Stochastic Processes

STAT-150 Lecture Series, Jim Pitman

(version 1.2, updated 09/25/2019)

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 (R.V.S) that is indexed by a *parameter set* $\mathcal{I}$. We shall use the following notation

$$(X_i, i \in \mathcal{I}) = (X_i)_{i\in\mathcal{I}} \tag{1.1}$$

The parameter set commonly represents a set of times, but can extend to e.g. space or space-time. Underlying (1.1) there is always a *probability measure* $\mathbb{P}$ on some outcome space Ω with $\mathbb{P}$ a function of subsets F of Ω, ranging over a suitable collection of subsets $\mathcal{F}$ called *events*. Use the command `\mathbb{P}` or `\mathds{P}` to produce $\mathbb{P}$ or $\mathbb{P}$ in LaTeX respectively.

In the *canonical* setup, Ω 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 σ-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`.

Here's a conversion table on notation used in the course text Durrett. *Essentials of Stochastic Processes* and that of these lecture notes.

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

$$(X_0, X_1, \dots, X_{n-1}, X_n) = (X_i)_{i \in \mathbb{N}_0} \tag{1.2}$$

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 $(X_i)_{i \in \mathbb{N}_0}$ 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

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

known as the *Markov Property.* In words,

Past and future are conditionally independent given the present.

So 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 $X_0, X_1, X_2 \in \mathbb{N}_0$

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

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

$$\begin{aligned}&\mathbb{P}(X_0 = x, X_1 = y, X_2 = z) = \\ &\qquad\mathbb{P}(X_0 = x)\mathbb{P}(X_1 = y \mid X_0 = x)\mathbb{P}(X_2 = z \mid X_1 = y, X_0 = x)\end{aligned}$$

For a Markov Chain, this reduces to

$$\begin{aligned}&\mathbb{P}(X_0 = x, X_1 = y, X_2 = z) = \\ &\qquad\mathbb{P}(X_0 = x)\mathbb{P}(X_1 = y \mid X_0 = x)\mathbb{P}(X_2 = z \mid X_1 = y)\end{aligned}$$

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

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

Then

$$\mathbb{P}(X_0 = x, X_1 = y, X_2 = z) = \mathbb{P}(X_0 = x)P_1(x, y)P_2(y, z) \tag{1.3}$$

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

For finite $\mathcal{S}$, we can build a *transition matrix*

$$P = P(x, y) \tag{1.4}$$

a "set of rules" or "mechanism" for moving between different states in $\mathcal{S}$. Rules of probability imply that (1.4) 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

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

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

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

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 λ for X_0. That is

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

Again, rules of probability require λ 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 a finite state space $\mathcal{S}$, assigned initial distribution λ, and transition probability matrix P. Then for sequences of length $n + 1$, the joint distribution

$$(X_0 = x_0, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = x_n)$$

has the following prescription

$$\mathbb{P}\left(\bigcap_{i=0}^{n}\{X_i = x_i\}\right) = \lambda(x_0)\prod_{i=0}^{n-1} P(x_i, x_{i+1}) \tag{1.5}$$

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

$$\sum_{x_0}\sum_{x_1}\cdots\sum_{x_{n-1}}\sum_{x_n}\lambda(x_0)\prod_{i=0}^{n-1}P(x_i,x_{i+1})=1 \tag{1.6}$$

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.

1.4.1 Notation

$\mathbb{P}_\lambda$ is used to show that $\mathbb{P}$ makes X_0 have distribution λ.

$\mathbb{P}_x$ 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 $(X_i)_{i\in\mathbb{N}}$, 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 λ, that is $X_0 \sim \lambda$, let

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

Now, if $X_0 = x_0$, define

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

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

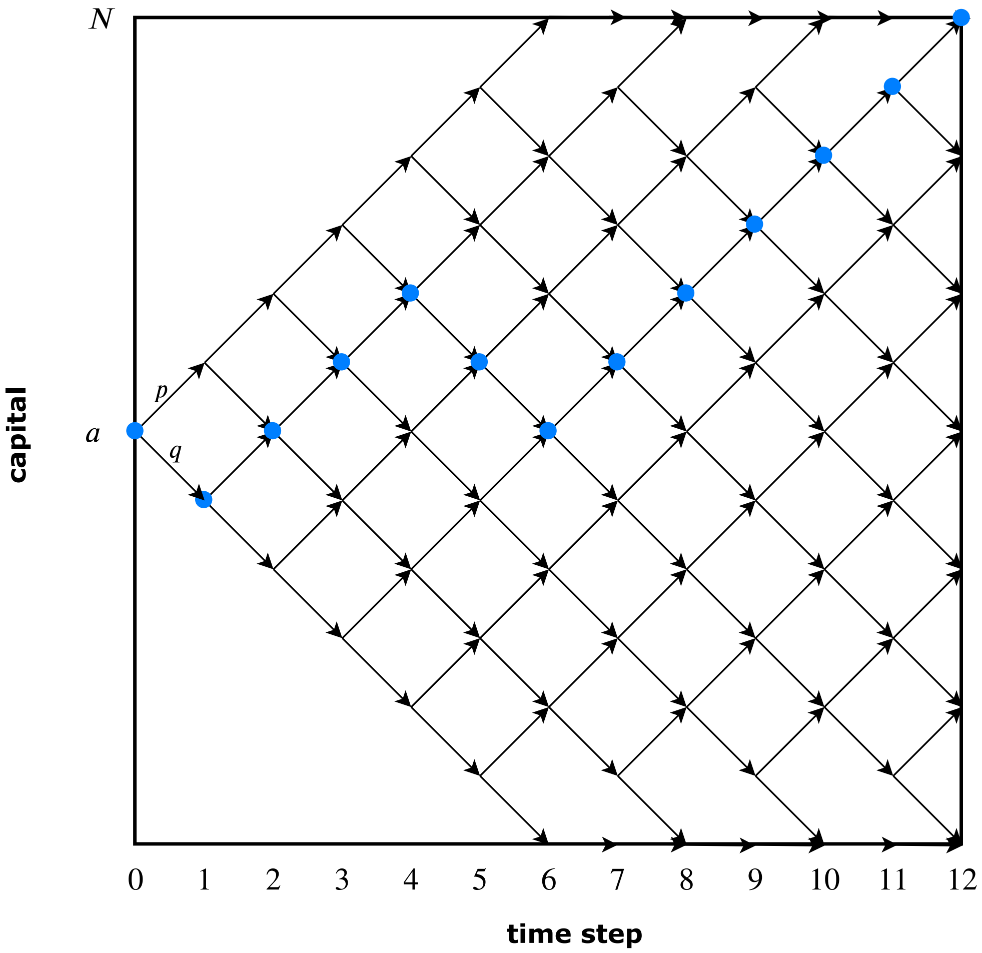
A gambler has a fortune of $\$a$, where $0 \leq a \leq N$. At each play the gambler wins a \$1 with probability p and loses \$1 with probability $q = 1 - p$. The gambler plays until $X_n = N$ (quitting with a gain) or until $X_n = 0$ (ruined). Here

$$X_n := \text{ the gambler's capital after } n \text{ plays}$$

The transition probabilities for the edge cases are

$$P(0,0) = 1 \quad \text{and} \quad P(N,N) = 1$$

So states 0 and N are *absorbing*. Reference. [1] Durrett's, *Essentials of Stochastic Processes* Section 1.1. Here's a depiction of the gambler's chain.



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 xth row and yth 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 λ. 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:

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

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 (nth 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 λ), 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 λP 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:

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

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 λ. But easy to interpret λP for a probability measure λ.

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,

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

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

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

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, zth 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 ≥ 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

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

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 nth 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 $X_1, X_2, \ldots$ are i.i.d. with the distribution $(p_0, p_1, p_2, \ldots)$ (perhaps all strictly positive) instead of $(q, p, 0, 0, 0, \ldots)$. 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:

$$\left[\begin{array}{cccccccc}
* & 0 & 1 & 2 & 3 & 4 & 5 & \cdots \\
0 & p_0 & p_1 & p_2 & \cdots & \cdots & \cdots & \cdots \\
1 & 0 & p_0 & p_1 & p_2 & \cdots & \cdots & \cdots \\
2 & 0 & 0 & p_0 & p_1 & p_2 & \cdots & \cdots \\
3 & 0 & 0 & 0 & p_0 & p_1 & p_2 & \cdots \\
4 & 0 & 0 & 0 & 0 & p_0 & p_1 & \cdots \\
5 & 0 & 0 & 0 & 0 & 0 & p_0 & \cdots \\
\vdots & & & & & & &
\end{array}\right]$$

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 $4 \leq k \leq 24$ 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 ∞ 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.

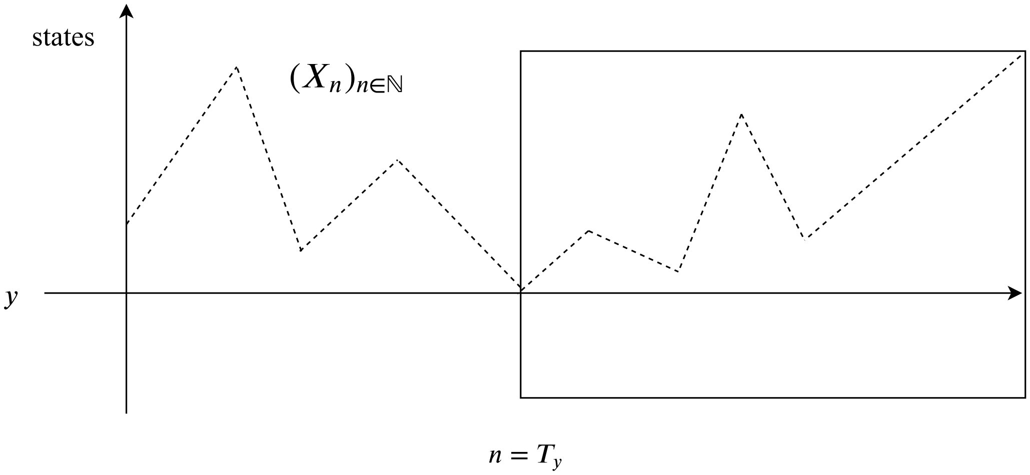
[1] sometimes $i, j \in \mathcal{S}$

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

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

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 $k \geq 1$

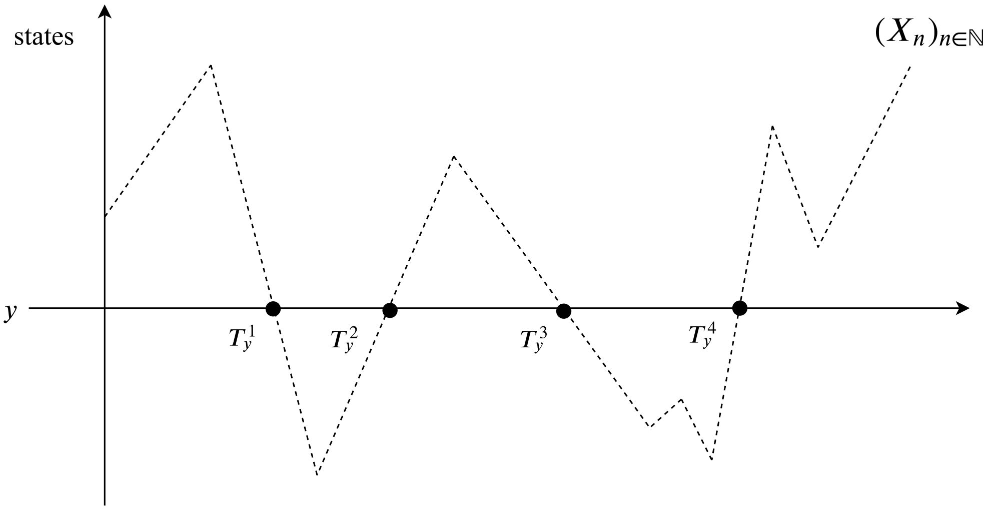
$$T_y^k := \min\{n > T_y^{k-1} : X_n = y\}.$$

Note T_y^k is a k^{th} iterate of the scheme for defining $T_y = T_y^1$, not the k^{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 $X_0 = y$:

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

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 ∞ as a part of $(N_y \geq 1)$. Hence

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

Recall $N_y := \sum_{n=1}^{\infty} \mathbb{1}(X_n = y)$ 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 ρ_y:

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

called the *first return probability* of state y. Now how to get from ρ_y to ρ_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 $N_y := \sum_{n=1}^{\infty} \mathbb{1}(X_n = y)$ 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 ρ_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 nth 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 ≤ 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

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

□

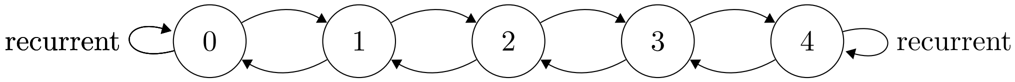
Pitman mentions Kai Lai Chung of Stanford who presents this Lemma in terms of a pedestrian repeatedly crossing the road. Depending on visibility and weather conditions, (current state), there is a varying chance of making it to the other side in k steps. But suppose that no matter how favorable the visibility and weather conditions, there is always at least a small chance α 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 $(X_1, \dots, X_N)$ 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 $((\bar{S}_n, S_n), 0 \leq n \leq N)$ is a Markov chain with transition matrix

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

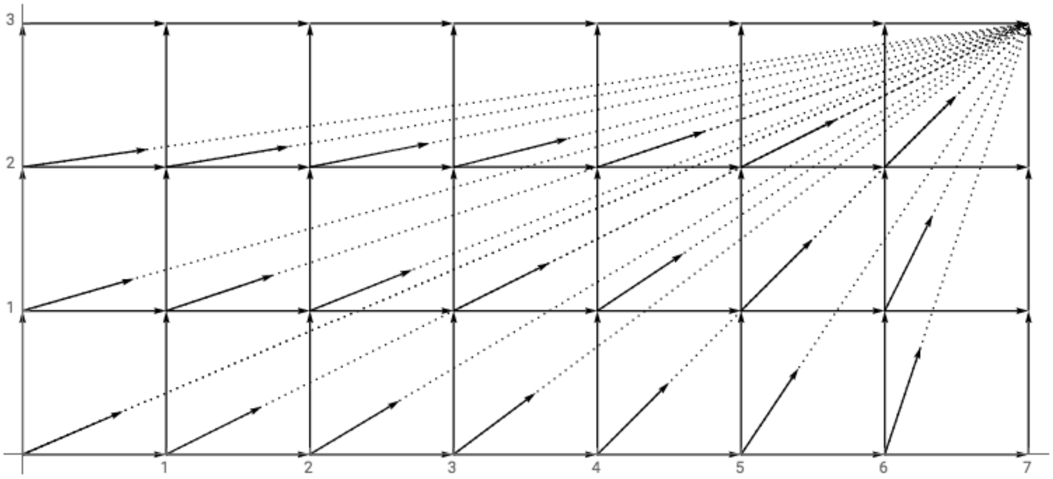
$$P((f,s),(f,s+1)) = \frac{A-s}{A+B-f-s} \tag{4.2}$$

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 $(X_1, \dots, X_N)$.
- 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 $(X_1, \dots, X_N)$.

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 $0 \leq f \leq 7$ the number of failures (0) and $0 \leq s \leq 3$ 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 $(X_1, \dots, X_N)$ that these random variables are *exchangeable*, meaning that for every permutation σ of $[N] := \{1, \dots, N\}$

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

where $\stackrel{d}{=}$ denotes equality in distribution. [2] See Pitman *Probability* Section 3.6. In particular, this holds for the sample $(X_1, \dots, X_N)$ of A ones and B zeros considered here. Except in degenerate cases, the sequence $(X_1, \dots, X_N)$ is not Markov: given $X_1, \dots, X_n$ 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 $X_1, \dots, X_n$ 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

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

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 σ 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 ω 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

$$\mathbb{P}(V_1 V_2) = \mathbb{E}(X_1 X_2) = (2p-1)_+$$

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 $\mathbb{P}(V_1 V_2)$ 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 π, 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 π 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,

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

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

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

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

when π 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 π 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 π 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

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

  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 π solves the detailed balance equations for P , then π also solves the balance equations for P;

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

- If X_1 and X_2 have three or more possible values, it is possible to have $X_1 \overset{d}{=} X_2$ without $(X_1, X_2) \overset{d}{=} (X_2, X_1)$. 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 π on S is P-invariant iff P is doubly stochastic. (Text Theorem 1.14)
- If π is P-invariant, then π is P^n-invariant for every positive integer n. Here P^n is the nth 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 π for P is reversible iff for every $N \geq 1$ there is the equality in distribution
$$(X_0, \ldots, X_N) \overset{d}{=} (X_N, \ldots, X_0).$$

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 π 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 $X_n \xrightarrow{d} \pi$ 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 $(X_1, \dots, X_N)$ 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 2×2 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 2×2 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 π 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 π 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 2×2 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 λP 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 λ of X_1, the line through λ in direction λP 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 λ and (B, A) and the downsloping line $\{(f, s) : f + s = 2\}$.

(iv) For $(B, A) \neq (0, 0)$, the unique stationary distribution π 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 λ 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 λP 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). □

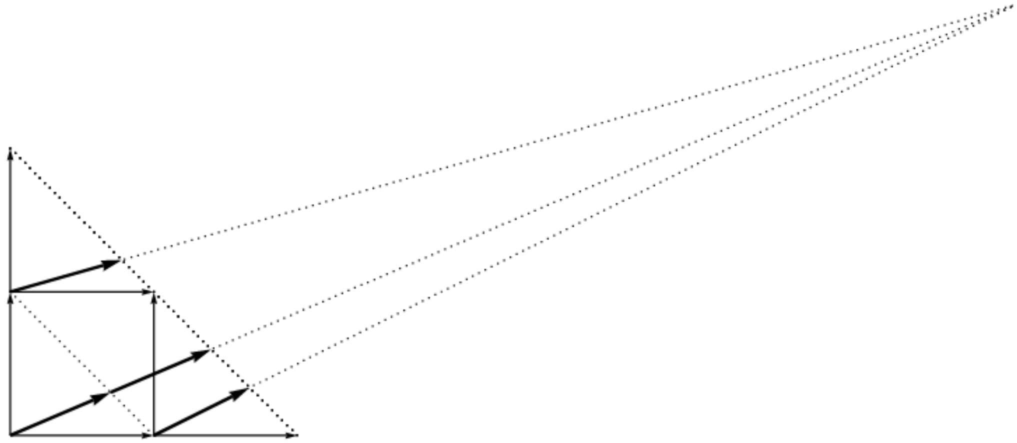
Remarks. Parts ((i)) and ((iv)) of the above corollary were pointed out by Bruno de Finetti in his study of exchangeable sequences of random variables. I do not know a reference for parts ((ii)) and ((iii)). It is a curious feature of this geometric construction of the vector $\lambda P(\,\cdot\,)$ from 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 λ_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.



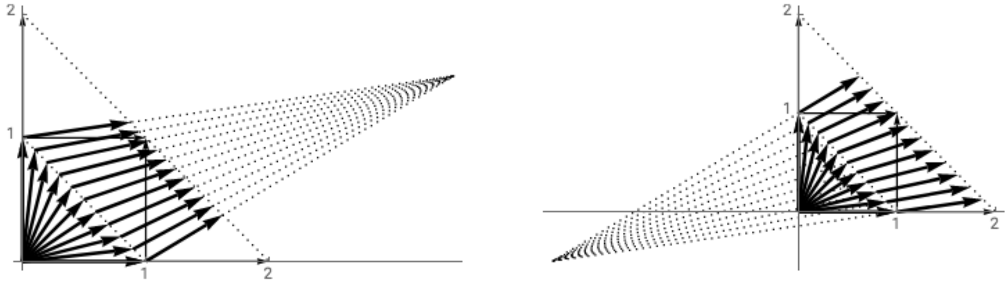
This diagram just amplifies the bottom left corner of the transition diagram of Figure 4.1 for sampling without replacement. The chain (W_0, W_1, W_2) makes 2 steps through 6 states (f, s) for non-negative counts f and s of failures and successes with $f + s \leq 2$, driven by (X_1, X_2) a sample of 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 2×2 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 λ is the corresponding probability vector λP, which always points from λ to (B, A). The stationary probability vector is $\pi = (7, 3)/10$, the unique vector such that both λ and λP 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 λP added to λ 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 λ_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 π 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 π 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 π. This is almost trivial, but of course we have to check that π 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 Δ_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^{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^{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,

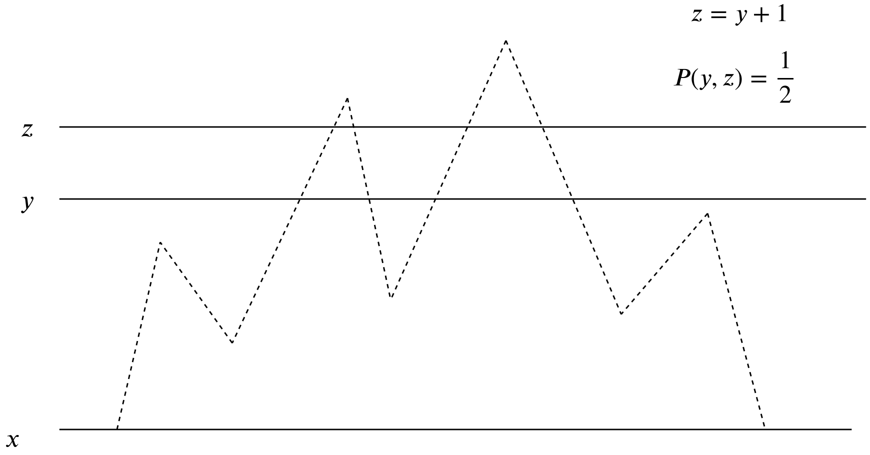
$$\mu(x,y) = \mathbb{E}_x(\#\text{ of hits on } y \text{ before } T_x)$$

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 π

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



5.7 Limit Theorem

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

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

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 π 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, τ 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

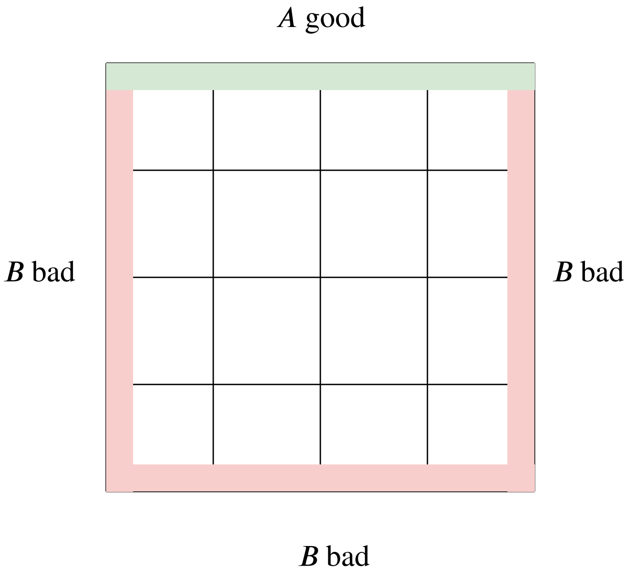
$$h(x) = \mathbb{P}_x(V_A < V_B) \tag{6.7}$$

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 $X_0, X_1, X_2, \dots$ 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 $X_1, X_2, \dots$ 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

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

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

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

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

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

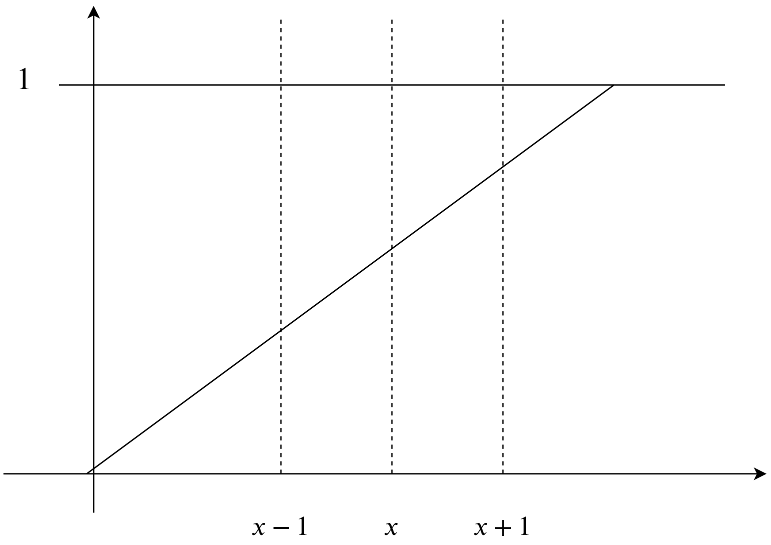
and we set the boundary conditions

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

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 $n \geq 0$

$$\begin{aligned} h(x) &= \mathbb{E}_x \left(\frac{q}{p}\right)^{X_n} \\ &= \left(\frac{q}{p}\right)^N \mathbb{P}_x(\text{hit } N \text{ before } n) + \left(\frac{q}{p}\right)^0 \mathbb{P}_x(\text{hit } 0 \text{ before } n) + \sum_{y \notin \{0,N\}} \cdots \end{aligned}$$

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

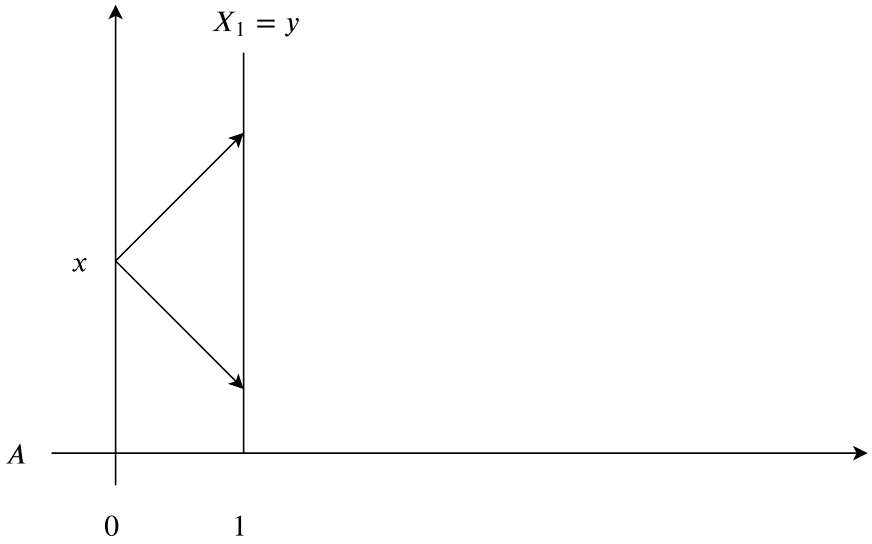
Suppose we have a set of states A (we can make them absorbing as a matter of technique, it makes no difference to the answer), and consider

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

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 ±\$1 gains with equal probability until we hit either \$0 or some

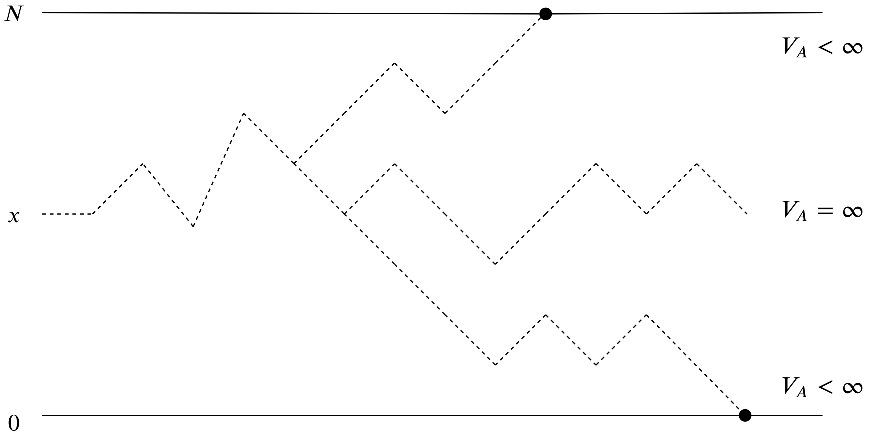
\$$N$. Last lecture, we showed

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

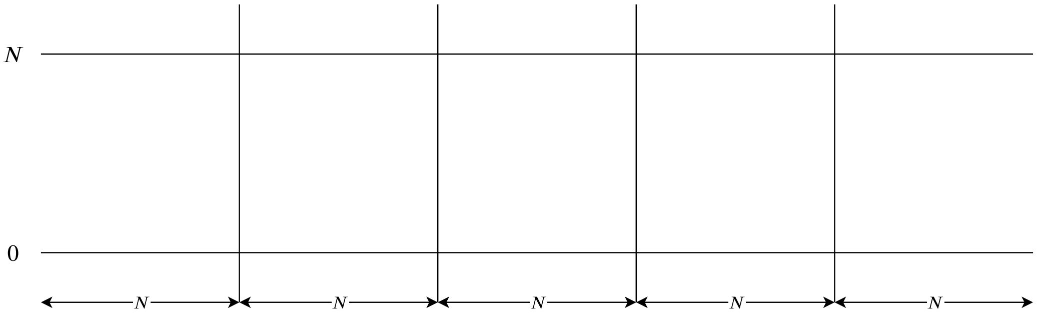
Set $A := \{0, N\}$ as our absorbing states and V_A the duration of the game, where

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

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 thi 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 τ_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 τ_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 τ_{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 τ_{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 μ_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 $G_0 \sim \textsf{\textbf{Geometric}}(q)$, 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 τ_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

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

meaning that conditional given $G_0 = g$) the remaining time $\tau_N - g$ has the same distribution as τ_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:

$$\begin{aligned} \mu_2 &= p^0 q(1+\mu_2) + pq(2+\mu_2) + p^2 \cdot 2 \\ \mu_2(1-q-pq) &= q + 2pq + 2p^2 \end{aligned}$$

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

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

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

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

Recall that $X_\lambda \sim$ **Poisson**(λ) 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,

$$\begin{aligned} \mathbb{E}X_\lambda &= \lambda \\ var(X_\lambda) &= \lambda \end{aligned}$$

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

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

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

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

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

$$\begin{aligned}
(r)_{n\uparrow} &:= r(r+1)\cdots(r+n-1) \\
\frac{(r)_{n!}}{n!} &= \binom{r+n-1}{n}
\end{aligned}$$

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 ES_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)

$$p_0, p_1, p_2, \cdots$$

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

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

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.

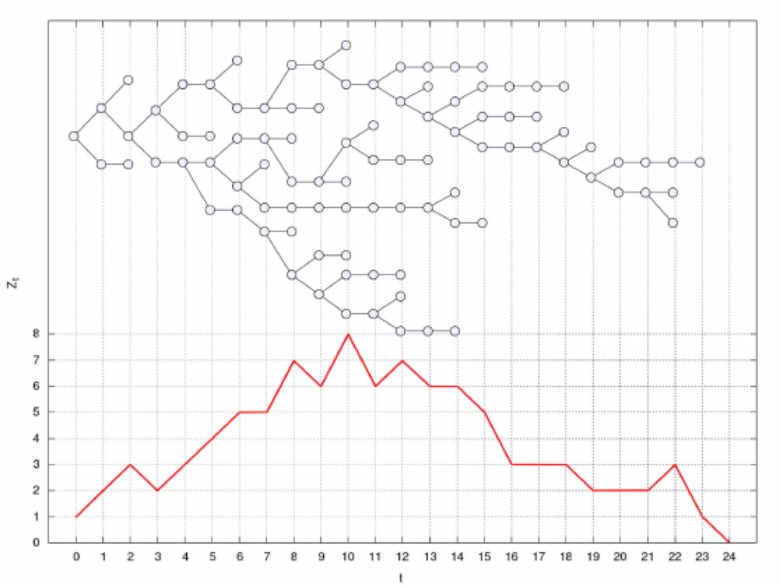


Figure 8.1: See [7] A realization of the Galton-Watson process. At the top, the tree associated to the process is shown, starting from the left ($Z_0 = 1$). At the bottom, the evolution of the number of elements originated in each generation t are displayed.

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 $0 \le s \le 1$. 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 ϕ_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 ϕ_{n-1} is the n th iterate of the offspring generating function ϕ. Because ϕ 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 ϕ, 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 ϕ 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.

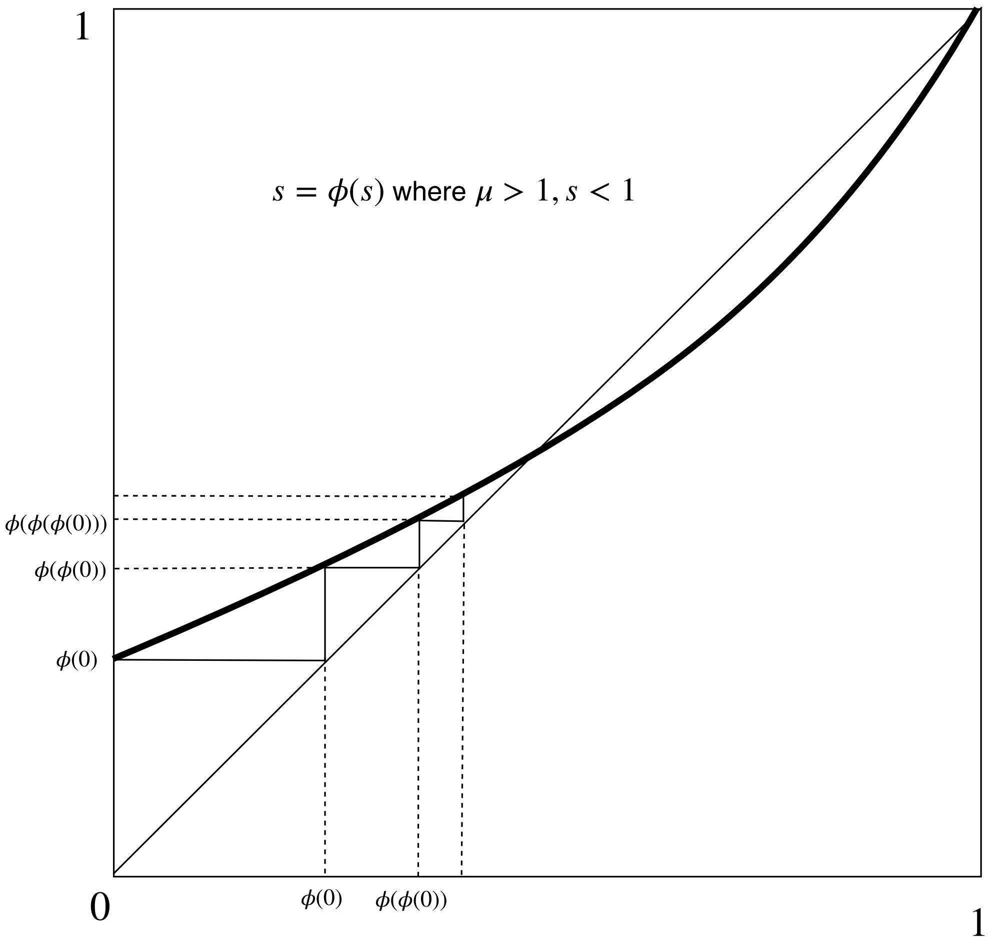


Figure 8.2: Supercritical. As an example, we sketch ($\phi(s)$ with respect to s for the generating function of **Poisson**$(3/2)$. This gives a fixed point iteration returning the unique root s of $s = \phi(s)$ with $s < 1$. Here the mean is larger than 1 ($\phi'(1) = \mu > 1$).

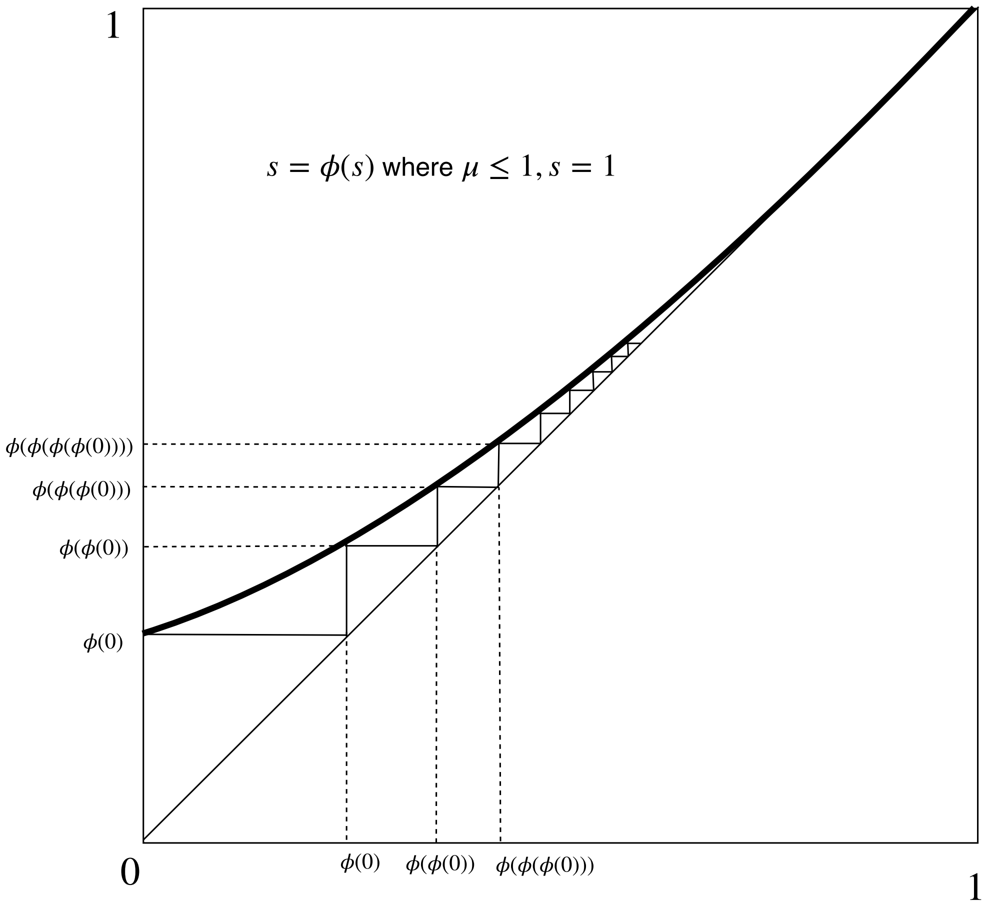


Figure 8.3: Subcritical or critical non-degenerate. We see that if $p_0 > 0$ and $\mu := \sum_n np_n \leq 1$, then the probability generating function is a convex curve with slope ≤ 1 at 1 and value $p_0 > 0$ at 0. So the curve cannot ever cross the diagonal s. There is a no root of $\phi(s) = s$ with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

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.