Stats 150, Summer 2019
Lecture 1, Thursday, 8/29/2019

CLASS SYLLABUS:
Course Description: Random walks, discrete time Markov chains, Poisson processes. Continuous time Markov chains, queueing theory, point processes, branching processes, renewal theory, stationary processes, Gaussian processes.

Prerequisite: Stat 134 or equivalent.

Text: Essentials of Stochastic Processes by Richard Durrett (available on Springer)

Grading:

$$\text{Overall grade} = \max\{\ 20\%\text{hw} + 30\%\text{mt} + 50\%\text{final},\ \ 20\%\text{hw} + 80\%\text{final}\ \}$$

Homework is due **weekly by 11:59pm each Thursday** on Gradescope. Lowest 2 homework scores will be dropped. No late homework.

Midterm: In-Class, Thursday Oct 17

Final: Tuesday, Dec 17, 8-11am

Supplemental Sections: Tues 4-6pm : 334 Evans

GSI: Jake Calvert

OH: T/Th 11am-12pm, 2-4pm: 303 Evans

Pitman notes that exams will require no calculations (will be theoretical) and mentions that should we have computational knowledge, he encourages that we practice simulations. We can compute 'the hell out of' problems to numerically get the correct answers, and Pitman encourages this.

Schedule of Topics:

- Week 0, Th Aug 29
  - 1.1 - 1.2 : Markov Chains, Multistep Transitional Prob (p. 1, 9)
- Week 1, T/Th Sept 3, 5
  - 1.3 Classification of States (p. 13)
  - 1.4 Stationary Distributions, Doubly Stochastic Chains (p. 21)
  - 1.5 Detailed Balance Condition, Reversibility, The Metropolis-Hastings Algorithm (p. 28)
- Week 2, T/Th Sept 10, 12
  - 1.6 Limit Behavior (p. 40)
  - 1.7 Returns to a Fixed State (p. 46)
  - 1.8 Proof of the Convergence Theorem (p. 50)
- Week 3, T/Th Sept 17, 19
  - 1.9 Exit Distributions (p. 53)
  - 1.10 Exit Times (p. 61)
- Week 5, T/Th Oct 1, 3
  - 2. Poisson Processes (p. 95)
  - 2.1 Exponential Distribution (p. 95)
  - 2.2 Defining the Poisson Process (p. 100)
  - 2.2.1 Constructing the Poisson Process (p. 103)
  - 2.2.2 More Realistic Models (p. 104)
- Week 6, T/Th Oct 8, 10
  - 2.3 Compound Poisson Processes (p. 106)
  - 2.4 Transformations: Thinning, Superposition, Conditioning (p. 108)
- Week 7, T/Th Oct 15, 17: Review, **Midterm Exam**
- Week 8, T/Th Oct 22, 24
  - 3.1 Renewal Processes, Laws of Large Numbers (p. 125)
  - 3.2 Applications to Queueing Theory (p. 130): GI/G/1 Queue, Cost Equations, M/G/1 Queue
  - 3.3 Age and Residual Life*, Discrete Case + General Case (p. 136)
- Week 9, T/Th Oct 29/31
  - 4.1 Continuous Time Markov Chains (p. 147)
  - 4.2 Computing Transitional Probability: Branching Processes (p. 152)
  - 4.3 Limiting Behavior, Detailed Balance Condition (p. 162)
- Week 10, T/Th Nov 5, 7
  - 4.4 Exit Distributions and Exit Times (p. 170)
  - 4.5 Markovian Queues, Single Server Queues, Multiple Servers, Departure Processes (p. 176)
  - 4.6 Queueing Networks* (p. 183)
- Week 11, T/Th Nov 12, 14
  - 5.1-2 Martingales, Conditional Expectations (p. 201)
  - 5.3-4 Gambling Strategies, Stopping Times, Applications (207)
- Week 12, T/Th Nov 19, 21
  - 5.4.1-2 Applications: Exit Distributions, Exit Times (p. 212)
  - 5.4.3-4 Extinction and Ruin Probabilities, Positive Recurrence of the GI=G=1 Queue* (p. 216)
- Week 13, T/Th Nov 26 : Gaussian Process and Brownian Motion (+ Thanksgiving)
- Week 14, T/Th Dec 3, 5 : Gaussian Processes and Brownian Motion
- Week 15, T/Th Dec 10, 12: RRR Week

1 Course Overview: Stochastic Processes

We may ask: What is a Stochastic Process? We can do a lot worse than go to Wikipedia. Most of our ideas and concepts are very mainstream, and we can simply Google them. The Wikipedia pages are mostly high-quality, according to Pitman. By the time we finish the course, we will likely cover the contents on those Wikipedia pages.

A pretty good start to answering this is to say that a Stochastic Process is a collection of random variables indexed by a parameter set I. That is,

$$(X_i,\ i \in I)\,,$$

and usually we usually take I as **time**, but it could be space (or even fancier, could be both).
Behind these random variables, there is some **probability measure** which we will call $\mathbb{P}$.

We sometimes denote the measure as $\mathbb{P}_\lambda$, and start with the friendly non-negative integers: $I = \{0, 1, 2, 3, \dots\}$.

$$\begin{aligned} X_0 &= \text{Initial State} \\ X_1 &= \text{State with Time 1} \\ &\vdots \\ X_n &= \cdots \end{aligned}$$

As the first interesting process, we want to think of the **Markov Chain** with a countable space S and a stationary transitional probability matrix P.

Notice there is a slight difference between Pitman's blackboard notation and that given in the text. When Pitman writes $\mathbb{P}$ as a probability measure on the blackboard, the text will simply write P. Pitman will write P where the text will write p for a **matrix** (although Pitman does not encourage using this lowercase letter to denote a matrix).

Remark: $\mathbb{P}$ denotes a probability measure.
P denotes a matrix.

To specify a Stochastic Process (S.P.), we must decide the joint distribution of its variables. That is, we want to know the probability of :

$$\mathbb{P}\left(X_0 \le 3, X_1 \le 5, X_2 \le 7\right) = \text{something}$$

This is a joint probability involving 3 variables X_0, X_1, X_2. The idea behind an S.P. is that this does not matter on how many variables at which we look. We simply use the probability measure $\mathbb{P}$.

A question is posed in class: What is the **sample space** (outcome space) Ω here? The quick answer given by Pitman is that it depends. The canonical answer is:
For state space S and Time set I, the Ω space is:

$$\begin{aligned} \Omega &= \prod_{i \in I} S_i \\ &= \text{all}\ \ (x_i,\ i \in I)\,, \end{aligned}$$

where S_i is a copy of S and $x_i \in S$. If we take an infinite horizon $I = \{0, 1, 2, \dots\}$, then we have:

$$\Omega = \text{all spaces } (x_0, x_1, x_2, \dots), \quad x_i \in S$$

2 Markov Chain

Markov Chains are a bit more interesting, where we have some dependence between variables, so we cannot simply multiply together measures or sample spaces.
How do we define probabilities for a Markov Chain (MC)? Let's assume S is the set of nonnegative integers, $\mathbb{N}_0 = \{0, 1, 2, \dots\}$. We call this a **counting variable**. Then we have:

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

where P is a matrix and we add up over all the cases. The **Markov Chain** specifies the joint probability $P(x, y, z)$ in a very simple way:
It gives the joint probability:

$$\begin{aligned} P(x, y, z) &= \mathbb{P}(X_0 = x) \cdot \underbrace{P(x, y)}_{\text{TPM}} \cdot P(y, z) \\ &= \mathbb{P}(X_0 = x) P(y|x) = P(z|y) \quad \text{(we do not use this notation)} \end{aligned}$$

where TPM stands for 'transition probability matrix', and

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

where

$$P(AB) = P(A \cap B) = P(A) P(B|A)$$

Notice the inversion of the variables here in our choice not to use $P(z|y)$ notation.

Now if we want to look at $P(x, y, z, w)$, to modify our formula, we take:

$$P(x, y, z, w) = \mathbb{P}(X_0 = x) P(x, y) P(y, z) P(z, w)$$

Remark: Notice that for any sequence of 3 random variables (RV) x_0, x_1, x_2, we can write the probability that we have the triple of values:

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

What is special about this formula is that the last factor only has y and is **independent of** x. This is the key to the Markov Chain probability:

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

Definition: Markov Property -

Generally $X_0, X_1, X_2, \dots$ has the **Markov Property** if

$$\begin{aligned}\mathbb{P}\left(X_{n+1} = x_n | X_0 = x_0, X_1 = x_1, \dots, X_n = x_n\right) \\ = \mathbb{P}\left(X_{n+1} = x_{n+1} | X_n = x_n\right)\end{aligned}$$

Or more simply, as a mantra: Past and future are **conditionally independent** given the present.

2.1 Homogeneous Transition Probabilities

One more point: We are assuming that our Markov Property has a **homogeneous transition probabilities**, where we use the **same** rule (matrix P) to go from one state to the next (as opposed to changing the rule at each step).
Now for any assigned initial distribution λ for X_0,

$$\mathbb{P}(X_0 = x_0) = \lambda(x_0), \qquad \sum_{x_0}^{\lambda(x_0)} = 1, \quad \lambda(x_0) \geq 0,$$

and for any transitional probability P,

$$P(x, y) \geq 0, \qquad \sum_y P(x, y) = 1,$$

which we say are the **rules of a transition matrix**.

From initial λ and a matrix P, we now have:

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

This is a concise way to say that we make a next move given the current state using the same matrix P. This is the prescription of the joint distribution of the first $n+1$ steps of a Markov Chain with initial distribution λ and a homogeneous TPM (Transition Probability Matrix) P. It is more or less obvious that this is a proper (rules of probability) assignment of a joint distribution.
To check this, we want to see that summing all outcomes gives 1 and that all probabilities are nonnegative (trivial). That is,

$$\text{Prob} \geq 0, \qquad \text{Probs sum to 1.}$$

To see the second part via some inductive or iterative argument, notice that we have:

$$\sum_{x_0} \sum_{x_1} \cdots \left(\sum_{x_n} \xi \right) = 1$$

due to our previous requirement above (pink) $\sum_y P(x, y) = 1$. If we are pedantic, we may want to formalize via induction; however, for our purposes it suffices to recognize the argument. This is saying that we can certainly make a probability assignment as well as that we can simulate this guy.

2.2 Simulations

We have a supply of independent uniform $[0,1]$ variables: $U_0, U_1, U_2, \dots$. (Of course, these random numbers are not truly random as given by pseudorandom numbers). Any random process we discuss in this course can be simulated by some uniform variables.

First of all, how do we make $X_0 \sim \lambda$? To simplify, let's say $S = \{0, 1, 2, \dots\}$. We can take U_0 to make X_0 (notice that the sample space need not be a sequence space). That is, set $X_0 := 0$ if $U_0 \in [0, \lambda(0)]$, and

$$X_0 := \begin{cases} 0, & U_0 \in [0, \lambda(0)] \\ 1, & U_0 \in (\lambda(0), \lambda(0) + \lambda(1)] \\ 2, & U_0 \in (\lambda(0) + \lambda(1), \lambda(0) + \lambda(1) + \lambda(2)] \\ \vdots, & \vdots \end{cases}$$

If $X_0 = x_0$, then:

$$X_1 = 0, \text{ if } U_1 \in [0, P(x_0, 0)],$$

and

$$X_1 = 1, \text{ if } U_1 \in (P(x_0, 0), P(x_0, 0) + P(x_0, 1)],$$

and so on. If $X_0 = x_0$ and $X_1 = x_1$, then make $X - 2$ from $P(x_1, \cdot)$. This would take about 10 lines of Python or R code, and Pitman says we should really be able to do this.

Example: Gambler's Ruin
Consider a gambler with an initial fortune $\$a$ with $0 \le a \le N$ and $a \in \{1, 2, \dots, N-1\}$.
Take $X_0 = a$, and with each play, gambler gains +\$1 with probability p and loses \$1 with probability q where $p + q = 1$.

Pitman notes this is fairly obviously a Markov Chain. X_n is the capital after n states. λ is degenerate with $\lambda(a) = 1$ and $\lambda(x) = 0$ else. We write:

$$\begin{aligned} P(x, x+1) &= p, \qquad 0 < x < N \\ P(x, x-1) &= q \end{aligned}$$

For mathematical convenience, we construct an **absorbing boundary** to say that the game goes on infinitely if the gambler hits 0. We say:

$$P(0,0) = 1, \quad P(N,N) = 1.$$

For $N = 3$, we can simply fill in the matrix:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ q & 0 & p & 0 \\ 0 & q & 0 & p \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

If we take $p = q = \frac{1}{2}$, then it is not too hard to prove that we will hit one of the boundaries in that we will walk out with 0 or N dollars (as opposed to oscillating forever). We reason in class that this case is **linear in** a. There is only one such function that is linear in a and satisfies our given boundary equations, so we conclude this is simply:

$$\frac{a}{N}.$$

For the case where $p \neq q \neq \frac{1}{2}$, then this is more complicated with writing equations with unknown probabilities.

Lecture ends here.

Stats 150, Fall 2019
Lecture 2, Tuesday, 9/3/2019

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

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

With many steps and homogeneous transition probabilities, we have:

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

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

0.1 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:

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

To have this fit with our convention of matrix multiplication, we 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 $P(x, y)$.

0.2 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:

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

where argument y is hidden. If we evaluate the row vector λP at entry y, we 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.

1 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 something very abstract). Pitman notes some applications to Google's PageRank with a Markov property and others. Without being scared about the potential size of the **state space**, we open to some abstraction in our immediate example.
Consider the Markov Chain from X_0 to X_1 and the conditional expectation:

$$\mathbb{E}\left(f(X_1) \mid X_0 = x\right) = \sum_y \underbrace{P(x,y)}_{\text{matrix}} \underbrace{f(y)}_{\text{col.vec.}}$$

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: f can be signed (there is no difficulty if we are losing money as opposed to gaining); it is only difficult to interpret if λ is signed.

2 Two Steps

Now consider two steps:

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

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, namely 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.

2.1 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, then Pitman may ask 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 'coinkidink' (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 (minimizing numerical error). Realize that every technique here has an **immediate application** to Markov chains (with very many steps). Note the Chapman-Kolmogorov equations:

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

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

Break time.

3 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 matrix 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 transition matrix P. This is truly a trivial MC. All rows of P are equivalent to some $p = (p_0, p_1, \dots)$. We consider:

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

(Pitman adds that we ignore costs or losses for convenience, so that natural state space of S_n is $\mathbb{N}_0$).

4 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:

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

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.

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

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:

$$P^2(0, 0) = q^2$$
$$P^2(0, 2) = p^2$$
$$P^2(0, 1) = 2pq,$$

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

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

Now being at an initial fortune i, we have:

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

4.1 More Challenging:

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

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

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

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

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

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

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

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

So in conclusion, we have found:

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

We may 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 termed 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. Pit-

man 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 underbraced expression.
Repeating this convolution, we move forward from $n = 2$:

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

Example: Pitman asks us to simulate 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 underbraced term (power of 5). Pitman credits the inventor 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 we can count the number of dice patterns on one hand; however, the evaluations of $P^4(0,k)$ for $4 \le k \le 24$ above are not so trivial (outside of this method by Laplace). This method can be used to prove all the familiar properties of sums of independent discrete variables (e.g. sums of Poissons is Poisson). We should try it for this purpose.

Lecture ends here.

Stats 150, Fall 2019
Lecture 3, Thursday, 9/5/2019

Administrative Book-keeping: Homeworks are due weekly on Thursday at midnight. The assignment for the following week will be released at worst (latest) Friday morning.
Pitman opens 4-6pm and 6-8pm for Office Hours.

Pitman notes that we've only really made it through to §1.2, and we need to catch up to our advertised speed. We will handle in-class questions but will omit most proofs. For this lecture and forward, we will assume the audience has read the relevant textbook sections.

1 §1.3 : Classification of States

Consider a Markov chain with transition matrix P. With the state space S, consider states $x, y, i, j \in S$. Take P, S to be fixed.
We are interested in 'hitting times':

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

Given boundary state B, we consider one path X_n to get from x to T_B. We call these **first passage (hitting) times**.
The immediate (pedantic) issue with this is if $X_n \notin B$ for all n. In this case, we need to define the convention:

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

Theorem 1.1. Strong Markov Property (SMP)

Given $T_y = n < \infty$ and $X_n = y$, then

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

is the original MC given that it starts at y.

To see this, we start with $X_0, X_1, X_2, \ldots$ which is a Markov chain. Here, X_0 has any initial distribution.

The distribution of $X_{n+1} \mid X_n = y$ is $\delta_y P(\cdot) = \mathbb{P}(y, \cdot)$, which is:

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

and so on, giving rise to a family of equations.

Proof. See Durett page 14. □

Remark: In discrete time (even with general state space), **all** Markov chains have the Strong Markov Property.

Now, we can use the SMP to discover and prove things about Markov chains.

2 Iterating:

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

Suppose we have a path that hits the state y a finite number of times, say four times: T_y, T_y^2, T_y^3, T_y^4. Then via our convention, we say that $T_y^5 = \infty$.
Random variable:
Consider the number of hits at y (not counting time 0):

$$N_y := \sum_{n=1}^{\infty} 1\,(X_y = y)$$

to be the total number of hits to y at n after time 1. We consider that the possible values of this is $\{0, 1, 2, \ldots, \infty\}$, an infinite time horizon.
Equivalently, by definition or logic, we have:

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

As another example, consider $(N_y \geq 1)$ is 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 that $N_y := \sum_{n=1}^{\infty} 1(X_n = y)$, simply counting the number of hits on y.
Pitman asks the audience to explicitly find:

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

Now let's discuss the probabilities. Let

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

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:

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

Now why is this true, as to get from ρ_y to ρ_y^2? Basically, this is by the SMP (Strong Markov Property).
Notice:

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

which tells us that the probability of hitting y at least k times is:

$$\mathbb{P}_y(N_y \geq k) = \rho_y^k \text{ for } k = 0, 1, 2, 3, \cdots = \mathbb{N}.$$

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

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

which tells us that the probability distribution (starting at y) of $N_y := \sum_{n=1}^{\infty} 1(X_n = y)$ is Geometric with

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

Pitman wants us to notice, using tail-sums:

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

3 States of y

Now there are two cases Pitman wants us 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.

Break time.

Pitman gives us a formula if we really want to have a more concrete understanding of ρ_y:

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

4 Lemma 1.3

Take B to be a set of states.

4.1 Hypothesis:

Suppose the probability starting at x that $T_y \leq k$ is at least $\alpha > 0$ for some k and all x. In other words,

$$\mathbb{P}_x(T_y \leq k) \geq \alpha > 0$$

for some k and all x.
As an example of this hypothesis, consider the Gambler's ruin chain with state 0 and n absorbing states. That is, $B = \{0, N\}$, with $P(i, i+1) = p$ and $P(i, i-1) = q$. Then this condition holds with $k = \lceil \frac{N}{2} \rceil$. Then

$$\alpha = \max\{p^k, q^k\}.$$

4.2 Conclusion:

Suppose the hypothesis is satisfied. Then

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

Proof. Consider the subset $U \subset \{0, 1, 2, \dots\}$. Surely, this is trivial for $n = 1$ in that $\mathbb{P}_y(T_B > k) \leq 1 - \alpha$ for all y.
Now,

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

as an upper bound.

□

Pitman mentions Kai Lai Chung of Stanford who paints this concept as the idea of a pedestrian crossing the road. This gives a geometric bound . If there is a certain chance of reaching a boundary state, eventually a Markov chain will hit such a boundary state.
Observe that in standard Real numbers,

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

implies

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

Further, notice:

$$(T_B = \alpha) \subseteq (T_B > nk),$$

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

Conclusion of Gambler's Ruin Example: Suppose $0 < p < 1$. In terms of transient and recurrent states, every state $x \notin \{0, N\}$ is **transient**! Moreover, $x \in \{0, N\}$ is recurrent.

Definition: Irreducible Matrix -

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

$$\forall_{x,y \in S}, \exists_n : P^n(x, y) > 0$$

In words, for every pair of states x, y, it is possible to get from x to y in some number n of steps.
Note that n is a function of x, y. That is, $n = n(x, y)$.

Theorem 4.1. If matrix P is irreducible, then either:
- all states are recurrent
- all states are transient

With sloppy language, we then say either the matrix P is recurrent or transient, respectively

Notice that the Gambler's Ruin problem exhibits a matrix that is NOT irreducible, which can be seen via the definition above and the requirement that there exists some n where $P^n(x, y) > 0$.
An easy fact:
Suppose S is finite and P is irreducible. Then in this language, P is 'recurrent'.

Lecture ends here.

On Tuesday we will cover stationary distributions and stationary measures.

Stats 150, Fall 2019

Lecture 4, Tuesday, 10/9/2019

CLASS ANNOUNCEMENTS: Pitman announces he added a Mathematica notebook on bCourses to illustrate some examples.

Topics Today:

- Symmetries in Distributions
  - Exchangeability
  - Reversibility
  - Stationarity

1 Sampling Without Replacement (SWOR)

To do this, suppose we have a population of tickets in a box, say 3 tickets labeled 1, and 7 tickets labeled 0. These tickets are all shuffled within the box, and we pull them out via 'random draws' in a sequence:

$$(X_1, X_2, \ldots, X_{10}) = (100|0110000),$$

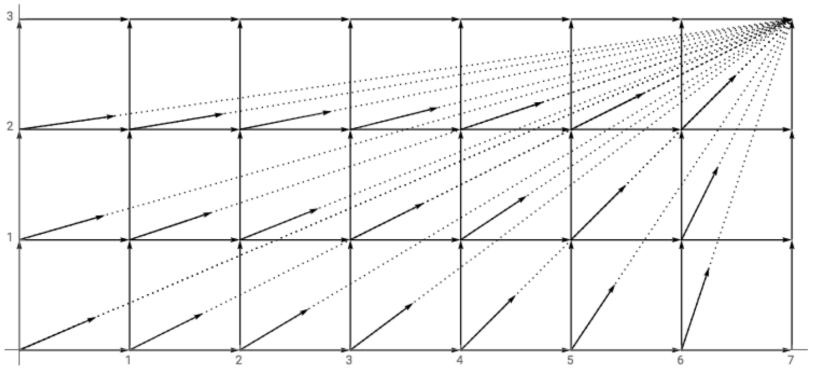
where writing the sequence like this (the bar can be placed anywhere) helps us think about conditioning, via our usual combinatorial counting. We can make a Markov chain out of this. Let S_n be the number of successes (1) in the first n draws. Let F_n be the number of fails (0) in the first n draws. By convention, we write the fails first:

$$W_n = (F_n, S_n),$$

and note obviously $F_n + S_n = n$. Now if we say this is a chain, then very easily,

$$(W_n, 0 \leq n \leq N)$$

is a Markov chain. For example, we could have a pair in the state space like $(B, A) = (7, 3)$.



We have a nice grid diagram starting at (0,0) and ending at (B, A). Now, this gives $\binom{10}{3} = \binom{10}{7}$ such paths via basic counting.

1.1 Transition Probabilities

Pitman wants us to think about vectors here. Consider the 2-state probability vector, denoted by (q, p) (or similarly denoted as (λ_0, λ_1) or (π_0, π_1)), where q, p are probabilities of the assignment (of 0 and 1, respectively). Consider the probability 'vectors' from the origin (0,0) to some point on the graph as a way to think about the transitional probability.
From each state, we draw the line connecting the current state with our final target, as can be seen in the earlier diagram.
Now, we look at (X_1, X_2) for $X_1, X_2, \ldots, X_N$ on sampling without replacement (SWOR). Consider, in terms of the probability of getting 1 in the first draw is:

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

where the left factor is the π part, and the right factor is the P part.

Definition: Reversible -

We say that the pair (X_1, X_2) 'reversible' if (and only if)

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

That is, they are equal in distribution.

More generally, for $N \geq 3$ and taking the single permutation σ which reverses the order of indices, we say the sequence of random variables $(X_1, \ldots, X_N)$ is **reversible** if

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

Remark: If we have X has the same distribution as Y, then $\psi(X) \stackrel{d}{=} \psi(Y)$. To see this, consider:

$$\begin{aligned}
\mathbb{P}(\psi(X) \leq 3) &= \mathbb{P}(\{x : \psi(x) \leq 3\}) \\
&= \mathbb{P}(X \text{ has property } \psi(X) \leq 3) \\
&= \mathbb{P}(Y \text{ has property } \psi(Y) \leq 3).
\end{aligned}$$

That is, equality in distribution pushes forward through functions.

Now, consider the projection function $(X_1, X_2) \stackrel{p}{\mapsto} X_1$ and $(X_2, X_1) \stackrel{p}{\mapsto} X_2$ (that is, take the first element). Now because $(X_1, X_2) \stackrel{d}{=} (X_2, X_1)$, via this projection mapping, we conclude $X_1 \stackrel{d}{=} X_2$.
Hence we proved that as we are along the 'line of symmetry', the two marginal probabilities are the same.

2 Exchangeability and Reversibility

With $N = 2$, the terms exchangeability and reversibility are equivalent. Take a distribution $X_1 \sim \pi$, and take a conditional distribution

$$X_2|X_1 \sim P(X_1, \cdot),$$

which is the blackboard shorthand for:

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

where we take the element of row x and column y of transition matrix P. Now this shows that the joint distribution of X_1, X_2 is

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

because intersection is a commutative operation. Notice:

$$(X_1, X_2) \overset{d}{=} (X_2, X_1) \iff \pi(x)P(x, y) = \pi(y)P(y, x), \forall_{x,y \in S},$$

where obviously these must have the same space of values. We say that X_1, X_2 are **reversible** (or exchangeable).

Remark: Consider the space of 0s and 1s. From the beginning of class, take (X_1, X_2) to be a sample of size 2 **without replacement** (the case with replacement is trivially true) is in fact, reversible.
To show that the pair of indicators is exchangeable, we need only show that the off-diagonals are equal.

Now, for a typical example of a two-step transition, first look at the matrix P for sampling without replacement. It has some form like:

$$\frac{1}{A + B - 1} \begin{bmatrix} * & 0 & 1 \\ 0 & B - 1 & A \\ 1 & B & A - 1 \end{bmatrix}.$$

Now to claim this is a transition matrix, we need to check that all entries are nonnegative, so take integers $A, B \geq 1$. Then we check that the row sums are equal to 1. These can be seen easily via inspection. Now for sampling without replacement for (B, A) for 0 and 1.

2.1 Generalizing

Pitman notes that a mathematician may wish to generalize this past the positive integers. That is, what pairs of reals (B, A) is this P a (valid) transition matrix? Well, of course we want $A + B - 1 \neq 0$ to avoid division by zero. Then, notice that the row sum requirement (identically equal to 1) is completely independent of our choices for A, B. Finally, we require that all entries are nonnegative. Considering a (B, A)-space, take $B \geq 1$ and $A \geq 1$, and notice that this will satisfy this nonnegativity requirement. Moreover, notice that if we take $A \leq 0$ and $B \leq 0$, then we have division of a nonpositive number by a negative denominator.

We claim that every 2×2 transition matrix $P = \begin{bmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{bmatrix}$ is of the form:

$$P = \frac{1}{A + B - 1} \begin{bmatrix} B - 1 & A \\ B & A - 1 \end{bmatrix}$$

for a unique pair of (B, A). Of course, to prove this, we need to play with a system of equations and unknowns. Pitman recalls the result is essentially:

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

Let $B := \frac{P_{10}}{P_{01}-P_{11}}$, so that

$$B - 1 = \frac{P_{10} - P_{01} + P_{11}}{P_{01} - P_{11}} = \frac{1 - P_{01}}{P_{01} - P_{11}} = \frac{P_{00}}{P_{01} - P_{11}}.$$

Now it only remains to check our result (more or less following from properties of a transition matrix). Generally, what we showed is that a two-state Markov chain is just like a pair of indicators and sampling without replacement, where we changed only a few things.
Notice that the stationary distribution π is

$$(\pi_0, \pi_1) = \left(\frac{B}{A+B}, \frac{A}{A+B} \right),$$

which in general gives that the stationary π for P: $\pi P = \pi$ is:

$$(\pi_0, \pi_1) = \frac{(P_{10}, P_{01})}{P_{10} + P_{01}},$$

which is a generalization of our A, B formula earlier for sampling without replacement.

Definition: Stationary Distribution, Reversible -

We say that π is a stationary distribution for P if and only if $\pi P = \pi$, which is precisely when we have the following '**balance equation**':

$$X \sim \pi \text{ and } X_2 \mid X_1 \sim P(X, \cdot) \implies X_2 \sim \pi.$$

We say the stationary distribution π is **reversible** if additionally

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

which we've showed is equivalent to the property:

$$\pi(x)P(x, y) = \pi(y)P(y, x), \forall_{x \neq y}.$$

We call this the '**detailed balance equations**'.

2.2 Does a solution π of the Detailed Balance Equation imply the solution π of the Balance Equation?

If we have found a solution π of the DBE, this means that

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

Now, having the solution π for BE is that:

$$X_1 \stackrel{d}{=} X_2.$$

It is completely trivial that if we can flip the pair across the diagonal, then the two entries must be equal.

3 Checking by Algebra

We know that $\pi(x)P(x,y) = \pi(y)\pi(y,x)$. We want to show

$$\sum_x \underbrace{\pi(x)P(x,y)}_{\pi(y)P(y,x)} = \pi(y).$$

Now within the summation, y is fixed, and so this equation is true for all states y.

4 Some Easy Facts:

(1) If π is stationary for P, then π is stationary for P^n. We can show this easily for $n = 2$:

$$\pi P^2 = \pi PP = \pi P = \pi.$$

Mathematical (strong) induction delivers the result similarly.
(2) If π is a reversible equilibrium, then for a Markov chain with initial distribution π and transition matrix p (π, P), then we have the same probability distribution. That is,

$$(X_N, X_{N-1}, X_{N-2}, \ldots, X_0) \stackrel{d}{=} (X_0, X_1, \ldots, X_N).$$

(3) Not all equilibriums are reversible. Consider states rotating around a circle, so that the transition matrix is

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

Notice that column sums of P also equal 1, so the transpose of P, P^{-1}, also has row sums 1. We say that P is a 'doubly-stochastic' matrix. This shows that if π is constant, then $\pi_i = \frac{1}{N}$, where N is the number of stationary vectors. We may ask: is this reversible?
We answer that clearly not, where the plot is not symmetric across the diagonal.
Pitman concludes the lecture with some interesting and much less trivial facts, consequences of the above definitions. These are accepted as known and quotable for homeworks.
In general, Pitman allows us to cite anything in the text.

5 Key Theorems

Recall that we've talked about irreducible states. If P has a finite number of states and P is irreducible (that is, $\forall_{i,j}\exists_n : P^n(i,j) > 0$.).
Then $\exists!$ stationary probability of $\pi : \pi P = \pi$. A formula for this is:

$$\pi_i = \frac{1}{\mathbb{E}_i T_i} > 0,$$

where $\mathbb{E}_i T_i$ is the mean return time. The same is true for any irreducible positive recurrent P, then $\mathbb{E}_i T_i < \infty$, for all i. Then we say that

$$\pi_i := \frac{1}{\mathbb{E}_i T_i}$$

is the unique stationary probability distribution.
Now define $d(x) :=$ gcd of all $n : P^n(x, x) > 0$ (where intuitively we are moving around a circle). If P is irreducible, then $d(x) \equiv d$ for all x, and if it is **aperiodic** (namely $d = 1$), then $P^n(i, j) \to \pi_j$ as $n \to \infty$. That is, when we take higher powers, we approach equilibrium with:

$$P^n \to \begin{pmatrix} \pi \\ \pi \\ \pi \\ \vdots \end{pmatrix}$$

Pitman urges us to check the claims in the lecture notes on our own as an exercise.

Lecture ends here.

Stats 150, Fall 2019

Lecture 5, Thurs, 9/12/2019

Topics Today:

- §1.4 - 1.8 of text
- The remaining of readings are assigned outside of lecture.

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.
(1) Recall the definition of an irreducible chain. That is,

$$\forall x, y \in S, \exists_n : P^n(x, y) > 0.$$

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.
(2) Fact: (See text). 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 recurrent (the chain is recurrent). (Theorem 1.7 in Durett).

Definition: Positive Recurrent -

We say that the chain is **positive recurrent** when, for some or for all x:

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

which is:

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

We should check that if $E_x T_x < \infty$ for some x and P is irreducible, then

$$E_x T_x < \infty, \quad \forall_x.$$

Definition: Null recurrent -

If a state is recurrent but not positive recurrent (for example $P_x(T_x < \infty) = 1$ but $E_x T_x = \infty$), then we say that x is **null recurrent**.

2 Review

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.

Remark: 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.

3 Example of Null Recurrence

Example: 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 has the value $+1$ with probability $\frac{1}{2}$ and -1 with probability $\frac{1}{2}$. Now 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 $(1 - \rho_0)$:

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

But we know that $\binom{2m}{m}\left(\frac{1}{2}\right)^{2m}$ 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 $\mathbb{E}_0 S_{2m} = 0$. Now 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 (or the Normal Approximation), that is,

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

and apply this to our earlier expression to show that:

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

where C is some constant.
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 that the definition of recurrent gives:

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

Also, we should know that positive recurrence implies recurrence, but not necessarily conversely.

Remark: Pitman summarizes that on our homework, we can quote the result that:
If we have a stationary measure, then the chain is **positive recurrent**.

4 Notion of x-blocks of a Markov chain

Start at x (for simplicity) or wait until we hit x. Then look the successive return times $T_x^{(i)}$ which is the ith copy of T_x. Now recall this has the Strong Markov Property, which gives us two things:
(1) Every $T_x^{(i)}$ has the same distribution as T_x.
(2) 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:

Example: Let $N_{xy}^{(2)} :=$ the # of visits to y in the ith 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'.

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

Now this implies that there is a formula involving expectations. Take $\mathbb{E}_x$, the expectation starting at x:

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

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:
(1) $N_{xy}^{(i)}$ has the same distribution as $N_{xy} := N_{xy}^{(1)}$.
(2) Further, the $N_{xy}^{(i)}$ are independent and identically distributed (iid).

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.

Break time.

5 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 from earlier, we write:

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

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 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:

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

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.

5.1 Explanation of the Key Formula

We may ask why we have:

$$\sum_y \mu(x,y)P(y,z) = \mu(x,z). \tag{1}$$

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:

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

6 Limit Theorems

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

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

We have this Cesàro mean convergence always for irreducible positive recurrent chains (these themselves do not 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.

6.1 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 random walk, we find that the expectated probability of every state 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 (and hence is unique as a limit must be unique).

Lecture ends here.

Stats 150, Fall 2019
Lecture 6, Tuesday, 9/17/2019

CLASS ANNOUNCEMENTS: Pitman notes that Homework 2 was intentionally hard to push us; however we may have relief at times like Homework 3.

Pitman opens to questions regarding irreducible, aperiodic, recurrent (including positive or null recurrent), or transient. For a nice transition probability matrix, there exists a stationary probability π so that:

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

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

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

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.

1 Transient Aspects of Chains

We want to discuss various transient aspects of chains, especially the distributions of hitting places and times. Pitman notes that both are interesting and we approach them with similar techniques. Hitting places is a bit easier to treat, so we start with those.

1.1 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:

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

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

Theorem 1.1. (Durett 1.28, p. 55) Consider a Markov chain with state space S. Take two (necessarily disjoint) $A, B \subseteq S$.
We 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).

1.2 Assumptions

Suppose we have $h : S \to \mathbb{R}$ so that:

$$h(a) = 1, \forall_{a \in A}$$
$$h(b) = 0, \forall_{b \in B}.$$

Then by definition, $C := S - (A \cup B)$, intuitively we think of C as an interior set and A, B as boundary sets.
Suppose that $h(x) = \sum_y \mathbb{P}(x, y)h(y)$. Very commonly, we'll write this in matrix notation, where h is a column vector with $h = Ph$ (unlike we have done with row vectors like π).
Additionally, for all $x \in S$, suppose:

$$\mathbb{P}_x(V_A < \infty \text{ or } V_B < \infty) = 1 \iff \mathbb{P}_x(V_{A \cup B} < \infty) = 1.$$

Further, assume that the set of interior states are finite.
Because we put 0 in the definition of V_A, then this is trivial for $x \in A \cup B$. The question is if we start inside the lattice, what conditions do we have to hit the boundaries?

We hold off on the Theorem's claim until after some discussion.

1.3 Side Example:

Find the distribution of $X_{V_{A \cup B}}$, where $V_{A \cup B}$ is the first hit of $A \cup B$. Notice that if we start at $x \in A \cup B$, we are there already, hence there is nothing to find. If we start at $x \notin A \cup B$, we may ask: What is $\mathbb{P}_x(X_{V_{A \cup B}} \in A)$? That is, starting at an interior state, what is the probability that we end up at the

1.4 Method of Solution (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 1, X_1. Generalizing, let Y be any nonnegative (for simplicity) random variable, and let $X_0, X_1, X_2, \ldots$ be a Markov chain with transition matrix P. Consider $\mathbb{E}_x$ as a function of Y, and notice we can write, by summing out all $y \in Y$:

$$\begin{aligned}
\mathbb{E}_x Y &= \mathbb{E}_x \underbrace{\sum_y \mathbb{1}(X_1 = z)\, Y} \\
&= \sum_z \mathbb{E}_x \left[\mathbb{1}(X_1 = z) Y\right] \\
&= \sum_z \mathbb{P}_x(X_1 = z)\mathbb{E}(Y \mid X_1 = z).
\end{aligned}$$

Notice that we haven't used the Markov property, so we use that now:

$$\mathbb{E}_x Y = \sum_z P(x,z)\mathbb{E}(Y \mid X_1 = z),$$

which is simply computing the expectation by conditioning for general Y. Commonly, take the case where Y is an indicator, for example $Y = \mathbb{1}(V_A < V_B)$. Then

$$\mathbb{E}_x Y = \mathbb{P}_x(V_A < V_B).$$

Pitman notes that 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:

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

where Pitman notes this boils down to simply being familiar with the formal notation for this to be clear.

Remark: Pitman poses a question: 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 (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 we arrive, we stick there, and we solve the problem under these circumstances.

Now, notice that we agreed that setting $h(x) := \mathbb{P}_x(V_A < V_B)$, for $x \notin A \cup B$, solves the equation:

$$\boxed{h(x) = \sum_y P(x,y)h(y)} \quad \text{(harmonic equation)}$$

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 finally arrive at the conclusion of the theorem.

2 Pitman's version of Durett's Theorem

Assume P has $A \cup B$ as absorbing states. Assume further that $\mathbb{P}_x(\text{hit } A \cup B \text{ eventually}) = 1, \forall_{x \in S}$.
Then $h(x) := \mathbb{P}_x(\text{hit } A \text{ before } B)$ is the **UNIQUE** solution of $h = Ph$, where $h = \mathbb{1}_A$ is the indicator function 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.

Break time.

Notice that $h = Ph$ is a very special equation, as a solution to certain problems. In order to understand this equation, it is important to understand what is Pf for a function (column vector) f (assume nonnegative and bounded so that we can make sense of the summations). Then the action of a matrix on a column vector simply gives us:

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

summing over all x in the state space. P gives the probability distribution, and f simply gives the evaluations. Hence directly by our notation, we have:

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

Hence

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

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

$$\mathbb{E}[h(X_1) \mid X_0] = (Ph)(X_0)$$

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:

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

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

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

then we can write:

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

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

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

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

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

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

Hence

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

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

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

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

because $\mathbb{P}_x(\text{hit } A \cup B \text{ eventually}) = 1$ via our assumption.

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 then is:

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

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. That is, $h = Ph$. Hence:

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

We set the 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, passing through 0 and 1. Hence $h(x) = \frac{x}{N}$ as the unique solution to this system of equations. The theory is a bit more clever in that if we are certain to hit a boundary (as we have shown in lecture 5), then:

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

which Pitman notes is a famous result as due to Abraham deMoivre around 1730.

3.1 Now what about a biased coin?

Pitman asks what are the harmonic equations? We reason that this results in the same equations, with slight modifications:

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

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 Martingale at the end of the text, as aspects of a hitting-time problem (we will revisit this at the end of the course). The idea is to say that $h(x) = x$ is no longer harmonic when $q \neq p$ (biased coin). Now we believe that

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

We check this:

$$\begin{aligned} Ph(x) &= P\left(\frac{q}{p}\right)^{x+1} + q\left(\frac{q}{p}\right)^{x-1} \\ &= \left(\frac{q}{p}\right)^x \left[\cancel{p}\frac{q}{\cancel{p}} + \cancel{q}\frac{p}{\cancel{q}}\right] \\ &= \left(\frac{q}{p}\right)^x. \end{aligned}$$

Now Pitman notes this is a bit clever, but it is not a bad idea to try that the harmonic equation is a power. As soon as we have found this, we can play this game again.
Pitman concludes that by $h = P^n h$, we have:

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

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

$$h(x) = \left(\frac{q}{p}\right)^X = \left(\frac{q}{p}\right)^N \mathbb{P}_x(\text{hit } N) + \left(\frac{q}{p}\right)^0 \mathbb{P}_x(\text{hit } 0),$$

and additionally

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

Now we have two equations and two unknowns, and we can solve as normal.

Lecture ends here.

Stats 150, Fall 2019
Lecture 7, Thursday, 9/19/2019

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 $(X_0, X_1, X_2, \dots)$ for some random variable $Y = Y(X_0, X_1, X_2, \dots)$.
If we know $E_x Y$ and we want to compute the expectation of probability distribution $\lambda = \lambda(x), x \in S$, namely $\mathbb{E}_\lambda Y$, 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 some X_0. That is,

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

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

Example: Suppose we have a set of states A (we will make them absorbing as a matter of technique), and consider:

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

and we want to find: $\mathbb{E}_x V_A$ for any initial x.
Now if $x \in A$, then we trivially have:

$$\mathbb{E}_x V_A = 0.$$

If $x \notin A$, then we define a function for mean, say $m(x) = m_A(x) := \mathbb{E}_x V_A$, 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 $X_0 = x$ and $X_1 = y$, let $x \notin A$. Then:

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

Notice especially that this is correct if $y \in A$. If we happen to hit an absorbing state, then the second term $\mathbb{E}_y(V_A)$ is zero. Additionally, this is correct if $y \notin A$, where $\mathbb{E}_y V_A \geq 1$ (it would take at least one step).
This means that we can write down a system of equations, relating to the mean times (for $x \notin A$):

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

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

Pitman notes that in the text, there is a theorem that states that as long as we can reach the boundary from the interior (in some number of steps) with positive probability, this system of equations will have a unique solution. Pitman notes that we can simply check this computationally via matrices.

1.1 Application to Simple Symmetric Random Walks

This is just the usual Gambler's ruin for a fair coin. We start with x dollars and play for $\pm\$1$ gains with equal probability until we hit either 0 or some $\$N$. Last lecture, we showed:

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

Now set $A := \{0, N\}$ as our absorbing states, and V_A here will be the duration of the game, where

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

Recall that there remains the scenario of never hitting the boundary A, but we have already found before that the probability assigned to this enormously 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. We use this argument to form the geometric bound as we have before in a previous lecture.
It remains to solve $m(x) := \mathbb{E}_x V_A$. To find the equations, we first write out the boundary conditions. That is,

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

Now the nontrivial cases, we again break into two parts:

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

for $0 < x < N$. Then we solve for this system of equations. Pitman notes it's a good idea to recall that $h(x) = \frac{1}{2}h(x+1) + \frac{1}{2}h(x)$, which implies that $h(x)$ is linear (affine). Now writing in terms of placeholder constants a, b, we have:

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

Consider now, if we insert another term:

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

Then we observe:

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

$$= \boxed{c(x^2+1)}.$$

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

1.2 Summarizing our findings:

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

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 our function need to be linear), finding the exact solution is not too tricky.

Break time.

We went slowly to look at first passage times in a particular example, and Pitman notes this is simply find the matrix, write down the equations, and compute the equations. Now we would like to consider that X_1 may not be the only variable on which we would like to condition. There may be more clever techniques, where we will want to use our imagination to find a better variable on which to condition.
According to Pitman, often in a Markov chain, we can commonly try X_0, X_1, X_n. In particular, to derive the recurrence of a mean (as on our homework), we would want to use X_n.
Now we consider a more special example:

2 Independent Bernoulli (p) Trials

We want to find the mean time until we see N successes in a row, for example $N = 3$.
For example, we illustrate one sequence:

$$X_n : 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1$$

Let τ_N be the number of trials required. The distribution itself is tricky, but we want the expectation, which is relatively simple.
Of course,

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

where neither the simple probability (first equality) nor the tail sum (second) has a simple formula for our purposes.
Hence we ask, what should we condition on? We try τ_{N-1}, which is to have $N-1$ of 1s in a row.

$$\tau_N = \tau_{N-1} + \Delta_N,$$

where

$$\Delta_N = \begin{cases} 1 \text{ with probability } p \\ 1 + \text{ a copy of } \tau_N \text{ otherwise} \end{cases}$$

Now this leaves us depressed as we must now start all over if we fail the last required trial; however, this expression above may give us enough to solve the problem!
Let $\mu_N := \mathbb{E}\tau_N$. We have:

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

where rearranging gives:

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

We test this:

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

by the mean of geometric (tail sums). Similarly,

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

and we guess:

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

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 reconsider the two states for the final step. Now Pitman gives his own solution.

2.1 Pitman's Solution

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 $G_0 \sim$ Geometric(q), where q is the failure probability.
We want to find $\mathbb{E}(\tau_N)$ by some suitable conditioning (which requires artistic thinking). Here, we want to condition on G_0, as G_0 is closely related to τ_N. If $G_0 > N$, then $\tau_N = N$. On the other hand, if $G_0 = g \leq N$, then

$$\tau_N = g + \hat{\tau}_N,$$

where equality in distribution gives:

$$(\hat{\tau}_N \mid G_0 = g) \stackrel{d}{=} \tau_N.$$

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 us:

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

which is another equation satisfied by μ_N. Pitman leaves it to us as an exercise to algebraically show the equality of these results in general.
We look at a simple $N := 2$ case. Here in this solution, we have:

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

and we should check this more closely.

Lecture ends here.

Stats 150, Fall 2019
Lecture 8, Tuesday, 9/24/2019

1 §1.11 Infinite State Space

This is starred in the text but is not optional for our course. We will discuss techniques for both finite and infinite state spaces, especially

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

2 Review of Math Background

Know the following by heart (we'll need to use them on the midterm).

2.1 Binomial Theorem

This is to write:

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

We should observe that $\binom{n}{k} = \frac{n!}{(n-k)!k!} = \boxed{\frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1}}$, and it is important 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 $n \to r \in \mathbb{R}$.
Take r real and look at

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

if $|x| < 1$ for real or complex x. Notice that the conventional r 'choose' k makes sense particularly through the polynomial definition of the binomial factor $\binom{r}{k}$.
This is the instance with $f(x) = x^r$. Now if we want to consider:

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

which is our familiar Taylor expansion, for $|x| < R$, where R is our radius of convergence. Usually for our purposes, $R \geq 1$.
Now of course, recall

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

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

3 Probability Generating Functions

Suppose we have a nonnegative integer-valued random variable X, which for simplicity will have nonnegative integer values $X \in \{0, 1, 2, \dots\}$.
We define the PGF (probability generating function) of X to be the function

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

Pitman says that we usually take $0 \leq z \leq 1$. When discussing PGFs, we may push z beyond this but we will keep it within this bound. Now we try to write the above in terms of a power series. Recall that $\mathbb{E}g(X) = \sum_{n=0}^{\infty} \mathbb{P}(X = n)g(X)$, so

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

We worked with PGFs very briefly in a previous lecture, namely taking X uniform on $\{1, 2, 3, 4, 5, 6\}$, and we looked at:

$$\phi_X(z) = \frac{1}{6}\left(z + \cdots + z^6\right).$$

Recall this is where Pitman asked us to look this expansion up in Wolfram Alpha.
Notice that by convention, $0^0 = 1$, so $\phi_X(0) = \mathbb{P}(X = 0)$.
Now for a poisson 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

$$\mathbb{E}X = \frac{d}{dz}\phi_X(3)|_{z=1^-},$$

where we approach from the left if we need to be pedantic.
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:

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

Recall that $X_\lambda \sim \text{Poisson}(\lambda)$ if and only if:

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

which via the generating function implies:

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

Now we go back to our above to derive (or simply recall):

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

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:

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

where G denotes the geometric density function. Then

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

In summary, we can think of a probability generating function as a probability, and we only need that G_{1-z} is idependent 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 $\{0, 1, 2, \dots\}$. Then

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

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

$T_r = T_{r,p} =$ number of 0s before rth 1 in indep. Bernoulli(p) 0/1 trials.

Looking at iid copies of G_p we use generating functions:

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

where we simply plug into Newton's binomial formula. Notice that this actually is equal to:

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

where we define:

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

From 134, we know this to be the negative binomial distribution.

Break time.

4 Probability Generating Functions and Random Sums

Suppose we have $Y_1, Y_2, \ldots$ iid nonnegative 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 akk Y_i). Now consider another random variable, $X \geq 0$, integer valued, and look at: $Y_1 + Y_2 + \cdots + Y_X =: S_x$, 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 S_X below is annoying, so Pitman tells us that we should condition on this.

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

which is a composition of generating functions. In the middle line, Pitman notices 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 $Y_1, Y_2, \ldots$.

5 Application to Galton-Watson Branching Process

Assume that we're given some probability distribution (offspring distribution) $p_0, p_1, p_2, \ldots$. Start with some fixed number k of individuals in generation 0, where each of these k individuals has offspring with distribution according to X. Our common notation is

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

and so we have the following equality in distribution:

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

where the X_i are iid $\sim p$.
Continuing the problem, given $Z_0, Z_1, \ldots, Z_n$ with $Z_n = k$, and $Z_{n+1} \sim X_1 + \cdots + X_k$. 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, this is a Markov chain on $\{0, 1, 2, \ldots\}$. Pitman notes that there is a conspicuous detail, in that $k = 0$ is absorbing, which happens to fit with the convention of empty sums (that summing 0 copies of nothing gives nothing).
Now, we should expect that generating functions should be helpful, as we are iterating random sums. We'll iterate the composition of generating functions. For simplicity, start with $z_0 = 1$. Let $\phi_n(s) = \mathbb{E}\left(s^{Z_n}\right)$ for $0 \leq s \leq 1$. We see that

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

Hence

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

which we define as the **offspring generating function**. To find 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:

$$\begin{aligned} \mathbb{P}_1(\text{extinction}) &= \mathbb{P}_1(Z_n = 0 \text{ for large } n) \\ &= \lim_{n\to\infty} \mathbb{P}_1(Z_n = 0). \end{aligned}$$

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. Pitman gives that there is a very nice picture we can draw for intuition here. As an example, we sketch ($\phi(s)$ with respect to s) the generating function of Poisson(3/2). This gives a fixed point iteration returning

the unique root s of $s = \phi(s)$ with $s < 1$. We draw the special case where the mean is large than 1 ($\phi'(1) = \mu > 1$).
We see that if $p_0 > 0$ and $\mu := \sum_n np_n > 1$, then the probability generating function is a convex curve with slope > 1 at 1 and value $p_0 > 0$ at 0. Then by elementary analysis, we have that there is a unique root $0 < s < 1$ of $\phi(s) = s$, and

$$\phi(\phi(\phi(\cdots(0)))) \xrightarrow{n\to\infty} \text{ the unique root .}$$

Now, even if we aren't a fan of generating functions, we should note that they are inescapable in the solution to the branching extinction problem.
(By the way, there is a very annoying case for branching processes that we should not forget, which is where $\mathbb{P}(X = 1) = 1$, just makes the population stay at 1, and extinction probability is 1. There is no random fluctuation in it.)
Pitman notes that there is another interesting case, where $\mu := 1$ and $p_0 > 0$ (to avoid the above boring function). Then the only root returned from fixed point iteration is precisely at 1.