However, he is interrupted by a PPP$(\lambda)$. Recall that in past lectures, when constructing a Poisson Point Process, we considered more than just Poisson points on a line; instead we consider points on a strip of the plane. We relate our repairman problem to a Poisson point process with rate 1 per area.

By our construction (and familiar properties) of Poisson point processes, we have a few facts to take from our depiction on the next page.

- The $\times \times \times \times \cdots \times \times$ marks on the timeline are a Poisson point process with rate $\lambda$. We drew the picture to make it so, projecting a $1 \times \lambda$ rectangle down to the timeline with width 1.
- The $\otimes \otimes \otimes \cdots \otimes \otimes$ similarly are a Poisson point process with rate $\mu$.
- The $\times \times \cdots$ and $\otimes \otimes \cdots$ points are independent because in the planar Poisson point process, the areas constructed from each of the $\lambda$ and $\mu$ strips will be disjoint. Precisely, these strips are $[0, \infty) \times [0, \mu]$ and $[0, \infty) \times [\mu, \lambda + \mu]$.
- Moreover, consider the discrete sequence of marks $\times \times \times \times \times \times \otimes \times \times \times \times \otimes$, by throwing away the timing and only looking at the sequence. These are independent **Bernoulli**$(\times/\otimes)$ trials, where $\mathbb{P}(\times) = \frac{\lambda}{\lambda+\mu}$ and $\mathbb{P}(\otimes) = \frac{\mu}{\lambda+\mu}$.
- In the continuous model, let $T_\times$ be the continuous time to the first $\times$, so that $T_\times \sim$ **Exponential** $(\lambda)$. Let $T_\otimes$ be the continuous time to the first $\otimes$, so that $T_\otimes(\mu)$. This is a race between independent exponentials (as they are disjoint

parts of a PPP). Either by integration, or because the vertical height of the first Poisson point is uniform on an interval of length $\lambda + \mu$, and the event $(T_\times < T_\otimes)$ occurs when this vertical height falls in a subinterval of length $\lambda$,

$$\mathbb{P}(T_\times < T_\otimes) = \frac{\lambda}{\lambda + \mu}.$$

In better form, let $T_\lambda$ and $T'_\mu$ be two independent **Exponential** $(\lambda)$ and **Exponential** $(\mu)$ variables. Then $\mathbb{P}(T_\lambda < T'_\mu) = \frac{\lambda}{\lambda+\mu}$, and the event $(T_\lambda < T'_\mu)$ is independent of $T_\lambda \wedge T'_\mu := \min(T_\lambda, T'_\mu)$.

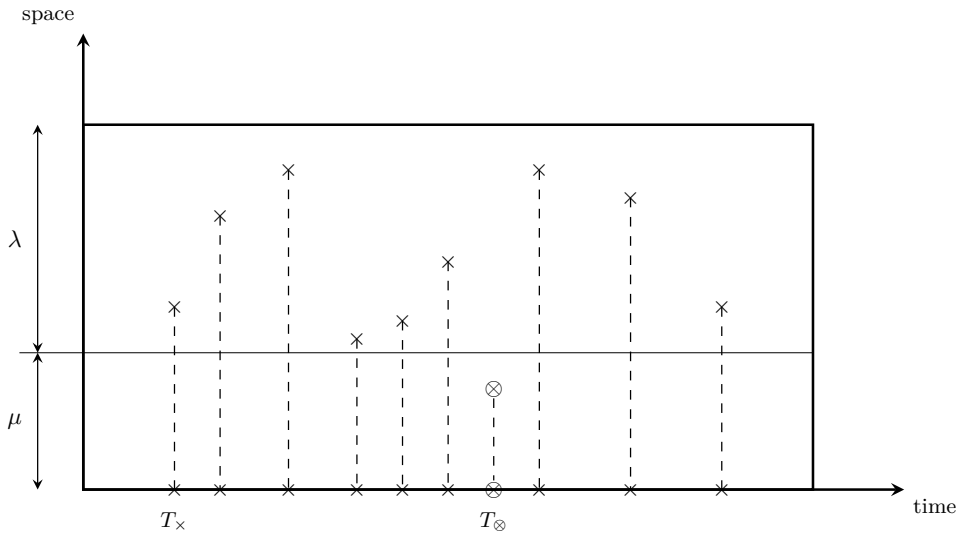



Figure 17.1: Realization of the repairman problem. On the time axis, we have a merged PPP with rate $\mu + \lambda$. This realization is a distilled rendition from the depictions presented in lecture.

Now, bringing these ideas back to our homework problem, given that the repairman is interrupted 6 times, the waiting time until the first completion is still the first point of a **Poisson** $(\mu)$ process. Pitman gives that this picture of the Poisson point process makes this question a lot more trivial than it may seem. The last spacing before $T_\otimes$ is **Exponential** $(\lambda + \mu)$. This is a race between exponentials, so given that the repairman finishes the task, he must have done so at a faster rate.
A good question asked in class: What happens in this model if interruptions are a renewal process? Would the rate of completions stay the same? The above argument can be adapted as follows. The answer is yes, the repair rate per unit time stays the same, no matter what the interruption process. In fact, you can condition on the times of interruptions, say

$$0 < t_1 < t_2 < \cdot$$

so these are now fixed times, e.g. any finite list of fixed times, or any infinite list increasing to infinity, and the set of times when the repairman completes a job is the set of points of a PPP with rate $\mu$. Here is one way to see this. By the memoryless property of the exponential, whenever there is an interruption, in terms of time to finishing the next job, the repairman is indifferent between continuing to work on the current job, or starting a new one. Either way, the time till the next completion will be **Exponential** $(\mu)$. Or, look at the time $T$ to first completed job and consider the probability $\mathbb{P}(T > t)$. If say $0 < t_1 < t_2 < t$ this event occurs iff the first job was not completed in time $t_1$ the next job was not completed in time $t_2 - t_1$, and the following job was not completed in time $t - t_2$. So

$$\mathbb{P}(T > t) = e^{-\mu t_1} e^{-\mu(t_2 - t_1)} e^{-\mu(t - t_2)} = e^{-\mu t}$$

hence $T$ is **Exponential** $(\mu)$. The same argument can be continued to show that regardless of the choice of interruption times $t_i$, the times of completion of jobs are the points of a PPP$(\mu)$. This problem is very instructive for understanding the implications of the memoryless property of the exponential spacings between arrivals in a PPP with constant rate.

## 17.2 Further Remarks

Moreover, looking at $T_\otimes$ as the time to first $\otimes$ as $T_\otimes = \mathcal{E}_1 + \mathcal{E} + \cdots + \mathcal{E}_N$, where the $\mathcal{E}_i$ are the **Exponential** $(\lambda + \mu)$ spacings before the points either $\times$ or $\otimes$. We know

$$\mathbb{P}(N = n) = \mathbb{P}(\overbrace{\times \times \times \cdots \times}^{n-1} \otimes) = \left(\frac{\lambda}{\lambda + \mu}\right)^{n-1} \left(\frac{\mu}{\lambda + \mu}\right)$$

which implies $N$ is **Geometric**$(\frac{\mu}{\lambda+\mu})$. In this model, we have that $N$ is independent of $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \ldots$. We see then that the **Exponential** $(\mu)$ time $T_\otimes$ is represented in this picture as the sum of a **Geometric**$(\frac{\mu}{\lambda+\mu})$ number of copies of iid **Exponential** $(\lambda + \mu)$. Therefore, the poisson strip model offers an explanation and proof of the basic fact that if $\mathcal{E}_1, \mathcal{E}_2, \ldots$ are iid **Exponential** $(\nu)$ and $N$ is independent of these with **Geometric**$(p)$ on $\{1, 2, 3, \ldots\}$, then

$$\mathcal{E}_1 + \mathcal{E}_2 + \cdots + \mathcal{E}_N \stackrel{d}{=} \textbf{Exponential}\,(\nu p)$$

where $\nu = \lambda + \mu$, and $p = \frac{\mu}{\lambda+\mu}$. You should also be able to check this by a density computation, by conditioning on $N$ and using the fact that the sum of $n$ i.i.d. exponentials $(\nu)$ has the gamma$(n, \nu)$ density.

## 17.3 Analogies Between Discrete and Continuous Time Markov Chains

Suppose $Y_n$ for $n = 0, 1, 2, \ldots$ is a discrete time Markov chain with transition matrix $P$ and $Y_0 = y$ for some $y$ in state space $S$. Let

$$H_y := \min\{n \geq 1 : Y_n \neq y\}$$

That is, $H_y$ is the *holding time* in state $y$.

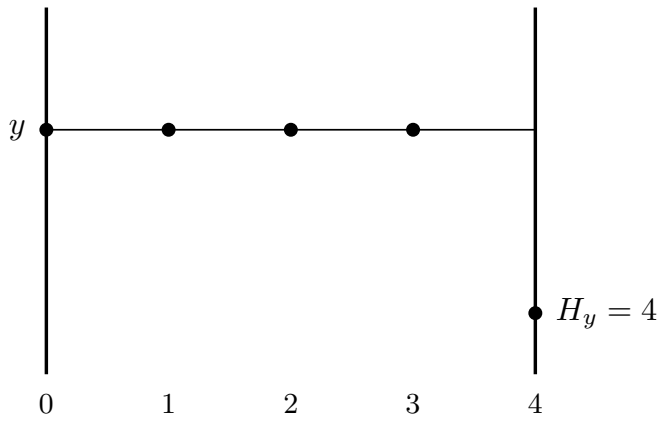



Figure 17.2: Depiction of $H_y = 4$, where $(Y_n)_{n\geq 0}$ is a discrete time Markov chain.

We notice that by definition, $H_y \in \{1, 2, 3, \dots\}$. Then except in trivial cases with $P(y, y) = 1$, we have

$$\mathbb{P}(H_y < \infty) = 1$$

Then assuming this is not absorbing (that is, $P(y, y) < 1$), we have that

$$\mathbb{P}(H_y = n) = [P(y, y)]^{n-1}\Big(1 - P(y, y)\Big), \quad \text{for } n = 1, 2, \dots$$

where this bracketed factor denotes the power of the single element and not the matrix power. That is, $H_y \sim \textbf{Geometric}(1 - P(y, y))$.

Now, let $J_y := Y_{H_y}$, which we know for sure is not equal to $y$. Then we may look into some other state $z$ which has

$$\mathbb{P}(J_y = z) = \frac{P(y, z)}{1 - P(y, y)}$$

which holds for $z \neq y$. Also, you easily check that $H_y$ and $J_y$ are independent.
Bringing this foward into continuous time, take $(X_t, t \geq 0)$ to be a continuous time Markov chain with the semigroup of transition matrices $(P_t, t \geq 0)$ where $P_t = e^{Qt}$. Recall from last lecture that we found

$$Q = \lim_{t \downarrow 0} \frac{(P_t - I)}{t}$$

which is the matrix of transition rates. We may ask similar things as before in terms of holding times. Let $y$ be a state. We write

$$H_y := \inf\{t > 0 : X_t \neq y\}$$

The idea here is that we are holding for a while and then jumping to another state. Then we have the following analogies from discrete time that carry over to this continuous time discussion. In continuous time, we want to think about $(1-P(y,y))$ in a small amount of time. Now because this converges to 0, we want to normalize (divide by) $t$. Looking at the formula for $Q$ above, we notice that

$$H_y \sim \textbf{Exponential}\,(\lambda_y) \text{ where } \lambda_y := -Q(y,y)$$

Let $J_y := X_{H_y}$, which is the state that we jump to in the picture above. From the interpretation of relative jump rates as probabilities, we have

$$\mathbb{P}(J_y = z) = \frac{Q(y,z)}{\lambda_y}$$

This is partitioned in a way that is completely analogous to the discete case. Moreover, $H_y$ and $J_y$ are independent in the continuous time setup, as in discrete time. These are basic intepretations of the transition rates of a continuous time Markov chain in terms of its successive holds and jumps.

## 17.4 Corollary

We have a very simple way of simulating a continuous time Markov chain. If we start at $y$, for each state $z \neq y$, we have an exponential variable $\mathcal{E}_{yz}$ with rate $Q(y,z) \geq 0$. We assume these are variables are independent, and let

$$H_y := \min_z \mathcal{E}_{yz}, \quad J_y := \arg\min_z \mathcal{E}_{yz}$$

Then $(X_t, t \geq 0)$ iid and $X_0 = y$ is constructed as $X_t = y$ for $0 \leq t < H_y$ and $X_{H_y} = J_y \neq y$. Then given $J_y = z$, we iterate, with a fresh set of independent exponentials $\mathcal{E}_{zw}$, and so on. It is a fact that in the race of exponentials, the time to completing the race is independent of the winner of the race.

## 17.5 Examples

A very simple example is the two-state Markov chain, which is also known as the alternating exponential renewal process. Let these states be 0 and 1 with

$$Q = \begin{bmatrix} Q(0,0) & Q(0,1) \\ Q(1,0) & Q(1,1) \end{bmatrix} =: \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}$$

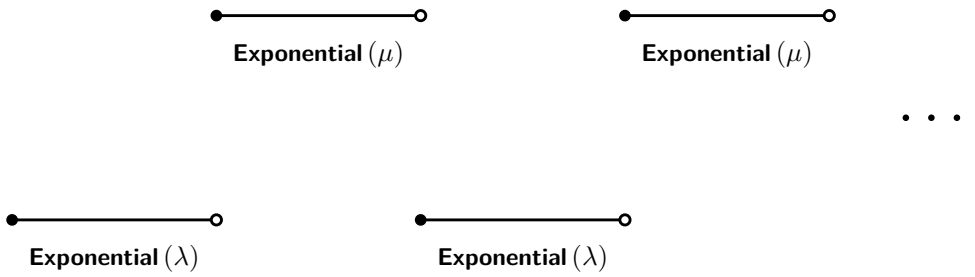



Figure 17.3: Realization of an alternating renewal process.

The picture illustrates why this is also called an alternating exponential renewal process, as we are simply flipping between the two rates. Now with our same terminology from before,

$$P_t(x, y) = \left[e^{Qt}\right](x, y)$$

which is given by an explicit formula in Durett. We have done this in our homework in the discrete case. There is a similarly simple formula which can be found via the differential equations quite easily. Now, we may ask what happens in the limit. Recall that a large rate means that an event happens fast. Imagine the case where $\lambda$ is small and $\mu$ is large. Then this gives a typiccal picture with short holds in state 1 and long holds in state 0:

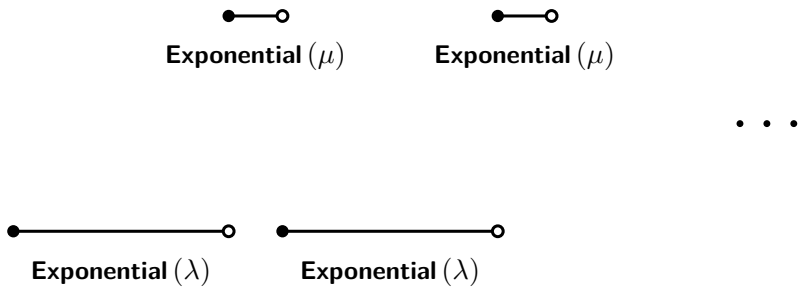



Figure 17.4: Realization of an alternating renewal process, where $\lambda > \mu$.

In terms of renewal theory, we have an iid cycle every time we drop back down to 0. Appealing to the Renewal Reward Theorem, we compare the mean time for the above state against the mean time for the cycle:

$$\lim_{t\to\infty} P_t(x, 1) = \frac{1/\mu}{1/\lambda + 1/\mu} = \frac{\lambda}{\lambda + \mu}$$

Similarly, for $P_t(x, 0)$, we have limit $\frac{\mu}{\lambda+\mu}$. When $\mu$ is much larger than $\lambda$, the chain

moves quickly from 1 to 0 and slowly from 0 to 1. So in the long run it is more likely to be found in state 0

# LECTURE 18

# Continuous Time Markov Chains, Part 3

### Class Announcements

We're currently on Durett's chapter 4, looking at continuous-time Markov chains. We have the following topics to cover (we have covered 4.1 and 4.2 mostly 4.2.1 is left for your reading):

- §4.1: Definitions and Examples
- §4.2: Transition Probability Functions (TPF) with $P_t = \exp(Qt)$
- §4.2.1: Branching Processes
- §4.3: Limit Behavior
- §4.3.1: Detailed balance
- §4.4: Exit distributions
- §4.5: Queues
- §4.6: Queueing Networks

Today we'll cover §4.3 and 4.3.1, and on Thursday we'll cover 4.4 and 4.5. We won't go over 4.6 explicitly as part of the syllabus. No final exam questions about 4.6.

## 18.1 Limit Behavior

We'll work with the theory for continuous-time Markov chains (CTMC). We've already done this for discrete time. For simplicity, today let's assume that the state space $S$ is finite.

### 18.1.1 Review of the Discrete Case

In the discrete case, for a transition probability matrix $P$, assuming that $P$ is irreducible, i.e.

$$\forall_{x,y}\exists_n : P^n(x,y) > 0$$

then

$$\exists \text{ a unique invariant } \pi : \pi P = \pi, \quad \pi\mathbb{1} = 1$$

Also, we always have that

$$\frac{1}{n}\sum_{k=1}^{n} P^k \to \Pi := \begin{pmatrix} \pi \\ \pi \\ \vdots \end{pmatrix} = \mathbb{1}\pi$$

Moreover, we have

$$\pi_x = \frac{1}{\mathbb{E}_x T_x}$$

by discrete renewal theory (Recall the Kac identity homework problem). We've generalized this very well in the continuous case via renewal theory. It is not immediately obvious that $\pi$ as above is the solution to the equations given by the quantifiers above, but this is proved in the text. Note this all works even in the periodic case. But to get ordinary convergence of the matrix powers $P^n$ to $\pi$ it is necessary to assume $P$ is aperiodic.

### 18.1.2 Extending to a Continuous Parameter Markov Chain

On a finite state space $S$, and Markov chains of the hold/jump type we have been dicussing, consider the following: **irreducible** means $\forall_{x,y} : P_t(x,y) > 0$, either for some $t > 0$, or for all $t > 0$, it makes no difference. To see why this is true, recall that we know $P_t = \exp(tQ)$ and that our chain may be constructed as $X_t = Y_{N(t)}$ where $Y$ is a suitable jumping chain in discrete time. Also, $N(t)$, the number of attempted jumps by time $t$, is a Poisson($\lambda$) process with some rate $\lambda \geq \max_i \lambda_x$, where $\lambda_x$ is the holding rate of state $x \in S$. To find the holding rates from the $Q$ matrix, we note

$$\lambda_x = -Q(x,x) = -\frac{d}{dt}P_t(x,x)_{t\to 0} = \lim_{t\downarrow 0}(1 - P_t(x,x))/t$$

and simply pick this off the diagonal entry of the $Q$ matrix. The holding rate is the initial rate of loss of probability from $x$. This $P_t(x,x)$ starts at $P_0(x,x) = 1$ and can only possibly decrease because probabilities are bounded above by 1. Hence the negative sign of the derivative at 0.

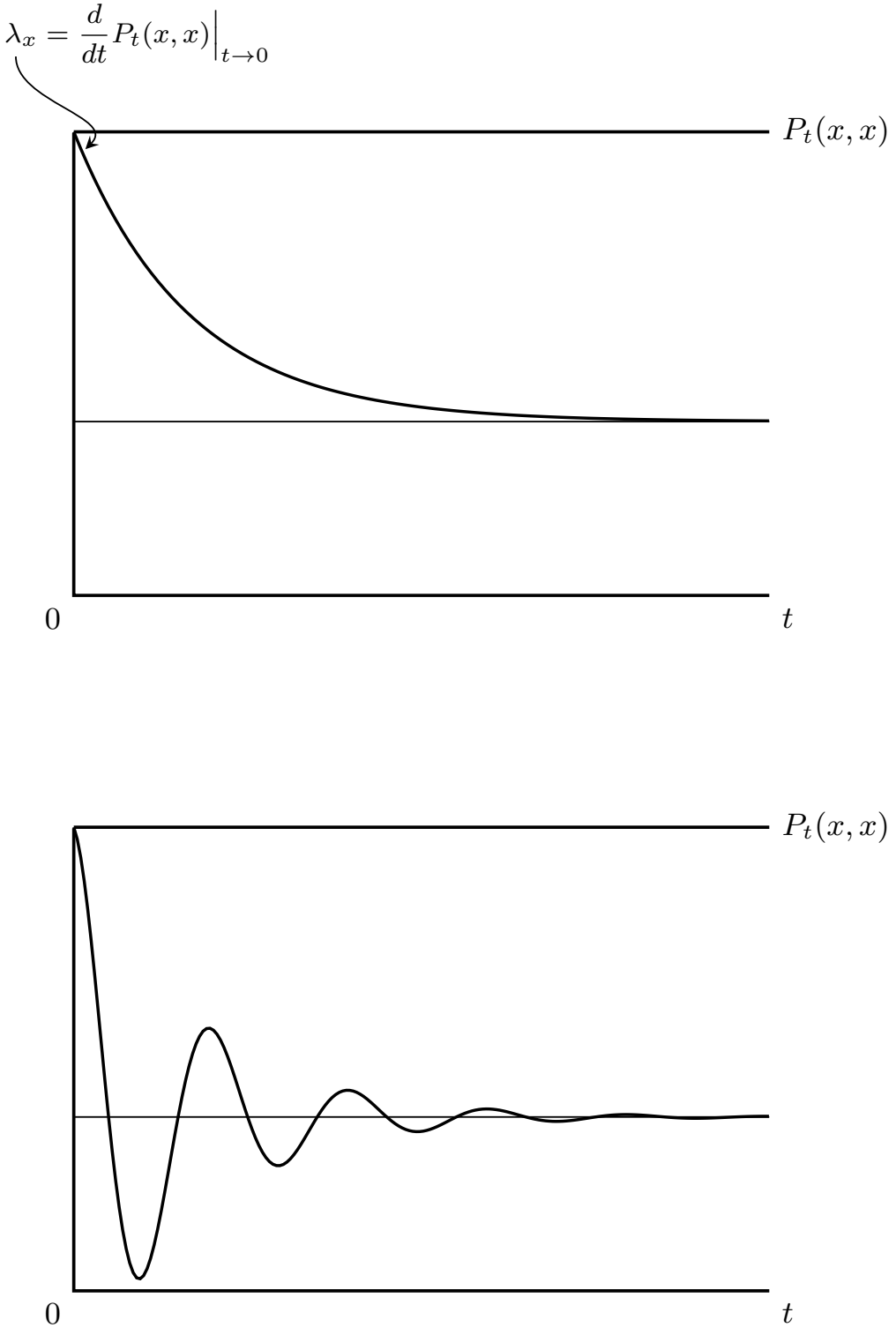



Figure 18.1: Two typical Realizations of $P_t(x,x)$ showing derivative at 0 and convergence to a limit $\pi(x)$ as $t \to \infty$.

Starting from $P_t(x,x) = 1$ at time $t = 0$, a common situation is to decrease to a limit, as in the first grap[h]. Alternatively, there may be dampening oscillation, but still with convergence to a limit. Imagine e.g. the jumping chain is clockwise rotation around $N$ points in a circle, with all rates $\lambda$ large. Initially $P_t(x,x)$ will decrease from 1 to quite close to 0 after time about $t = (N/2)/lambda$, when you expect the chain to be on the opposite side of the circle, then it may increase again to time $t = N/\lambda$ when you expect the chain to have done one rotation, and so on cyclically, but with damping caused by gradual diffusion and loss of memory of where the chain started, with limit $1/N$ as $t \to \infty$ as the unique invariant probability is obviously uniform. Notice that

$$\boxed{\lambda_x = -\frac{d}{dx} P_t(x,x)\Big|_{t\to 0}}$$

is also the **rate of exponential** hold at $x$. From the Poisson construction, we have $P_t = \exp(Qt)$ for a suitable $Q$, and more or less by definition (Taylor's theorem), we have

$$P_t = \exp(Qt) := I + Qt + \frac{Q^2t^2}{2!} + \frac{Q^3t^3}{3!} + \cdots$$

$$\implies \frac{d}{dt}\exp(Qt) = Q\exp(Qt)$$

$$\implies \frac{d}{dt}\exp(Qt)\Big|_{t\to 0} = Q.$$

### 18.1.3 An Old Example

Recall the silly matrix example

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

which makes the chain simply alternate deterministically between two states. It cannot be that this $P = P_t$ for a continuous time transition probability function $P_t$. In the Poisson jumping construction, take $U$ to be the transition probability matrix for $Y_0, Y_1, Y_2, \ldots$. Then from $X_t = Y_{N(t)}$, we have

$$\mathbb{P}_x(X_t = y) = \sum_{n=0}^{\infty} e^{-\lambda t}\frac{(\lambda t)^n}{n!} U^n(x,y) \tag{18.1}$$

Suppose $P_t(x,y) > 0$ for some $t > 0$, depending on $x$ and $y$. Then in the representation (18.1), we have that if the sum is $> 0$, then there exists some $n$ so that the $n$th term is $> 0$. Then $U^n(x,y) > 0$, because all the other facgtors are $> 0$. Hence $U$is irreducible. Essentially, being able to transfer between states in some amount of time implies that it is possible to do the same in any other amount of time, due to the independent exponential holds in various states. In summary, we learned that

$(P_t)$ irreducible implies $P_t(x, y) > 0$ for all $x, y$. Then by the discrete MC theory recalled above this implies that

$$P_{n\delta}(x, y) \to \pi(y)$$

for some $\pi(y)$ that does not depend on $x$. We can easily check this for e.g. $\delta = 1, \frac{1}{2}, 2, \frac{1}{4}, 4, \ldots$. As $\delta$ gets smaller, the limit assertion gets stronger, by consideration of subsequences. With some real analysis, we can generalize this further, to conclude that in fact $P_t(x, x)$ must have this same limit $\pi(x)$ as $t \to \infty$ in the usual way, not just along multiples of $2^{-k}$ for some $k$ as above. See text Theorem 4.8 on page 163. Now there remains the issue of how to find the limit. Pitman would like to emphasize one point. We do not have to have our hands on a formula for the solution of the differential equations. We have a theory to show that these limits exist and can be found even if we cannot solve the differential equations any more explicitly than by the matrix formula $P_t = \exp Qt$. That is, we know that for a finite irreducible chain, there is a unique stationary probability $\pi$. For each fixed $t > 0$ this is the unique solution of the system of linear equations

$$\pi P_t = \pi, \qquad \pi \mathbb{1} = 1 \tag{18.2}$$

which solution is necessarily such that $\pi(x) > 0$ for all $x$, hence a probability vector. In practice, usually we don't have a useful formula for $P_t$. So the expression for $\pi$ in terms of $Q$ is very useful. Noticing that (18.2) has no $t$ on the RHS, we differentiate this vector equation. Then we have

$$\frac{d}{dt} \pi P_t = \vec{0}$$

We can push the differential operator through the rows of the matrix

$$\begin{aligned} &\implies \pi \frac{d}{dt} P_t = 0 \\ \pi Q P_t = 0, &\forall\, t \geq 0. \end{aligned}$$

Now letting $t \downarrow 0$, we see that $P_t \to I$ (where almost no time has elapsed). Then this implies

$$\pi Q = 0$$

and conversely. So it is enough to solve the above equation with $\pi 1 = 1$. If there are $N$ states, this is $N + 1$ equations in $N$ unknowns. But $Q\mathbb{1} = \vec{0}$ is a linear relation among columns of $Q$ which shows $Q$ has rank at most $N - 1$. Replacing the last column of $Q$ by $\mathbb{1}$ as in the display above (4.26) of the text on page 165 is a trick for converting this into a system of $N$ equations in $N$ unknowns with a unique solution. Hence the modified $Q$ matrix is invertible, by linear algebra. In any case, to find $\pi$ you must solve the system of linear equations one way or another.

## 18.2 Detailed Balance

Recall that solving $\pi P = \pi$ in discrete case is often simplified greatly by solving the detailed balance equations ( related to reversible equilibrium). This is

$$\pi(x)P(x,y) = \pi(y)P(y,x), \forall_{x,y}$$

Detailed balance implies $\pi P = \pi$ in discrete time. The same notion works in continuous time. For the analog, look at

$$\pi(x)P_t(x,y) = \pi(y)P_t(y,x), \forall_{t\geq 0}$$

If we repeat the previous argument, taking derivatives with rwspect to $t$, this is true if and only if

$$\pi(x)Q(x,y) = \pi(y)Q(y,x), \forall_{x\neq y}$$

This has a very intuitive meaning. Interpret $\pi(x)Q(x,y)$ to be the long-run rate of transitions from $x \to y$. The factor $\pi(x)$ is the long run fraction of time in state $x$, and while in state $x$ the chain is jumping to $y$ according to the points of a PPP with rate $Q(x,y)$. See this week's homework for further discussion: These Poisson jump processes are even independent as $y$ varies. From the above equation, reversible equilibriun holds iff for all $x \neq y$ the long run rate of jumps per unit time from $x$ to $y$ is balanced by the long run rate of jumps from $y$ to $x$. Note that each of these point processes is a renewal counting process, because each time the process jumps into a particular state, it starts afresh from that state. Hence all the exponential holding times, and the independence of various holds and jumps.

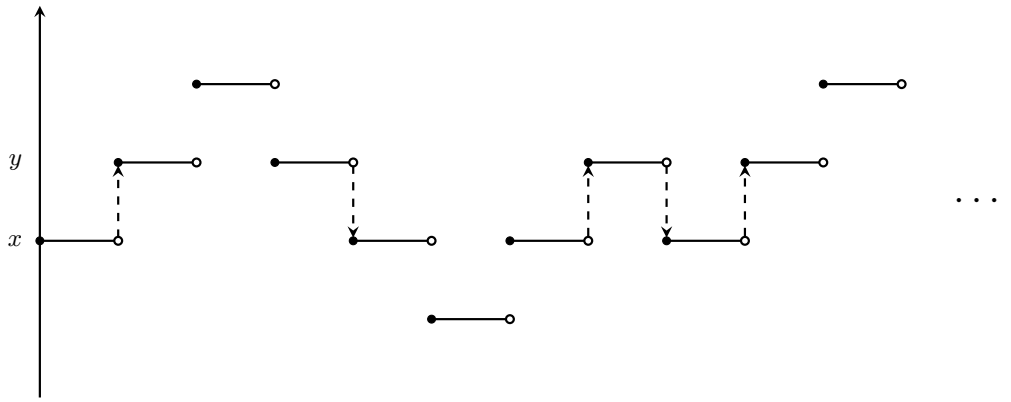



Figure 18.2: Realization of a Markov chain highlighting the first instances of moving from state $x$ to $y$ and from state $y$ to $x$.

Pitman notes that in a birth/death process, there is an alternation: every up step from $x$ to $y = x + 1$ is followed by an down step from $y$ to $x$ and vice versa. This is not always true. For example, consider a degenerate jumping chain on a circle $x \to y \to z \to x$. Then there are no $y \to x$ transitions, but still there is the

uniform equilibrium distribution. In the reversible case, the rate of $x \to y$ transitions perfectly balances the rate of $y \to x$ transitions.

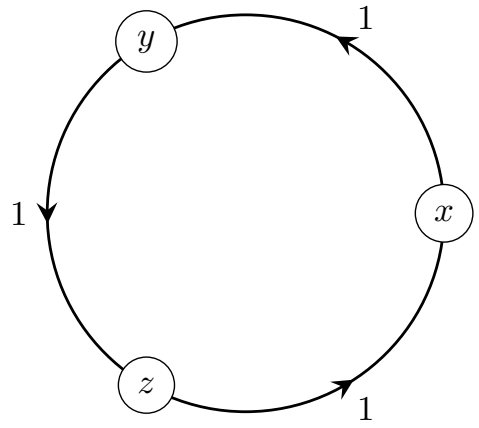



Figure 18.3:

## 18.3 Example: Birth and Death Chain

Consider the birth/death chain on $S = \{0, 1, 2, \ldots, N\}$ for some $N$. As definition of a birth/death chain if

$$P(x, y) > 0 \iff y = x \pm 1.$$

We have a key fact:

> The only possible stationary distributions for a birth and death (B.D.) chain are reversible.

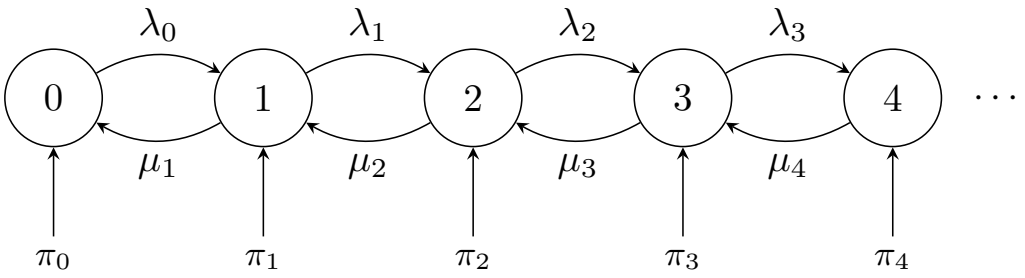



Figure 18.4: Depiction the state-space of a birth and death chain with birth rates $\lambda_0, \lambda_1, \ldots$, death-rates $\mu_1, \mu_2, \ldots > 0$, and stationary weights $\pi_0, \pi_1, \ldots$.

We can seek a reversible solution. Take the set of equations

$$\pi_0 \lambda_0 = \pi_1 \mu_1$$
$$\pi_1 \lambda_1 = \pi_2 \mu_2$$
$$\vdots$$

Quite trivially, we have $\pi_1 = \pi_0 \frac{\lambda_0}{\mu_1}$. Similarly,

$$\pi_2 = \pi_1 \frac{\lambda_1}{\mu_2} = \pi_0 \frac{\lambda_0}{\mu_1} \frac{\lambda_1}{\mu_2}$$

and so on. We can ask: under which conditions on $\lambda_i$ and $\mu_i$ does there exist a stationary distribution? We seek $\pi_0$, so take

$$\pi_0 \underbrace{\left(1 + \frac{\lambda_0}{\mu_1} + \frac{\lambda_0}{\mu_1}\frac{\lambda_1}{\mu_2} + \cdots\right)}_{\text{must be } <\infty} = 1$$

by the normalization condition. Now $\pi_0 > 0$, so the underbraced portion must be finite. As a fact, if the sum is finite, then the chain is *positive recurrent and irreducible*, and

$$P_t(x, y) \xrightarrow{t\to\infty} \pi(y)$$

from the above equations. The earlier discussion is the case with $\lambda_N := 0$ and $\lambda_0, \ldots, \lambda_{N-1} > 0$ and $\mu_1, \ldots, \mu_N > 0$. This makes for a finite state space $\{0, 1, \ldots, N\}$. But the result as stated above is correct also for state space $\{0, 2, 2, \ldots\}$ as in typical queuing models e.g. $M/M/1$ and $M/M/s$ for $s < \infty$ and $M/M/\infty$, the last of which has a Poisson equilibrium or limit distribution, as shown directly by Poisson tools and homework (number of satellites in orbit at time $t$ if they are lunched at Poisson rate $\lambda$ and stay up for $\exp(\mu)$ lifetimes).

## 18.4 Example: M/M/1 Queue

Consider customers arriving at a rate $\lambda$ via a Poisson point process. Service times are iid exponential($\mu$). Now $\lambda_n \equiv \lambda$ and $\mu_n \equiv \mu$. Then with $\lambda$ as arrival rate and $\mu$ service rate, we have

$$\lambda < \mu \iff 1 + \frac{\lambda}{\mu} + \left(\frac{\lambda}{\mu}\right)^2 + \left(\frac{\lambda}{\mu}\right)^3 + \cdots < \infty$$

as the condition for stable equilibrium. This implies that $X_t \xrightarrow{d}$ **Geometric**$(1 - \frac{\lambda}{\mu})$ on $\{0, 1, 2, \ldots\}$ as $t \to \infty$. Then with $p := 1 - \frac{\lambda}{\mu}$, we have:

$$\mathbb{E}X_t \to \frac{q}{p} = \frac{\lambda/\mu}{1 - \lambda/\mu} = \frac{\lambda}{\mu - \lambda}$$

which is the **limit mean queue length**. Recall that $L = \lambda W$ as discussed in the text. Pitman notes there are slight modifications to this example in the text including networks of queues, where we can largely apply detailed balance and usually get the solution.

### Summary

We have essentially developed the core theory on discrete and continuous Markov

chains, and any further lectures will be working through examples. We can work through more examples until we are comfortable with the material. Following, we will do a bit of Brownian motion and Martingales to finish off the semester.

# LECTURE 19

# Continuous Time Markov Chains, Part 4

### Class Announcements

Pitman mentions that we may find part b) of the last problem to be challenging. If we find ourselves stuck in a fairly general scenario, we can give up, or we can break this down into a specific example. Pitman recommends "How to Solve It" by Polya, which is to take a special case of some interest. Then if we understand the structure of the special case, we may get an idea for the general case. In this problem, Pitman suggests to try the case if there are only two ways out of the initial state $x$. That is, from $x = 0$, we can only move to $\pm 1$ (this is the typical setup for a B/D process). Even in this simple scenario, the problem is not trivial.

## 19.1 Review of Hold-Jump Description

As this is tremendously important, we'd like to look deeper into the Hold-Jump description of a Continuous-time Markov chain. Take a finite state space $S$ and a Markov chain $(X_t, t \geq 0)$ and $X_t \in S$. We typically make the implicit assumption (and we make this explicit now) that the paths of $X$ are right-continuous step functions, as in the following figure.
Then, (as we have covered in §16.1),

$$P_t(x, y) := \mathbb{P}_x(X_t = y)$$

gives a semigroup of transition matrices, meaning $P_{s+t} = P_s P_t$ for $s, t \geq 0$, with $P_0 = I$, the identity matrix. In matrix form, we have

$$P_{s+t}(x, z) = \sum_{y \in S} P_s(x, y) P_t(y, z)$$

Then in probability form, by conditioning on $X_s$ and the time-homogeneous Markov property, the above is equivalent to

$$\mathbb{P}_x(X_{s+t} = z) = \sum_{y \in S} \mathbb{P}_x(X_s = y, X_{s+t} = z) \tag{19.1}$$

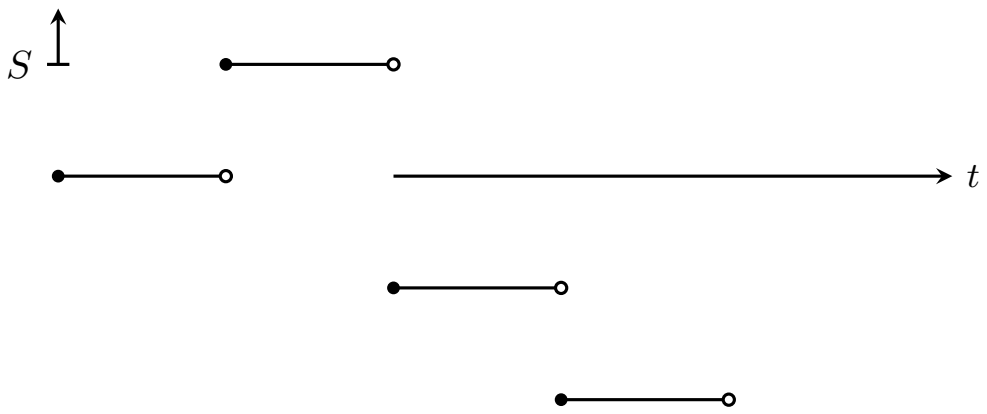



Figure 19.1: Depiction of the hold-jump description.

We know by analysis that every such semigroup $(P_t, t \geq 0)$ is of the form

$$P_t = \exp(Qt), \text{ where } Q = \frac{d}{dt} P_t \Big|_{t=0+} = \lim_{t \to 0} \frac{P_t - I}{t}$$

Then to get the backwards equation (essentially that the $Q$ part comes first), we differentiate to get

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

Alternatively, we can get the forwards equation

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

Each of these is actually a system of differential equations, written in matrix form. As discussed in the text,

- the backwards equation comes from a limit of (19.1) for $(s, t)$ replaced by $(\delta, \delta + t)$, as $\delta \downarrow 0$. (an ininitesimal first step analysis)

- the forwards equation comes from a limit of (19.1) for $(s, t)$ replaced by $(t, t + \delta)$, as $\delta \downarrow 0$. (an ininitesimal last step analysis)

The corresponding equations in discrete time are the more obvious

$$P^{n+1} - P^n = (P - I)P^n = P^n(P - I)$$

Most results in continous time can be at least formally understood by replacing matrix powers by matrix exponentials, and $P - I$ by $Q$.

### 19.1.1 Probability Interpretation

Now the probability interpretation from this is how we may make such a Markov chain from suitable random choices. We have constructed this before by using a PPP, which implies a "hold-jump" description. Suppose that we start in state $x$, so that we must hold there for $H_x$ amount of time. Then $H_x \sim$ **Exponential** $(\lambda_x)$ for $\lambda_x = -Q(x,x)$. Observe that for any Markov chain

$$P_t(x,x) = \mathbb{P}_x(X_t = x) \geq \mathbb{P}_x(H_x > t) = e^{-\lambda_x t}$$

because one way to get $X_t = x$ is if $H_x > t$, meaning the chain has never left its starting state $x$ by time $t$. Also, by the jump hold description, the difference

$$P_t(x,x) - e^{-\lambda_x t} = \mathbb{P}_x(X_t = x, H_x \leq t) = o(t) \text{ as } t \downarrow 0$$

meaning the difference is negligible compared to $t$ as $t \downarrow 0$. This is because the event $(X_t = x, H_x \leq t)$ implies that the chain has both jumped out of state $x$ to some other state, and jumped back again to $x$, all in time $t$, and by the Poisson construction this has a small probability of order $t^2$ as $t \downarrow 0$ Since by calculus

$$1 - e^{-\lambda_x t} = \lambda_x t + o(t) \text{ as } t \downarrow 0$$

we see that also

$$P_t(x,x) = 1 - \lambda_x t + o(t) \text{ as } t \downarrow 0$$

which identifies the rate $\lambda_x$ of the exponential hold at $x$ with minus the derivative of $P_t(x,x)$ at $t = 0+$ So you should think of $P_t(x,x)$, for fixed $x$, as a function of $t$ which

- starts at 1 at $t = 0$,
- decreases from 1 with slope $-\lambda_x$ at $t = 0+$,
- lies above the exponential curve $e^{-\lambda_x t}$ at every $t \geq 0$,
- is tangent to this curve at $t = 0+$

Note that $P_t(x,x)$ is not necessarily decreasing for all $x$, but that this function always has a limit as $t \to \infty$ for any continuous time Markov chain.

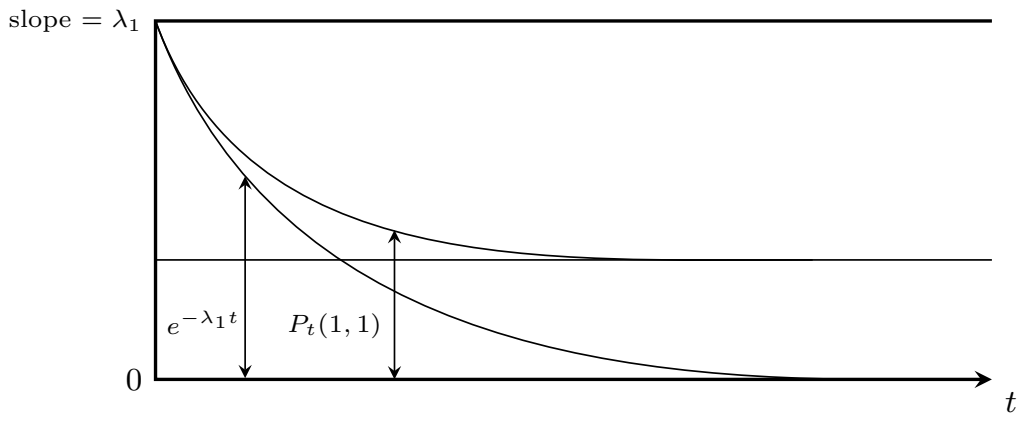



Figure 19.2: Realization $e^{-\lambda_1 t} < P_t(1,1)$ .

Let's take a state-space $x \in \{1,2,3\}$ and consider a chain with limit distribution $\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. If we want to think of all elements of the transition matrix, it may be helpful to graph the following picture of all three curves $P_t(1,x)$ for $x = 1,2,3$. These are determined by the forwards differential equations.

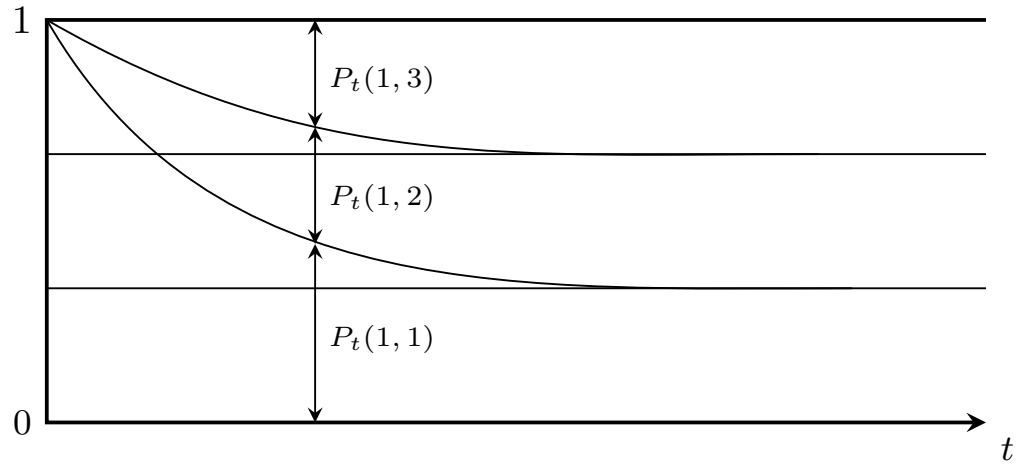



Figure 19.3: Realization $P_t(1,3) + P_2(1,2) + P_t(1,1) = 1$

Continuing the general discussion, for a chain started in state $x$, at time $H_x$, the chain jumps to some other state that is not $x$. That is, we go to state $J_x$, with:

$$\mathbb{P}(J_x = z) = \frac{Q(x,z)}{\lambda_x} \qquad (z \neq x)$$

where $\lambda_x = -Q(x,x) = \sum_{z \neq x} Q(x,z)$ is the total rate of transitions out of state $x$. This uses the fact that $\sum_z P_t(x,z) = 1$, and therefore differentiating gives

$$\sum_z \frac{d}{dt} P_t(x,z) = 0 \implies \sum_z Q(x,z) = 0, \forall_x$$

Essentially, if one curve 'goes down', another must 'go up' to compensate. That is, $Q\mathbb{1} = \vec{1}$. Next, we condition on the case $H_x = t$ and $J_x = z$. We've held and jumped, so the next thing is to hold: we hold in state $z$ for an **Exponential** $(\lambda_z)$ time, where $\lambda_z = -Q(z,z)$. Then at this time, because we conditioned on arriving at $z$, we jump according to the jump distribution $\frac{Q(z,\cdot)}{\lambda_z}$, where $\cdot \neq z$. We introduce the following notation.

**Embedded Jumping Chain**

Let $Z_0 = x$, $Z_1 = J_x$, $Z_2 =$ next state after $J_x$, and so on, where $Z_0 \neq Z_1 \neq Z_2 \neq Z_3 \neq \dots$. We call this the *embedded jumping chain*. This is a discrete time Markov chain with transition matrix

$$(x,y) \to \frac{Q(x,y)}{\sum_{z \neq x} Q(x,z)}$$

where we use only the off-diagonal part of $Q$, and the denominator is $\lambda_x = -Q(x,x)$

Then in terms of holds and jumps, we summarize, with slightly different notation,

- $X$ holds for time $H_0$ in the initial state $Z_0 = X_0$, then
- $X$ holds for time $H_1$ in state $Z_1 = X_{H_0} \neq Z_0$, then
- $X$ holds for time $H_2$ in state $Z_2 = X_{H_0+H_1} \neq Z_1$,

and so on, where the assignment of values of $X$ at the jump times is made to assure right continuous step function paths. Then given $Z_0, Z_1, \dots, Z_n$ and $H_0, H_1, \dots, H_n$ with $Z_n = z$,

$$\begin{aligned} H_{n+1} &\sim \exp(\lambda_z) \\ Z_{n+1} &\sim Q(z,\cdot)/\lambda_z, \text{ where } \cdot \neq z. \end{aligned}$$

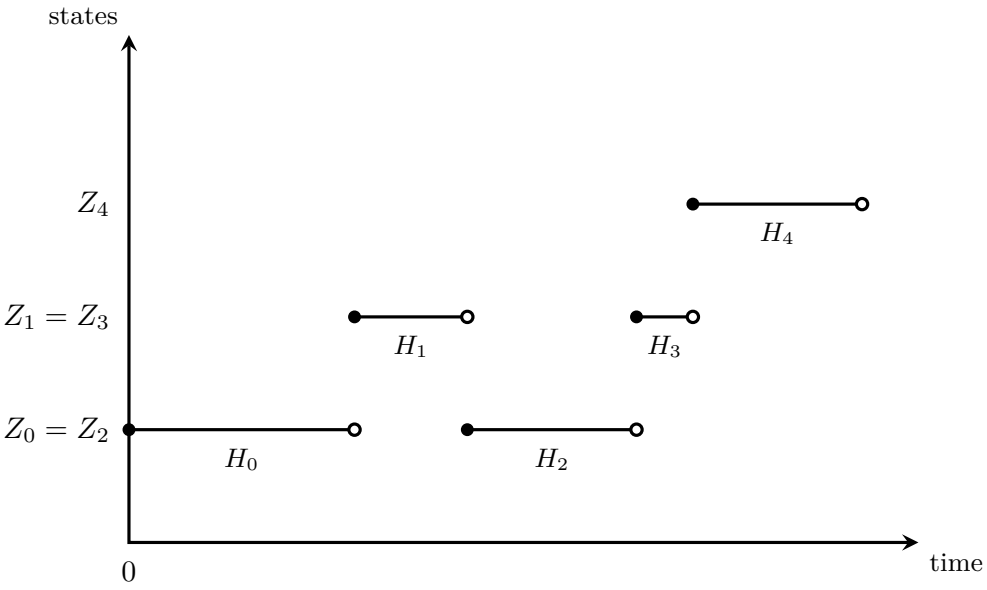



Figure 19.4: Realization $Z_0, Z_2, Z_3, \ldots$ and $H_0, H_2, H_3, \ldots$.

## 19.2 Exit Distributions/Times (Durett §4.4)

This is just to sketch some ideas. See text for details. Suppose that we have an irreducible chain, and that we have some set of states $C$ (the 'interior') and some target sets $A, B$, where $A, B, C$ are disjoint. We are interested in two things

(1) $\mathbb{P}_x(\text{hit } B \text{ before } A)$, the probability of starting at $x$ of hitting $B$ before $A$, and

(2) the expectation $\mathbb{E}_x(\text{time to hit } A \cup B)$.

We discussed and solved this problem before in the discrete case, and now we would like to generalize this to continuous-time. Notice that (1) is easy, because there is no new difficulty from working in continuous-time. Recall that our $Q$ matrix tells us how long we need to hold and where to jump when we need to jump. Then (1) is simply a question about what the jumping chain (which we previously called $Z_0, Z_1, Z_2, \ldots$) does. That is,

$$\mathbb{P}_x(X \text{ hits } B \text{ before } A) = \mathbb{P}_x(Z \text{ hits } B \text{ before } A)$$

We can solve this with the same old methods, applied to the jumping chain $Z$ instead of the continuous time chain $X$, and as you see in the text, we can wrap this up with a solution in terms of the $Q$ matrix.

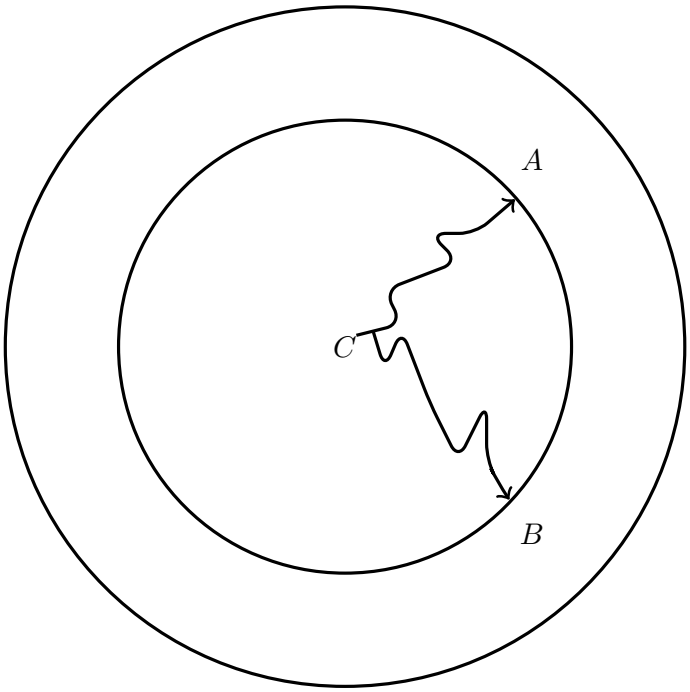



Figure 19.5: Add something here

Recall that for a discrete-time Markov chain the method was to consider a harmonic function, that is a solution of $h = Ph$, with some boundary conditions. Notice this is equivalent to $(I - P)h = 0$. Now in this case, we can neatly write

$$\frac{(I - P_t)}{t} h = 0 \implies \boxed{Qh = 0}$$

by taking the limit as $t \downarrow 0$. So the harmonic functions for a continuous time chain are the solutions $h$ of $Qh = 0$.

Let's look at problem (2), which is a little more interesting. Recall from discrete time in that if we take a set $C$ of interior states, we can write the matrix $P$ as:

$$P = \begin{bmatrix} P_{C\times C} & P_{C\times(A\cup B)} \\ P_{(A\cup B)\times C} & P_{(A\cup B)\times(A\cup B)} \end{bmatrix}$$

where e.g. $P_{C\times C}(x, y) := P(x, y)$ for $x, y \in C$. This is simply the $C \times C$ restriction of $P$, let us abbreviate it to $P_C$. So $P_C$ is a sub-stochastic matrix indexed by $C$, as in discussion in our midterm exam.

The fact that some of the the row sums are less than 1 (by irreducibility of the original matrix $P$) tells us that we have a convergent Neumann series of matrices

$$I + P_C + P_C^2 + P_C^3 + \cdots = (I - P_C)^{-1}$$

Recall this is the *Green matrix*

$$(I - P_c)^{-1}(x, y) = \mathbb{E}_x \sum_{n=0}^{\infty} \mathbb{1}(X_n = y, T > n)$$

$$= \sum_{n=0}^{\infty} \underbrace{\mathbb{P}_x(X_n = y, T > n)}_{=P_C^n(x,y)}$$

where $T = \min\{n \geq 0 : X_n \notin C\}$. If $\mathbb{E}_x T$ is the expected hitting time, we can extract this from our Green matrix by 'covering up' the states in $C$ with the indicators of $X_n = y$ and summing out the cases. That is,

$$\mathbb{E}_x T = \sum_y (I - P_C)^{-1}(x, y)$$

Let's do the same thing but in the continuous parameter case. Take

$$T := \min\{t \geq 0 : X_t \notin C\}$$

Let $C$ be fixed. We want something similar to our discrete case, where for $n = 1, 2, \ldots,$ we have

$$P_C^n(x, y) = \mathbb{P}_x(X_n = y, T > n)$$

We should expect the continuous-time analog to be

$$P_{C,t}(x, y) = \mathbb{P}_x(X_t = y, T > t)$$

Then

$$\mathbb{E}_x \int_0^{\infty} \mathbb{1}(X_t = y, T > t) = \mathbb{E}_x(\text{ time in } y \text{ before } T)$$

$$= \int_0^{\infty} P_{C,t}(x, y)\ dt$$

We should guess, expect, or hope that

$$P_{C,t} = \exp(Q_c t)$$

where $Q_C$ is simply $Q$ restricted to $C$. Once guessed, this is quite easily justified by consideration of the hold-jump description of the original chain run until the random time $T$ when it first hits $C$.

Now suppose we trust that $P_C = \exp(Qt)$, and that $\int_0^{\infty} \exp(Q_C t)\ dt < \infty$. In discrete time, the matrix power sum formula is based on the sum of a geometric series of real or complex numbers:

$$\sum_{n=0}^{\infty} r^n = \frac{1}{1 - r} = (1 - r)^{-1} \qquad (|r| < 1)$$

A familiar continous analog of this formula is

$$\int_0^\infty e^{-\lambda t}\, dt = \frac{1}{\lambda} \qquad (\lambda > 0)$$

which strongly suggests that for suitable matrices $Q$ such that the integral converges

$$\int_0^\infty e^{Qt}\, dt = \frac{1}{-Q} = (-Q)^{-1}$$

In particular, this formula can be justified for $Q = Q_C$ obtained by restriction to $C$ as above of the rate matrix of an irreducible chain. See text for further details and examples. Note that Durett uses the notation $R$ (a good mnemonic for a *restriction* of $Q$) instead of $Q_C$.

# LECTURE 20

# Laplace Transforms

### Class Announcements

We have a worksheet this week with a lot of problems, where some of these are due for homework. Pitman emphasizes that our time should be applied to problems and exercises. Today we'll start with some remarks regarding the last problem of last week's homework. This is a very conceptual problem that gets us to think about Markov chains in the manner that Pitman would like us to view them.

**Homework Problem.** Suppose you have a recurrent chain in continuous time with transition rate $Q(0, j)$ from state 0 to state $j > 0$. Starting in state 0, let $H(0, j)$ be the total time spent in state 0 before the first jump from 0 to $j$.

a) What is the distribution of $H(0, j)$?

b) What is the joint distribution of $H(0, j)$ and $H(0, k)$ for different $j$ and $k$?on of $H(0, j)$ and $H(0, k)$ for different $j$ and $k$?

## 20.1 Problem Discussion

Recall that the problem looks at a fixed state 0 of a chain in continuous time, and we look for each $y \neq 0$ at transitions from 0 to state $y$. Assume (otherwise this does not work) that the chain is recurrent. To remind us, we have $P_t = e^{Qt} = \cdots$, and $Q(0, j)$ is the rate of transition from 0 to $j$. We have two meanings:
1) The small time meaning of transition rate, in words, is that if we start in state 0 and we run for a *small* time $t$,

$$P_t(0, j) = \mathbb{P}_0(X_t = j) = Q(0, j)t + o(t)$$

as our first-order approximation for the very small chance that we change states. This is,

$$\frac{P_t(0, j)}{t} = Q(0, j) + \underbrace{\frac{o(t)}{t}}_{\to 0}$$

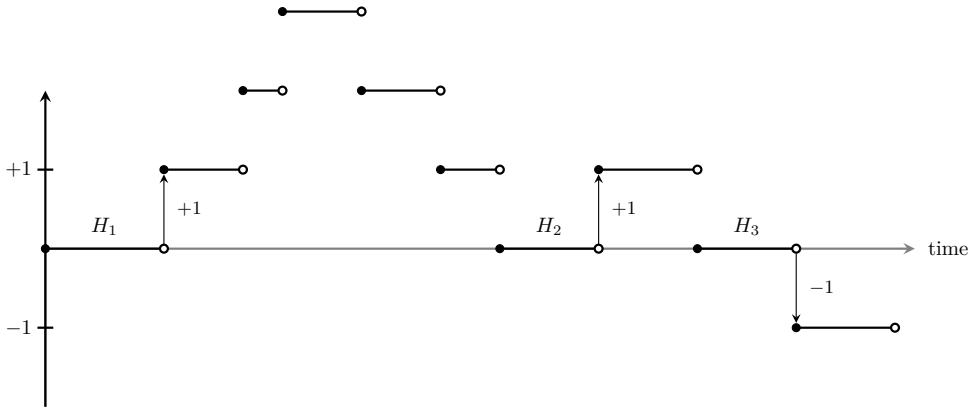



Figure 20.1: Successive holds in state 0, denoted $H_i \sim$ **Exponential** $(\lambda_0)$. Transitions $0 \to j$, where $j \in \{-1, +1\}$.

We go to 1, float around in our state space, then come back to 0, and so on. The entire discussion is about these "successive" holds in state 0. Let $H_1, H_2, H_3, \ldots$ be the successive holds in state 0. We have a few observations

#### Observation #1

Note that there are infinitely many of these because the problem gives that the chain is recurrent (we assume 0 was recurrent). That is, we keep coming back to 0 and have more opportunities to make transitions out of state 0.

#### Observation #2

The holding times $H_i$ are iid **Exponential** $(\lambda_0)$ where $\lambda_0 = \sum_{j \neq 0} Q(0, j)$, the off-diagonal row sum.

#### Observation #3

Surely, $H_{0j} \geq H_1$, so we can write

$$H_{0j} = H_1 + H_2 + H_3 + \cdots + H_{N(j)}$$

where $N(j)$ is the number of holds in state 0 before the first (direct) $0 \to j$ transition. In our diagram, we have $N(+1) = 1$ and $N(-1) = 3$. Then

$$N(j) \sim \textbf{Geometric}(p = \frac{Q(0, j)}{\lambda_0})$$

because of the hold-jump description of a Markov chain.

#### Observation #4

Let $J_1, J_2, J_3, \ldots$ be the locations of jumps (that is, $J_k$ is the state jumped to after the hold $H_k$). Then $J_1, J_2, J_3, \ldots$ are iid with $\mathbb{P}(J_i = j) = \frac{Q(0,j)}{\lambda_0}$.

#### Observation #5

By the Markov property, it is intuitive that $H_1, H_2, \ldots$ and $J_1, J_2, \ldots$ are independent (this was a part of the hold-jump description).

#### Observation #6

In terms of $H_i$ and $J_i$, we have that

$$N(j) := \min\{k \geq 1, J_k = j\}$$

where we start our indexing at 1.

#### Observation #7

We have the familiar fact that

$$\mathbb{P}(N(j) = k) = q^{k-1}p$$

where $p = \frac{Q(0,j)}{\lambda_0}$ and $q := 1 - p$.

## 20.2 Problem Summary

In summary, we have

$$H_{0j} = H_1 + \ldots + H_{N(j)}$$

where $N(j)$ is **Geometric**$(Q(0,j)/\lambda_0)$ independent of $H_1, H_2, \ldots$. This implies that

$$H_{0j} \sim \textbf{Exponential}\,(Q(0,j))$$

with

$$Q(0,j) = \lambda_0 \cdot \frac{Q(0,j)}{\lambda_0}$$

Either by computation (e.g. of densities) or by analogy or instance of previous setup, we have a picture of the thinning of a Poisson Point Process.

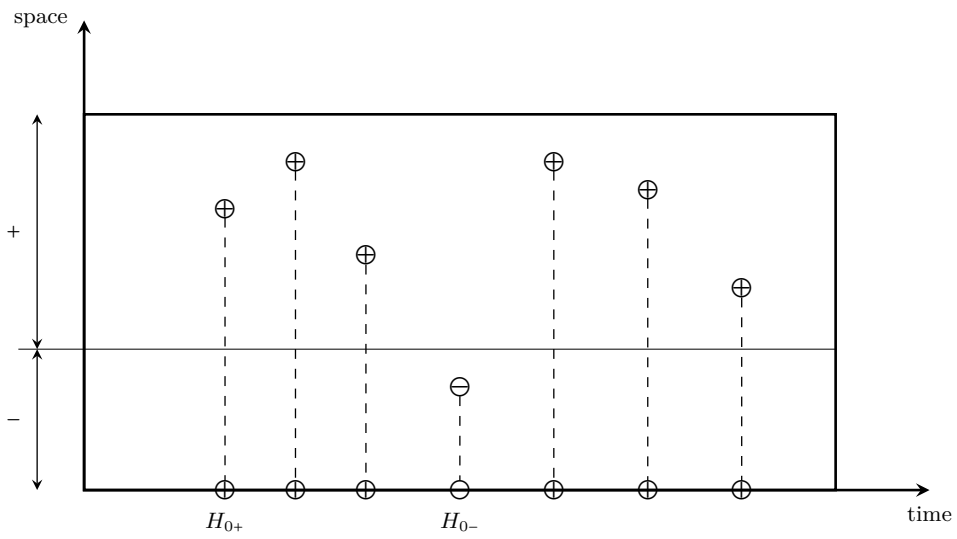



Figure 20.2: PPP realization.

In this picture, we see that the time until the first $\oplus$ is exponential. Now from the same picture, we can do part (b) of this problem. Because we have two independent (disjoint) regions in our PPP, $H_{0+}$ and $H_{0-}$ are independent by Poisson marking or thinning facts. This is also true for $H_{0,j}$ as $j$ varies. To generalize this from 2 outcomes to 3, we simply another strip in our picture. Continue this as necessary.

## 20.3 The Laplace Transform

There was a question in the audience regarding how to show that the sum of a **Geometric**$(p)$ number of **Exponential** $(\lambda)$ is **Exponential** $(\lambda p)$ without the PPP picture we have drawn before. Pitman gives two solutions. 1. Compute the density by conditioning on some random variable $N$, but we will not pursue this because it is slightly boring. 2. Compute a suitable transform.

Recall that if we are adding a random number $N$ of iid positive-integer valued random numbers, we could use PGFs (e.g. as we have for branching processes). We need a variant of PGFs for continuous variables, especially for nonnegative random variables $X \geq 0$. We have two related concepts

1) MGF$(\theta)$ (Moment Generating Function)

2) Laplace transform, which is nothing more than MGF$(-\theta)$. The Laplace transform is nice for $X \geq 0$ for two reasons.

**Laplace Transform**

For a random variable $X \geq 0$, define the *Laplace Transform* of $X$ as

$$\phi_X(\theta) := \mathbb{E}e^{-\theta X}$$

A leading example for which we can easily find this function is if $X \sim$ **Exponential** $(\lambda)$ and $X \stackrel{d}{=} \mathcal{E}/\lambda$. In this case, then

$$\begin{aligned}
\phi_X(\theta) &= \mathbb{E}e^{-\theta\mathcal{E}/\lambda} \\
&= \mathbb{E}e^{-\frac{\theta}{\lambda}\mathcal{E}} \\
&= \int_0^\infty e^{-\frac{\theta}{\lambda}t} e^{-t} t dt \\
&= \left(\frac{\theta}{\lambda} + 1\right)^{-1} \\
&= \frac{\lambda}{\lambda + \theta}
\end{aligned}$$

Notice that this is the (same answer as and hence exactly is) the race between two exponentials. In general, we have

$$\phi_X(\lambda) := \mathbb{E}\left(e^{-\lambda X}\right) = \mathbb{P}\left(X < \frac{\mathcal{E}}{\lambda}\right) = \mathbb{P}(\lambda X < \mathcal{E})$$

We arrive at this by first considering if $X$ is constant. Now if we assume that (1) $\mathcal{E} \sim$ **Exponential** $(1)$ where $\mathbb{P}(\mathcal{E} > t) = e^{-t}$ and (2) $\mathcal{E}$ and $X$ are independent.

In summary, we can always interpret the Laplace transform $\phi_X(\theta)$ of a random variable as $\mathbb{P}(X < \frac{\mathcal{E}}{\lambda})$ for $\mathcal{E}$ independent of $X$. Now for $X \sim$ **Exponential** $(\lambda)$, we have

$$\begin{aligned}
\phi_X(\theta) &= \mathbb{E}e^{-\theta X} \\
&= \mathbb{P}\left(X < \frac{\mathcal{E}}{\theta}\right) \\
&= \mathbb{P}\left(\frac{\mathcal{E}'}{\lambda} < \frac{\mathcal{E}}{\theta}\right), \quad \mathcal{E}, \mathcal{E}' \text{ iid} \\
&= \frac{\lambda}{\lambda + \theta}
\end{aligned}$$

## 20.4 Properties of Laplace Transforms

Recognizing that this is a way to morph PMFs to a way of looking at continuous variables, once we have found the Laplace transform, we should have some unique-

ness to the probability distribution.

**Properties of Laplace Transforms**

- **Uniqueness.** For $X, Y \geq 0$, if $\phi_X(\theta) = \phi_Y(\theta)$ for all $\theta \geq 0$, then $X \stackrel{d}{=} Y$. That is, $\mathbb{E}g(X) = \mathbb{E}g(Y)$ for any $g$ such that either side is defined. Pitman defers this proof for later in the course.
- **Sums of independent random variables.** Provided $X, Y$ independent,

$$\phi_{X+Y}(\theta) = \phi_X(\theta)\phi_Y(\theta)$$

*Proof.*

$$\begin{aligned}\phi_{X+Y}(\theta) &= \mathbb{E}e^{-\theta(X+Y)} \\ &= \mathbb{E}\underbrace{e^{-\theta X}e^{-\theta Y}}_{\text{independent}} \\ &= \phi_X(\theta)\phi_Y(\theta).\end{aligned}$$

□

Another thing that was quite helpful (as we have seen with Branching processes) is that this helps us with random sums. Suppose that $X_1, X_2, \ldots$ are iid and that they all have the same $\phi_{X_i}(\lambda) = \lambda_X(\lambda)$ as a consequence of iid. Let $N$ be a random index in $\{0, 1, 2, \ldots\}$ independent of $X_1, X_2, \ldots$. Let's try to calculate the Laplace transform. Condition to get

$$\mathbb{E}e^{-\theta(X_1+X_2+\cdots+X_N)} = \sum_{n=0}^{\infty} \mathbb{E}\left[e^{-\theta(X_1+\cdots+X_n)}\mathbb{1}(N=n)\right] \quad \text{(true with no assumptions)}$$

$$= \sum_{n=0}^{\infty} \mathbb{P}(N=n)[\underbrace{\phi_X(\theta)}_{z}]^n$$

which is the same as $G_N(\phi_X(\theta))$ where $G_N(z) := \sum_n \mathbb{P}(N=n)z^n$.

## 20.5 Conclusion

The Laplace transform of $X_1 + \ldots + X_N$ of $\theta$ is $G_N(\phi_X(\theta))$. To apply this, let's go back to the geometric sum of exponentials with parameter $\lambda$.
From before, we have

$$\phi_X(\lambda) = \frac{\lambda}{\lambda+\theta} = \frac{1}{1+\theta/\lambda}$$

as the Laplace transform of an exponential. Then the generating function is:

$$G_N(z) = \frac{pz}{1-qz} = \sum_{n=1}^{\infty} pq^{n-1}z^n = pz\sum_{n=1}^{\infty}(qz)^{n-1}$$

Of course, this does not work very nicely if we have a possibility of 0. Finishing the calculation, we see that

$$G_N(\phi_X(\theta)) = \frac{p\frac{\lambda}{\lambda+\theta}}{1-q\frac{\lambda}{\lambda+\theta}}$$

simply substituting in the Laplace transform into the argument of the probability generating function. Then simplifying the above gives:

$$G_N(\phi_X(\theta)) = \frac{p\theta}{\lambda+\theta-q\lambda} = \frac{p\lambda}{p\lambda+\theta}$$

which is precisely the Laplace transform at $\theta$ of **Exponential** $(()\, p\lambda)$. This calculation is much easier and sweeter than having to work with the densities. Although this is a bit indirect, this is very helpful. There are concepts like the Stone-Weierstrass theorem that justify these concepts.

# LECTURE 21

# Laplace Transforms and Introduction to Martingales

## 21.1 Convergence of Random Variables and Their Distributions

As a sort of preamble into closing our discussion on Laplace transforms, we want to formalize our discussion of this with more care and precision in our language. Recall that we have a notion of convergience in distribution from before. That is, if $X_n$ is a sequence of discrete random variables, each with values in the same countable set $S$, then $X_n \xrightarrow{d} X$ means

$$\mathbb{P}(X_n = s) \to \mathbb{P}(X = s), \forall_{s \in S}$$

The main example for this that we have seen many times before is via the main limit theorem of Markov chain theory, which is that if $X_n \sim MC(\lambda, P)$ and $P$ is *irreducible, periodic, positive recurrent*, then

$$\mathbb{P}(X_n = s) = \lambda P^n(s) \to \pi(s).$$

where $\pi$ is the unique stationary probability distribution for $P$. So under these conditions, we have convergence in distribution to a limit distribution of $X$, with $\mathbb{P}(X = s) = \pi(s)$. This is for the discrete case; we want to have a version for real random variables $X_n$, whose distribution might be discrete, or continuous, meaning $x \mapsto \mathbb{P}(X_n \leq x)$ is a continous function of $X$. Typically such distributions will have a density, meaning that you can differentiate the CDF to get the density, say $f_n(x)$, with $\int_\infty^x f_n(t)dt = \mathbb{P}(X_n \leq x)$. Or the $X_n$ might be discrete, with a continous limit distribution of $X$.

We write that $X_n \xrightarrow{d} X$ to mean that the CDF of $X_n$ converges to the CDF of $X$ to the extent possible, which means at *all continuity points* of the limit. This is a slightly technical condition, so we will draw pictures and discuss mainly for continuous CDFs as limit. Famous examples of CDFs of limit distributions are:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \, dz$$

which is the standard normal. Also, another example is

$$F(x) = \int_0^x e^{-t}\, dt, \quad x \geq 0$$

which is the standard exponential. In general, convergence in distributions means

$$\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x) \text{ at all continuity points } x \text{ of the limit CDF.} \tag{21.1}$$

If $X$ is normal or exponential, that means all points $x$.

From our probability course, we have the famous *Central Limit Theorem* which is to say that for $X_1, X_2, \ldots$ iid, with $\mathbb{E}X_i = \mu, \mathbb{V}\text{ar}(X_i) = \sigma^2$, then we can either add the observations or take their means. After correct scaling:

$$Z_n := \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \frac{(X_1 + X_2 + \cdots + X_n)/n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z.$$

Most statisticians simply call this standard normal variable $Z$ with

$$\mathbb{P}(Z \leq x) = \Phi(x)$$

If we write this out, we have the exact same thing as in (21.1) above, just with $Z_n$ instead of $X_n$. The reason for using random variables in the formulation of $\xrightarrow{d}$ is that it is so easy to describe operations on random variables, like the scaling in the CLT, which lead to a limit distribution. But if you look carefully at the definition $\xrightarrow{d}$ you see it is really the sequence of probability distributions which is converging, not the random variables. As a trivial example to keep in mind, an i.i.d. sequence always converges in distribution. But there is no meaningful sense in which the random variables themselves converge, unless the common distribution is that of a constant.

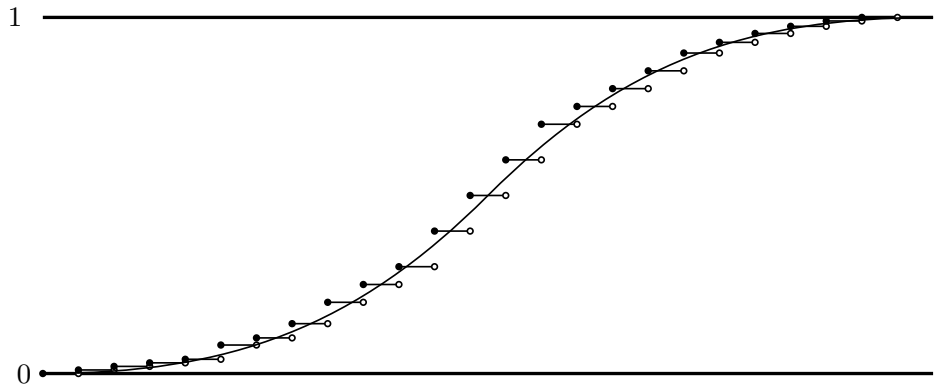



Figure 21.1: Discrete CDF $Z_n$. As $n \to \infty$, $Z_n$ converges to the smooth curve.

The question arises in the audience as to why our definition uses the CDF, not densities for example. The issue is that when we have a convergence in distribution, as in the CLT, it has to be that all the point probabilities go to zero. It is possible sometimes to get normalized discrete probabilities to converge to densities, e.g. point probabilities for binomial $(n, p)$ for fixed $p$ and large $n$ can be scaled to converge to the normal density curve, but not under the general conditions of the CLT. Also, we want to include the kind of degenerate limit involved in the law of large numbers, where the limit distribution is a constant, typically with probability one at the mean $\mu$.

Here is a fact from analysis: if for each $n$ a function $F_n(x)$ is a CDF (which means it is non-decreasing, right-continuous, with left limit 0 at $-\infty$ and right limit 1 at $\infty$), and $F_n(x) \to F(x)$ for all $x$, and $F$ is continuous, then

$$\sup_x \Big| F_n(x) - F(x) \Big| \to 0.$$

Basically, you get uniform convergence for free because of the non-decreasing property of CDFs and continuity of the limit. Obviously, this cannot be true if the limit has a jump somewhere but all the $F_n$ are continous. But in cases like the LLN we easily get continuous distributions converging to a discrete limit with only one jump at some $\mu$.

Now back to our present setting with the Law of Large Numbers, consider the degenerate case where the limit is discrete. With this same setting, consider:

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow{d} \mu = \mathbb{E}X$$

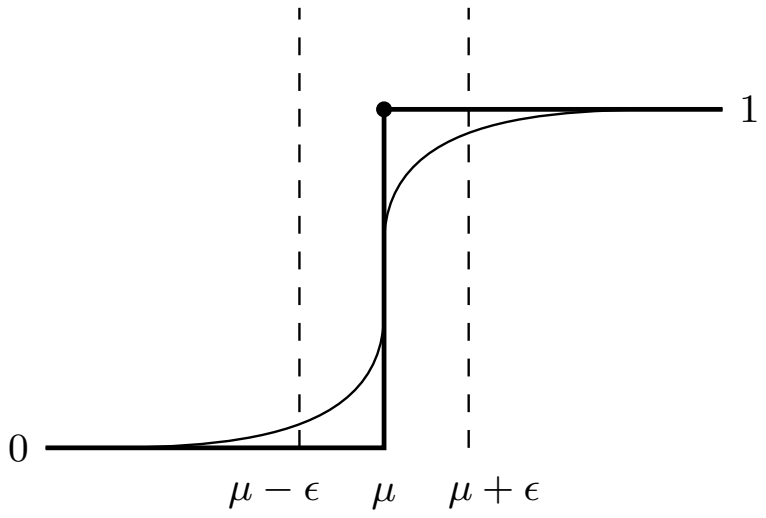



Figure 21.2: Realization of the limit CDF with point mass 1 at the limit $\mu$. Weak Law of Large Numbers.

To see what this says, we can write this out in cases:

$$\mathbb{P}\left(\frac{X_1 + \cdots + X_n}{n} \leq x\right) \to \begin{cases} 0, & x < \mu \\ 1, & x > \mu \end{cases}.$$

Now notice that this says nothing at $x = \mu$ because the limit has a jump of 1 at this point. We have to look closer to get anything useful, e.g. make assumptions for the CLT and then the probability that the average is less than $\mu$ will converge to $1/2$, from the standard normal probability $\Phi(0) = 1/2$

Consider the equivalent statement:

$$\mathbb{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \epsilon\right) \to 0$$

We introduce these concepts only as a unifying language; we have dealt with these in previous lectures, and we would like to have convergence in distribution as a way to bring our ideas together.

## 21.2 Mean Square

We can have another sort of convergence, namely "Mean square" convergence. We say that $X_n \to X$ in mean square ($L^2$). This means that

$$\mathbb{E}(X_n - X)^2 \to 0, \quad \text{as } n \to \infty$$

This implies, via Chebyshev's inequality, that

$$\mathbb{P}(|X_n - X| \geq \epsilon) \leq \frac{\mathbb{E}(X_n - X)^2}{\epsilon^2} \to 0, \quad \forall_{\epsilon>0}.$$

This via analysis implies $X_n \xrightarrow{d} X$. Now, one of the reasons for these ideas is to bring them into an application.

## 21.3 Application of Convergence in Distribution ($\xrightarrow{d}$)

Let $X \geq 0$, $\lambda > 0$, and define $X_\lambda := N_\lambda(X)$, where $N_\lambda$ is a Poisson process with rate $\lambda$ independent of $X$. We have a random variable $X$ with whatever value, and along a timeline, we have a PPP with rate $\lambda$. We like to think that $\lambda$ is under our control, so that we can draw the following diagrams, based on our old friend, a PPP in the plane with rate one point per unit area:

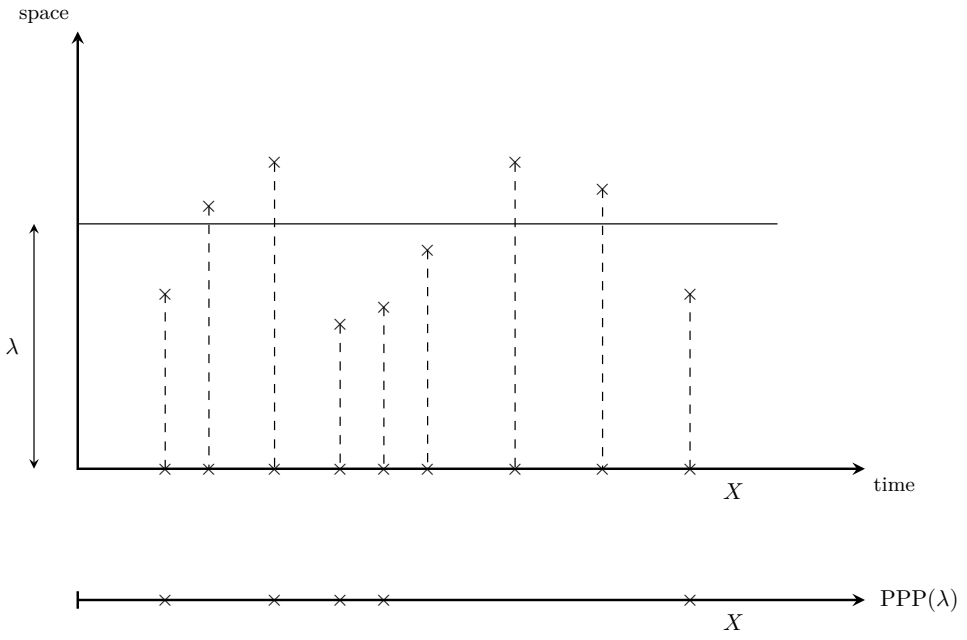



Figure 21.3: Realization of a PPP($\lambda$) with level $\lambda$ and random time $X$. Observe $N_\lambda(X) = 5$.

From our picture, we can count $N_\lambda(X) = 5$ points under $\lambda$ before $X$. Fix arbitrary $X$ in our minds so that we can tweak $\lambda$. What happens when $\lambda \uparrow \infty$? First of all, suppose $X = t$ is fixed. Then, $N_\lambda(X) = N_\lambda(t) \sim$ **Poisson** $(\lambda t)$, like it always is with this notation. This implies

$$\mathbb{E}N_\lambda(t) = \lambda t, \text{ and } \mathbb{V}\text{ar}(N_\lambda(t)) = \lambda t$$

Recall the critical fact about the Poisson distribution that its variance is the same as its mean. Now if we take $N_\lambda(t)$ and divide by $\lambda t$, taking a Poisson variable and dividing it by its mean, via the Weak Law of Large Numbers (WLLN), we should be getting the rate per unit area in the plane of the PPP. That is,

$$\frac{N_\lambda(t)}{\lambda t} \xrightarrow{d} 1 \text{ as } \lambda \to \infty$$

Essentially, we have this convergence in distribution via convergence in Mean square, from the fact that

$$\mathbb{E}\left(\frac{N_\lambda(t)}{\lambda t} - 1\right)^2 = \frac{\lambda t}{(\lambda t)^2} = \frac{1}{\lambda t} \to 0 \text{ as } \lambda \to \infty$$

This comes to no surprise; these computations are simply for familiarity. However, this was for a fixed time $X = t$. What about a random $X$?
Let's first notice that

$$\frac{N_\lambda(t)}{\lambda} \xrightarrow{d} t, \text{ as } \lambda \to \infty,$$

which holds for every fixed $t \geq 0$. Now, let's make $t$ random, replace $t$ by $X$ independent of $(N_\lambda(t), t \geq 0)$. We may guess, hope, or imagine that

$$\frac{N_\lambda(X)}{\lambda} \xrightarrow{d} X$$

In fact this convergence holds with probability one (as discussed in previous lecture), but you can easily check by computing the mean square difference that the convergence holds in mean square, at least if $\mathbb{E}X < \infty$. With more effort, you can show this assumption is unnecessary, and in fact the convergence holds both almost surely, and in distribution, without any additional assumptions on $X$.
Now, consider the exact distribution of $N_\lambda(X)$. This is a discrete distribution, called a mixed Poisson distribution. In particular, consider that we have a picture previously where we had 5 points. We can ask, what is the probability of getting exactly $n$ points? To get this, we have to condition on $X$, which means: write down the answer as if we knew what $X$ is, then take expectations. If $X$ is constant, then the probability is

$$\mathbb{P}(N_\lambda(X) = n) = \frac{e^{-\lambda X}(\lambda X)^n}{n!}$$

which is a trivial formula for fixed $X$. Now if we do not know the value of $X$, we take expectations (law of total probability). If we know the conditional probability, we take the expectation. So in the general case (we do not know the value of $X$), we simply write:

$$\mathbb{P}(N_\lambda(X) = n) = \mathbb{E}\left[\frac{e^{-\lambda X}(\lambda X)^n}{n!}\right]$$

That is, for the discrete and density case, we respectively have

$$= \sum_{x=0}^{\infty} \frac{e^{-\lambda x}(\lambda x)^n}{n!} \mathbb{P}(X = x)$$

$$= \int_0^{\infty} \frac{e^{-\lambda x}(\lambda x)^n}{n!} f_X(x)\ dx.$$

This combines

$$\mathbb{P}(A) = \mathbb{E}[\mathbb{1}_A]$$

$$\mathbb{E}(Y) = \mathbb{E}\left[\mathbb{E}(Y|X)\right]$$

Now, we have another fact:

$$\frac{N_\lambda(X)}{\lambda} \xrightarrow{d} X \text{ as } \lambda \to \infty$$

Notice that $N_\lambda(X)$ is a discrete random variable with values $\{0, 1, 2, \dots\}$, so that $\frac{N_\lambda(X)}{\lambda}$ is a discrete random variable with values $\{0, \frac{1}{\lambda}, \frac{2}{\lambda}, \dots\}$. The grid is geting finer as $]lambda$ increases. Additionally,

$$\mathbb{P}\left(\frac{N_\lambda(X)}{\lambda} \leq x\right) \to \mathbb{P}(X \leq x)$$

at all continuity points of limit CDF. We may ask why bother with these considerations?

- First of call, this provides a simple illustration for convergence in distribution.
- Second it shows that if we know the distribution of $N_\lambda(X)$ for all $\lambda > 0$, we can find (by limit) the distribution of $X$. We essentially have a formula for working out the probabilities for the possibly continuous limit object from the discrete mixed Poisson probabilities.

In fact, we have just shown how to invert a Laplace transform!

## 21.4 Review of Laplace Transform

Recall this is simply

$$\varphi_X(\lambda) := \mathbb{E}e^{-\lambda X}.$$

If we know $\phi_X(\lambda)$, we know

$$\mathbb{P}(N_\lambda(X) = 0) = \mathbb{E}e^{-\lambda X}$$

from the case $n := 0$ of the equation (2) above. Let's look at what happens when we differentiate the Laplace transform: for $\lambda > 0$

$$\begin{aligned}\frac{d}{dx}\varphi_X(\lambda) &= \frac{d}{d\lambda}\mathbb{E}e^{-\lambda X} \\ &= \mathbb{E}\frac{d}{d\Lambda}e^{-\lambda X} \\ &= \mathbb{E}(-X)e^{-\lambda X} \\ &= -\mathbb{E}Xe^{-\lambda X},\end{aligned}$$

Where the swap of $d/d\lambda$ and $\mathbb{E}$ can be justified by analysis beyond our present scope. Comppare this against:

$$\begin{aligned}\mathbb{P}(N_\lambda(X) = 1) &= \lambda X\frac{e^{-\lambda X}}{1!} \\ &= \lambda\left(-\frac{d}{dx}\varphi_X(X)\right)/1!\end{aligned}$$

because each successive differention brings down another factor of $-X$ Then we can continue for the $n$ case:

$$\mathbb{P}(N_\lambda(X) = n) = \frac{\lambda^n}{n!}\left(-\frac{d}{d\lambda}\right)^n\varphi_X(\lambda)$$

which we can get from repeated differentiation of the Laplace transform! Pitman cites "An Introduction to Probability Theory and Its Applications, Volume II" by Feller and suggests that we look into this if we have a background in analysis. The argument here is extremely clever in the combination of ideas in inverting a Laplace transform, using convergence in distribution and conditioning to bring together small and intuitive steps, to exploit properties of a Poisson point process to prove what is essentially a theorem of real analysis, that the probability distributon of a random variable $X \geq 0$ is determined by its Laplace transform.

## 21.5 A First Look at Martingales

Let's paint the setting. Consider some background process $X_0, X_1, X_2, \ldots$ which carries some information that is evolving over time. We think of the stretch as a vector $(X_0, X_1, \ldots, X_n)$ to be a vector of the history up to stage $n$. We caution that the $X$s might not be numerical, as they may be letters, permutations, trees, or anything that can encode the available data up to time $n$.

We have a numerical process which we call $M_0, M_1, M_2, \ldots$, and the idea is that $M_n$ is our accumulated fortune from some gambling game. For a *Martingale*, the game is fair. We also have a notion of a "sub martingale" where the game is favorable, or a "super martingale" where the game is unfavorable. All of these

definitions are conditionally given the past. Following Durett, we do not assume that $M_0$ is a function of $X_0$. That is, we make no assumptions on the starting condition $(M_0, X_0)$ except that $\mathbb{E}|M_0| < \infty$. Now, we can talk about the Martingale property.

Let $M_n$ be a function of $(M_0, X_0, X_1, \ldots, X_n)$ in some general way. The key property is that if we look at the conditional expectation of the next variable, if we know the history, this expectation should be equal to the current variable: this is the fairness condition:

$$\mathbb{E}(M_{n+1} \mid M_0, X_0, X_1, \ldots, X_n) = M_n, \quad \forall\, n = 0, 1, 2, \ldots$$

an obviously equivalent condition is in terms of increments:

$$\mathbb{E}(M_{n+1} - M_n \mid M_0, X_0, X_1, \ldots, X_n) = 0, \quad \forall_{n=0,1,2,\ldots}$$

These equalities become inequalities $\leq$ in a super martingale (unfair) and $\geq$ in a sub martingale (favorable). This inequality is the same for every step. To make sure our expectation makes sense, we need to make sure that $\mathbb{E}|M_n| < \infty$. To be pedantic, this is a Martingale with respect to the $(X_n)$ sequence. If no other sequence is mentioned, it is taken that $X_n = M_n$. There is just one game being played in the casino, and there is no opportunity to use additional randomization for e.g. building a gambling strategy.

### 21.5.1 Examples

There are many old friends we can revisit here.

#### 21.5.1.1 Example 1

Take $X_1, X_2, \ldots$ iid with $\mathbb{E}|X_i| < \infty$ and $\mathbb{E}X_i = \mu$. Take $S_n = S_0 + X_1 + \cdots + X_n$, where $S_0$ is assumed to be independent of $X_1, X_2, \ldots$.
Then we can say $(S_n)$ is a Martingale (MG) relative to $(X_n)$ if and only if $\mathbb{E}X = 0$.
Similarly, $(S_n)$ is a super martingale relative to $(X_n)$ if and only if $\mathbb{E}X \leq 0$.
Finally, $(S_n)$ is a sub martingale if and only if $\mathbb{E}X \geq 0$.
Notice that these inequalities are weak (not strict) so that a martingale (fair) is both a super martingale and sub martingale.

#### 21.5.1.2 Example 2

Consider the same setup but now ask how we may make $S_n^2$ into a Martingale? Let's write this as:

$$\begin{aligned} S_{n+1}^2 - S_n^2 &= (S_n + X_{n+1})^2 - S_n^2 \\ &= S_n^2 + 2S_nX_{n+1} + X_{n+1}^2 - S_n^2 \\ &= 2S_nX_{n+1} + X_{n+1}^2 \end{aligned}$$

Then this equality gives that:

$$\begin{aligned}
&\mathbb{E}\left(S_{n+1}^2 - S_n^2 \mid S_0, X_1, \ldots, X_n\right) \\
&= \mathbb{E}(2S_n X_{n+1} + X_{n+1}^2 \mid S_0, X_1, \ldots, X_n) \\
&= 2\mathbb{E}(S_n X_{n+1} \mid S_0, X_1, \ldots, X_n) + \mathbb{E}(X_{n+1}^2 \mid S_0, X_1, \ldots, X_n) \\
&= 2(\mathbb{E}X)S_n + \mathbb{E}X^2 \\
&= \mathbb{E}X^2 \text{ if } \mathbb{E}X = 0,
\end{aligned}$$

where $S_n$ is a function of the past history. In conclusion, we get this result, if $\mathbb{E}X = 0$ and $\mathbb{E}X^2 < \infty$, then

- $S_n$ is a Martingale relative to the history of $(X_n)$ and
- $(S_n^2 - n\mathbb{E}X^2)$ is a Martingale relative to $(X_n)$.

This is because

$$\begin{aligned}
&\mathbb{E}(S_{n+1}^2 - (n+1)\mathbb{E}X^2 \mid S_0, X_1, \ldots, X_n) \\
&= \mathbb{E}(S_{n+1}^2 \mid S_0, X_1, \ldots, X_n) - (n+1)\mathbb{E}X^2 \\
&= S_n^2 - n\mathbb{E}X^2.
\end{aligned}$$

That is, $M_n := S_n^2 - n\mathbb{E}X^2$ is a Martingale.

#### 21.5.1.3 Example 3

Take $X_0, X_1, X_2, \ldots$ to be a Markov chain with parameters $(\lambda, P)$. Take $h$ to be a real function. Then

$$\mathbb{E}(h(X_{n+1}) \mid X_0, \ldots, X_n) = (Ph)(X_n)$$

Now because

$$(Ph)(x) = \sum_y P(x,y)h(y).$$

So if $Ph = h$ (that is, $h$ is harmonic), then

$$\overbrace{\mathbb{E}(h(X_{n+1}}^{M_n} \mid X_0, \ldots, X_n)) = \overbrace{h(X_n)}^{M_n}$$

Then $M_n := h(X_n)$ is a Martingale. Thus and the idea of a Martingale is a **generalization** or abstraction of the idea of a harmonic function of a Markov chain. In a way, martingales have been a part of our past discussions. Nearly all the things that have worked out nicely in our past lectures work out due to the Martingale process e.g. all features of the gambler's ruin problem. We'll see this more in our readings in the text.

# LECTURE 22

# Conditional Expectation and Martingales

### Class Announcements

Reading on Martingales: all of chapter-5, a nice short chapter, about 20 pages or so. Moreover, you'll find a lot of repetition of ideas that we've covered; introduced in the context of specific applications. Chapter-5 provides a more general, abstract, and intuitive treatment of these ideas.

## 22.1 Food for Thought

Nothing at least obviously to do with Martingales. We present an adaptation of the Poisson Process (PP) on a line that we've seen all semester. In particular, a PP with rate 1. We denote the sequence of arrivals by

$$\gamma_1 < \gamma_2 < \cdots$$

where as a reminder

$$\gamma_k = \text{ sum of } k \text{ iid } \textbf{Exponential}\,(1) \text{ random variables}$$

and hence of course $\gamma_k$ is gamma$(k, 1)$ distributed with mean $k$. This implies that $\gamma_1, \gamma_2, \ldots$ are points of a Poisson Point Process with rate 1, denoted PPP(1). We now present a graphical construction by generating semi-circles who's diameters are the intervals $[\gamma_{k-1}, \gamma_k]$ whose lengths are the i.i.d. **Exponential** (1) variables. Here $\gamma_0 = 0$. See Figure 22.1. Pitman asserts that having drawn schematic pictures like this in the past, this should remind us of *Renewal Theory*.

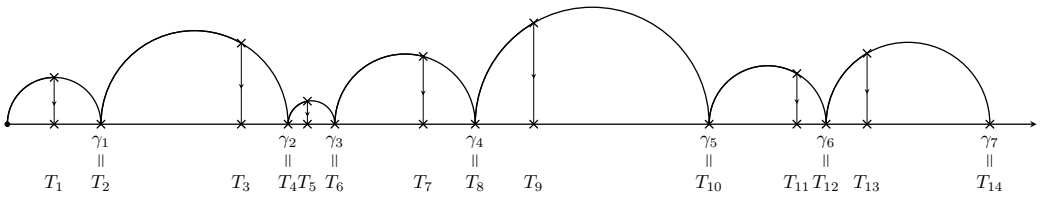



Figure 22.1: Realization of a new PPP, resulting from marking the horizontal location for a point chosen uniformly on a semi-circle with diameter corresponding to an **Exponential**(1) random variable. By design $T_{2n} = \gamma_n$.

Now from here, we select a point uniformly at random by length of the arc or if you like by the arc angle relative to the center of each semicircle. Sometimes these points come early, sometimes later, and other times appear towards the middle. From there, we "rain the points down." That is, we project each point onto the horizontal axis; marking it's horizontal coordinate. This in turn produces a *new* point process with arrivals, which we denote by

$$0 < T_1 < T_2 < \cdots$$

where by design

$$\gamma_n = T_{2n}$$

As an exercise, which is completely in the scope of this course and may appear on your final exam:

*What sort of point process are the $T$'s?*

As a harder exercise. Discuss what happens if you replace the scheme of picking from arc length on the diameter by some other scheme, e.g. picking the extra points uniformly and independently from $[\gamma_{k-1}, \gamma_k]$.

## 22.2 Conditional Expectation

Pitman asserts that Martingales are all about conditional expectation. You cannot understand Martingales without first understanding conditional expectation. Hence, you must get your head around what

$$\mathbb{E}(Y \mid X)$$

is. Which is an amazingly good notation for a very important mathematical idea. We must ask

- What is $X$? Answer: $X$ can be any random object including a vector. That is, $X$ is encoding a list of things. For example, in the context of Martingales

  $$X = (M_0, X_0, \ldots, M_n, X_n)$$

  It's imperative to understand that $X$ is *coding* information.

- What is $Y$? Answer: Usually for us it is a real valued numerical random variable. It is okay for $Y$ to be a random vector, where everything reduces component-wise into random variables. Random vectors are nothing to be intimidated by, the interesting thing of course is there may be some dependency between the random variables. We'll stick with 'real valued random variable.'

$X$ and $Y$ are rather abstract objects and of course we're working in the context of some probability space or probability modeling. These are primitive concepts that we must understand in various levels of generality. We *always* assume[1] that $X$ and $Y$ are defined some *common* probability space

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

where $\Omega$ being the set of outcomes, $\mathcal{F}$ the set of events that we can measure the probability of, and $\mathbb{P}$ the probability measure and especially $\mathbb{P}$ = probability and $\mathbb{E}$ = expectation, being in the background.

### 22.2.1 Defining $\mathbb{E}(Y \mid X)$

Back to the question: What is $\mathbb{E}(Y \mid X)$? Answer: $\mathbb{E}(Y \mid X)$ is a *random variable*. It is created from the joint distribution of $X$ and $Y$ with special properties.[2]

- $\mathbb{E}(Y \mid X)$ is a *function of* $X$.
- $\psi(x) = \mathbb{E}(Y \mid X = x)$, where the argument $x$ is a suitable real number.

If for simplicity, we assume that $X$ and $Y$ are discrete. Then $\mathbb{P}(X = x) > 0$ for some countable list of values $\mathcal{S}$ with $\sum_{x \in \mathcal{S}} \mathbb{P}(X = x) = 1$. Which means that all the probability is accounted for by a list of possible values, then

$$\mathbb{E}(Y \mid X = x) = \sum_{\text{values } y \text{ of } Y} y \, \mathbb{P}(Y = y \mid X = x)$$

where of course, by the usual Bayes rule

$$\mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(Y = y, X = x)}{\mathbb{P}(X = x)}$$

In other words, the function value for a particular $x$ is just the average of the values of $y$, if you know that $x$. We now present a "silly" example.

### 22.2.2 Example: Heights of Students

Let $\Omega$ be the students in this room. Hence, each student is an $\omega$. We have two variables:

- $Y(\omega)$ = height of student $\omega$

[1] often implicitly

[2] Pitman emphasizes that we must *parse* the notation, making sense of it. Look at the symbols on the page and understand what the person writing these symbols was intending for them to mean.

- $X(\omega)$ = row number of student $\omega$

These are attributes associated for each $\omega$. Conceptually, there's no difference between labeling rows by numbers or by letters. That is, there's no need for $X$ to be numerical. There is a need for $Y$ to be numerical, since attribute of height is clearly numerical. We establish a probability model

$$P(\omega) = \frac{1}{N}$$

where $N$ = the number of students in the room. We have each student fill out a ticket following a configuration. See Figure 22.2 below.

| $\omega$ | $X(\omega)$ | $Y(\omega)$ |
|---|---|---|

Figure 22.2: Typical ticket configuration.

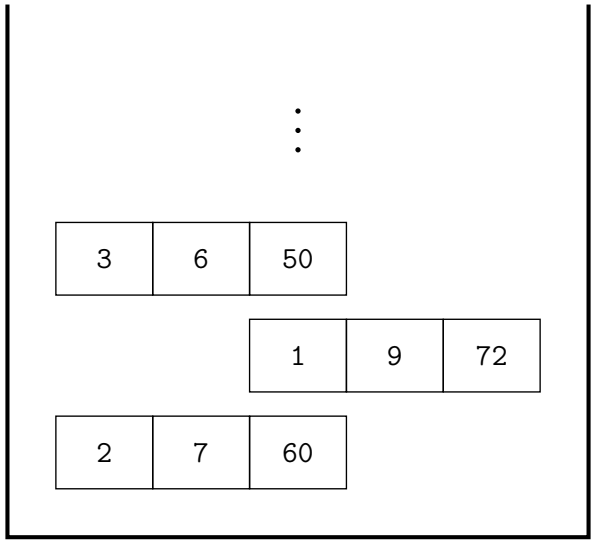



Figure 22.3: Box of $N$ tickets.

We collect the $N$ tickets into a box. See Figure 22.3. $\mathbb{E}(Y)$ is clearly the expected height of a student chosen uniformly in the class. Also known as the mean or average of the height of a student in the class. That is,

$$\mathbb{E}(Y) = \frac{1}{N} \sum_{\omega \in \Omega} Y(\omega)$$

We won't discuss $\mathbb{E}(X)$, as it's uninteresting in this context. However, we observe for $N(x) =$ the number of students in the $x$th row

$$\mathbb{E}(Y \mid X = x) = \frac{1}{N(x)} \sum_{\{\omega : X(\omega) = x\}} Y(\omega) \tag{22.1}$$

Recall $\mathbb{E}(Y \mid X)$ is a random variable; a function of $X$. In particular and from earlier

$$\psi(x) = \mathbb{E}(Y \mid X = x) \tag{22.2}$$

which is the average height of the students in row $x$. Each $\omega$ acquires this attribute by sitting in row $x$. Being pedantic, we could also express the above as

$$\mathbb{E}\big(Y \mid X = x\big)(\omega)$$

**Single Most Important Thing about $\mathbb{E}(Y \mid X)$**

The following is guaranteed to be on the final.

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

The order of equality is intentional, as we compute the left hand-side by computing the right hand-side. In the theory of Markov chains, we've seen this over and over again, finding the right $X$ to condition on.

We realize the boxed construction is obvious from the example, recalling (22.2) and (22.1) and

$$\begin{aligned}
\mathbb{E}\big[\mathbb{E}(Y \mid X)\big] &= \sum_x \mathbb{P}(X = x)\mathbb{E}(Y \mid X = x) \\
&= \sum_x \frac{N(x)}{N} \frac{1}{N(x)} \sum_{\{\omega : X(\omega) = x\}} Y(\omega) \\
&= \frac{1}{N} \sum_{\omega \in \Omega} Y(\omega) \\
&= \mathbb{E}(Y)
\end{aligned}$$

This "box of tickets" image is enormously powerful. Any discrete set-up with a finite number of outcomes can be visualized or imagined as a box of tickets. The key idea here is $\mathbb{E}(Y \mid X)$ is an effort to predict a value of $Y$ given you know $X$. If you know $X$, use the row rather than the class average. It's obvious you'll get better results.

Pitman mentions the Netflix competition. A generic competition sponsored by Netflix to promote machine learning. The goal of predicting which movies to recommend to customers based on the movies they've already seen. The prediction problem is everywhere and its treated in a conceptual way in a conceptual framework, that being the framework of conditional expectation.

### 22.2.3 Example: Predicting $Y$, Knowing Only its Distribution

A toy problem. What is the best predictor of $Y$ if you know only the distribution of $Y$. For this, we impose a penalty or loss for this to make sense. Thanks to Gauss, we utilize a *quadratic loss*. Observe,

$$\mathbb{E}(Y-y)^2 = \mathbb{E}(Y-\mu+\mu-y)^2$$

where

$$\begin{aligned}\mu = \mathbb{E}(Y) &= \mathbb{E}\big[(Y-\mu)+(\mu-y)\big]^2 \\ &= \mathbb{E}(Y-\mu)^2 + (\mu-y)^2 + 2\mathbb{E}(Y-\mu)(\mu-y) \\ &= \mathbb{E}(Y-\mu)^2 + (\mu-y)^2\end{aligned}$$

Observe $2\mathbb{E}(Y-\mu)(\mu-y)=0$, since

$$\begin{aligned}2\mathbb{E}(Y-\mu)(\mu-y) &= 2(\mu-y)\mathbb{E}(Y-\mu) \\ &= 2(\mu-y)(\mathbb{E}Y-\mu) \\ &= 2(\mu-y)(\mathbb{E}Y-\mathbb{E}Y) \\ &= 0\end{aligned}$$

Pitman mentions that when you solve a quadratic calculus problem, that minimizing or maximizing a quadratic is all about completing the square. In fact we just did it above. Hence, we minimize by taking $y=\mu$ with the minimum being

$$\mathbb{E}(Y-\mu)^2 = \mathbb{V}\mathrm{ar}(Y)$$

If you are given $X$, how now to predict $Y$ the best? (Of course for mean squared error.) On average you making sure you make a small mistake, where the penalty is a a square. The penalty is interesting as it's very expensive to make big mistakes and does not cost so much to make small mistakes. It's obvious, given that you know $X$, you condition on it. Getting into a homework exercise, prove

$$\psi(X) = \mathbb{E}(Y \mid X)$$

does minimize

$$\mathbb{E}(Y-\psi(X))^2$$

over all (measurable) functions of $X$

### 22.2.4 Basic Properties of $\mathbb{E}(Y \mid X)$

All of these properties are easily derived from the definitions and they possess a strong intuitive content.

- $\mathbb{E}(Y + Z \mid X) = \mathbb{E}(Y \mid X) + \mathbb{E}(Z \mid X)$
- $\mathbb{E}(cY \mid X) = c\mathbb{E}(Y \mid X)$, where $c$ is constant
- $\mathbb{E}\big[g(X)Y \mid X\big] = g(X)\mathbb{E}(Y \mid X)$
- $\mathbb{E}\big[g(X) \mid X\big] = g(X)$ ($Y = 1$ from the above property)

provided both $Y$ and $g(X)Y$ are integrable and $\mathbb{E}|Y| < \infty$, $\mathbb{E}|g(X)Y| < \infty$.

## 22.3 Exponential Martingales

Recall that $M_n$ is your accumulated fortune at time $n$ in some gambling game. We reiterate the definition of a Martingale.

**Martingale**

$M_0, M_1, M_2, \ldots$ is a *Martingale* with respect to $X_0, X_1, X_2, \ldots$, if

$$\mathbb{E}(M_{n+1} \mid X_0, M_0, X_1, M_1, \ldots, X_n M_n) = M_n$$

which is equivalent to

$$\mathbb{E}(M_{n+1} - M_n \mid X_0, M_0, X_1, M_1, \ldots, X_n M_n) = 0$$

### 22.3.1 Example: Exponential Martingale

Let $X_1, X_2, \ldots$ are iid. Look at $Y_n = X_1 X_2 \cdots X_n$. Let's compute

$$\begin{aligned}\mathbb{E}(Y_{n+1} \mid X_1 X_2 \cdots X_n) &= \mathbb{E}(X_1 X_2 \cdots X_n X_{n+1} \mid X_1 X_2 \cdots X_n) \\ &= \mathbb{E}(Y_n X_{n+1} \mid X_1 X_2 \cdots X_n) \\ &= Y_n \mathbb{E}(X_{n+1} \mid X_1 X_2 \cdots X_n)\end{aligned}$$

We'd like this to be Martingale. Assume the $X_i > 0$ and from there, we clearly see

$$\mathbb{E}(X_{n+1} \mid X_1 X_2 \cdots X_n) = 1$$

will make a Martingale. Hence all we require is $\mathbb{E}X_i = 1$. A variation to this problem is as follows, consider

$$Z_n = \exp\left[\theta(X_1 + X_2 + \ldots + X_n)\right] = \prod_{i=1}^{n} \exp(\theta X_i)$$

will be Martingale if and only if $\mathbb{E}e^{\theta X_i} = 1$.

### 22.3.2 What to Know about Martingales

If $M_n$ is a Martingale relative to anything, then

$$\mathbb{E}(M_n) = \mathbb{E}M_0$$

which is almost obvious and intuitive from the fair game interpretation.

*Proof.* By mathematical induction. The claim is trivial for $n = 0$. Suppose the claim holds for $n$, then

$$\mathbb{E}(M_{n+1}) = \mathbb{E}\big[\mathbb{E}(M_{n+1} \mid M_n)\big] = \mathbb{E}(M_n)$$

□

# LECTURE 23

# Martingales and Stopping Times

11/21/2019.

## 23.1 Warm-up Examples of Martingales

As warm-up, let's consider these examples. Suppose that $\mathbb{E}(Y|X) = \psi(X)$. We'll look at two things:

- Suppose that we don't know $X$, but we know $\psi(X)$. Then what is
$$\mathbb{E}(Y \,|\, \psi(X))?$$
- If $b$ is a bijection, e.g. $b(X) = 2X + 3$ or $b(X) = 5 - 2X$ or $b(X) = X^3$ for a real random variable $X$, or $b(X) = X^2$ for a positive random variable $X$, then what is
$$\mathbb{E}(Y \,|\, b(X))?$$

These may be on our final exam. There are several things like this on the midterm practice problems. Pitman suggests that we think of this in terms of the height $Y(\omega)$ of student $\omega$ picked uniformly at random from a class, with $X(\omega)$ their row number, and $\psi(X(\omega))$ the average height of all students in the same row as student $\omega$. Then the answers to such questions should be intuitive, and once guessed, they are easily checked using any one of the many characterizations of $\mathbb{E}(Y|X)$ that we now know.

## 23.2 Stopping Times for Martingales

There's quite a nice discussion of this in Durett. The setup here is to take
$$(\Omega, \mathcal{F}, \mathbb{P})$$
as our probability space, and
$$X_0, X_1, X_2, \ldots$$

as some background process and $X_n = (X_n(\omega), \omega \in \Omega)$. This may or may not be numerical, as this is unimportant (e.g. $X_n$ could be a vector in $\mathbb{R}^p$ or a permutation of the first $n$ integers, so the length of the vector grows with $n$).

**Martingale Property**

We say $(M_n, n = 0, 1, 2, \dots)$ is a *martingale*, relative to $X_0, X_1, \dots)$ *if and only if* for $\mathbb{E}|M_n| < \infty$ and $M_n$ is a function of $X0, X_1, \dots, X_n$ for all $n$, with

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

or equivalently, which is typically easier to check:

$$\mathbb{E}(M_{n+1} - M_n \mid X_n, \dots, X_0) = 0$$

Note that Durrett has a slightly different definition, allowing $M_0$ not to be a function of $X_0$, and letting $M_n$ be a function of $(M_0, X_0, \dots, X_n)$. That is just the above definition with the initial information $X_0$ expanded to $(X_0, M_0)$. It is cleaner for the following definition just to have one history sequence $(X_n)$, with everything in sight a determininstic function of variables in this sequence. If you want extra randomization you stick it in $X_0$. Recall the notion of a *stopping time* $T$, relative to $(X_0, X_1, \dots)$. That is a random variable which takes on values in $\{0, 1, 2, \dots, \infty\}$ (we typically don't have a discussion if we allow $T$ to take on value 0), subject to any one of the following (equivalent) conditions:

- $(T = n)$ is determined by $(X_0, X_1, \dots, X_n)$ for all $n = 0, 1, 2, \dots$, meaning that the indicator of the event $(T = n)$ is a Boolean function of $(X_0, \dots, X_n)$;
- $(T \le n)$ is determined by $(X_0, X_1, \dots, X_n)$ for all $n = 0, 1, 2, \dots$.
- $(T > n)$ is determined by $(X_0, X_1, \dots, X_n)$ for all $n = 0, 1, 2, \dots$.

We need only manipulate the Booleans by addition and subtraction to show the equivalence. Intuitively, we decide the value of $T$ sequentially in a way that depends on $X_0, X_1, \dots$, but the decision that $(T = n)$ is only allowed to use the past values of the sequence, not to anticipate future vales. It is no problem to allow $M_0$ to play a role, which is to expand $X_0$ to $(X_0.M_0)$ and/or to replace $X_n \mapsto (X_n, M_n)$ for all $n$. This is just changing the basic history. The concept of stopping time relative to a history remains the same. Let's recall something we know about stopping times, in this fairly large level of generality. Special to the structure of Markov chains, we have the Strong Markov Property. There's another special rule that applies to i.i.d. sequences $X_1, X_2, X_3, \dots$.

### 23.2.1 Case: IID Sequence

Assume that $X_0$ is anything, independent of $X_1, X_2, X_3$ which are iid real-valued random variables like $X$, with $\mathbb{E}|X| < \infty$. Let $S_n := \sum_{k=1}^{n} X_k$ with $S_0 = 0$ according to an empty sum. The $X_0$ is just there to allow stopping times involving additional randomization independent of the $X_1, X_2, \dots$. Then we have a few key points:

- $S_n - n\mathbb{E}X$ (after centering) is a martingale relative to $(X_0, X_1, X_2, \dots)$.

- If $T$ is a stopping time and $\mathbb{E}(T) < \infty$, then $\mathbb{E}|S_T| < \infty$ and

$$\mathbb{E}S_T = (\mathbb{E}T)(\mathbb{E}X)$$

  which is *Wald's identity*. We've proved in this in Lecture 12.

Notice that if we set $M_n := S_n - n\mathbb{E}X$, which is our martingale of the moment, then Wald's identity gives

$$\mathbb{E}M_T = \mathbb{E}M_0 = \mathbb{E}0 = 0$$

Unpacking this a bit, assuming that $\mathbb{E}T < \infty$, we have $M_T = S_T - T\mathbb{E}(X)$, and so $\mathbb{E}M_T = 0$. This implies

$$\mathbb{E}S_T = (\mathbb{E}T)\mathbb{E}(X)$$

We have preservation of the expectation that works at the stopping time.
Notice that $S_0 = 0$ is to make empty sums ($\varnothing$). For example, we would like to include the case where $T$ is independent of $(X_1, X_2, \dots)$. To see this, we need to go back to the definition of stopping times and realize that $X_0$ is included to allow that $T$ is not only a function of $(X_1, X_2, \dots)$, and hence can be independent. Alternatively, we can even let $X_0 := T$ as an artificial way. This is why we are deliberately not including $X_0$ in our sum, including this only to allow for generality and ad hoc randomness.

### 23.2.2 Generalizing Our Example

Now the question arises: Suppose $(M_n, n = 0, 1, 2, \dots)$ is a martingale, and $T$ is a stopping time with $\mathbb{E}T < \infty$. We may ask: Is it true that $\mathbb{E}M_t = \mathbb{E}M_0$? This worked for $M_n = S_n - n\mathbb{E}X$ in the iid case, but does this work more generally? Pitman answers that **no**, we do not have this necessarily if $T$ is *unbounded.* However, we do have this property if $T$ is bounded. amd with care about switching limits and $\mathbb{E}$ it is possible to establish it in some unbounded cases. Let's define what this means. We say that $T$ is bounded if there is some *fixed* number $b$ such that

$$\mathbb{P}(T \leq b) = 1$$

That is, it is certain that $T$ is less than or equal to the bound.

### 23.2.3 Issues When $T$ is Unbounded

From the previous class, we consider an exponential or multiplicative martingale. Here is an instance of that. Let $X_1, X_2, \dots$ be i.i.d. **Bernoulli**$(\frac{1}{2})$ random variables. We are simply doing fair coin tosses to decide our $X_i$, so that $X$ has value 0 with probability $\frac{1}{2}$, and value 1 with probability $\frac{1}{2}$. In formulas, we'll define our game

$$M_0 := 1 \quad (\varnothing \text{ product } = 0)$$

$$M_n := 2^n X_1 X_2 \cdots X_n = \prod_{k=1}^{n} (2X_k),$$

which we know commonly as the "double-or-nothing" gambling game. This is more or less historically the first major use of the martingale term in a gambling context, which preceded the general mathematical definition, due to Jean Ville, and adopted by Joe Doob.

Clearly 0 is an absorbing state in this game. We should check that this is a martingale by a prior discussion. A very simple check would be to see that

$$\begin{aligned}\mathbb{E}M_n &= \mathbb{E}\left(\prod_{k=1}^{n} 2X_k\right)\\ &= \prod_{k=1}^{n} \underbrace{(\mathbb{E}2X_k)}_{2\cdot\frac{1}{2}=1}\\ &= 1,\end{aligned}$$

Also, we should double-check[1] that

$$\mathbb{P}(M_n = 2^n) = \left(\frac{1}{2}\right)^n = 2^{-n}$$

So there is a very small chance that we end up with a very large fortune. On the other hand, we have the other path

$$\mathbb{P}(M_n = 0) = 1 - 2^{-n}$$

Hence again

$$\mathbb{E}M_n = 2^{-n}\cdot 2^n + (1-2^{-n})\cdot 0 = 1$$

We looked at this example to see that there can be issues (trouble) at particular stopping times. Recall that we've defined that $T := \min\{n : M_n = 0\}$. Notice that $\mathbb{P}(T > n) = 2^{-n}$ so that $T$ is unbounded. Then

$$\begin{aligned}\mathbb{E}T &= \sum_{n=0}^{\infty} \mathbb{P}(T > n)\\ &= \sum_{n=0}^{\infty} 2^{-n}\\ &= 2 < \infty.\end{aligned}$$

and so $T \sim \textbf{Geometric}(\frac{1}{2})$ on $\{1, 2, 3, \dots\}$. This is good because

$$\mathbb{P}(T < \infty) = 1$$

However, $\mathbb{P}(M_T = 0) = 1$. So in summary, we found:

$$1 = \mathbb{E}M_0 = \mathbb{E}M_1 = \mathbb{E}M_2 = \cdots, \qquad M_n \to M_T = 0,$$

[1] Pitman jokes as this is a double-or-nothing game.

where the convergence of random variables holds with probability one, but it turns out $\mathbb{E}M_T \neq 1$. Concusion, if we have unbounded stopping times $T$, the preservation of expectations can fail even if $\mathbb{P}(T < \infty) = 1$, and for a general martingale even the stronger condition $\mathbb{E}T < \infty$ is not enough to deduce $\mathbb{E}M_T = \mathbb{E}M_0$, as was the case in Wald's identity for the centered i.i.d. sums.

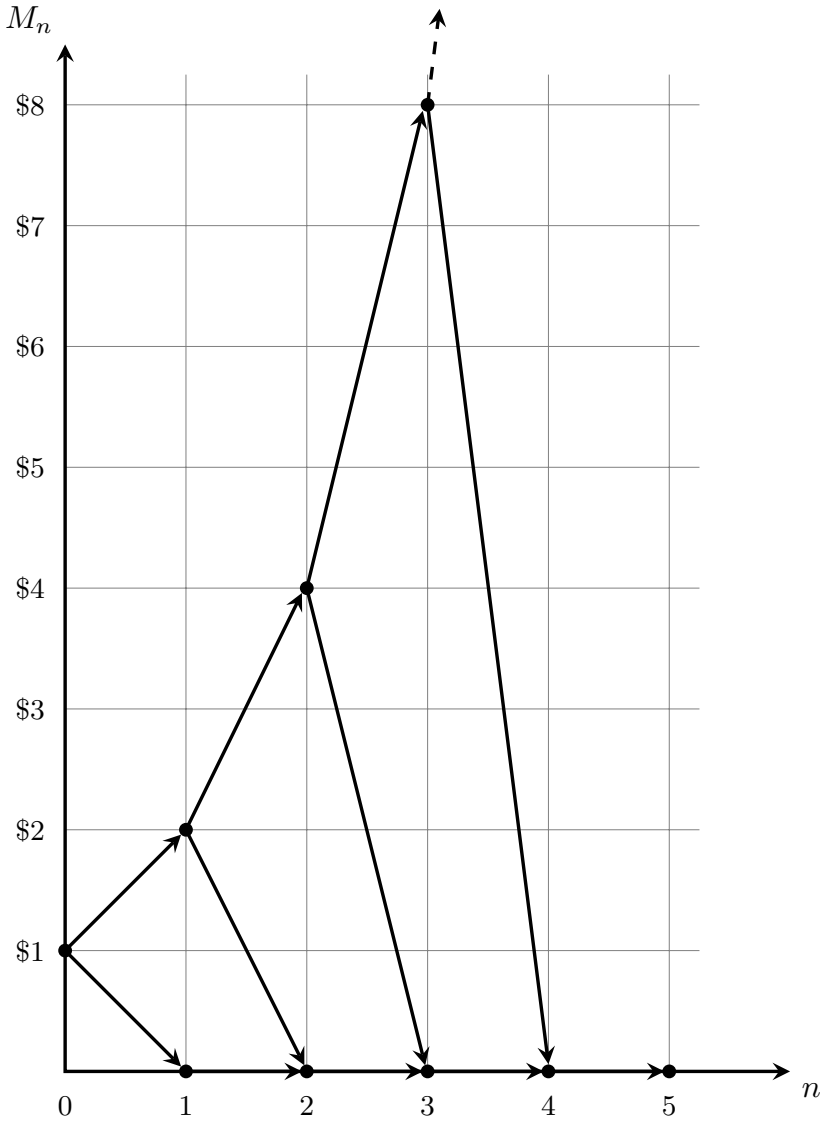



Figure 23.1: Double or nothing game. Note the probability of $M_n \to \infty$ is zero.

Let's look at a generic picture of a CDF, and assume that $X \geq 0$ for simplicity. The shaded region above the CDF is $\mathbb{E}(X)$. One way to understand this is to think of creating a uniform random variable $U$ on $[0, 1]$ and pick the value $X = X(U)$ to be the inverse CDF.

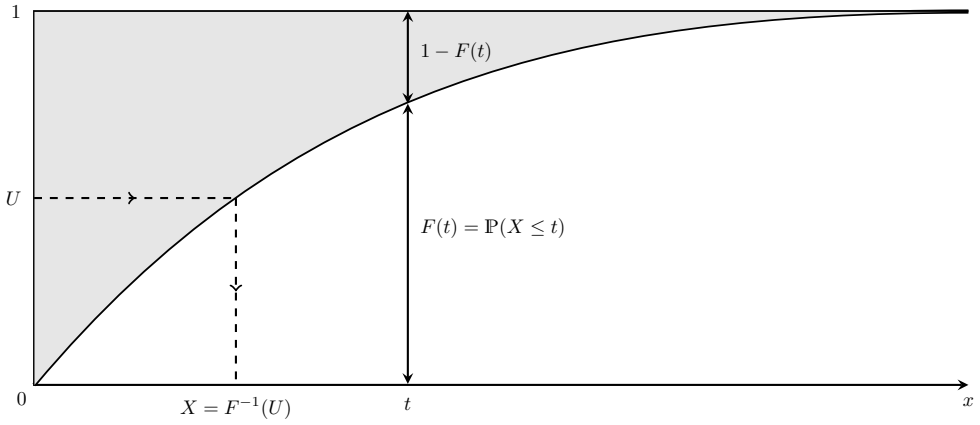



Figure 23.2: For a non-negative $X$, with CDF $F(x) = \mathbb{P}(X \leq x)$, the expectation $\mathbb{E}X = \int_0^\infty (1 - F(x))dx$ is the shaded area between the graph of $F(x)$ and level 1 as shown above. For a uniform random variable $U$ along the vertical, we can create $X$ as the inverse CDF of $U$.

#### 23.2.3.1 Expectation of $M_n$

It is instructive to visualize the expectation of $M_n$ for different $n$. Notice that every one of these areas is 1, but the shaded region (rectangle) is getting much longer. In the limit, we are guaranteed to lose our money (with probability 1) even in this fair gamble.

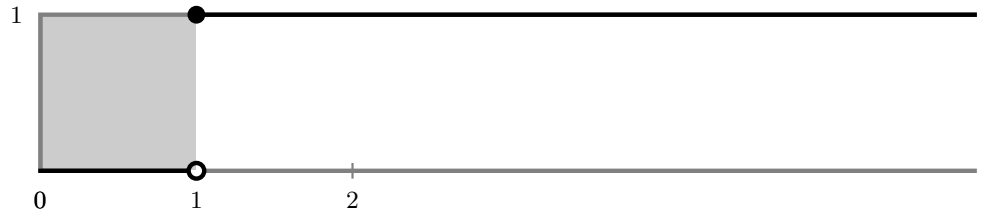



Figure 23.3: Realization of $\mathbb{E}M_0$

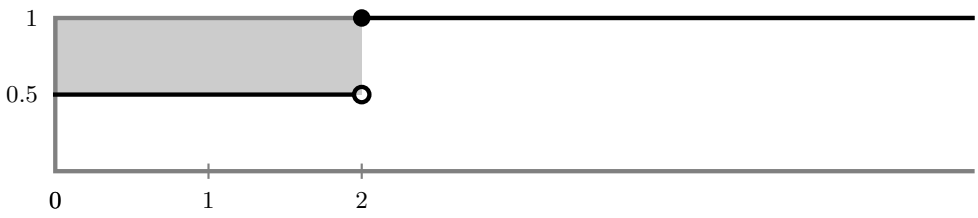



Figure 23.4: Realization of $\mathbb{E}M_1$

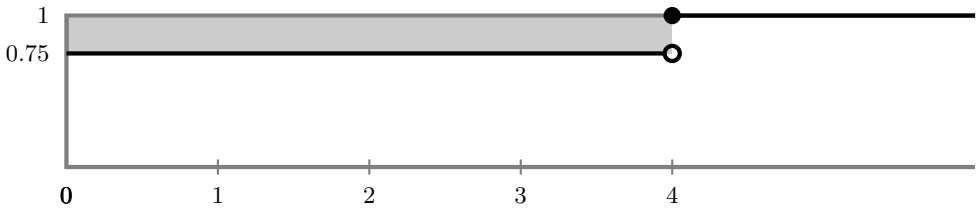



Figure 23.5: Realization of $\mathbb{E}M_2$

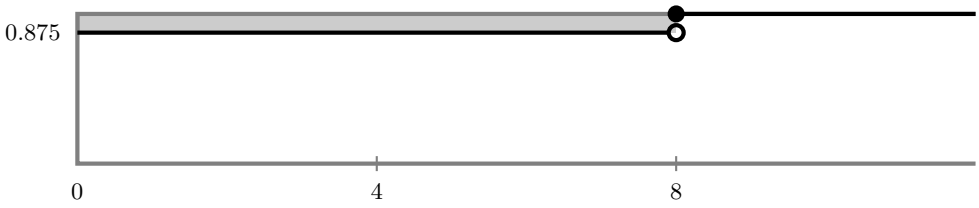



Figure 23.6: Realization of $\mathbb{E}M_3$

## 23.3 Expectations In Bounded Stopping Times

If we assume that $(M_n)$ is a martingale relative to $(X_0, X_1, \dots)$, and $T$ is a **bounded** stopping time, then

$$\mathbb{E}M_0 = \mathbb{E}M_T$$

First, it is natural to try to prove this by conditioning on $T$. If $T \leq b$, then

$$M_T = \underbrace{\sum_{n=0}^{b} \mathbb{1}(T=n)}_{1} M_T = \sum_{n=0}^{b} M_n \mathbb{1}(T=n)$$

Then this tells us that we can certainly compute that

$$\begin{aligned}\mathbb{E}M_T &= \sum_{n=0}^{b} \mathbb{E}\left[M_n \mathbb{1}(T=n)\right] \\ &= \sum_{n=0}^{b} \underbrace{\mathbb{E}(M_n|T=n)} \cdot \mathbb{P}(T=n) \quad \text{(equivalently by conditioning)}\end{aligned}$$

This underbraced part can be problematic. Now if we assume that $T$ is independent of $(M_n)$, since we know that $\mathbb{E}M_0 = \mathbb{E}M_n = \mathbb{E}(M)n \mid T = n)$, we can continue in the obvious way to get

$$\begin{aligned}\mathbb{E}M_T &= \sum_{n=0}^{b} (\mathbb{E}M_n)\mathbb{P}(T=n) \\ &= \sum_{n=0}^{b} (\mathbb{E}M_0)\mathbb{P}(T=n) \\ &= \mathbb{E}(M_0) \sum_{n=0}^{b} \mathbb{P}(T=n) \\ &= (\mathbb{E}M_0) \cdot 1 \\ &= \mathbb{E}M_0\end{aligned}$$

Pitman suggests that we formulate examples where $\mathbb{E}(M_n|T=n) \neq \mathbb{E}M_0$. For such examples, the above approach fails to get the desired result, which might lead us to think the result is incorrect. However, what happens is that if there are terms that are larger than $\mathbb{E}M_0$, then they must be compensated by some terms that are smaller than $\mathbb{E}M_0$. This is not obvious but can be proved as follows.

Now we have a key idea for how to proceed without the assumption that $T$ is independent of $(M_n)$. We'll take $T$ to be a stopping time, even allowing the case where $\mathbb{P}(T=\infty) > 0$. The key is to show that if $(M_n)$ is a martingale relative to $(X_n)$ and $T$ is a stopping time relative to $(X_n)$, then

$$(M_{n\wedge T}, n = 0, 1, 2, \dots)$$

is a martingale relative to $(X_0, X_1, \dots)$. In particular,

$$\mathbb{E}M_{n\wedge T} = \mathbb{E}M_0$$

for every $n$, and

$$\mathbb{E}M_T = \mathbb{E}M_0 \text{if } \mathbb{P}(T \leq b) = 1$$

for some $b < \infty$, by taking $n \geq b$ so that $n \wedge T = T$.

Recall that we've defined:

$$n \wedge m := \min\{n, m\}$$

$$n \vee m := \max\{n, m\}$$

So we have:

$$M_{T\wedge n} := \begin{cases} M_T & , T < n \\ M_n = M_T & , T = n \\ M_n & , T > n. \end{cases}$$

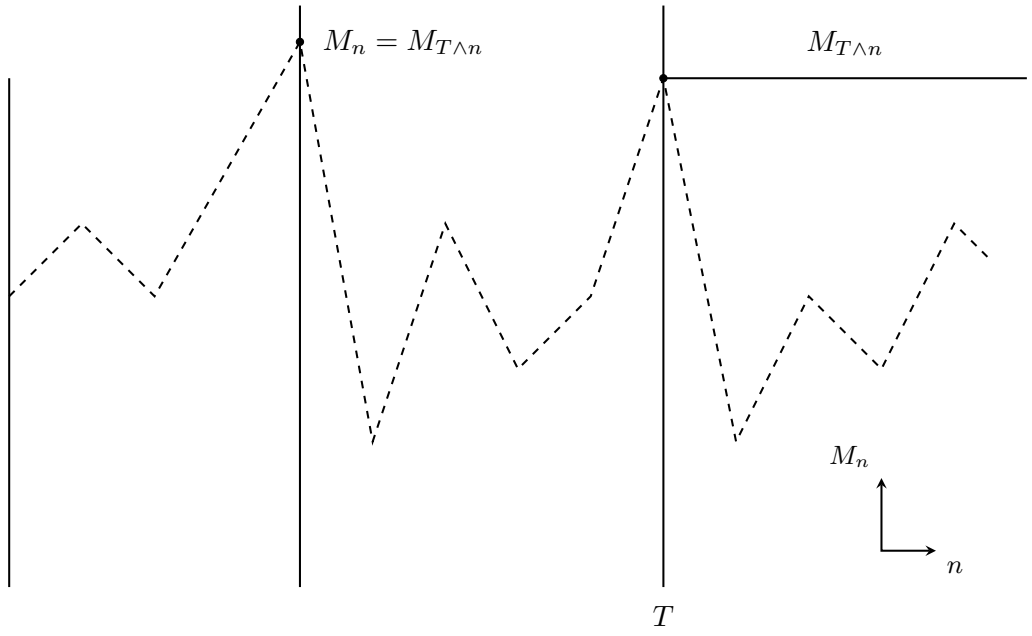



Figure 23.7: Realization of cases of a stopping martingale $M_n$ and stopping time $T$.

Intuitively, in the fair gambling game, we can choose to stop and walk away with our money at any (stopping) point in time, and the net result after $n$ steps should be fair. Now let's look at the mathematical proof. As a matter of technique, we will check the following

- $M_{n\wedge T}$ is a function of $(M_0, X_0, M_1, X_1, \ldots, M_n, X_n)$.

$$M_{n\wedge T} = \sum_{k=0}^{n-1} \underbrace{\mathbb{1}(T = k)}_{f(X_0,\ldots,X_k)} M_k + \underbrace{\mathbb{1}(T \geq n)}_{f(X_0,\ldots,X_n)} M_n,$$

  where we can evaluate these indicators as shown in the underbraces by definition of stopping time.

- We need to check the martingale property, and this is most easily accessible via the difference form. That is, we check that

$$\mathbb{E}(M_{(n+1)\wedge T} - M_{n\wedge T} \mid X_0, \ldots, X_n) = 0.$$

Notice that after we've stopped, the difference is trivial and there is nothing to check. It remains to check for the case when we have not yet stopped. Via the method of indicators, we have:

$$M_{(n+1)\wedge T} - M_{n\wedge T} = \mathbb{1}(T \leq n) \cdot 0 + \mathbb{1}(T > n) \cdot (M_{n+1} - M_n),$$

which directly implies that

$$\mathbb{E}\left[M_{(n+1)\wedge T} - M_{n\wedge T}\right] \mathbb{1}(A_v) = \mathbb{E}(M_{n+1} - M_n)\mathbb{1}(T > n)\mathbb{1}(A_v) = 0,$$

for all $A_v$ functions of $(X_0, M_0, \ldots, X_n, M_n)$.

## 23.4 Life Lesson

Pitman closes today's lecture with a life lesson.

**Ultimate Rule for Cooking Perfect Toast**

Cook until you see the toast smokes and cook for 10 seconds less.

This is one intricate example of what is invalid as a stopping time. As another example, we cannot play in the stock market and stop just before we lose. We cannot see into or condition on the future.

## 23.5 Notes on Martingales (Adhikari)

This section and the ones following it are from Professor Adhikari's STAT-150, Spring 2014 course. There is some duplication of content, but also some novel examples. The waiting for patterns examples are of special interest.
We're going to step back from Markov Chains and re-examine a few familiar questions using a different approach. The techniques that we will develop are not only beautiful, they also give us computational power we don't yet have.

**Martingale Property: First Definition**

Be patient for a while about why we won't stop at this definition. $\{M_0, M_1, M_2, \ldots\}$ is a *martingale* if for all $n$

- $E(|M_n|) < \infty$
- $E(M_{n+1} \mid M_0, M_1, \ldots, M_n) = M_n$ (martingale property)

The first condition allows us to talk about expectations without getting into trouble. The second is the "net gain in a fair game" condition: if you think of $M_n$ as your net gain at time $n$ in a gambling game, then the condition says that your expected net gain tomorrow, given the history of your net gains through today, is simply your net gain today. You don't expect to make any more or lose any. Martingales have elegant properties, and the "fair game" formulation can simplify calculations. Before we get to that, here are some examples of martingales.

**Simple symmetric random walk.** The increments are i.i.d., $P(Y_i = 1) = 1/2 = P(Y_i = -1)$, the starting point is a fixed $m_0$, and the process is defined by $M_n = m_0 + Y_1 + Y_2 + \cdots + Y_n$ for all $n$.

$$E(M_{n+1} \,|\, M_0, M_1, \ldots, M_n) \;=\; E(M_{n+1} \,|\, M_n) \;=\; M_n + E(Y_{n+1}) \;=\; M_n$$

The first equality is by the Markov property.

**Sums of independent mean 0 random variables.** The argument above only used the facts that the increments are independent and have mean 0. They don't have to be identically distributed.

**Branching process.** Let $\{X_n\}$ be a Galton-Watson branching process with $P(X_0 = 1) = 1$, and let $\mu$ be the mean of its offspring distribution. We know that the process is a Markov Chain and that $E(X_{n+1} \,|\, X_n) = \mu X_n$. Let $M_n = X_n/\mu^n$. Then $\{M_n\}$ is a martingale, because

$$E(M_{n+1} \,|\, M_0, M_1, \ldots, M_n) \;=\; E(M_{n+1}|M_n) \;=\; \frac{\mu X_n}{\mu^{n+1}} \;=\; M_n$$

Again the first equality is by the Markov property. Frequently we will start with a "base" process and work with another process that is a function of the base process. For this it is convenient to define martingales more generally

**Martingale: Second Definition**

Let $\{B_n\} = \{B_0, B_1, \ldots\}$ be a stochastic process. Then $\{M_n\}$ is a *martingale with respect to* $\{B_n\}$ if the following three conditions are true for all $n$

- $M_n$ is a function of $(B_0, B_1, \ldots, B_n)$
- $E(|M_n|) < \infty$
- $E(M_{n+1} \,|\, B_0, B_1, \ldots, B_n) = M_n$ (martingale property)

**Example.** Let $\{B_n\}$ be the simple symmetric random walk, and let $M_n = B_n^2 - n$. Then $\{M_n\}$ is a martingale with respect to $\{B_n\}$. The first two conditions are easy to verify. For the third, start with $B_{n+1} = B_n + Y_{n+1}$ where $Y_n$ is the $(n+1)$st increment of the simple symmetric random walk.

$$\begin{aligned} E(B_{n+1}^2 \,|\, B_0, B_1, \ldots, B_n) &= E(B_{n+1}^2 \,|\, B_n) \\ &= E([B_n^2 + 2B_n Y_{n+1} + Y_{n+1}^2] \,|\, B_n) \\ &= B_n^2 + 1 \end{aligned}$$

because $E(Y_{n+1}) = 0$ and $E(Y_{n+1}^2) = 1$. So

$$E(M_{n+1} \,|\, B_0, B_1, \ldots, B_n) \;=\; E(B_{n+1}^2 - (n+1) \,|\, B_n) \;=\; B_n^2 + 1 - (n+1) \;=\; B_n^2 - n \;=\; M_n$$

**Expectation at a fixed time.** Use the martingale property to calculate $E(M_{n+1})$ by iterated conditional expectation. Your calculation will show that

$$E(M_n) \;=\; E(M_0) \text{ for every fixed time } n$$

For example, if you start with \$5 and gamble \$1 repeatedly on tosses of a fair coin (that is, if you run a simple symmetric random walk starting at level 5), then at time $n = 100$ you expect to have \$5, and the same is true at any other fixed time $n$.

**Expectation at a random time.** If a result is true for all fixed times, then a natural question is whether it's also true for random times. So let $T$ be a random time. Then is it true that $E(M_T) = E(M_0)$? To answer this, we will start with an important class of random times; most of the random times you've come across in this course belong to this class.

**Stopping Time**

$T$ is a stopping time relative to $\{B_n\}$ if for every $n \geq 1$ the event $\{T \leq n\}$ is determined by $(B_0, B_1, \ldots, B_n)$. More formally, $I(T \leq n)$ is a function of $(B_0, B_1, \ldots, B_n)$, for every $n \geq 1$.

The definition says that in order to decide whether the time has already come, all you need is the history of the process upto today. You don't need to look into the future.

**The first hitting time of a level** is a stopping time. In order to decide whether the process has hit a specified level by today, all you need is its history up to today.

The time of the eventual maximum of a process is **not** a stopping time. In order to decide whether the process has reached its eventual maximum by today, you need not only its history up to today, you also need the entire future so that you can check whether the process ever gets higher than it has been so far.

**"Theorem."** If $\{M_n\}$ is a martingale relative to $\{B_n\}$, and $T$ is a stopping time relative to $\{B_n\}$, then, **if conditions are nice,** $E(M_T) = E(M_0)$.

It's not much of a theorem until we're clear about what "nice" means, but bear with me for now and let's see what we get out of such a result. I hope you'll see that it's well worth investigating.

**For now, assume things are nice and the "theorem" holds.**

**Gambler's ruin probabilities.** Let $\{X_n\}$ be the simple symmetric random walk starting at level $a > 0$. Let $T$ be the first time at which the walk hits 0 (ruin) or $a + b$. Then the "theorem" says $E(X_T) = E(X_0) = a$. And by the definition of $T$,

$$E(X_T) \;=\; 0 \cdot P_a(\text{ruin}) + (a+b)(1 - P_a(\text{ruin}))$$

So

$$a \;=\; (a+b)(1 - P_a(\text{ruin}))\; P_a(\text{ruin}) = \frac{b}{a+b}$$

That's a lot simpler than solving the system of equations that you developed earlier in the course by first step analysis.

**Gambler's ruin process: Expected time till absorption.** Continue with the process above, and recall that $\{M_n = X_n^2 - n\}$ is a martingale with respect to $\{X_n\}$. The "theorem" says

$$E(M_T) = E(M_0) = a^2$$

And by the definition of $T$

$$E(M_T) \;=\; E(X_T^2) - E(T) \;=\; (a+b)^2 \cdot \frac{a}{a+b} - E(T)$$

So

$$a^2 \;=\; (a+b)a - E(T)\; E(T) = ab$$

That's a whole lot simpler than our previous method to derive this result, which was to use a first step analysis to set up a recursion and then solve the recursion.

**Expected waiting time till a pattern appears.** You have used Markov Chain methods to find these; that involves developing and solving a system of equations. Instead, let's see if we can set the process up as a fair game and use the "theorem." I'll do this informally in an example, without setting up a lot of notation. Let's toss a $p$-coin and let $T$ be the time when HHH first appears. Here's a betting strategy that makes it remarkably easy to find $E(T)$. At each toss, bet that the three upcoming tosses will be HHH. Tell the bank to give you \$1 if you win your bet, and nothing if you lose your bet. The bank won't be too happy: it expects to pay you $\$p^3$, but it's not getting anything in return. So, to make things fair, you should pay the bank $\$p^3$ every time you make this bet. Then the process of your net gains will be a martingale.

**Stop betting at time $T$.** According to our "theorem," the process is fair at time $T$, so the amount you expect to have paid the bank upto that time will equal what

the bank expects to pay you. You're paying $p^3$ for each of $T$ tosses, so your expected payout is:

• **What goes out:** $p^3E(T)$.

What does the bank expect to pay you? It has to pay you \$1 at time $T$. But that's not all. At time $T$ there are two unresolved bets: the ones you made at times $T-1$ and $T-2$. You might win those too, so the bank has an expected payout for those as well. You'll win the bet you made at time $T-2$ if the $(T+1)$st toss is H. You'll win the bet you made at time $T-1$ if the $(T+1)$st and $(T+2)$nd tosses are HH. So the total the bank expects to pay you is:

• **What comes in:** $1+p+p^2$.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **bank expects to pay:** | | | | | | | | | | | 1 | $p$ | $p^2$ |
| | | | | | | | | | | | ↓ | ↓ | ↓ |
| tosses | _ | | _ | … | | _ | H | | H | | H | ? | ? |
| | ↑ | | ↑ | … | ↑ | | ↑ | | ↑ | | | | |
| **you expect to pay:** | $p^3$ | | $p^3$ | … | $p^3$ | | $p^3$ | | $p^3$ | | | | |

So

$$p^3E(T) \;=\; 1+p+p^2 \;\; E(T) \;=\; \frac{1+p+p^2}{p^3}$$

Here's the diagram for $E(T_{HTH})$, just for fun.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **bank expects to pay:** | | | | | | | | | | | 1 | 0 | $pq$ |
| | | | | | | | | | | | ↓ | ↓ | ↓ |
| tosses | _ | | _ | … | | _ | H | | T | | H | ? | ? |
| | ↑ | | ↑ | … | ↑ | | ↑ | | ↑ | | | | |
| **you expect to pay:** | $p^2q$ | | $p^2q$ | … | $p^2q$ | | $p^2q$ | | $p^2q$ | | | | |

So

$$p^2qE(T_{HTH}) \;=\; 1+0+pq \;\; E(T_{HTH}) \;=\; \frac{1+pq}{p^2q}$$

Beautiful! Moreover, the answer simplifies to

$$E(T_{HTH}) \;=\; \frac{1}{p^2q}+\frac{1}{p}$$

The denominators are the chance of the pattern, and the chance of the overlap between the beginning and the end of the pattern: the first element H (chance $p$) is the same as the last element. You should check that

$$E(T_{HHH}) \;=\; \frac{1}{p^3}+\frac{1}{p^2}+\frac{1}{p}$$

Once again, the denominators are the chance of the pattern and the chances of the overlaps between the beginning and the end. The first two elements HH (chance $p^2$) are the same as the last two elements. The first element H is the same as the last element.

You can use the same simple method to find the expected waiting time till any pattern of any length from any alphabet. If letters are drawn independently at random

with replacement from the English alphabet of 26 letters, then the expected waiting time till ABRACADABRA appears is $26^{11} + 26^4 + 26$.

We now have plenty of motivation to examine the "theorem" and try to work out what's "nice." First, here's an example that's clearly not "nice" as far as the "theorem" is concerned.

**Example in which "theorem" doesn't hold.** Let $\{X_n\}$ be the simple symmetric random walk starting at 0, and let $T$ be the first hitting time of level 1. We know that $T$ is finite with probability 1. So $X_T = 1$ with probablity 1. But we started at $X_0 = 0$ with probability 1, so $E(X_T)$ is clearly not equal to $E(X_0)$.

What could not be "nice" about the simple symmetric random walk hitting level 1? Recall that while $T$ is indeed finite with probability 1 (the walk is recurrent), $E(T) = \infty$ (the walk is null recurrent). That's where the trouble starts.

The goal now is to get a sense of the properties that a martingale and a stopping time must have in order for the "theorem" to hold. To do this, we'll first examine *martingale differences*:

Let $\{M_n\}$ be a martingale with respect to $\{B_n\}$. For $n \geq 1$, let $D_n = M_n - M_{n-1}$ be the $n$th martingale difference. Then $\{D_n\}$ is the process of increments of $\{M_n\}$:

$$M_n \;=\; M_0 + \sum_{i=1}^{n} D_i$$

Since $E(M_n) = E(M_0)$ for all $n$, we have $E(D_n) = 0$ for all $n$, so $M_n$ is the sum of $M_0$ and random variables that have mean 0. Now recall one of our earliest examples of a martingale: a process with independent increments that have mean 0. We have just shown that every martingale has increments with mean 0; are the increments also independent? Not quite. But close.

**Lemma.** Let $\{M_n\}$ be a martingale with respect to $\{B_n\}$, and suppose $E(M_n^2) < \infty$ for all $n$. For every $n \geq 1$,

$$E(D_n \,|\, B_0, B_1, \ldots, B_{n-1}) = 0$$

and

$$E(D_n D_{n-1} \,|\, B_0, B_1, \ldots, B_{n-1}) = 0.$$

**Proof.** The first equality follows from the definition of $D_n$ and the martingale property. For the second, first note that $E(M_n^2) < \infty$ allows us to talk about variances and covariances without getting into trouble. Because $D_{n-1}$ is a function of $(B_0, B_1, \ldots, B_{n-1})$,

$$E(D_n D_{n-1} \,|\, B_0, B_1, \ldots, B_{n-1}) \;=\; D_{n-1} E(D_n \,|\, B_0, B_1, \ldots, B_{n-1}) \;=\; D_{n-1} \cdot 0 \;=\; 0$$

You should show that $E(D_n D_m \,|\, B_0, B_1, \ldots, B_m) = 0$ for all $m < n$. Then, by iterated conditional expectations, you will have proved:

**Theorem.** Let $\{M_n\}$ be a martingale with $E(M_n^2) < \infty$ for all $n$, and let $\{D_n\}$ be the process of differences. Then the differences are uncorrelated and have mean 0.
What does this have to do with working out when $E(M_T)$ is equal to $E(M_0)$? Notice that

$$M_T \;=\; M_0 + \sum_{i=1}^{T} D_i \text{ and so } E(M_T) \;=\; E(M_0) + E\Big(\sum_{i=1}^{T} D_i\Big)$$

So

$$E(M_T) \;=\; E(M_0) \text{ if and only if } E\big(\sum_{i=1}^{T} D_i\big) = 0$$

So to work out when our "theorem" $E(M_T) = E(M_0)$ holds, we have to examine whether or not

$$E\Big(\sum_{i=1}^{T} D_i\Big) \;=\; 0$$

We're working with the expectation of a random sum. To use what we know about expectations and limits, it helps to write the random sum as

$$\sum_{i=1}^{T} D_i \;=\; \sum_{n=1}^{\infty} D_n I(T \geq n)$$

That means

$$E(M_T) \;=\; E(M_0) \text{ if and only if } E\Big(\sum_{n=1}^{\infty} D_n I(T \geq n)\Big) \;=\; 0$$

$T$ is a stopping time relative to $\{B_n\}$. So for every $n$, $I(T \geq n) = I((T \leq n-1)^c)$ is a function of $(B_0, B_1, \ldots, B_{n-1})$. So for every $n$,

$$E(D_n I(T \geq n) \,|\, B_0, B_1, \ldots, B_{n-1}) \;=\; I(T \geq n) E(D_n \,|\, B_0, B_1, \ldots, B_{n-1}) \;=\; 0$$

by our Lemma. So $E(D_n I(T \geq n)) = 0$ for every $n$, which means

$$\sum_{n=1}^{\infty} E(D_n I(T \geq n)) \;=\; 0$$

So

$$E(M_T) \;=\; E(M_0) \text{ if and only if } E\Big(\sum_{n=1}^{\infty} D_n I(T \geq n)\Big) \;=\; \sum_{n=1}^{\infty} E(D_n I(T \geq n))$$

The truth of our "theorem" $E(M_T) = E(M_0)$ has come down to whether or not we can switch expectation and and an infinite sum.

That's not an easy condition to check directly, so let's look for simpler conditions that make it true.

**Theorem 1.** (No quotation marks!) If $T$ is a bounded stopping time, then $E(M_T) = E(M_0)$.
**Proof.** If $T$ is bounded, then the infinite sums in our argument are in fact finite sums, and there is no problem interchanging expectation and a finite sum.

But very few natural stopping times are bounded. None of the stopping times we've encountered so far have been bounded. We have to see what is needed to deal with unbounded stopping times.

**Forcing a stopping time to be bounded.** Let $T$ be a stopping time, and for fixed $n$ consider the time $T \wedge n = \min\{T, n\}$. Informally, you're deciding to watch the process up to time $T$ or time $n$, whichever comes earlier. You should check that $T \wedge n$ **is a stopping time**. And of course it's bounded. Put this together with Theorem 1 to get

**Theorem 2.** (also no quotation marks) If $T$ is a stopping time, then for each fixed $n$, $E(M_{T\wedge n}) = E(M_0)$.

Suppose $P(T < \infty) = 1$. Then with probability 1,

- $\lim_n (T \wedge n) = T$. In fact, $T \wedge n = T$ for all sufficiently large $n$.
- $\lim_n M_{T\wedge n} = M_T$.

So

- $E(\lim_n M_{T\wedge n}) = E(M_T)$, and
- $\lim_n E(M_{T\wedge n}) = E(M_0)$ by Theorem 2

And so

**Theorem 3: Optional stopping.** Let $\{M_n\}$ be a martingale and $T$ a stopping time such that $P(T < \infty) = 1$. If $E(\lim_n M_{T\wedge n}) = \lim_n E(M_{T\wedge n})$, then $E(M_T) = E(M_0)$.

And once again there's a question of whether it's OK to interchange expectation and limit. As you can see, we're hitting the boundaries of what we can prove without more real analysis and measure theory. Still, here are some conditions that allow the switch in Theorem 3. Some have been proved already, some are easily believable, and some you have to take on faith — until you take a serious analysis course.

- $T$ is bounded
- $\{M_n\}$ is bounded, or is bounded upto time $T$ (e.g. the gambler's ruin probability calculation)
- $E(T) < \infty$ and the differences $D_n$ are bounded (e.g. waiting till a pattern)

While we're thinking about that last condition, here's a famous and closely related result about random walks; the increments don't have to have mean 0.

**Wald's Identity.** Let $X_1, X_2, \ldots$ be i.i.d. with $E(|X_1|) < \infty$, and let $E(X_1) = \mu$. Let $T$ be a stopping time relative to $\{X_n\}$, and suppose $E(T) < \infty$. For $n \geq 1$, let $S_n = X_1 + X_2 + \ldots + X_n$. Then $E(S_T) = E(T)\mu$.

**Proof.** By now the moves are very familiar.

$$\begin{aligned}
E(S_T) &= E\left(\sum_{i=1}^{T} X_i\right) \\
&= E\left(\sum_{n=1}^{\infty} X_n I(T \geq n)\right) \\
&= \sum_{n=1}^{\infty} E(X_n I(T \geq n)) \text{ we hope!} \\
&= \sum_{n=1}^{\infty} E(X_n) P(T \geq n) \text{ by conditioning on } (X_1, X_2, \ldots, X_{n-1}) \\
&= \mu \sum_{n=1}^{\infty} P(T \geq n) \;=\; \mu E(T)
\end{aligned}$$

# LECTURE 24

# Martingales and Introduction to Brownian Motion

### Class Announcement

- We have 3 lecture left and then RRR week, and review sessions in RRR week will be at the normal class location and times.
- Regarding material, we will begin a brief introduction to Brownian motion and related continuous parameters and continuous space processes.

## 24.1 Martingales Continued

From Durett, let $S_n := X_1 + \cdots + X_n$ as the sum of i.i.d. copies of $X$. Assume $\mathbb{E}X = 0$ and $\mathbb{E}X^2 = 1$ (with $\sigma^2 = 1$ by scaling). Thanks to Wald, we know that for a stopping time $T$ of a sequence $X_1, X_2, \ldots$, we have that

$$\mathbb{E}S_T = (\mathbb{E}T)(\mathbb{E}X) \qquad \text{provided } \mathbb{E}T < \infty$$

(Wald's first identity). According to Wald's second identity (Durrett Problem 5.7 on page 221, part of last week's homework)

$$\mathbb{E}S_T^2 = (\mathbb{E}T)(\mathbb{E}X^2) \qquad \text{again if } \mathbb{E}T < \infty. \tag{24.1}$$

For bounded $T$ this is straightforward. We know that $M_n : S_n^2 - n$ is a martingale implies Wald's second identity, provided that $T$ is *bounded*, as we have found from last class. Taking $0 = \mathbb{E}M_0 = \mathbb{E}M_{T\wedge n}, \forall n = \mathbb{E}M_T$, if $\mathbb{P}(T \leq b) = 1$. Then take $n \geq b$ which gives that $T \wedge n = T$. This implies $M_{T\wedge n} = M_T$, and so $\mathbb{E}M_T = 0$. Then $\mathbb{E}(S_T^2 - T) = 0$ which implies $\mathbb{E}S_T^2 = \mathbb{E}T$.

The issue arises as to how we can push this to unbounded $T$. This is much harder than it looks. The issue is that we have a sequence of random variables $M_{n\wedge T}$ with $\mathbb{E}M_{n\wedge T} \equiv 0$ and $\mathbb{P}(M_{n\wedge T} \to M_T) = 1$. Then $M_{n\wedge T} = M_T$ for all large $n$ on $(T < \infty)$. In general, we know that

$$\mathbb{P}(Y_n \to Y) = 1$$

and $\mathbb{E}Y_n$ has a limit does *not* imply $\mathbb{E}Y = \lim Y_n$.

Our key example of this was the "double-or-nothing" example with

$$\mathbb{E}Y_n \equiv 1, \mathbb{P}(Y_n \geq 0) = 1$$

so that

$$\mathbb{P}(Y_n = 0 \text{ for all large } n) = 1 \implies \mathbb{P}(Y_n \to 0) = 1$$

but $\mathbb{E}Y_n \equiv 1 \not\to 0$. For rigorous switching of limits and integrals like this (swapping the operations lim and $\mathbb{E}$), we would need Math 202 or Math 105 for analysis and Math 218 / Stats 205 for the probability portion to properly address these limits. From any of the texts for these courses, there are the following results.

Let us assume the Monotone Convergence Theorem, that if $0 \leq X_n \uparrow X$, then $0 \leq \mathbb{E}X_n \uparrow \mathbb{E}X$, allowing $+\infty$ as a value for $\mathbb{E}X$. In fact, this is the definition of $\mathbb{E}X$ for $X \geq 0$ in an advanced course. This implies

$$\mathbb{E}\sum_n Y_n = \sum_n \mathbb{E}Y_n,$$

for $Y_n \geq 0$ where provided we have nonnegative variables, we can perform these swaps. We have done this many times before, in the study of Markov chains, and in the proof of Wald's first identity. It's important to note that in general, as shown by the double or nothing game:

$$X_n \geq 0, \mathbb{P}(X_n \to X) = 1 \text{ and } \mathbb{E}X_n \to L \text{ does and imply } \mathbb{E}X = L.$$

However, there is one general fact:

**Fatou's Lemma**

$$X_n \geq 0, \mathbb{P}(X_n \to X) = 1 \text{ and } \mathbb{E}X_n \to L \implies 0 \; leq \mathbb{E}X \leq L$$

In particular, if the limit $L$ of $\mathbb{E}X_n$ is finite, then so is $\mathbb{E}X$, and $\mathbb{E}X \leq L$. For positive random variables converging to a limit, you can only lose expectation in the limit (like in the double or nothing game) you cannot gain expectation.

**Dominated Convergence Theorem**

If $\mathbb{P}(|X_n| \leq Y) = 1$ for all $n$ and $\mathbb{E}Y < \infty$ and $\mathbb{P}(X_n \to X) = 1$, then

$$\mathbb{E}X_n \to \mathbb{E}X, \text{ and } |\mathbb{E}X| \leq \mathbb{E}Y,$$

where $Y$ is called the *dominating variable*.

Finally, with the dominated convergence theorem, we recall the notion of $X_n \xrightarrow{L^2} X$ means $\mathbb{E}(X_n - X)^2 \to 0$ and $\mathbb{E}X_n^2 < \infty$ for all $n$. Take $L^2 = \{\text{all } X : \mathbb{E}X^2 < \infty\}$ which is almost like Euclidean $\mathbb{R}^n$ just infinite-dimensional if there are an infinite number of possible outcomes.

The key fact here is that $L^2$ is complete. That is, if

$$\lim_{m,n\to\infty} \mathbb{E}(X_m - X_n)^2 = 0,$$

then

$$\exists X \in L^2 \text{ and } X_n \xrightarrow{L^2} X.$$

If we have a Cauchy sequence in $\mathbb{R}^n$, then a Cauchy sequence converges. This fact works for a sequence of random variables in the $L^2$ space. The fact that $L^2$ is complete is shown using the DCT. Accepting this, then the problem is easy. $L^2$ has an inner product $(X, Y) := \mathbb{E}(XY)$ and note that $|(X, Y)| \leq \sqrt{(X, X)}\sqrt{(Y, Y)}$ via Cauchy-Schwarz, so

$$|(X, Y)| = |\mathbb{E}(XY)| \leq \sqrt{\mathbb{E}X^2}\sqrt{\mathbb{E}Y^2}.$$

Then if $M_n$ is a martingale, then $M_n$ has orthogonal increments, meaning:

$$\mathbb{E}M_m \underbrace{(M_n - M_m)}_{} = 0, \forall m < n,$$

which follows immediately from the martingale property by conditioning. Now it follows that if $(M_n)$ is a MG that is bounded in $L^2$, then $(M_n)$ is Cauchy in $L^2$, hence convergent in $L^2$. Apply this to $M_n = S_{n\wedge T}$ in the setting of Wald's identity, and you learn easily that if $\mathbb{E}T < \infty$ then $S_{n\wedge T}$ is convergent in $L^2$ to $S_T$, hence that $\mathbb{E}S_T^2 = \mathbb{E}T$ (assuming $\mathbb{E}X = 0$ and $\mathbb{E}X^2 = 1$. Unfortunately that is a lot of trouble just to switch a limit and an expectation. If you can find a simpler proof, let me know!

## 24.2 Introduction to Brownian Motion

Let's continue with $S_n = X_1 + X_2 + \cdots + X_n$ as a random walk with mean 0 and variance 1 increments (simply for convenience). Then $\mathbb{E}X = 0$ and $\mathbb{E}X^2 = 1$. We can go a long way with this model. Consider the simple random walk (SRW) with $-1$ with probability $\frac{1}{2}$ and $+1$ with probability $\frac{1}{2}$.

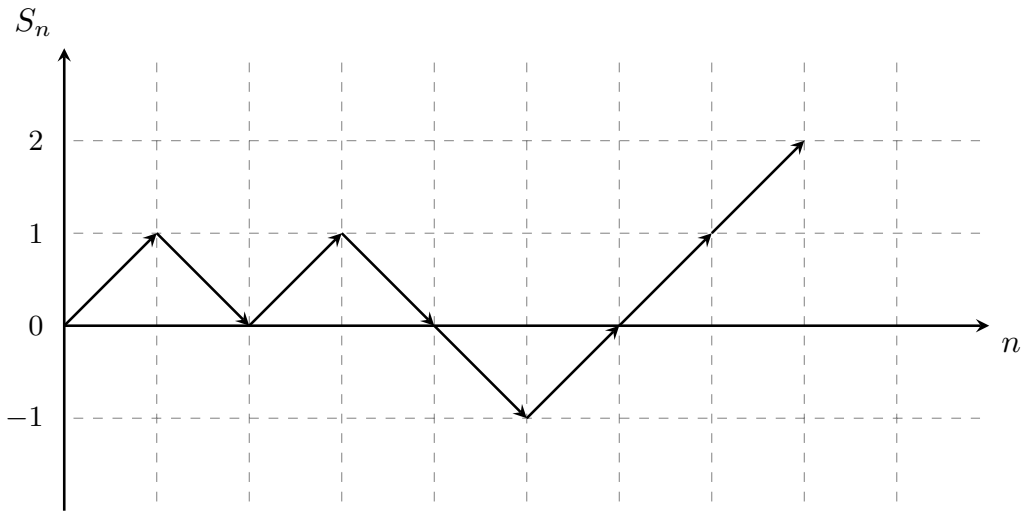



Figure 24.1: Simple random walk. $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = \frac{1}{2}$.

We have the two "martingale" facts $\mathbb{E}S_n = 0$ and $\mathbb{E}S_n^2 = n$. We're now interested in

$$\mathbb{E}S_n^3$$

Pitman jokes that if we do not notice symmetry, we will be punished on the final. Symmetry means that there is some equality in distribution. Our homework is designed to show how clever we can be with exploiting symmetry. This is much easier. Notice that we have

$$X \stackrel{d}{=} -X$$

where we can swap roles of $-1$ and $+1$ and get the same simple random walk. This implies that

$$(X_1, X_2, \ldots, X_n) \stackrel{d}{=} (-X_1, -X_2, \ldots, -X_n)$$

Then applying any function $\psi$ to the coordinates gives

$$\psi(X_1, X_2, \ldots, X_n) \stackrel{d}{=} \psi(-X_1, -X_2, \ldots, -X_n)$$

Take $\psi(X_1, \ldots, X_n) := X_1 + \cdots + X_n$, and we see that

$$S_n \stackrel{d}{=} -S_n$$

Then we can get

$$(S_n)^3 \stackrel{d}{=} (-S_n)^3 = -S_n^3$$

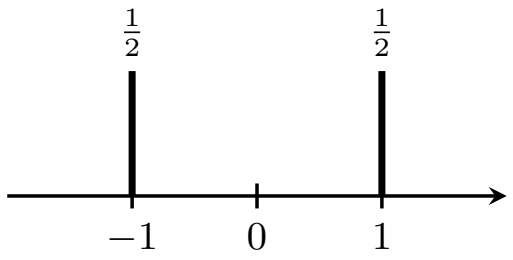



Figure 24.2: Depiction of symmetry. $X \stackrel{d}{=} -X$.

We start with an $n$-dimensional vector, we apply a function (sum) and take the expectation. Of course, we have the obvious bound $|S_n| \leq n$, so taking the expectation is legal, and we have

$$\mathbb{E}S_n^3 = \mathbb{E}(-S_n^3) = -\mathbb{E}S_n^3 \implies \mathbb{E}S_n^3 = 0.$$

We can continue inductively that by symmetry, we have that all odd moments of $S_n$ are 0.
We have the parity issue that for $n \in \mathbb{Z}$,

$$\mathbb{P}(S_{2n} \text{ is even}) = 1, \text{ and } \mathbb{P}(S_{2n-1} \text{ is odd}) = 1$$

Now we should ask, what is $\mathbb{E}S_n^4$? We can evaluate and expand

$$\begin{aligned}\mathbb{E}S_n^4 &= \mathbb{E}(X_1 + X_2 + \cdots + X_n)^4 \\ &= n(\mathbb{E}X^4) + \binom{4}{2}\binom{n}{2}\left(\mathbb{E}X^2\right)^2 + \underbrace{\binom{4}{3}(\mathbb{E}X)(\mathbb{E}X^3) + \cdots}_{=0}.\end{aligned}$$

Then in our example of the SRW,

$$\mathbb{E}S_n^4 = n + 3n(n-1).$$

We've all been taught the Central Limit Theorem, which says that

$$\frac{S_n}{\sqrt{n}} \stackrel{d}{\to} B_1 \sim N(0,1)$$

where statisticians often use $Z$ for $N(0,1)$ and $B_1 \stackrel{d}{=} Z$, where we take the variable $B$ for Brownian motion and $B_1$ to be Brownian motion at time 1.

### 24.2.1 Proof Sketch of Central Limit Theorem

We'll talk about the heart of the argument to show that

$$\frac{S_n}{\sqrt{n}} \stackrel{d}{\to} B_1$$

*Proof.* (Sketch only! ) For our simple random walk, we have

$$\mathbb{E}\left(\frac{S_n}{\sqrt{n}}\right)^4 = \mathbb{E}\frac{S_n^4}{n^2} = \frac{n+3n(n-1)}{n^2} \to 3,$$

as $n \to \infty$. □

Notice that $Y_n \xrightarrow{d} Y$ means

$$\mathbb{P}(Y_n \leq y) \to \mathbb{P}(Y \leq y)$$

for all continuity points of the RHS, as we have defined in a previous lecture. We have this if and only if

$$\mathbb{E}g(Y_n) \to \mathbb{E}g(Y)$$

for all suitably 'nice' $g$ (e.g. bounded and continuous or bounded and differentiable). Here, we see that

$$\mathbb{E}g\left(\frac{S_n}{\sqrt{n}}\right) \to 3, \text{ for } g(x) = x^4.$$

Then from our findings earlier of the moments, we have

$$\mathbb{E}\left(\frac{S_n}{\sqrt{n}}\right) = 0$$

$$\mathbb{E}\left(\frac{S_n}{\sqrt{n}}\right)^2 = 1$$

$$\mathbb{E}\left(\frac{S_n}{\sqrt{n}}\right)^3 = 0$$

$$\mathbb{E}\left(\frac{S_n}{\sqrt{n}}\right)^4 = \cdots \to 3$$

and so on. Notice that $B_1$ has the limit distribution with

$$\mathbb{E}\left(\frac{S_n}{\sqrt{n}}\right)^{2m} \to (2m-1)(2m-3)\cdots 5\cdot 3\cdot 2\cdot 1 = (2m-1)!!,$$

which is the double factorial (product of the first $m$ odd numbers).

### 24.2.2 Conclusion

We learn by computing moments to expect there is a limit distribution of $B_1$ with these moments, given by

$$\mathbb{E}(B_1^n) = \begin{cases} 0, & n \text{ is odd} \\ (2m-1)!!), & m = 2n \text{ is even} \end{cases}$$

It remains to show that there is in fact a unique limit distribution of $B_1$ with these moments. This is not easy, nor is it easy to show that convergence in distribution is implied by this convergence of moments. But that is the case, as shown in more advanced courses.

### 24.2.3 A Better Technique

Thanks to Euler and Laplace, we have a better technique and we can handle all moments at once using moment generating functions (MGFs).
Notice that

$$\begin{aligned}\mathbb{E}e^{\theta X} &= \frac{e^{+\theta}+e^{-\theta}}{2}\\ \implies \mathbb{E}e^{\theta S_n} &= \left(\frac{e^{\theta}+e^{-\theta}}{2}\right)^n\\ &= 1+\frac{\theta^2}{2}+\cdots\end{aligned}$$

which then implies, by replacing $\theta$ with $\theta/\sqrt{n}$

$$\mathbb{E}e^{\theta\frac{S_n}{\sqrt{n}}} = \left(\frac{e^{\theta/\sqrt{n}}+e^{-\theta/\sqrt{n}}}{2}\right)^n = \left(1+\frac{\theta^2}{2n}+\frac{\theta^4}{8n^2}+\cdots\right)^n$$

which converges to **Exponential** $\left(\theta^2/2\right)$. You see this by recalling that if $x_n \to x$ then $\left(1+\frac{x_n}{x}\right)^n \to e^x$. Now check the moments found earlier by expanding

$$\begin{aligned}e^{\theta^2/2} &= 1+\frac{\theta^2}{2}+\frac{\left(\frac{\theta^2}{2}\right)^2}{2!}+\frac{\left(\frac{\theta^2}{2}\right)^3}{3!}+\cdots\\ &= 1+\underbrace{(\mathbb{E}B_1)}_{=0}\theta+\underbrace{\left(\mathbb{E}B_1^2\right)}_{=1}\frac{\theta^2}{2}+\underbrace{\mathbb{E}(B_1^3)}_{=0}\frac{\theta^3}{3!}+\frac{\mathbb{E}(B_1^4)\theta^4}{4!}+\cdots\end{aligned}$$

Take $n = 10,000$ steps and consider the plot of $t \to \frac{S_{nt}}{\sqrt{n}}$.
At time 1, we have

$$\frac{S_{10,000}}{100} \overset{d}{\approx} B_1.$$

Then at time 2, we have the sum of two independent copies of something that is $\mathcal{N}(0,1)$, which is $\mathcal{N}(0,2)$ via the basic additive property of normal distribution. Look at the process:

$$\left(\frac{S_{nt}}{\sqrt{n}}, t \geq 0\right) \overset{d}{\to} (B_t, t \geq 0),$$

where $\overset{d}{\to}$ means convergence in distribution of finite discrete distributions, and where we take $S_{nt}$ for $nt$ not integer to be defined by linear interpolation (usual sawtooth path). In particular,

$$\left(\frac{S_{nt_i}}{\sqrt{n}}, 1 \leq i \leq n\right) \overset{d}{\to} (B_{t_i}, 1 \leq i \leq n).$$

Then the limit process $(B_t, t \geq 0)$ must have certain properties.

Notice that $\mathbb{E}B_t^2 = t \cdot \mathbb{E}B_1^2$ for positive integers $t$, so that

$$B_t \stackrel{d}{=} t^{1/2} B_1$$

which we call *Brownian Scaling*. This gives us a 1-dimensional distribution of $B_t$. To get finite discrete distributions (FDD's), we can take $0 < t_1 < t_2 < \cdots < t_n$ and look at

$$B_t, B_{t_2} - B_{t_1}, \ldots, B_{t_n} - B_{t_{n-1}},$$

which are independent copies of $B_t, B_{t_2 - t_1}, B_{t_3 - t_2}, \ldots$. We scale it so that the variance increment is the interval length in time (we've hidden the $\sigma$ at the start).

### 24.2.4 Conclusion

Then for all bounded continuous $g$ and some unbounded $g$ (i.e. product of powers), we have convergence in distribution ($\stackrel{d}{\to}$) defined as

$$\mathbb{E}g\left(\frac{S_{nt_1}}{\sqrt{n}}, \frac{S_{nt_2}}{\sqrt{n}}, \ldots, \frac{S_{nt_m}}{\sqrt{n}}\right) \to \mathbb{E}g(B_{t_1}, \ldots, B_{t_m})$$

So in terms of Brownian motion, we have one final fact. It is possible to construct the limit process $B$ (Brownian motion) with *continuous paths*. This is a theorem due to Weiner in the 1920s, and Pitman warns this is not easy.
Recall that our Poisson process holds and jumps; on the other hand, Brownian motion 'wiggles' and are incredibly variable (with probability 1, the paths are *everywhere* continuous but *nowhere* differentiable). In the next lectures, we'll see how various properties of random walks can be understood as Brownian limits.

# Bibliography


[1] Richard Durrett. *Essentials of Stochastic Processes*. Springer, Reading, Massachusetts, 2012.

[2] Jim Pitman. *Probability*. Springer, Berkeley, California, 1992.

[3] Geoffrey R. Grimmett and David R. Stirzaker. *Probability and Random Processes*. Oxford University Press, New York, third edition, 2001.

[4] Sø ren Asmussen. *Applied Probability and Queues*, volume 51 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 2003. Stochastic Modelling and Applied Probability.

[5] J. R. Norris. *Markov Chains*, volume 2 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998. Reprint of 1997 original.

[6] William Feller. *An Introduction to Probability Theory and its Applications. Vol. I.* John Wiley and Sons, Inc., New York; Chapman and Hall, Ltd., London, 1957. 2nd ed.

[7] Alvaro Corral and Francesc Font-Clos. Criticality and self-organization in branching processes: application to natural hazards. 07 2012.