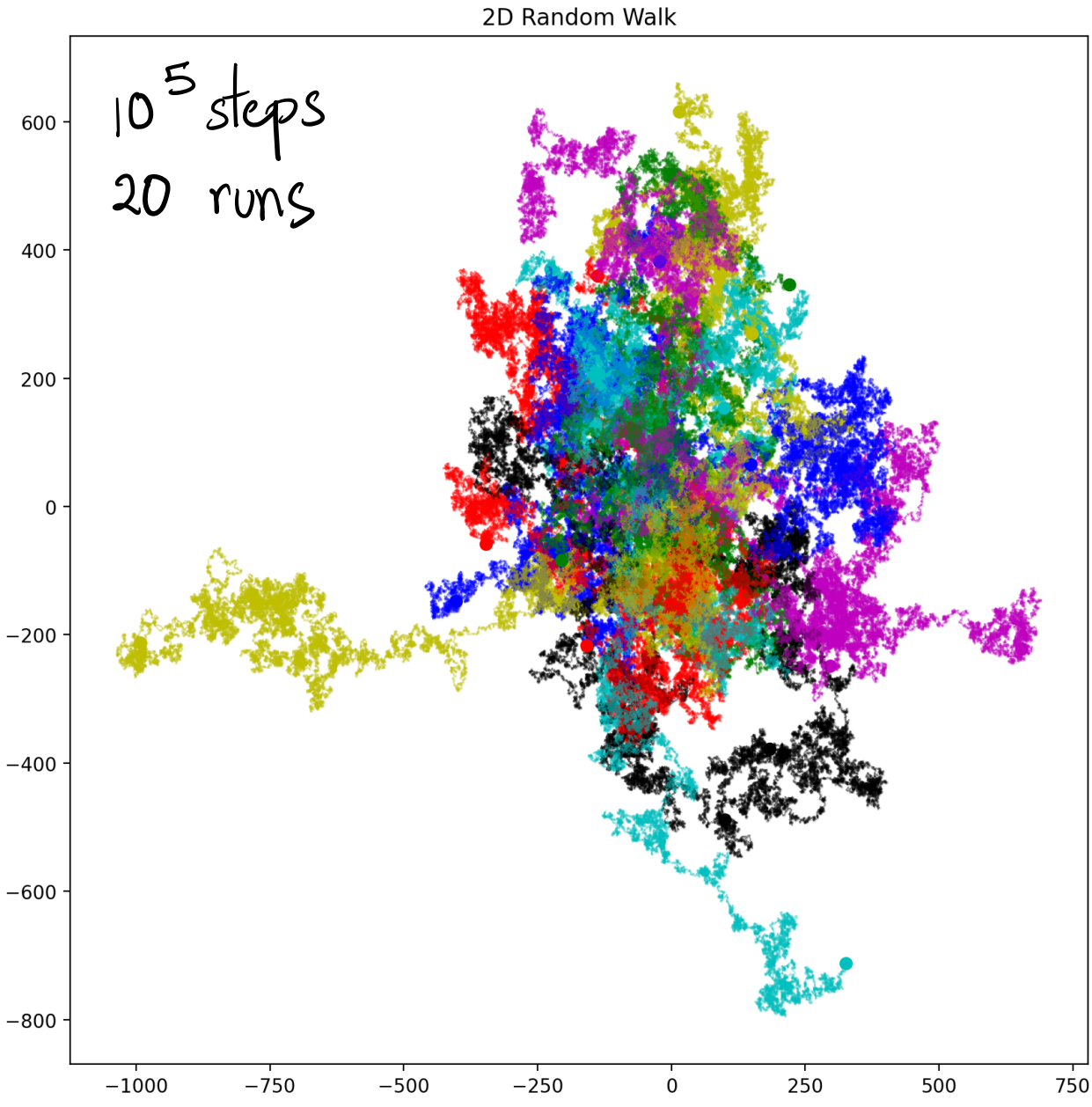
Semester: Winter 2023
Instructor: Paquette, Elliot

Stochastic Processes



Course Content. Markov chains. Random walks. Martingales.

Summary of Contents

Foreword and Acknowledgements

These notes are for a Honours BSc/MSc level course in probability, covering discrete time stochastic processes (especially Markov chains and martingales). The intent in creating these notes is to do martingale style analysis earlier, as it has a large and fruitful overlap with Markov chain theory, and moreover is an essential toolkit for dealing with inhomogeneous time processes and non-Markov processes.

Very little in this set of lecture notes is in any way original, although as the goal was to be self-contained, some material had to be represented in an original way. In no particular order, the textbooks which I consulted while preparing this material were:

1. Rick Durrett. *Probability: theory and examples*. Vol. 49. Cambridge university press, 2019. URL: https://services.math.duke.edu/~rtd/

2. Sheldon M Ross. *Stochastic processes*. John Wiley & Sons, 1995

3. Robert P Dobrow. *Introduction to stochastic processes with R*. John Wiley & Sons, 2016

4. David A Levin and Yuval Peres. *Markov chains and mixing times*. Vol. 107. American Mathematical Soc., 2017

5. Richard M Dudley. *Real analysis and probability*. CRC Press, 2018

My own experiences taking similar courses also largely influenced these notes: especially, I took similar courses with Christopher Hoffman and Eyal Lubetzky, and my notes and recollections from these courses played a big role.

I am also indebted to Jana Kurrek, who wrote class notes for Math 447, a majors-version of this course which I taught in the previous two years. Those notes were written in the same latex style as these, and Jana shared that latex with me. Some sections began with those notes and evolved from there. I'll add that I probably would not have written these notes, had it not been for the example Jana set.

Finally, these notes certainly are not perfect, and the class of 547 in Winter term 2023 helped improve them. Noah Marshall, who was TA for this course, especially contributed quite a few corrections.

Elliot Paquette
April 3, 2023

Background material

Probability formalism

This course is on stochastic process theory, which concerns, in brief, the (measure-theoretic) probability theory of sequences of random variables. In fact, the most fundamental sequence of random variables – the independent, identically distributed real valued random variables (*iid sequences*) – is usually exluded from this course. The properties of iid sequences are generally covered in a first semester course, such as MATH356/587.

In this section, we develop briefly the background material on which this course rests. First, we will always have in the background a probability triple $(\Omega, \mathscr{F}, \Pr)$, of a (hidden) state space Ω which is just some set, a σ-algebra $\mathscr{F}$, and a probability measure $\Pr$. In standard probability theory, we do not put any further assumptions on this probability space. The cost of this choice is that we need to enforce assumptions on the *random variables* that are defined on this probability space.

The most important random variables are the *real-valued random variables*, which are functions X from Ω to $\mathbb{R}$ with the property that $X^{-1}(E) \in \mathscr{F}$ for all Borel subsets E of $\mathbb{R}$. In this course, we will also want to deal with random variables living in other state spaces. The natural extension are *standard Borel* random variables. A measurable space $(S, \mathscr{A})$ is standard Borel if there exists a metric d on S which makes it into a complete separable metric space and so that $\mathscr{A}$ is the Borel σ-algebra generated by this metric.

A space S that satisfies this is called a *Polish space*.

Essentially every random variable we construct in this course will be standard Borel. A non-exhaustive list of such spaces are below:

1. Real-valued random variables, i.e. those mapping to $(\mathbb{R}, \mathscr{B})$ where $\mathscr{B}$ is the Borel σ-algebra on $\mathbb{R}$.
2. Countably-valued random variables, i.e. those mapping to $(S, 2^S)$ for a countable set S, and where 2^S is the power set.
3. Sequence spaces built over other Borel spaces, which is to say that $(X_j : j \in \mathbb{N})$ is a sequence of standard Borel random variables, we can consider the whole sequence $Y := (X_j : j \in \mathbb{N})$ *itself* as a random variable. It maps to the countable product space, equipped with the σ-algebra given by the product σ-algebra of the associated σ-algebras. This product σ-algebra turns out to be the Borel σ-algebra associated to the product space, which is itself again a Polish space.

The core background for this course from MATH356/587 are *(weak/strong) laws of large numbers* and the *central limit theorem*. All

of these, in their simplest form, concern sequences $(X_j : j \in \mathbb{N}_0)$ of iid real valued random variables. Each of these three theorems is equipped with a different notion of convergence of sequences of random variables, which will all play a role in this course.

We write $\mathbb{N}$ for the natural numbers $\{1, 2, 3, \cdots\}$ and $\mathbb{N}_0$ for the numbers $\{0, 1, 2, 3, \cdots\}$.

It will be convenient to generalize these convergences slightly to the standard Borel space setting. So we will suppose that $(X_j : j \in \mathbb{N}_0)$ are a sequences of random variables taking values in a standard Borel space $(S, \mathscr{A})$, with an associated metric d.

The simplest form of stochastic convergence is *in-probability* convergence, which we recall:

Definition 1 (In-probability convergence): The sequence $(X_j : j \in \mathbb{N}_0)$ converges in-probability to X_0 if for all $\epsilon > 0$

$$\lim_{j\to\infty} \Pr(d(X_j, X_0) > \epsilon) = 0,$$

in which case we write $X_j \xrightarrow[j\to\infty]{\Pr} X_0$.

The weak law of large numbers then asserts that the *time-average* of an iid sequence is its ensemble average (which is to say its expectation):

Theorem 1: Weak law of large numbers

Suppose $(X_j : j \in \mathbb{N})$ are iid real valued random variables, and suppose that $\mathbb{E}|X_1| < \infty$. Then

$$\frac{1}{n}\sum_{j=1}^{n} X_j \xrightarrow[n\to\infty]{\Pr} \mathbb{E}X_1.$$

In fact, under the assumptions given above, much more is true. With the same setup, we can replace the in-probability convergence with the stronger almost sure convergence.

Definition 2 (Almost sure convergence): The sequence $(X_j : j \in \mathbb{N}_0)$ converges almost surely to X_0 if

$$\Pr(\limsup_{j\to\infty} d(X_j, X_0) > 0) = 0,$$

in which case we write $X_j \xrightarrow[j\to\infty]{\text{a.s.}} X_0$.

Almost sure convergence implies in-probability convergence. Conversely, it is generally the case that in-probability convergence is strictly weaker than almost sure convergence.

In a finite probability space (i.e. where there is a finite set E such that $\Pr(E^c) = 0$), almost sure convergence and in-probability convergence are actually equivalent.

There is another partial converse of in-probability convergence and almost sure convergence: if $X_j \xrightarrow[j\to\infty]{\Pr} X_0$ then there is a (deterministic) subsequence j_n so that $X_{j_n} \xrightarrow[n\to\infty]{\text{a.s.}} X_0$.

The strong law of large numbers states:

Theorem 2: Strong law of large numbers

Suppose $(X_j : j \in \mathbb{N})$ are iid real valued random variables, and suppose that $\mathbb{E}|X_1| < \infty$. Then

$$\frac{1}{n}\sum_{j=1}^{n} X_j \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}X_1.$$

Remark. As formulated, the weak law of large numbers is a little sad (it's just worse!). There are other formulations of the weak law which make it more interesting. For example, under the assumption that $(X_j : j \in \mathbb{N}_0)$ have the same mean, satisfy $\sup_j \mathbb{E}X_j^2 < \infty$ and are uncorrelated, then the weak law still holds. Furthermore, in this case, this is just a straight application of Chebyshev's inequality, which is much simpler than the strong Strong law.

Finally, the other main result on iid sequences is the central limit theorem (CLT). The central limit theorem gives in a sense the "next order term" in the law of large numbers, which is to say it quantifies how close is $\frac{1}{n}\sum_{j=1}^{n} X_j$ to $\mathbb{E}X_1$ as a function of n, under the assumption $\mathbb{E}X_1^2 < \infty$.

To simply get the order, the variance of $Y_n := \frac{1}{n}\sum_{j=1}^{n} X_j - \mathbb{E}X_1$ is easily checked to be $1/n$. Hence from Chebyshev's inequality $\Pr(|Y_n|\sqrt{n} > t) \le \mathrm{Var}(X_1)/t^2$. This shows that, in order of magnitude, Y_n is $1/\sqrt{n}$.

However, unlike with the laws of large numbers, it is not possible to characterize this convergence as either weak or strong convergence. This leads to the final basic notion of convergence:

Definition 3 (Weak convergence): The sequence $(X_j : j \in \mathbb{N}_0)$ converges in law to X_0 if for all bounded continuous functions $\phi : S \to \mathbb{R}$

$$\lim_{j\to\infty} \mathbb{E}\phi(X_j) = \mathbb{E}\phi(X_0)$$

in which case we write $X_j \xrightarrow[j\to\infty]{\text{law}} X_0$.

This type of convergence has a long list of equivalent formulations (and an equally long list of equivalent names: weak convergence, weak-* convergence, and convergence in distribution are all common alternative names). For working with this type of convergence, it is convenient to be able to change between these different formulations, which goes by the name Portmanteau lemma:

Lemma 1 (Portmanteau Lemma): The following are equivalent for random variables $(X_j : j \in \mathbb{N}_0)$ on Polish space S:

1. The sequence converges in law: $X_j \xrightarrow[j\to\infty]{\text{law}} X_0$.

2. Let BL be all functions f from $S \to \mathbb{R}$ which are bounded above by 1 and which satisfy $|f(x) - f(y)| \leq d(x,y)$. Then

$$\sup_{\phi \in \mathrm{BL}} |\mathbb{E}\phi(X_j) - \mathbb{E}\phi(X_0)| \to 0.$$

   This also defines a metric for weak-convergence (the "bounded-Lipschitz" or "Dudley" metric).

3. For all open sets $A \subset S$,

$$\liminf_{j\to\infty} \Pr(X_j \in A) \geq \Pr(X_0 \in A).$$

4. For all sets $A \in \mathcal{A}$ for which $\Pr(X_0 \in \partial A) = 0$, where ∂A is the boundary of A,

$$\lim_{j\to\infty} \Pr(X_j \in A) = \Pr(X_0 \in A).$$

In the important case of real-valued random variables, we can add a few extra bulletpoints to this list, which are especially useful for the development of the theory of weak convergence.

Lemma 2 (Portmanteau Lemma, real case): The following are equivalent for real-valued random variables $(X_j : j \in \mathbb{N}_0)$ on Polish space S:

1. The sequence converges in law: $X_j \xrightarrow[j\to\infty]{\text{law}} X_0$.

2. For all $t \in \mathbb{R}$ so that $\Pr(X_0 = t) = 0$,

$$\lim_{j\to\infty} \Pr(X_j \leq t) = \Pr(X_0 \leq t).$$

   This can also be formulated as saying the distribution functions of X_j converge to the distribution function of X_0 at all its points of continuity.

3. The quantile functions $Q_j(p) := \inf\{x \in \mathbb{R} : p \leq \Pr(X_j \leq x)\}$ converge at all points of continuity.

4. The characteristic functions $\psi_j(\xi) = \mathbb{E}e^{i\xi X_j}$ converge pointwise, i.e.

$$\lim_{j\to\infty} \psi_j(\xi) = \psi_0(\xi) \quad \forall\, \xi \in \mathbb{R}.$$

   This also generalizes to $\mathbb{R}^d$–valued random variables $(X_j : j \in \mathbb{N}_0)$ via $\psi_j(\xi) = \mathbb{E}e^{i\langle \xi, X_j\rangle}$ for the real inner-product $\langle \cdot, \cdot \rangle$.

Finally, the central limit theorem states that the deviations in the strong law of large numbers, appropriately rescaled converge to a standard normal random variable.

Theorem 3: CLT

Suppose $(X_j : j \in \mathbb{N})$ are iid real valued random variables, and suppose that $\mathbb{E}|X_1|^2 < \infty$. Then

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} (X_j - \mathbb{E}X_1) \xrightarrow[n\to\infty]{\text{law}} \sqrt{\text{Var}(X_1)} N(0,1).$$

Conditioning

Conditioning is the action of changing a probability space by *revealing* part of the randomness. In the context of a stochastic process $(X_j : j \in \mathbb{N}_0)$, one can consider the index j as a measurement of time. In that way, at a time j, there are outcomes which have been observed $(X_1, X_2, \cdots, X_j)$ and there are outcomes which have not yet been observed $(X_k : k > j)$. In the case of iid sequences, the law of the future $(X_k : k > j)$ has no dependency on the outcomes $(X_1, X_2, \cdots, X_j)$ (hence the nomenclature independent). To move away from the case of independent sequences, we would instead like to have probability laws where in some reasonable way, the law of $(X_k : k > j)$ *can* depend on the outcomes of $(X_1, X_2, \cdots, X_j)$.

Conditioning is substantially simpler in discrete probability spaces, or similarly, when the random variables on which we condition take on finitely many values. This however will not be sufficient for what we need to do, and so we need to develop conditional expectation and conditional probability a little more generally. For concreteness, however, it is helpful to develop all the definitions in the case of discrete probability spaces.

The starting point is simply:

> **Definition 4 (Conditional Probability):** The conditional probability of A given B, defined for $\Pr(B) > 0$, is
>
> $$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Recall that this gives an intuitive way to define independence:

> **Definition 5 (Independence):** Events A and B are independent if $\Pr(A \cap B) = \Pr(A)\Pr(B)$. If $\Pr(B) > 0$, then this can be equivalently formulated as $\Pr(A \mid B) = \Pr(A)$.

The conditional probability $\Pr(\cdot \mid B)$ is *another* probability measure on the space $(\Omega, \mathscr{F})$, and hence it is possible to define expectations with respect to this measure. The conditional expectation $\mathbb{E}[X \mid B]$, can be uniquely defined by $\mathbb{E}[\mathbf{1}_A \mid B] = \Pr(A \mid B)$. [1]enerally, having defined expectations for indicator functions, one can extend the expectation *simple random variables*, which are finite linear combinations of indicator functions and then to general random variables by requiring that the expectation satisfies the monotone convergence theorem.

[1] G

Example 1: Dice roll

Let X be a random variable with $X \overset{\text{law}}{=} \text{Unif}(\{1,2,3,4,5,6\})$. Let B be the event $\{X \in \{4,5,6\}\}$ and let A be the event $\{X \in \{3,4\}\}$. Then

$$\Pr(A \mid B) = \Pr(X = 4)/\Pr(B) = 1/3.$$

Since $\Pr(A) = 1/3$, we even have A is independent of B. From linearity,

$$\mathbb{E}[X \mid B] = \sum_{j=1}^{6} j \Pr(X = j \mid B) = \sum_{j=4}^{6} j/3.$$

Now, we need to go beyond conditioning on events and condition on random variables. In the case that the random variable X has a finite number of outcomes, we can do this building on Definition .

Definition 6 (Conditioning on a simple RV): Suppose that X is a simple random variable (meaning there is a finite set U so that $\Pr(X \in U) = 1$), define for nonnegative random variables Y and for random variables Y with $\mathbb{E}|Y| < \infty$

$$\mathbb{E}[Y \mid X] = \sum_{u \in U} \mathbb{E}[Y \mid \{X = u\}] \mathbf{1}\{X = u\}.$$

This defines a *random probability measure* $\Pr(\cdot \mid X)$ by $\Pr(A \mid X) = \mathbb{E}[\mathbf{1}_A \mid X]$, which allows us to conceptually *do* probability theory, having revealed the outcome of the experiment (provided we can describe the whole family of laws $\{\Pr(\cdot \mid X = u) : u \in U\}$).

The conditional expectation $\mathbb{E}[\,\cdot \mid X]$ can be considered as a partial expectation, in which X is "revealed" and the remainder of the randomness is averaged over. This conditional expectation remains random, and if we take the expectation of it, we take the total expectation. This gives us the law of total expectation:

Example 2: Dice roll

Continuing with a dice roll X and B the event $X \geq 4$, let $Y = \mathbf{1}_B$.
Then

$$\Pr(X = 4 \mid Y) = (1/3) \cdot \mathbf{1}_B + 0 \cdot \mathbf{1}_{B^c}.$$

On the other hand with A the event $\{X \in \{3,4\}\}$.

$$\Pr(A \mid Y) = (1/3)\mathbf{1}_B + (1/3)\mathbf{1}_{B^c} = 1/3.$$

Theorem 4: Law of Total Expectation: discrete case

For nonnegative random variables Y and for random variables Y with $\mathbb{E}|Y| < \infty$,

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

Proof.

$$\begin{aligned}\mathbb{E}(\mathbb{E}[Y \mid X]) &= \sum_x \mathbb{E}[Y \mid X = x] \cdot \Pr(X = x) \\ &= \sum_x \mathbb{E}[Y \mathbf{1}_{X=x}] \\ &= \mathbb{E}\left[Y \sum_x \mathbf{1}_{X=x}\right] \\ &= \mathbb{E}(Y)\end{aligned}$$

□

More generally, we can consider iterated conditional expectations, in which we condition on partial information, and then take partial expectations revealing even more.

Theorem 5: Tower property of conditional expectation

For nonnegative random variables Y and for random variables Y with $\mathbb{E}|Y| < \infty$, and for random variables X, Z

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

Example 3: Dice roll

Continuing with a dice roll X, the event B that $X \geq 4$, and the event A that $X \in \{3,4\}$ let $Y = \mathbf{1}_B$ and $Z = \mathbf{1}_A$.
Then, partitioning the space into the various outcomes of (Y,Z),

$$\mathbb{E}(X \mid (Y,Z)) = 1.5\mathbf{1}_{\{1,2\}}(X) + 3\mathbf{1}_{\{3\}}(X) + 4\mathbf{1}_{\{4\}}(X) + 5.5\mathbf{1}_{\{5,6\}}(X).$$

Taking expectation over everything gives

$$\mathbb{E}[\mathbb{E}(X \mid (Y,Z))] = 1.5\tfrac{1}{3} + 3\tfrac{1}{6} + 4\tfrac{1}{6} + 5.5\tfrac{1}{3} = \mathbb{E}(X).$$

Taking conditional expecation

$$\mathbb{E}[\mathbb{E}(X \mid (Y,Z)) \mid Y] = 2\mathbf{1}_{\{1,2,3\}}(X) + 4\mathbf{1}_{\{4,5,6\}}(X).$$

Abstract conditional expectation

To generalize beyond conditioning on simple random variables, we need the notion of conditioning on a σ-algebra $\mathscr{G} \subset \mathscr{F}$. This will be a direct generalization of the conditioning considered above by taking $\mathscr{G} = \sigma(X)$.

Definition 7 (Abstract conditional expectation): Let Y be either a non-negative random variable Y (or a random variable Y with

$\mathbb{E}|Y| < \infty$). For a σ-algebra $\mathcal{G} \subset \mathcal{F}$, the conditional expectation of Y, $\mathbb{E}[Y \mid \mathcal{G}]$ is a random variable that satisfies

1. $\mathbb{E}[Y \mid \mathcal{G}]$ is $\mathcal{G}$–measurable.
2. For any event $G \in \mathcal{G}$

$$\mathbb{E}[\mathbf{1}_G Y] = \mathbb{E}[\mathbf{1}_G \mathbb{E}(Y \mid \mathcal{G})].$$

In the case that $\mathcal{G} = \sigma(X)$, we write $\mathbb{E}[Y \mid X] := \mathbb{E}[Y \mid \mathcal{G}]$.

An important piece of context, which helps justify the notation that $\mathbb{E}[Y \mid X] = \mathbb{E}[Y \mid \sigma(X)]$, is a structure theorem for $\sigma(X)$–measurable random variables:

Lemma 3: Suppose X is a random variable taking values in $(S, \mathcal{A})$. If Y is a real valued random variable and Y is $\sigma(X)$–measurable, then there is measurable function $h : S \to \mathbb{R}$ so that $Y = h(X)$ almost surely.

Hence, the conditional expectation $\mathbb{E}[Y \mid X]$ is a function of X.

Conditional expectation exists and is unique (see for example Durrett, *Probability: theory and examples*[Chapter 4]):

Theorem 6: Uniqueness of CE

Conditional expectation exists and is unique in the following sense: if X_1 and X_2 both satisfy the definition of conditional expectation, then $X_1 = X_2$ a. s.

Since in fact there can be multiple nonequal random variables which satisfy the definition of conditional expectation, we say any random variable that safisfies Definition 7 is a *version* of the conditional expectation.

Example 4: Discrete case

In the case that X takes values in a finite set U and $\mathcal{G} = \sigma(X)$ we can check that Definition 6 gives a consistent answer with Definition 7, and hence gives an explicit construction of the conditional expectation.
From Definition 6,

$$\mathbb{E}[Y|X] = \sum_{u \in U} \mathbb{E}[Y \mid \{X = u\}]\mathbf{1}\{X = u\} =: Y_1.$$

Then this random variable Y_1 is $\mathcal{G}$–measurable. Any event $G \in \mathcal{G}$ can be expressed as $\{X \in E\}$ for some subset $E \subseteq U$.

Then

$$\begin{aligned}\mathbb{E}[\mathbf{1}_G Y] &= \sum_{u\in E} \mathbb{E}[Y\mathbf{1}\{X=u\}] \\ &= \mathbb{E}\Big(\sum_{u\in U} \mathbb{E}[Y \mid \{X=u\}]\mathbf{1}\{X=u\}\,\mathbf{1}_G\Big) \\ &= \mathbb{E}[\mathbf{1}_G Y_1].\end{aligned}$$

In practice, finding the conditional expectation of a random variable has no simple recipe, but there are two other essential cases where conditional expectation can actually be computed. Verifying that something is the conditional expectatiotn is relatively simple, as conditional expectation is unique, and it just needs to satisfy Definition 7.

Lemma 4: Suppose that (X, Y) are independent random variables on some space $(S, \mathscr{A})$. Let $h : S^2 \to \mathbb{R}$ be a bounded measurable map:

$$\mathbb{E}[h(X,Y) \mid X] = \int_S h(X,y)\,\mathrm{Pr}^Y(dy).$$

Here Pr^Y is the law of Y.

Proof. This is a direct application of Fubini's theorem, and the definition of conditional expectation. □

The second case is when one has joint densities:

Lemma 5: Suppose that (X, Y) is a random vector in $\mathbb{R}^\ell \times \mathbb{R}^d$ having a joint density f with respect to Lebesgue measure. Then conditionally on X, Y has a density on $\mathbb{R}^d$, $f_{Y|X}$ given by

$$f_{Y|X}(y) = \frac{f(y,X)}{\int_{R^\ell} f(y,X)\,dy}.$$

Moreover, for bounded measurable $\psi : \mathbb{R}^\ell \to \mathbb{R}$

$$\mathbb{E}[\psi(Y) \mid X] = \int_{\mathbb{R}^\ell} \psi(y) f(y,X)\,dy.$$

This is again Fubini's theorem.

Besides these cases where it is easy to find the conditional expectation, there are a few general properties worth recording about conditional expectation. First, the law of total expectation generalizes:

Theorem 7: Law of Total Expectation

For nonnegative random variables Y and for random variables Y with $\mathbb{E}|Y| < \infty$ and for a sub-σ-algebra $\mathscr{G} \subseteq \mathscr{F}$

$$\mathbb{E}(\mathbb{E}[Y \mid \mathscr{G}]) = \mathbb{E}Y.$$

Proof. Use the definition of conditional expectation with $G = \Omega$. □

Going a step further, the nesting property generalizes:

Theorem 8: Tower property

For nonnegative random variables Y and for random variables Y with $\mathbb{E}|Y| < \infty$, and for two σ-algebras $\mathscr{H} \subseteq \mathscr{G}\mathscr{F}$,

$$\mathbb{E}(\mathbb{E}[Y \mid \mathscr{H}] \mid \mathscr{G}) = \mathbb{E}[Y \mid \mathscr{H}] = \mathbb{E}(\mathbb{E}[Y \mid \mathscr{G}] \mid \mathscr{H})$$

This follows with a similar proof. Note the smaller σ-algebra always wins, or, said differently, having averaged over more randomness, it cannot be undone.

We also summarize two other important special, trivial cases where the conditional expectation is trivial

Theorem 9: S

ppose Y is either nonnegative or Y has $\mathbb{E}|Y| < \infty$ and suppose $\mathscr{G}$ is a sub-σ-algebra $\mathscr{G} \subseteq \mathscr{F}$.

1. If Y is independent of $\mathscr{G}$ then $\mathbb{E}(Y \mid \mathscr{G}) = \mathbb{E}(Y)$.
2. If Y is $\mathscr{G}$–measurable, $\mathbb{E}(Y \mid \mathscr{G}) = Y$.

Regular conditional probability

Conditional expectation has many useful properties, but it does not quite function the way that we would like to do conditioning, which is to say that we "freeze" some random variables and then work on a probability space that depends on those frozen variables. (In fact, in all 3 of the 3 easy examples Lemma 5, Lemma 4 and Theorem 4, we actually did construct a random probability measure). This extension of conditional expectation is called a regular conditional probability law:

> **Definition 8 (Regular conditional law):** Let $\mathscr{G} \subseteq \mathscr{F}$ be a sub-σ-algebra and let Y be a random variable taking values in $(S, \mathscr{A})$. A regular conditional probability law $\mathcal{P} : (\mathscr{A} \times \Omega) \to [0,1]$ is a function so that

1. For each $B \in \mathscr{A}$, $\mathcal{P}(B, \cdot)$ is a version of the conditional expectation $\mathbb{E}(\mathbf{1}_B(Y) \mid \mathscr{G})$.
2. There is a $\mathscr{G}$–measurable set E having $\Pr(E) = 1$ so that for every $\omega \in E$, $\mathcal{P}(\cdot, \omega)$ is a probability measure.

In other words, a regular conditional probability law allows us to do conditioning in the intuitive way – first conditioning on part of the probability space and then working with a new "random" probability law of some random variable.

A little bit of care is needed: regular conditional probability laws do not always exist. However, when Y is standard Borel, they do:

Theorem 10: Regular conditional probabilities

If Y is standard Borel, and $\mathscr{G}$ is a sub-σ-algebra, a regular conditional probability law $\Pr^{Y|\mathscr{G}}$ exists.

As a consequence, all the properties of expectations transfer to conditional expectations.

Theorem 11: Conditional expectation props

The following general properties of conditional expectations hold:

1. If X, Y are real valued random variables and $X \leq Y$ almost surely, then
$$\mathbb{E}[X \mid \mathscr{G}] \leq \mathbb{E}[Y \mid \mathscr{G}] \quad \text{a. s.}$$
If further $X = 0$ a. s. and $\mathbb{E}[Y \mid \mathscr{G}] = 0$ then $Y = 0$ a. s.
2. (Monotone convergence) If $(X_j : j \in \mathbb{N})$ are real-valued random variables and $0 \leq X_j \leq X_{j+1}$ for $j \geq 1$ then
$$\lim_{j\to\infty} \mathbb{E}[X_j \mid \mathscr{G}] = \mathbb{E}[\lim_{j\to\infty} X_j \mid \mathscr{G}] \quad \text{a. s.}$$
Dominated convergence and Fatou's lemma also follow.
3. (Jensen's inequality) If $\varphi : \mathbb{R} \to \mathbb{R}$ is convex, $\mathbb{E}|Y| < \infty$,
$$\varphi(\mathbb{E}[Y \mid \mathscr{G}]) \leq \mathbb{E}[\varphi(Y) \mid \mathscr{G}] \quad \text{a. s.}$$

Proof. Using the existence of the regular conditional probability law, (which exists for single rvrvs Y, pairs of rvrvs (X, Y), or sequences $(X_j : j \in \mathbb{N})$), we simply apply the associated statement for the deterministic expectation. □

Markov chains

Stochastic processes

We start by setting some nomenclature about (discrete time) stochastic processes.

> **Definition 9 (Stochastic process):** A *stochastic process* $(X_j : j \geq j_0)$ is a sequence of random variables taking values in a *state space* $(S, \mathscr{A})$, which we will take to be a standard Borel space.

We refer to the indexing sequence $(j \in \mathbb{Z} : j \geq j_0)$ as time and j_0 is the initial time. Where it is not important, we will just take the initial time $j_0 = 0$. Stochastic processes are natural frameworks for prediction and and uncertainty. They could describe a natural process, such as the state of a physical object (such as a coin or a dice) or system. It could represent a financial asset, such as a stock price. It could be the state of a stochastic algorithm or an algorithm that evolves in a random environment.

[2]

Hence at a given time j, the process has a *present* state X_j. It also has a *past* $(X_k : j_0 \leq k \leq j)$ and a *future* $(X_k : k > j)$. It will be helpful to be able to condition on the history $(X_k : j_0 \leq k \leq j)$ and to discuss the probability distribution of the future. So, we define $\mathscr{F}_j = \sigma(X_k : j_0 \leq k \leq j)$, which is informally all the information that can be learnt from the history of the process.

The sequence $(\mathscr{F}_j : j \geq 0)$ is a naturally increasing in j, in that $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \cdots$. Such an increasing sequence is just referred to as a *filtration*:

> **Definition 10:** A *filtration* $(\mathscr{F}_j : j \geq j_0)$ is a sequence of σ-algebras with the property that they are increasing, so for all $j \geq j_0$, $\mathscr{F}_j \subseteq \mathscr{F}_{j+1}$. A stochastic process $(X_j : j \geq j_0)$ is *adapted* to a filtration if X_j is $\mathscr{F}_j$–measurable for all $j \geq j_0$. Any stochastic process also gives rise to a filtration, its *natural filtration*, just by setting $\mathscr{F}_j = \sigma(X_k : j_0 \leq k \leq j)$.

There are a few measure-theoretic aspects of stochastic processes, which are helpful to understand for proofs Stochastic process takes values in $(S^\infty, \otimes_1^\infty \mathscr{A})$, which remains a standard Borel space. Hence, in complete generality (provided the state space $(S, \mathscr{A})$ is standard Borel), the future $(X_k : k > j)$ is a standard Borel random variable. Hence by the existence of regular conditional probability laws, conditionally on the past $\mathscr{F}_j$, there is a probability law describing its future. Furtheremore, the product σ-algebra $\otimes_1^\infty \mathscr{A}$ is generated

[2] We may also wish to have the indexing sequence be finite. Formally, we can embed finite chains $(X_j : j_0 \leq j \leq n)$ into infinite chains by just taking $X_k = X_n$ for all $k > n$, and in this way assume wlog that all stochastic processes we consider have infinite time horizons.

by cylinder sets, which depend on only finitely many coordinates. Hence, we have

Theorem 12: FD Marginals

The law of a stochastic process $(X_j : j \geq j_0)$ is determined by its *finite-dimensional marginals* meaning the (infinite family) of laws of the finite–dimensional vectors $(X_j : k \geq j \geq j_0)$ where k runs over all $\mathbb{N}$.

The Markov property

A Markov chain is a stochastic process $(X_j : j \geq j_0)$ that restricts the amount of dependence the law of the future can have on the past. Specifically, it satisfies the *Markov property*:

Definition 11 (Markov property): A stochastic process $(X_j : j \geq j_0)$ satisfies the *Markov property* if for $j \geq j_0$, the law of X_{j+1} given $\mathscr{F}_j$ equals the law of X_{j+1} given X_j, almost surely. If a stochastic process has the Markov property, it is called a *Markov chain*.

In a Markov chain, there are therefore conditional probability laws $\Pr^{X_{j+1}|X_j}$, describing the law of the next step of the Markov chain given the present step. Moreover, it turns out that to define these Markov chains, and to work with them, we just need to define these conditional probabilitiy laws. So, we define:

Definition 12: A *Markov kernel* $K : S \times \mathscr{A} \to \mathbb{R}$ is a function that satisfies:

1. For every $x \in S$, $K(x, \cdot)$ is a probability measure.
2. For every $A \in \mathscr{A}$, $K(\cdot, A)$ is measurable.

The Markov kererne encodes precisely the same data as the regular conditional probability law, which is to say that there is a kernel K_j so that $\Pr^{X_{j+1}|X_j} = K_j(X_j, X_{j+1})$ a. s. The kernels may depend on time, in which case the Markov chain is time inhomogeneous. However, the more important case is when the all the kernels are the same:

Definition 13: A Markov chain $(X_j : j \geq j_0)$ is time homogeneous if there is a single Markov kernel K so that for all $j \geq j_0$ and all $A \in \mathscr{A}$

$$\Pr[X_{j+1} \in A \mid \mathscr{F}_j] = K(X_j, A) \text{ a. s.}$$

Hence for a time homogeneous Markov chain, its probability law

is completely determined by the law of its initial state and its Markov kernel. It is frequently helpful to change the law of the initial state, or moreover to consider the law of the Markov chain started from a fixed initial condition. So we set:

> **Definition 14:** In the special case that Pr is simply the law of some Markov chain $(X_j : j \geq j_0)$, we use the notation $\Pr_x$ (for a state $x \in S$) to refer to the law of the Markov chain with the same Markov kernels but with $X_{j_0} = x$.

This notation is a helpful shorthand for switching between different starting conditions. Note that this notation is a little bit dangeroous in the case that multiple Markov chains are in consideration, or there is additional randomness in play.

Example 5: Random walk

One of the most fundamental Markov chains, this is the process of partial sums of independent random variables. That is, suppose that $(X_j : j \geq 1)$ are independent real valued random variables (or $\mathbb{R}^d$–valued, or even taking values in some group). Now, define $S_j = \sum_{k=1}^{j} X_k$. Then $(S_j : j \geq 0)$ is a Markov chain. If $(X_j : j \geq 1)$ are identically distributed, then this is a time-homogeneous Markov chain.

Example 6: Sampling without replacement

Suppose that S is a finite set of size n. Define a sequence $(X_j : 1 \leq j \leq n)$ by letting $X_1 \overset{\text{law}}{=} \text{Unif}(S)$ and then inductively letting X_{j+1} be sampled uniformly from all elements of S not yet chosen in the set $\{X_1, X_2, \cdots, X_j\}$.
Then the sequence $(X_j : 1 \leq j \leq n)$ is *not* a Markov chain, as the law of X_{j+1} does not just depend on X_j but on the entirety of the history of the process.
On the other hand, if we set $A_j = \{X_1, X_2, \cdots, X_j\}$ for all $1 \leq j \leq n$, then sequence $(A_j : 1 \leq j \leq n)$ *is* a Markov chain on 2^S. Morevoer, it is even time-homogeneous, in that we can use the Markov kernel

$$K(A, \{B\}) = \begin{cases} \frac{1}{n-|A|} & \text{if } |B \setminus A| = 1, \\ 1 & \text{if } A = B = S, \\ 0 & \text{else.} \end{cases}$$

This can then be extended uniquely to a Markov kernel.

Example 7: Discrete Ornstein-Uhlenbeck process

Suppose that $(Z_j : j \geq 1)$ are independent identically distributed real valued random variables, and let $\alpha \in (0,1)$ be fixed. Let X_0 have any real-valued distribution. Define inductively $X_{j+1} = \sqrt{1-\alpha}X_j + \sqrt{\alpha}Z_{j+1}$. Then $(X_j : j \geq 0)$ is a Markov chain.

As a stochastic process is determined by its finite dimensional marginals (Theorem 12), it is helpful to be able to describe the finite dimensional marginals of a Markov chain.

Theorem 13: Chapman Kolmogorov

A stochastic process $(X_j : j \geq 0)$ is a Markov chain if and only if there are are Markov kernels $\{K_j : j \geq 0\}$ and an initial law μ so that for any $k \in \mathbb{N}$ and any $\{E_j \in \mathscr{A}\}$ for $j \geq 0$

$$\Pr(\cap_{j=j_0}^{k}\{X_j \in E_j\}) = \int_{E_0} \cdots \int_{E_k} K_{k-1}(x_{k-1}, dx_k)K_{k-2}(x_{k-2}, dx_{k-1}) \cdots K_0(x_0, dx_1)\mu(dx_0).$$

The proof in one direction is just induction and the Markov property. In the other direction, from Theorem 12, the finite-dimensional marginals determine the law and moreover it should just be checked that the Markov property follows from the claim.

Exercise 1 (Markov): Show that for a stochastic process $(X_j : j \geq 0)$ on a standard Borel space, the following are equivalent characterizations of a Markov chain:

1. For any $j \geq 0$, the law of the future $(X_k : k > j)$ conditioned on the past $(X_k : k \leq j)$ is the same as the law of the future conditioned on the present X_j.
2. For any $j \geq 0$, conditioned on the present, the future and the past are independent.

THF/CS Markov chains

The case of time homogeneous Markov chains with a finite state space *THFS* are especially important, and they are a good starting place for developing the theory of Markov chains. Some of this also extends more generally to time homogeneous *countable* state Markov chains (*TFCS*). In these cases the Markov kernel K can be encoded

entirely in a square matrix (which will be infinite in the countably infinite case, but otherwise have the dimension of the cardinality of the state space)

Definition 15: The *transition probability matrix* or *tpm* $\boldsymbol{P}$ indexed by the elements of S) of a THFS $(X_j : j \geq j_0)$ is given by

$$\boldsymbol{P}_{a,b} = \Pr(X_{j+1} = b \mid X_j = a).$$

The transition probability matrix $\boldsymbol{P}$ has some structure, owing to the fact that it must be a probability. These properties together define:

Definition 16: A matrix $\boldsymbol{P}$ indexed by the elements of S is called *stochastic* if:

1. All entries are non-negative.
2. The row-sums of the matrix are all 1.

Example 8: Lazy coin

The state space $S = \{H, T\}$ is a two-element space. The simplest probabilistic model for a sequence of coin flips is to take $(X_j : j \geq 0)$ iid Unif$(\{H, T\})$. However, if you actually ask a group of 100 students to do this, you observe the following. The first time people flip the coins, it is generally pretty close to uniformly distributed. However, if you direct them to hold then coin, and (without turning it over) flip it again, you find that a substantial majority get a different side the second time than the first. (For a well-researched contrasting point of view, see 3[o]).
You can model this with a Markov chain having tpm:

$$\boldsymbol{P} = \begin{array}{c|cc|c} & H & T & \\ \hline & 0.49 & 0.51 & H \\ & 0.51 & 0.49 & T \end{array}$$

With the transition matrix, we can turn probabilistic questions into simple matrix computations.

Definition 17 (Row vector formalism): For Markov chains, it is convenient to identify the pmfs of random variables on the state space S with row vectors of non-negative numbers summing to 1. We will overload the notation, relying on context whether a law on S is being treated as a measure, a row vector

[3] See Persi Diaconis, Susan Holmes, and Richard Montgomery. "Dynamical Bias in the Coin Toss". In: *SIAM Review* 49.2 (2007), pp. 211–235. DOI: 10.1137/S0036144504446436. eprint: https://doi.org/10.1137/S0036144504446436. URL: https://doi.org/10.1137/S0036144504446436

or a pmf. For a Markov chain $(X_j : j \geq j_0)$ its *initial distribution* is the law of X_{j_0}.

Theorem 14: Matrix formalism

We formulate the below with $j_0 = 0$ without loss of generality, and we set μ to be the initial distribution of a THFS Markov chain $(X_j : j \geq 0)$ having tpm $\boldsymbol{P}$:

1. (Chapman-Kolomogorov) For any $n \in \mathbb{N}$ and any states $a_j \in S$ for $0 \leq j \leq n$

$$\Pr(X_j = a_j, \quad \forall\, 0 \leq j \leq n) = \mu_{a_0} \boldsymbol{P}_{a_0,a_1} \boldsymbol{P}_{a_1,a_2} \cdots \boldsymbol{P}_{a_{n-1},a_n}.$$

2. (n-step transitions) For any $n \in \mathbb{N}$, the process $(X_{jn} : j \geq 0)$ is again a Markov chain on S, and its tpm is given by $\boldsymbol{P}^n$ (meaning matrix multiplication).

3. (marginal law) For any $j \in \mathbb{N}$, the law of X_j is given by $\mu \boldsymbol{P}^j$, meaning the vector matrix multiplication.

Proof. For the first claim, by Bayes' law

$$\begin{aligned}\Pr(X_j = a_j, \quad \forall\, 0 \leq j \leq n) = \Pr(X_n = a_n \mid X_j = a_j, \quad \forall\, 0 \leq j \leq n-1) \\ \times \Pr(X_j = a_j, \quad \forall\, 0 \leq j \leq n-1).\end{aligned}$$

By the Markov property

$$\Pr(X_n = a_n \mid X_j = a_j, \quad \forall\, 0 \leq j \leq n-1) = \Pr(X_n = a_n \mid X_{n-1} = a_{n-1}).$$

Finally by the definition of the tpm, we conclude

$$\Pr(X_j = a_j, \quad \forall\, 0 \leq j \leq n) = \boldsymbol{P}_{a_{n-1},a_n} \Pr(X_j = a_j, \quad \forall\, 0 \leq j \leq n-1).$$

The proof now follows by induction.

For the second claim, on applying the first claim and summing over all possible intermediate states

$$\Pr(X_n = a_n \mid X_0 = a_0) = \sum_{(a_j)} \boldsymbol{P}_{a_0,a_1} \boldsymbol{P}_{a_1,a_2} \cdots \boldsymbol{P}_{a_{n-1},a_n}.$$

By the definition of matrix multiplication, we conclude

$$\Pr(X_n = a_n \mid X_0 = a_0) = \boldsymbol{P}^n_{a_0,a_n}.$$

For the final claim, this follows similarly from the first claim. □

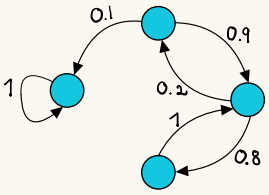
We can also give a graphical representation of THFS Markov chains, which can be helpful in understanding the behavior of small chains.

Definition 18 (Transition graph): A *transition graph*, associated to stochastic matrix $\boldsymbol{P}$ is a directed graph having vertex set S and edge set

$$\{(ab) : \boldsymbol{P}_{a,b} > 0\}.$$

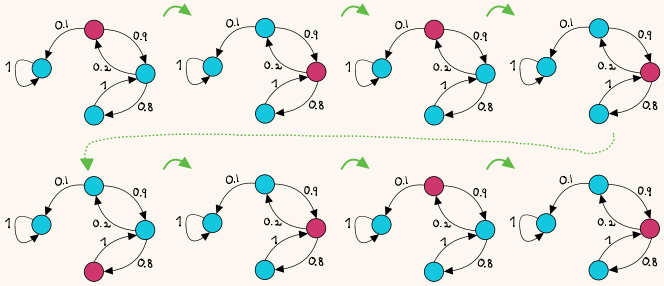
We further weight these edges by the value of $\boldsymbol{P}_{a,b}$.

Example 9: Simple transition graph



$$\boldsymbol{P} = \begin{array}{c|cccc} & L & U & R & D \\ \hline L & 1.0 & 0 & 0 & 0 \\ U & 0.1 & 0 & 0.9 & 0 \\ R & 0 & 0.2 & 0 & 0.8 \\ D & 0 & 0 & 1.0 & 0 \end{array}$$

The states in the graph are labeled L, U, R, D for left, up, right, down. One can visualize the state of a Markov chain as a sequence of transitions on the transition graph:



The highlighted states represent $(X_0, X_1, X_2, \cdots, X_7)$.

Stopping times and the Strong Markov property

A fundamental tool for the analysis of Markov chains, but also for all stochastic processes, is the idea of a stopping time.

Definition 19: For a stochastic process $(X_j : j \geq j_0)$ adapted to filtration $(\mathscr{F}_j : j \geq j_0)$ (such as the natural filtration generated by the stochastic proceses $(X_j : j \geq j_0)$), a *stopping time* τ is a random variable taking values in $\{j \in \mathbb{Z} : j \geq j_0\} \cup \{\infty\}$ with the property that for all $j \geq j_0$ the event $\{\tau = j\}$ is in $\mathscr{F}_j$.

Informally, a random variable τ is a stopping time if we can tell if τ has happened with the information available so far. Even more informally, a criterion that helps you get off the bus at the right time in a strange place is a stopping time – you can't choose the last stop

before it gets sketchy, if you don't know when it gets sketchy.

The most important example of a stopping time is the following:

Definition 20 (Hitting Time): Let $(X_j : j \geq 0)$ be a stochastic process. The *hitting time* or ("first passage time") of the $A \subseteq S$ is

$$\tau_A = \inf\{j \geq 0 : X_j \in A\}.$$

It is sometimes helpful to discard the case that $\tau_A = 0$, and so we also define

$$\tau_A^+ = \inf\{j > 0 : X_j \in A\}.$$

In the case that $X_0 \in A$, this is called the first return time of the process to A.

This is also the prototype of how to define a stopping time: it is the first time something happens (in contrast, say, to the last time something happens).

Exercise 2 (Stopping times): Show that the maximum and minimum of two stopping times is again a stopping time.

Time homogeneous Markov chains are probabilistic state machines: the law of their future depends only their current state, and neither how they got there (their past) nor even *how long it took* to get there (since their law has no time dependence). So, if you run a time homogeneous Markov chain up to the hitting time τ_x for some $x \in S$, the law of the process $(X_{k+\tau_x} : k \geq 0)$ should again be a Markov chain started from x. This generalizes to any stopping time, and is the content of the Strong Markov property.

Theorem 15: Strong Markov Property

Suppose that $(X_j : j \geq j_0)$ is a time homogeneous Markov chain adapted to a filtration $(\mathscr{F}_j : j \geq j_0)$ and that τ is a stopping time. Then conditioned on $\{\tau < \infty\}$, the law of $(X_{k+\tau} : k \geq 0)$ is again a time homogeneous Markov chain with initial distribution given by the law of X_τ under the conditional probability $\Pr(\cdot \mid \tau < \infty)$ and with the same markov Kernel as $(X_j : j \geq j_0)$.

Proof. It suffices to check the finite dimensional marginals of the process $(Y_k : k \geq 0)$ where $Y_k = X_{k+\tau}$ (on the event $\tau < \infty$), which is to say we should verify Theorem 13 holds. So let $\{E_j \in \mathscr{A} : j \geq j_0\}$ be some events and let $k \in \mathbb{N}$ be fixed. The key idea is to decompose

$\{\tau < \infty\} = \cup_{\ell=j_0}^{\infty}\{\tau = \ell\}$. Then on the event $\tau = \ell$, using Theorem 13

$$\begin{aligned}
&\Pr(\cap_{j=0}^{k}\{Y_j \in E_j\} \cap \{\tau = \ell\}) \\
&= \Pr(\cap_{j=0}^{k}\{X_{\ell+j} \in E_j\} \cap \{\tau = \ell\}) \\
&= \int_{E'} \cdots \int_{E_k} K(x_{k-1}, dx_k)K(x_{k-2}, dx_{k-1}) \cdots K(x_0, dx_1)\Pr(dx_0).
\end{aligned}$$

The event $E' = \{X_\ell \in E_0\} \cap \{\tau = \ell\}$. Define a measure μ on $\mathscr{A}$ by

$$\mu(E_0) = \sum_{\ell=j_0}^{\infty} \Pr(\{X_\ell \in E_0\} \cap \{\tau = \ell\}).$$

Summing over all ℓ and using monotone convergence, we have

$$\begin{aligned}
&\Pr(\cap_{j=0}^{k}\{Y_j \in E_j\} \cap \{\tau < \infty\}) \\
&= \int_{E_0} \cdots \int_{E_k} K(x_{k-1}, dx_k)K(x_{k-2}, dx_{k-1}) \cdots K(x_0, dx_1)\mu(dx_0).
\end{aligned}$$

Now observe that $\mu(\cdot)/\Pr(\tau < \infty)$ is nothing but the law of X_τ under the conditional measure $\Pr(\cdot \mid \tau < \infty)$, and so we have shown the claim. □

Classification of states of a THCS chain

Throughout this section, we suppose S is a countable set, and $\boldsymbol{P}$ is a transition probability matrix. The properties we develop here do not depend on the initial distribution.

Definition 21 (Communication): Say that a state j is *accessible* from a state i if there exists an $m \in \mathbb{N}$ such that $\boldsymbol{P}_{i,j}^m > 0$. Say that two states $i, j \in S$ of a THCS Markov chain communicate, written $i \leftrightarrow j$, if both are accessible from one another, i.e. if there exist $m, n \in \mathbb{N}$ such that,

$$\boldsymbol{P}_{i,j}^m > 0 \text{ and } \boldsymbol{P}_{j,i}^n > 0.$$

Equivalently, two states communicate if and only if each state has a positive probability of eventually being reached by a chain starting in the other state.

Theorem 16: Communication

The relation $\leftrightarrow$ is an equivalence relation on the state space.

Proof. The relation $\leftrightarrow$ is reflexive, symmetric, and transitive.

- (Reflexivity) $i \leftrightarrow i$ since $p_0(i, i) = 1 > 0$

- (Symmetry) $i \leftrightarrow j \implies j \leftrightarrow i$ by definition [4] The vector with $|\pi|_i = |\pi_i|$ is still an eigenvector with eigenvalue 1.
- (Transitivity) $i \leftrightarrow j$ and $j \leftrightarrow i \implies i \leftrightarrow k$ since,

$$\begin{aligned} \boldsymbol{P}_{i,k}^{m_1+m_2} &= \Pr\left(X_{m_1+m_2} = k \mid X_0 = i\right) \\ &\geq \Pr\left(X_{m_1+m_2} = k, X_{m_1} = j \mid X_0 = i\right) \\ &= \Pr\left(X_{m_1} = j \mid X_0 = i\right) \cdot \Pr\left(X_{m_1+m_2} = k \mid X_{m_1} = j\right) \\ &= \boldsymbol{P}_{i,j}^{m_1} \boldsymbol{P}_{j,k}^{m_2} \\ &> 0 \end{aligned}$$

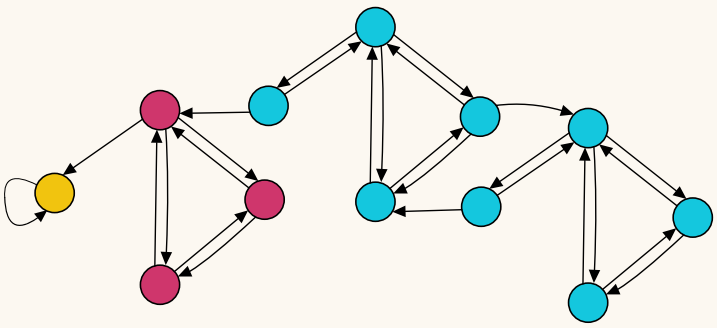
□

As a consequence, a tpm gives rise to a partition of the states of the space into classes.

Definition 22 (Irreducibility): The relation $\leftrightarrow$ partitions the state space into disjoint sets called communication classes. If there is only one communication class, then the chain is called *irreducible*.

Example 10: A partition transition graph with 3 classes

Example of Communication Classes: G has 3 communication classes,



These are our first peak at describing the eventual behavior of a Markov chain:

Definition 23: A state $x \in S$ is *recurrent* if $\Pr_x(\tau_x^+ < \infty) = 1$. Otherwise, the state is *transient*.
Recurrent states are further characterized as follows. A recurrent state $x \in S$ is *absorbing* if $\boldsymbol{P}_{x,x} = 1$. A recurrent state $x \in S$ is *positive recurrent* if $\mathbb{E}_x \tau_x^+ < \infty$. If $\mathbb{E}_x \tau_x^+ = \infty$ then the state is *null recurrent*.

Thus if a Markov state in state x almost surely returns to state x, the state x is recurrent.

Transience and recurrence have an equivalent characterization in terms of the number of returns:

Theorem 17: Fundamental theorem of recurrence

For a state x, let N_x be the number of visits the Markov chain makes to state x. The following are equivalent for a Markov chain:

1. $\Pr_x(\tau_x^+ < \infty) < 1$ (i.e. the state x is transient)
2. $\mathbb{E}N_x = \sum_{k=0}^{\infty} \boldsymbol{P}_{x,x}^k < \infty$
3. $N_x < \infty$ a. s.

Moreover, under $\Pr_x$ N_x is geometrically distributed: $\Pr_x(N_x = k) = (1-p)p^{k-1}$ for $k \in \mathbb{N}$ where $p = \Pr_x(\tau_x^+ < \infty)$ (or identically ∞ when $p = 1$).

Proof. The final distributional claim implies the equivalence the three alternatives. From the Strong Markov property, on the event $\tau_x^+ < \infty$, the law of $(X_{k+\tau_x^+} : k \geq 0)$ is once more $\Pr_x$. Thus under $\Pr_x$, $N_x - 1$ has the law of the number of coin flips required to see a 0 in a sequence of iid Bernoulli(p) random variables, which is the geometric random variable described. □

As a consequence, transience and recurrence are *class properties*.

> **Definition 24:** A property P of a state $x \in S$ is a *class property* if whenever x has P and $x \leftrightarrow y$, then y has P as well.

This implies that all states in a communication class share the same class properties.

Theorem 18: Recurrence classes

Transience and recurrence are class properties.

Proof. As these properties are negations of one another, if suffices to show that recurrence is a class property. Suppose that x is recurrent and y communicates with x. We should show that y is recurrent. From communication, there are numbers $m, n \in \mathbb{N}$ so that we can access x from y in m steps and vice versa in n steps. Fix an $\ell \in \mathbb{N}$. Let $\tau_x^{(\ell)}$ be the time of the ℓ-th visit to x. Let E_ℓ be the event that the Markov chain visits y between $\tau_x^{(\ell)}$ and $\tau_x^{(\ell+1)}$. Let M_y be the number of E_ℓ that occur.

Now by the Strong Markov property

$$\Pr_y(E_\ell \mid \tau_x^{(\ell)} < \infty) = \Pr_x((X_k) \text{ visits } y \text{ before returning to } x) =: q > 0.$$

Then from recurrence $\Pr_x(\tau_x^{(\ell)} < \infty) = 1$. On the other hand, it could be that it is not possible to get from y to x the first time. It does have positive probability, and since by the Strong Markov property

$$\Pr_y(\tau_x^{(\ell)} < \infty \mid \tau_x^{(1)} < \infty) = 1,$$

we have that for all $\ell \in \mathbb{N}$,

$$\begin{aligned}
&\Pr_y(\tau_x^{(\ell)} < \infty) \\
&= \Pr_y(\tau_x^{(\ell)} < \infty \mid \tau_x^{(1)} < \infty) \Pr_y(\tau_x^{(1)} < \infty) \\
&=: p > 0.
\end{aligned}$$

Putting everything together, we have shown

$$\begin{aligned}
\Pr_y(E_\ell) &= \Pr_y(E_\ell \mid \tau_x^{(\ell)} < \infty) \Pr_y(\tau_x^{(\ell)} < \infty) \\
&\geq pq > 0
\end{aligned}$$

As this holds for all $\ell \in \mathbb{N}$, we have

$$\mathbb{E}_y N_y \geq \mathbb{E}_y M_y = \infty.$$

From Theorem 17, the claim follows. □

> **Exercise 3 (Positive recurrence):** Show that positive recurrence is a class property.

> **Exercise 4 (Reachability):** Show that if a state x is recurrent, then $N_x = \infty$ a. s.. Show furthermore that if the chain is irreducible, then for any states x, y $\Pr_y(\tau_x < \infty) = 1$.

> **Exercise 5 (Finite implies Positive):** Show that if a recurrent class is finite then it is positive recurrent.

As a consequence null recurrence is exclusively in the domain of infinite chains. We will delay more discussion of null-recurrence versus positive-recurrence to after we have introduced martingales.

> **Exercise 6 (Transients):** Suppose that B is a recurrent class. Let $B' \subset S$ be the union of all states that can access B. Show that all states in $B' \setminus B$ are transient.

A further class property is that of periodicity.

Definition 25 (Period): The *period* of a state x is the greatest common divisor of the set

$$\{n \in \mathbb{N} : \boldsymbol{P}^n_{x,x} > 0\}.$$

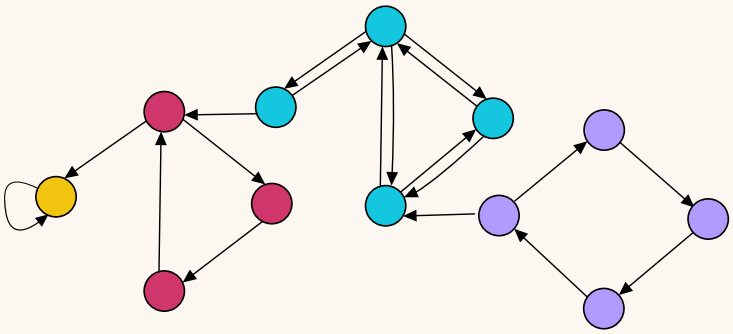
A state is *aperiodic* if it has period 1.

Theorem 19: Periodic classes

The period of a state is a class property, which is to say all states in a communication class have the same period.

Proof. Suppose a state x has period p and a state y has period q. If $x \leftrightarrow y$ then there are $m, n \in \mathbb{N}$ so that the chain access y from x in m steps and x from y in n steps. It follows by travelling from $x \to y$, from $y \to y$ ℓ times and then $y \to x$ in n steps. Hence p divides $m + n + \ell q$ for all $\ell \in \mathbb{N}$. It follows that p divides $m + n$ (taking $\ell = p$). Hence p also divides q (taking $\ell = 1$). By a symmetric argument, q divides p. So $p = q$. □

Example 11: Periods



This Markov chain has four classes, with periods 1,3,1 and 4, going from left to right.

Exercise 7 (Lazy FTW): One way to break periodicity is to introduce laziness. For a THCS Markov chain with tpm P, and a laziness parameter p, we can define a new tpm $Q = p\,\mathrm{Id} + (1-p)P$, which describes, in words, at each step flipping a coin with success probability p. If heads, stay put, if tails, take a step from the original chain.
If $p \in (0,1)$, the resulting chain will always be aperiodic. Show that the lazy chain $(Y_n : n \geq 0)$ has the following alternative description in terms of the original chain $(X_n : n \geq 0)$.
Let $(T_j : j \geq 1)$ be iid Geom(p) random variables on $\{1, 2, 3 \cdots\}$

(so $\Pr(T_j = k) = p^{k-1}(1-p)$) independent of $(X_n : n \geq 0)$. For any $n \geq 0$ let $N_n := \max\{k : \sum_1^k T_j \leq n\}$, where if the set is empty we take $N_n = 0$. Then $(X_{N_n} : n \geq 0)$ is THCS Markov chain with with tpm Q and the same initial distribution as Y. *Hint: think of a Markov chain on state space $S \times \{0,1\}$ which is just iid* Bernoulli(p) *in the second coordinate, and only makes P-transitions in the first coordinate when the second is* 1*, and use the strong Markov property.*

Stationary distributions

Our first major result on Markov chains will concern their convergence in distribution as time tends to infinity of the distribution of a Markov chain. The limit distribution of the chain will be a *stationary* distribution.

Definition 26 (Stationary Distribution): A probability measure π is a *stationary distribution* for a THCS Markov chain with tpm $\boldsymbol{P}$ if $\pi\boldsymbol{P} = \pi$, which is to say that π is an eigenvector of $\boldsymbol{P}$ of eigenvalue 1. In the case $|S| = \infty$, we reserve *eigenvector* for vectors v satisfying $\|v\|_1 := \sum_s |v_s| < \infty$ and $v\boldsymbol{P} = v$.

Note that beyond being eigenvectors, stationary distributions must furthermore be non-negative vectors which sum to 1.

Hence the set of stationary distributions of a THCS chain form a polytope (meaning a convex hull of a set of a finite set). The number of extreme points of this convex hull are in some sense the number of non-equal stationary distributions.

Theorem 20: Stationary distributions

For any positive recurrent class B, the formula

$$\pi_B(x) := \begin{cases} \frac{1}{\mathbb{E}_x \tau_x^+} & \text{if } x \in B. \\ 0 & \text{otherwise.} \end{cases}$$

defines the unique stationary distribution with support contained in B, and the set

$$\{\pi_B : B \text{ is a positive recurrent class}\}$$

are the extreme points of the set of stationary distributions. This is also a basis of left eigenvectors of eigenvalue 1 of $\boldsymbol{P}$.

Note that 1 is always a right eigenvector of $\boldsymbol{P}$, as the all-1 vector

is a right eigenvector of $\mathbb{P}$ of eigenvalue 1, and hence in the $|S| < \infty$ case, there is always a stationary distribution (and a recurrent class).

> **Corollary 1:** If $|S| < \infty$ then there is always a stationary distribution, and moreover the dimension of the space of stationary distributions is equal to the number of recurrent classes.

Note that infinite chains, which can be null recurrent, do not need to have stationary distributions. Moreover

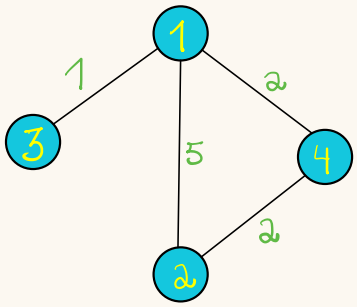
> **Corollary 2:** An irreducible THCS Markov chain is positive recurrent if and only if it has a stationary distribution π.

Example 12: Edge weighted graphs

In general, finding the stationary distribution of a chain is complicated. It can be helpful to have a family of examples where there is a simple rule to find the stationary distributions.
One way to do this is to take an undirected, connected graph (V, E), and then choose edge weights $w : E \to (0, \infty)$. Extend this to a vertex weight function $w(x) := \sum_{y \sim x} w(\{x, y\})$ where $\sim$ denotes adjacency (i.e. $x \sim y$ iff $\{x, y\} \in E$). Then define a tpm by

$$\boldsymbol{P}_{x,y} = \frac{w(\{x,y\})}{w(x)}.$$

The (unique) stationary distribution of such a Markov chain is always $\pi(x) := \frac{w(x)}{\sum_y w(y)}$.



So for example:
has stationary distribution $(\frac{8}{20}, \frac{7}{20}, \frac{1}{20}, \frac{4}{20})$.

Example 13: Doubly stochastic

A *doubly stochastic* matrix M is one for which both M and M^t are stochastic (recall Definition 16). For a finite doubly stochastic matrix, the all-1 vector is both a left and a right

eigenvector, and hence $\text{Unif}(S)$ is a stationary distribution.

Example 14: Reflected biased RW on $\mathbb{Z}$

Let $S = \mathbb{N}_0$. Let $p \in (0,1)$.

$$X_j = \begin{cases} 1, & \text{if } X_{j-1} = 0, \text{ else:} \\ 1 + X_{j-1}, & \text{with } \Pr(\cdot \mid \mathscr{F}_{j-1}) = p. \\ -1 + X_{j-1}, & \text{with } \Pr(\cdot \mid \mathscr{F}_{j-1}) = 1 - p. \end{cases}.$$

In words, the process jumps to the right with probability p, left with probability $(1-p)$, and jumps to 1 from 0 deterministically (this is the reflected part). The "biased" refers to the fact that p may not be $\frac{1}{2}$.
We shall show that when $p > \frac{1}{2}$, this Markov chain is transient, when $p = \frac{1}{2}$ the process is null-recurrent and when $p < \frac{1}{2}$ this process is positive recurrent.
We can check the positive recurrent part here, as from Theorem 20, it suffices to find a stationary distribution. To do this, we just try to solve for a left eigenvector v of $\boldsymbol{P}$ of eigenvalue 1. Set, arbitrarily $v_0 = 1 - p$. We now write the eigenvector equation. At 0, these equations are exceptional:

$$v_1 \boldsymbol{P}_{1,0} = v_0,$$

and so $v_1 = 1$. Generally, for $k \geq 1$

$$v_{k-1}\boldsymbol{P}_{k-1,k} + v_{k+1}\boldsymbol{P}_{k+1,k} = v_k.$$

When $k = 1$, these equations are exceptional, and we get

$$v_2(1-p) = 1 - (1-p),$$

and so $v_2 = \frac{p}{1-p}$. By induction, we can check for all larger k, $v_{k+1} = \frac{p^k}{(1-p)^k}$. As this series is summable for $p < \frac{1}{2}$, we have constructed a stationary distribution (after dividing by its sum).

To prove Theorem 20, we need a few general lemmas about stationary distributions.

Lemma 6: Suppose v is a left eigenvector of tpm $\boldsymbol{P}$ with eigenvalue 1. Then

1. For any transient state x, $v_x = 0$.

2. The restriction of v to any communication class is again an eigenvector of eigenvalue 1.
3. The absolute value $|v|$ (meaning the vector in which we have taken the absolute value of every entry) is an eigenvector of eigenvalue 1.
4. If v is a stationary distribution, then for any recurrent class B such that $v(B) > 0$, v does not vanish on B.
5. For a recurrent class B, there is at most 1 stationary distribution v that is supported on B.

Proof. 1. By the eigenvector property, we have that $v\boldsymbol{P}^n = v$ for all n, or in other words for all states x

$$v_x = \sum_{s \in S} v_s \Pr_s(X_n = x).$$

If x is transient, then $\Pr_x(X_n = x) \to 0$ as $n \to \infty$ as by Theorem 17, the whole sum in n is finite. Hence it follows $\Pr_s(X_n = x)$ also tends to 0 as $n \to \infty$ for any state s which is accessible from x (as otherwise we could lower bound $\Pr_x(X_{n+m} = x)$ for some fixed m by first bounding below the probability of traveling $x \to s$ in m steps and then $s \to x$ in n steps.

2. For the second point, let B be any communication class. If B is transient, there is nothing to show by the first point. Otherwise if B is recurrent, then if we let B' be all states which can access B, $B' \setminus B$ are all transient states (see Exercise 6). Let w be the restriction of v to B. Let x be any state in S. Then

$$\begin{aligned} w_x &= v_x \\ &= \sum_{s \in S} v_s \boldsymbol{P}_{s,x} && \text{by the eigenvector property of } v \\ &= \sum_{s \in B'} v_s \boldsymbol{P}_{s,x} && \text{as elements } s \notin B' \text{ have } \boldsymbol{P}_{s,x} = 0 \\ &= \sum_{s \in B} v_s \boldsymbol{P}_{s,x} && \text{by } v \text{ vanishes for transient states} \\ &= \sum_{s \in B} w_s \boldsymbol{P}_{s,x}. \end{aligned}$$

Thus w is again an eigenvector.

3. If $v = 0$ the claim is trivial. By renormalizing the eigenvector v, we may assume that $\sum_x |v_x| = 1$. Then for any $x \in S$

$$|v_x| = \left|\sum_s v_s \boldsymbol{P}_{s,x}\right| \leq \sum_s |v_s| \boldsymbol{P}_{s,x}. \tag{1}$$

Suppose there is strict inequality for any x. Then summing in x

$$1 = \sum_x |v_x| < \sum_{x,s} |v_s| \boldsymbol{P}_{s,x}.$$

Using that $\boldsymbol{P}$ is stochastic, if we peform the sum over x first, we conclude

$$\sum_{x,s} |v_s| \boldsymbol{P}_{s,x} = \sum_s |v_s| = 1.$$

This is a contradiction, and so we must have equality in (1) for all $x \in S$.

4. As $v(B) > 0$, there is a state $x \in B$ with $v_x > 0$. Now for any other $s \in B$, there is an $m \in \mathbb{N}$ so that $\boldsymbol{P}^m_{x,s} > 0$. Hence

$$v_s = \sum_y v_y \boldsymbol{P}^m_{y,s} \geq v_x \boldsymbol{P}^m_{x,s} > 0.$$

5. Suppose that v, w were two stationary distributions on B. Then both must be supported on all of B, by the previous point. Hence for any state $x \in B$, we can choose a nonzero linear combination of v, w that vanishes at x. This linear combination u is an eigenvector of eigenvalue 1. If $|u|$ is not identically distributed, we could renormalize to make a stationary distribution supported properly within B, but this is impossible, and so $|u| = 0$ identically. Thus u and v are proportional to each other. As they are stationary distributions, they must be equal.

□

> **Exercise 8 (Null recurrence):** Show that if a state x is null-recurrent and π is a stationary distribution, then $\pi_x = 0$. *Hint: use the occupation time idea from the next proof. The following application of the strong law of large numbers might be helpful: if $(X_j)_1^\infty$ are iid non-negative random variables having $\mathbb{E}X_1 = \infty$, then $\frac{1}{n}\sum_{j=1}^n X_j \xrightarrow[n\to\infty]{\text{a.s.}} \infty$.*

We can now show the proof of Theorem 20.

Proof. The main point here is to show that π_B is indeed a stationary distribution. Let $y \in B$ be fixed and define, for any m and any x

$$v_x^{(m)} = \frac{1}{m} \sum_{j=1}^m \Pr_y(X_j = x),$$

which is the expectation fraction of time that the Markov chain spends in state x. Observe that if we let $\tau_x^{(\ell)}$ be the ℓ-th arrival times of the chain to x (with $\tau_x^{(1)} = \tau_x^+$) then we also have for all $m \geq \tau_x^{(1)}$

$$\sum_{j=1}^m \mathbf{1}_{X_j = x} = \max\{\ell : \tau_x^{(\ell)} \leq m\}.$$

By the Strong Markov property, $\{\tau_x^{(\ell+1)} - \tau_x^{(\ell)} : \ell \in \mathbb{N}\}$ are iid. Hence, by the strong law of large numbers

$$\frac{\tau_x^{(\ell)}}{\ell} \xrightarrow[\ell\to\infty]{\text{a.s.}} \mathbb{E}_x \tau_x^+,$$

and so we have $\Pr_y$–almost surely,

$$\frac{1}{m}\max\{\ell : \tau_x^{(\ell)} \le m\} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{1}{\mathbb{E}_x \tau_x^+}.$$

By dominated convergence,

$$v_x^{(m)} \to \frac{1}{\mathbb{E}_x \tau_x^+}.$$

Note that $v^{(m)}$ is clearly a probability distribution as

$$\sum_x v_x^{(m)} = \mathbb{E}_y \frac{1}{m}\sum_{j=1}^m \sum_x \mathbf{1}_{X_j=x} = 1.$$

Furthermore,

$$\begin{aligned}(v^{(m)}\boldsymbol{P})_x &= \frac{1}{m}\sum_{j=1}^m \sum_w \Pr_y(X_j = w)\boldsymbol{P}_{w,x} \\ &= \frac{1}{m}\sum_{j=1}^m \Pr_y(X_{j+1} = x),\end{aligned}$$

and so

$$\sum_x |(v^{(m)}\boldsymbol{P})_x - v_x^{(m)}| \le \frac{2}{m}.$$

It follows on taking the limit that π_B is a stationary distribution.

The remainder of the claims now follow from Lemma 6:

1. By part 5, for every positive recurrent class B, there is exactly 1 stationary distribution supported on B.

2. An extreme point of the set of stationary distributions must be supported on a single positive recurrent class: if not, every restriction of it to a positive recurrent class is (after renormalizing) a stationary distribution, and hence it would be proper convex combination of other stationary distributions (we have implicitly used Exercise 8. Conversely, there is exactly 1 stationary distribution for every class, and so the extreme points are exactly the stationary distributions.

3. The claim for the geometric multiplicity is similar.

□

Example 15: Ehrenfest Urn

Imagine two closed chambers L and R containing n particles in total. Let $(X_j : j \geq 0)$ denote the number of particles in the L chamber. At every moment in time, a particle will either move from L to R or from R to L. The probability the particle moves from L to R in the j-th step is proportional to X_{j-1}, which is to say

$$X_j = \begin{cases} 1 + X_{j-1}, & \text{with } \Pr(\cdot \mid \mathscr{F}_{j-1}) = \frac{X_{j-1}}{n}, \\ -1 + X_{j-1}, & \text{with } \Pr(\cdot \mid \mathscr{F}_{j-1}) = \frac{n - X_{j-1}}{n}. \end{cases}$$

Then this chain has stationary distribution $\text{Binom}(n, \frac{1}{2})$ as can be checked by the following computation:
for any $k \in \{0, 1, 2, \cdots, n\}$,

$$\begin{aligned} &2^{-n}\binom{n}{k+1}\frac{k+1}{n} + 2^{-n}\binom{n}{k-1}\frac{n-(k-1)}{n} \\ &= 2^{-n}\left(\binom{n-1}{k} + \binom{n-1}{k-1}\right) = 2^{-n}\binom{n}{k}. \end{aligned}$$

Furthermore, as the chain is irreducible, this is unique.
This chain has its origins in the theory of statistical mechanics. If you view the chain as describing particles of gas bouncing around the room, you can ask what is the probability the gas were to entirely travel to one side of the room, hence suffocating all its inhabitants by pure spiteful randomness. In a purely random theory of gases (such as in the toy model of the Ehrenfest Urn), that can and does happen, provided we wait long enough. The fraction of time the system spends in that state however, is 2^{-n}. If n is large enough (say like 10^{23}), you're going to be waiting for a long time...

Convergence to stationarity

The first main highlight of this course is the Markov Chain convergence theorem. Returning to one of our first examples:

Example 16: Lazy coin

(Continuing on Example 8) The state space $S = \{H, T\}$ is a two-element space. We will suppose that due to low effort

flipping, the transition probability is

$$\boldsymbol{P} = \begin{array}{c|cc|c} & H & T & \\ \hline & 0.1 & 0.9 & H \\ & 0.9 & 0.1 & T \end{array}.$$

If one starts with initial distribution δ_H, then the law of the first step is $0.1\delta_H + 0.9\delta_T$, which is therefore very far from a fair flip. However, if lazy-flipper continues,

$$\boldsymbol{P}^{10} \approx \begin{array}{c|cc|c} & H & T & \\ \hline & 0.554 & 0.446 & H \\ & 0.446 & 0.554 & T \end{array} \quad \text{and}$$

$$\boldsymbol{P}^{100} \approx \begin{array}{c|cc|c} & H & T & \\ \hline & 0.5000000001 & 0.4999999999 & H \\ & 0.4999999999 & 0.5000000001 & T \end{array}$$

So high powers are converging (and in fact are converging exponentially quickly) to a 2×2 matrix which is the constant $\frac{1}{2}$. Hence, regardless of whether we start in an H configuration or in a T configuration, the distribution of the chain after 100 steps is wihin 8 digits of accuracy to a perfect coin flip. Note that $(\frac{1}{2}, \frac{1}{2})$ is the stationary distribution.

So this example showed that raising a certain tpm to a high power produced a matrix whose every row is the same, which is to say the distributions of the Markov chain from any initial distribution is always the same. Now there are some obstructions to a Markov having this behavior. The class of chains for which the same behavior shown in Example 16 still holds.

Definition 27 (Ergodic): Say a THCS Markov chain is *ergodic* if it is irreducible, aperiodic, and positive recurrent.

Theorem 21: Markov chain convergence

In an ergodic THCS Markov chain $(X_j : j \geq 0)$, there is a unique stationary distribution π, and for any initial distribution on X_0, $X_j \xrightarrow[j\to\infty]{\text{law}} \pi$.

To formulate this convergence, it is helpful to use the *total variation metric*.

Definition 28 (Total variation): The total variation metric between two laws μ, ν on S $d_{TV}(\mu, \nu)$ is

$$d_{TV}(\mu, \nu) = \frac{1}{2} \sum_{x \in S} |\mu_x - \nu_x|.$$

For random variables X, Y taking values on S, we set $d_{TV}(X, Y)$ to be the total variation distance between the laws of X and Y.

The total variation metric is the ideal way to measure the distance between distributions on countable spaces (in contrast, it tends to be too strong of a metric outside of discrete contexts). It admits many different representations:

Theorem 22: TV metric

The total variation metric on a countable space S admits the following representations:

1. For any laws μ, ν,

$$d_{TV}(\mu, \nu) = \sup_{A \in \mathscr{A}} |\mu(A) - \nu(A)|.$$

2. For any laws μ, ν,

$$d_{TV}(\mu, \nu) = \inf_{(X,Y)} \Pr(X \neq Y)$$

Here the infimum is over all random variables (X, Y) taking values in $S \times S$ such that X has law μ and Y has law ν. Such a construction of a joint law is called a *coupling* of the laws μ, ν.

Furthermore, on a countable space, convergence in total variation metric is equivalent to weak convergence.

The main tool that we need is the following:

Lemma 7: In an aperiodic, irreducible THCS Markov chain, for any pair of states x, y there is an $n \in \mathbb{N}$ so that for all $m \geq n$, $\boldsymbol{P}^m_{x,y} > 0$.

Proof. It suffices to show the claim for the case that $x = y$, as this then leads to the claim for $x \neq y$ by decomposing the path from $x \to y$ of length $m + k$ into a path $x \to x$ of length m and a path $x \to y$ of length k (which exists by irreducibility). For the case $x \to x$, by

definition of aperiodicity, the greatest common divisor of the set

$$R := \{m : \boldsymbol{P}^m_{x,x} > 0\}$$

is 1. Now note R has the property that if $\ell, r \in R$, so is $\ell + r \in R$.

The remainder of the proof requires a little bit of number theory. We need that given the greatest common divisor of R is 1 and that R is closed under addition, it actually follows that there is an n sufficiently large so that for all $m > n$, $m \in R$, which completes the proof.

Since R has greatest common divisor 1, there is some finite list of numbers $\{a_1, a_2, \cdots, a_k\} \subseteq R$ with greatest common divisor 1. By Bézout's identity, there are therefore integers $\{b_1, b_2, \cdots, b_k\}$ (at least one of which is negative) so that

$$a_1 b_1 + \cdots + a_k b_k = 1.$$

Hence letting $\bar{b} = \max\{-b_j : 1 \leq j \leq k\}$,

$$a_1(b_1 + \bar{b}) + \cdots + a_k(b_k + \bar{b}) = 1 + \bar{b}(\sum_k a_k) \in R$$

We also have $r = \bar{b}(\sum_k a_k) \in R$. And so we have shown there is an $r \in R$ so that $r + 1 \in R$ as well.

Now every integer $m > r^2$, when divided with remainder by r has $m = k(r-1) + \ell$ for some $\ell \in \{0, \cdots, r-2\}$ and $k \geq (r-1)$. Then $m = (k - \ell)(r-1) + \ell r$ is a positive linear combination of r and $(r-1)$, and so we have shown every m larger than r^2 is contained in R. □

The next idea concerns building a Markov chain out of two independent Markov chains. Let $\big(X_j : j \geq 0\big)$ and $\big(Y_j : j \geq 0\big)$ be two independent copies of a Markov chain with a tpm $\boldsymbol{P}$. Then the process $((X_j, Y_j) : j \geq 0)$ is still a Markov chain, on $S \times S$ and so has a tpm $\boldsymbol{P} \otimes \boldsymbol{P}$ (the Kronecker product of the two tpms), which has entries given by

$$(\boldsymbol{P} \otimes \boldsymbol{P})_{(x,y),(a,b)} = \boldsymbol{P}_{x,a} \boldsymbol{P}_{y,b}.$$

Note that for any $n \in \mathbb{N}$

$$(\boldsymbol{P} \otimes \boldsymbol{P})^n = \boldsymbol{P}^n \otimes \boldsymbol{P}^n, \tag{2}$$

as they describe the transitions in two independent chains.

> **Lemma 8:** If $\boldsymbol{P}$ is ergodic, then $\boldsymbol{P} \otimes \boldsymbol{P}$ is irreducible and recurrent.

Proof. We need to show that it is possible for any pair of states (x, y) to access any other pair of states (a, b). By Lemma 7, there is an n sufficiently large that for all $m > n$

$$\boldsymbol{P}^m_{x,a} > 0 \quad \text{and} \quad \boldsymbol{P}^m_{y,b} > 0.$$

Now note that by (2) $(\boldsymbol{P} \otimes \boldsymbol{P})^m_{(x,y),(a,b)} > 0$. Hence $\boldsymbol{P} \otimes \boldsymbol{P}$ is irreducible.

Now to check recurrence, it suffices to show that for any state x

$$\begin{aligned}
&\sum_{n=1}^{\infty} (\boldsymbol{P} \otimes \boldsymbol{P})^n_{(x,x),(x,x)} \\
&= (\boldsymbol{P}^n_{(x,x)})^2 \\
&= \infty.
\end{aligned}$$

Now in fact from positive recurrence,

$$\frac{1}{n} \sum_{j=1}^{n} \boldsymbol{P}^j_{(x,x)} \to \frac{1}{\mathbb{E}_x \tau_x^+}$$

(see the proof of Theorem 20), which implies that the sequence $\{\boldsymbol{P}^j_{(x,x)} : j \geq 1\}$ is larger than $\frac{1}{2\mathbb{E}_x \tau_x^+}$ infinitely often. Hence the recurrence follows. □

The proof of the Markov chain convergence theorem now follows from a clever trick, known as the Doeblin coupling argument:

Proof. From Theorem 20, there is a unique stationary distribution π for the ergodic chain $(X_j : j \geq 0)$. Thus define an inependent Markov chain $(Y_j : j \geq 0)$ with initial distribution π, and note that by stationarity $Y_j \overset{\text{law}}{=} \pi$ for all $j \geq 0$. Let $A \subset S \times S$ be the diagonal (i.e. the set of all (x, x) for $x \in S$). By Lemma 8, the chain $((X_j, Y_j) : j \geq 0)$ is irreducible and recurrent. Hence the stopping time $\tau_A < \infty$ almost surely (see Exercise 4).

Now define a process

$$Z_j := \begin{cases} (X_j, Y_j) & \text{if } j < \tau_A \\ (Y_j, Y_j) & \text{if } j \geq \tau_A. \end{cases}$$

Then the first coordinate of Z_j is a Markov chain with tpm $\boldsymbol{P}$ and the same initial distribution as X_0 (this takes some reflection – consider computing the finite dimensional marginals, in time). Then we have

$$\Pr((Z_j)_1 \neq Y_j) \leq \Pr(\tau_A > j).$$

Since $(Z_j)_1$ has the same law as X_j, we have shown that

$$d_{TV}(X_j, Y_j) \leq \Pr(\tau_A > j),$$

(see Theorem 22) which tends to 0 as $j \to \infty$ by the almost sure finiteness of τ_A. As Y_j has law π, this completes the proof. □

Remark 1 (T): e general principle used here was to construct a probability space on $(S \times S)^\infty$, with the properties that:

1. the stochatic process in the first coordinate had the law of Markov chain $(X_n : n \geq 0)$;
2. the second coordinate $(Y_n : n \geq 0)$ had the law of a stationary Markov chain with the same transition probability matrix;
3. and after the two chains collide, i.e. $X_n = Y_n$, they remain together for all time.

Then if τ is the hitting time of the chain to the diagonal $\{(x, y) \in S \times S : x = y\}$,

$$d_{TV}(X_n, Y_n) \leq \Pr(\tau > n).$$

There can be many ways to construct this probability space, and some are better than others, in the sense that τ happens faster. It's actually *not* necessary that the pair $((X_n, Y_n) : n \geq 0)$ form a Markov chain – if they do, we would call the coupling of the two chains Markovian. It is easier, however. It's also not necessary that the chains be independent before their coupling time, and to make faster-coupling Markov chains, we might actually prefer to do something non-independent. We'll do this with the card shuffling example below.

Exercise 9 (Strong Law of Large Numbers for Markov Chains): If $(X_n : n \geq 0)$ is a finite state, time-homogeneous, aperiodic, irreducible Markov chain and r is a bounded and real-valued function, then

$$\lim_{n\to\infty} \frac{r(X_1) + \cdots + r(X_n)}{n} = \mathbb{E}[r(X)] \quad \text{a.s.}$$

where $\mathbb{E}[r(X)] = \sum_j r(j)\pi_j$. *Hint: the chain is not i.i.d, but successive excursions between visits to the same state are independent.*

Example 17: Card shuffling

A rich class of markov chains with interesting mixing properties are card shuffling Markov chains. The state space in such a chain is the set of all permutations of an underlying finite set $\mathfrak{S}_m$. Formally we can represent these as bijections from $\{1, 2, \cdots, m\}$ to itself. Concisely, we can represent these in

"one-line" notation as a list of numbers, for example:

$$937452618 \quad \longleftrightarrow \quad \pi(1) = 9, \pi(2) = 3, \cdots, \pi(9) = 8,$$

just recording in the j-th location the image of j.
This has a group structure: for two permutations $A, B \in \mathfrak{S}_m$ the permutation AB is given by first applying B and then A (i.e. it is the composition of the two permutations). So we can define a Markov chain on $\mathfrak{S}_m$ by choosing a step distribution μ on $\mathfrak{S}_m$, and then forming an iid sequence $(U_k : n \geq 0)$ sampled from μ. The Markov chain is given by (for all $n \geq 0$)

$$X_{n+1} = U_{n+1} X_n.$$

For example, if we let μ be the uniform measure over the set of *transpositions*, which is to say permutations that swap two elements and leave the rest fixed, then this is called the *random transposition chain*. In one-line notation, this just means we randomly swap two numbers at each step.
Tying it back to card shuffling, this would have the interpretation as the Markov chain which at each step selects two cards from the deck uniformly at random and then interchanges them.
If that sounds painful to implement, you might try the *top-to-random* shuffle, in which you take the top card and then insert it in a random location in the deck.
Now it turns out that for any choice of μ, this Markov chain will have stationary distribution $\text{Unif}(\mathfrak{S}_m)$. If in addition the support of μ generates $\mathfrak{S}_m$ – which is to say using the steps which have positive probability under μ, can be composed to generate any permutation – then the Markov chain is irreducible.
So the question is here is not really do these chains converge, but how fast?

Exercise 10 (Always uniform): Show that for any choice of μ, the Markov chain defined this way is stationary with respect to uniform measure on $\mathfrak{S}_m$.

Remark 2 (T): quantify the rate of convergence, we might look at the *mixing time*. For finite chains S

$$\text{Mixing-Time} = \max_{x \in S} \inf\{n \in \mathbb{N} : d_{TV}(\delta_x P^n, \pi) \leq \tfrac{1}{2}\}.$$

Here δ_x is the row vector with mass 1 at x and 0 elsewhere. So $d_{TV}(\delta_x P^n, \pi)$ is the distance to stationarity of a Markov chain with initial state x and tpm P. We then look at the worst-case starting point. The $\frac{1}{2}$ is arbitrary, but it also does not matter so much for a coarse understanding of the rate. One can check (using properties of total variation) that replacing $\frac{1}{2}$ by $(\frac{1}{2})^k$ leads to a mixing time which is bounded above by k-times the definition above.

Lemma 9: Given two copies of the random transposition chain $(X_n : n \geq 0)$ and $(Y_n : n \geq 0)$, there is a coupling so that the first time τ that $X_n = Y_n$ satisfies

$$\mathbb{E}\tau \leq m(m-1)(\log m + 1).$$

Proof. In the one-line representation, we can represent the chain as choosing two positions and swapping the numbers in those positions at each step. We say the two permutations are aligned in a position i if they have the same value in that position. Now we divide the set of transpositions into the set $\mathcal{A}$ contain an aligned entry and the set $\mathcal{U}$ for which both are unaligned.

Now we choose the transpositions U, V to be applied to X, Y respectively by doing the following:

1. Flip a coin with probability $p = |\mathcal{A}|/(|\mathcal{A}| + |\mathcal{U}|)$.
2. If heads, we sample uniformly at random one transposition from the set $\mathcal{A}$ and set $\mathcal{U} = \mathcal{V}$.
3. If tails, we sample the U and V independently and uniformly at random from $\mathcal{U}$.

Using this strategy, we never decrease the number of alignments. Moreover, when we sample the from $\mathcal{U}$ we have a chance to add 1 or 2 alignments. If we have k unaligned spots, the probability that we add an aligned spot is at least $1/(k-1)$ (having applied the first transposition, and chosen the first step of the second transposition, there's always 1 of $(k-1)$ ways to choose the second spot which creates an alignment).

So if there are currently k out of m unaligned slots, we have at each run of the algorithm a probability $\binom{k}{2}/\binom{m}{2}$ of succeeding in the initial coinflip and a probability at least $1/(k-1)$ of reducing an alignment. Hence the number of steps we need to wait is, in expectation, at most

$$(k-1)\binom{m}{2}/\binom{k}{2}$$

Thus, if we sum this over k, we get the claim. □

Exercise 11 (Doing better): We were wasteful above: it is possible do something better than drawing two independent permutations U and V and praying. Show it is possible to improve the coupling strategy above to one which creates an alignment with at least probability $c > 0$ uniformly in the number of unaligned spots.

Exercise 12 (Mixing time lower bound):
Suppose we consider the random transposition chain $(X_n : n \geq 0)$ started from the identity permutation. One way to tell that our distribution is not close to uniform is that we have not even moved some digits.

1. Let τ be the first time that every number $k \in \{1, 2, \cdots, m\}$ has been moved at least once. Suppose that n is such that $\Pr(\tau > n) \geq \frac{9}{10}$. Prove that for such n,

$$d_{TV}(X_n, \text{Unif}(\mathfrak{S}_m)) \geq \tfrac{1}{100},$$

   by considering the event where the permutation has no fixed points (i.e. j so that $\pi(j) = j$).

2. Let A_n be the number of $k \in \{1, 2, \cdots, m\}$ which have not yet been moved after n steps. Find $\alpha > 0$ so that by computing first and second moments of A_{t_m}, when $t_m \approx \alpha m \log m$, $\Pr(A_{t_m} > 0) \to 1$ as $m \to \infty$.

(Note)[4]

[4] Combining this with the previous Lemma, we have on the one hand the time to stationarity is more than $\mathcal{O}(m \log m)$ steps and less than $\mathcal{O}(m^2 \log m)$. The truth is closer to the first one, and it leverages that after you have randomized a location, it remains in a sense uniformly random (this leads to the idea of "Strong stationary times", see (Levin and Peres, *Markov chains and mixing times*)).

Time reversal and reversibility

An alternative characterization of a Markov chain $(X_j : j \geq j_0)$ is that given a state X_k, the past and present of the chain are independent (see Exercise 1). Such a description has no obvious definition of time, and so it must be that you if reverse time in a Markov chain, it remains a Markov chain. [5]

Definition 29 (Time reversal): The time–reversal of $\boldsymbol{P}$ with respect to stationary measure π is

$$\boldsymbol{Q}_{x,y} := \boldsymbol{P}_{y,x} \frac{\pi(y)}{\pi(x)}.$$

This is a new transition probability matrix defined on the space $S' = \{x \in S : \pi(x) > 0\}$.

[5] In general, the time-reversal will not be a time-homogeneous Markov chain. Imagine for example, one takes an ergodic chain started from a deterministic initial condition x at time 0. The chain reversed from time 100 still ends almost surely at x after 100 steps!

Exercise 13 (Reversed): Use the Radon–Nikodym theorem to construct the time–reversed Kernel of a general Markov chain.

Theorem 23: Time–reversal

Let $(X_j : j \geq 0)$ be a stationary Markov chain with initial distribution π. Then for any $k \in \mathbb{N}$, the process $Y_j := X_{k-j}$ for $0 \leq j \leq k$ is a Markov chain with stationary distribution π and tpm $\boldsymbol{Q}$.

It suffices to check the finite dimensional marginals of the chain. As a consequence, it is also possible to extend a stationary Markov chain to a 2-sided Markov chain:

Theorem 24: 2-sided chains

For a stationary Markov chain with initial distribution π, it is possible to extend it to a 2-sided Markov chain $(X_j : j \in \mathbb{Z})$ which for any desired time $j_0 \in \mathbb{Z}$ has that

$$(X_{j+j_0} : j \geq 0) \quad \text{and} \quad (X_{-j+j_0} : j \geq 0)$$

are stationary Markov chains with tpms $\boldsymbol{P}$ and $\boldsymbol{Q}$ respectively.

This leads to the idea of *reversibility*.[6]

[6] This is one of the worst pieces of nomenclature in mathematics. A much better name would be time-reversal symmetry

Definition 30 (Detailed Balance Equations): An irreducible tpm $\boldsymbol{P}$ satisfies the detailed balance equation with respect to stationary distribution π if

$$\pi_i \boldsymbol{P}_{ij} = \pi_j \boldsymbol{P}_{ji} \quad \text{for all } i, j \in S.$$

A Markov chain whose tpm satisfies the detailed balance equations is *reversible*.

Equivalently $\boldsymbol{P} = \boldsymbol{Q}$.[7] [8]

[7] If a distribution π satisfies the detailed balance equations, it also follows that π is a stationary distribution for $\boldsymbol{P}$.

[8] There is an extension of detailed balance to σ-finite measures which is also useful for non-positively-recurrent chains, but this is a bigger dive into Markov chain theory than we will cover.

MCMC

Markov chain convergence is not just philosophically important. It also gives a useful tool for solving difficult problems, and one technique for leveraging this is called Markov chain Monte Carlo (MCMC). MCMC is a method of defining a Markov chain to sample from a desired distribution π. At first sight, the problem of sampling from a given distribution may not seem like an interesting problem. However, many difficult problems can be posed as *sampling problems*, and this can be an effective way to relax difficult optimization problems.

The Metropolis–Hastings Algorithm, explicitly, exploits that for many distributions π, it is possible to compute the ratio π_i/π_j, while the actual stationary distribution π itself is inaccessible (usually due to the inability to compute the normalizing constant). [9]

Definition 31 (Metropolis-Hastings Algorithm): Let π be a probability distribution on the countable space S. The Metropolis-Hastings Algorithm (MHA) is a stochastic process, defined as follows. As input, we suppose that we are given $\mathbf{T}$ a transition matrix for an irreducible Markov chain with the same state space as π. The chain with this transition matrix is known as the *proposal chain*.
Repeat the following, until a decided termination condition:

1. Let i be the current state of the MHA chain X_n. Choose a new state j, the proposal state, according to $\mathbf{T}_{i,j}$
2. Let $U \sim \text{Unif}(0,1)$. Define an acceptance function,

$$a(i,j) = \frac{\pi_j \boldsymbol{T}_{ji}}{\pi_i \boldsymbol{T}_{ij}} \quad \text{and let } X_{n+1} := \begin{cases} j & \text{if } U \leq a(i,j) \\ i & \text{otherwise} \end{cases}$$

[9] It is also assumed that there is some easy input chain on the state space – the proposal chain $\mathbf{T}$ in what follows – for which the sampling problem is easy.

Theorem 25: Metropolis Hastings

The Metropolis-Hastings Algorithm is a Markov chain which is reversible with respect to π.

Proof. The sequence $(X_n)_{n\geq 1}$ constructed by the Metropolis-Hastings Algorithm is a Markov chain, as each X_{n+1} only depends on X_n. If an irreducible Markov chain has a stationary distribution, then the chain is recurrent.

$\mathbf{P}$ be its transition matrix. We need to show that (X_n) is reversible with stationary distribution is $\vec{\pi}$. Given $X_0 = i$, then,

$$P(U \leq a(i,j)) = \begin{cases} a(i,j) & \text{if } a(i,j) \leq 1 \\ 1 & \text{otherwise} \end{cases}$$
$$= \begin{cases} a(i,j) & \text{if } \pi_j \boldsymbol{T}_{ji} \leq \pi_i \boldsymbol{T}_{ij} \\ 1 & \text{otherwise} \end{cases}$$

and for $i \neq j$,

$$\boldsymbol{P}_{ij} = \begin{cases} \boldsymbol{T}_{ij} \cdot a(i,j) & \text{if } \pi_j \boldsymbol{T}_{ji} \leq \pi_i \boldsymbol{T}_{ij} \\ \boldsymbol{T}_{ij} & \text{otherwise} \end{cases}$$

The diagonal entries of $\mathbf{P}$ are determined by the fact that the rows of $\mathbf{P}$ sum to 1. There are two cases,

- If $\pi_j \mathbf{T}_{ji} \leq \pi_i T_{ij}$

$$\pi_i \boldsymbol{P}_{ij} = \pi_i \boldsymbol{T}_{ij} a(i,j) = \pi_i \boldsymbol{T}_{ij} \left(\frac{\pi_j \boldsymbol{T}_{ij}}{\pi_i \boldsymbol{T}_{ij}} \right) = \pi_j \boldsymbol{T}_{ji} = \pi_j \boldsymbol{P}_{ji}$$

- If $\pi_j \mathbf{T}_{ji} < \pi_i \boldsymbol{T}_{ij}$

$$\pi_i \boldsymbol{P}_{ij} = \pi_i \boldsymbol{T}_{ij} a(i,j) = \pi_j \boldsymbol{T}_{ji} \left(\frac{\pi_i \boldsymbol{T}_{ij}}{\pi_j \boldsymbol{T}_{ji}} \right) = \pi_j \boldsymbol{T}_{ji} a(j,i) = \pi_j \boldsymbol{P}_{ji}$$

Hence, the detailed balance equations are satisfied. □

MCMC is quite difficult in general to analyze, but it is simple to implement.

Example 18: Statisical decipher

A beautiful example of Metropolis Hastings in action is the following deciphering algorithm. Suppose one has an alphabet $\mathcal{A}$, for example, all lower case Roman characters and the

space character. A *cipher* is just a permutation $\sigma : \mathcal{A} \to \mathcal{A}$. Now suppose we have a string S which is just some sequence from the alphabet $\mathcal{A}$, and suppose we observe $P = \widehat{\sigma}(S)$, where the map $\widehat{\sigma}$ is simply applied character-by-character. Our goal is to decipher this message. Suppose we at least know that the message S is in English. Now we can relax the problem of finding $\widehat{\sigma}$ to the problem of *sampling* a random σ so that $\sigma(S)$ looks like English.

Now we can invent a distribution on permutations that approximately solves this problem. One simple way to do that is to look at adjacent letter frequencies. So suppose we go to large English-language text, and compute the frequencies $F(a,b)$ with which an adjacent letter pair "ab" appears in the text. Now we for a given cipher σ we compute

$$L(\sigma) := \prod_j F(\sigma(P_j), \sigma(P_{j+1})),$$

where the index j runs over the length of P (less 1) and P_j is the character in the j-th position of the string P.

Now we can use Metropolis-Hastings to sample from the distribution π which is proportional to L. We let the proposal chain be one which at each step chooses two characters $a, b \in \mathcal{A}$ uniformly at random and then swaps $\sigma(a)$ and $\sigma(b)$. We then implement the Metropolis-Hastings algorithm as above. See: `https://github.com/elliotpaquette/Math447stuff/blob/main/AustenDecoder.ipynb`

(Bibliographic note) [10]

While Theorem 20 shows that MCMC (and specifically the Metropolis Hastings algorithm) converges, actually estimating the rate of convergence is a much more difficult (and moreover knowing when to stop).

There are many general methods for bounding the statistical distance to stationary. [11] We show a simple example which illustrates that there are problems for which MCMC can work great and also ones for which it fails hard.

[10] This example originates in (Persi Diaconis. "The markov chain monte carlo revolution". In: *Bulletin of the American Mathematical Society* 46.2 [2009], pp. 179–205), which is a beautiful invitation to MCMC. Credit is also due to Robert Dobrow (Dobrow, *Introduction to stochastic processes with R*) for compiling the data used for this example.

[11] See (Levin and Peres, *Markov chains and mixing times*) for a development of various methods.

Example 19: Glauber dynamics on the Ising Model

The Ising Model is a simple model of magnetization. It models the spins of nearby electrons in a lattice as $\{\pm 1\}$ states, say a finite subset of $\mathbb{Z}^2$, but mathematically we just need any undirected graph (V, E). The statespace of the lattice is $S = \{\pm 1\}^V$.

In the Ising model, spins interact with their nearest neighbors in that lattice. Spins prefer to be aligned, and so for a spin configuration $\sigma \in S$, we define its energy

$$\mathcal{H}(\sigma) = -\sum_{\{v,w\}\in E} \sigma_v \sigma_w.$$

So spins that agree reduce the energy of the configuration and spins that disagree increasse the energy.

The Ising model is a statistical spin distribution given by $\pi(\sigma) \propto e^{-\beta \mathcal{H}(\sigma)}$, where β is the inverse-temperature of the system (at high temperatures, the thermal energy overwhelms the structure of the system and all spin configurations are almost equally likely, while at low temperatures, the system freezes around low-energy configurations).

To model the effect of external interactions, one can put boundary conditions around the lattice region, freezing the spins to have some definite structure: in the figures these are arranged as half $+1$, half -1 in the outline of a bar magnet.

There is a theorem due to Peierls/Griffiths12[o] which show that when the system is sufficiently cold ($\beta > \beta_c$), this planar magnetism model "magnetizes", which in effect shows that the spin configuration concentrates around a deterministic configuration ($+1$ on the left, -1 on the right – which minimizes the energy).

There is a natural dynamics associated to this spin-system, the *Glauber dynamics* or *heat-bath dynamics*, which can be understood as a model of heat-induced random perturbations of the spin system. In this case, we choose a (non-boundary) site $v \in V$ uniformly at random, and then we update the spin at site v according to the distribution of $\sigma(v)$ when drawn from π and conditioned on all spins $(\sigma(w) : w \in V \setminus \{v\})$.

In short, we swap the spin v to 1 with probability

$$p(\sigma) = \frac{e^{\beta \tilde{H}}}{e^{\beta \tilde{H}} + e^{-\beta \tilde{H}}} \quad \text{where} \quad \tilde{H}(\sigma) = \sum_{w \sim v} \sigma_w,$$

with $\sim$ meaning adjacency. We set it to -1 otherwise. This

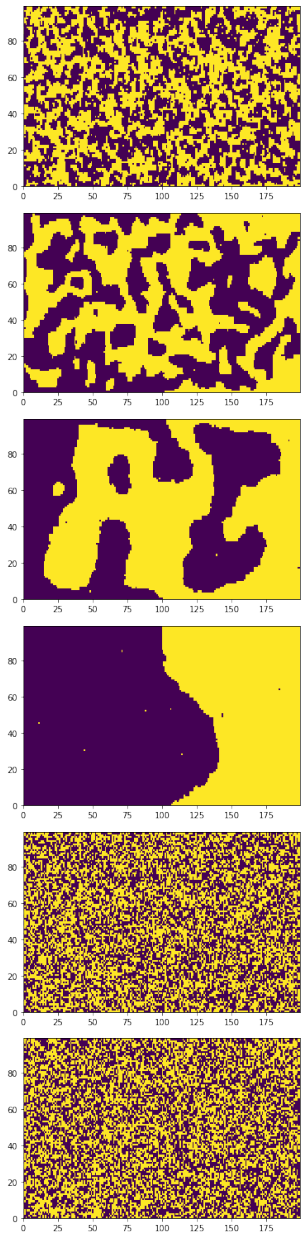


Figure 1: $\beta = 0$ (top-4) versus $\beta = 0.1$ (bottom-2). Boundary conditions are ± 0 on the left/right half. 5×10^j, for $j = 4, 5, 6, 7$ iterations on top; 5×10^j, for $j = 5, 6$ iterations on bottom.

is similar, but not equal, to the Metropolis-Hastings dynamics for the same chain.

Lemma 10 (Mixing time of the high-temperature Ising model): We suppose that $(X_n : n \geq 0)$ and $(Y_n : n \geq 0)$ are two copies of the Ising model Glauber dynamics. We suppose for simplicity that all interior vertices of the graph (on which the Glauber dynamics is active) have degree D and N vertices. Then if

$$\tanh(\beta) < \frac{D+1}{2D},$$

the chains couple in $\mathcal{O}_\beta(\log N)$ time.

Proof. Now we actually can produce a *grand coupling*, which is to say we build a simultaneous realization of all Markov chains from all states at all starting times. For a given initial state Σ, we set $X_n^{(\Sigma)}$ as the version of this chain started from Σ.

We define the dynamics by always choosing the same sites v_n to flip across all chains. Then we sample iid Unif$([0,1])$ random variables $(U_n : n \geq 0)$, which we will use to drive the flipped spins. This means that if we set

$$p_v(\sigma) = \frac{e^{\beta \tilde{H}_v}}{e^{\beta \tilde{H}_v} + e^{-\beta \tilde{H}_v}} \quad \text{where} \quad \rho(X_{n+1}, Y_{n+1})\tilde{H}_v(\sigma) = \sum_{w \sim v} \sigma_w,$$

then we update each spin configuration according to

$$X_{n+1} = \begin{cases} +1 & \text{if } U_n \leq p_{v_n}(X_n) \\ -1 & \text{if } U_n > p_{v_n}(X_n). \end{cases}$$

Let ρ be the *Hamming distance*, which is to say $\rho(X_n, Y_n)$ is the number of nonequal spins of X_n and Y_n.

If $m = \rho(X_n, Y_n)$ is the number of coordinates that do not match, then we can find a sequence of spins $\Sigma^{(j)}$ of length $m+1$ each differing from the previous by exactly 1 spin. Let $X^{(j)}$ be the Markov chain (from time n onwards) in the grand coupling started from the j-th spin $\Sigma^{(j)}$.

Then by the triangle inequality for all times beyond n,

$$\rho(X_n, Y_n) = \sum_{j=1}^{m} \rho(X_n^{(j)}, X_n^{(j+1)})$$

Now for each j, at time n, $(X_n^{(j)}, X_n^{(j+1)})$ differ in exactly 1 spin, at a site $v^{(j)} \in V$. If the site $v^{(j)}$ is selected by the Glauber dynamics, the two chains couple, and $\rho(X_n^{(j)}, X_n^{(j+1)}) = 0$. If a neighbor of $v^{(j)}$

[12] See Rudolf Peierls. "On Ising's model of ferromagnetism". In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 32. 3. Cambridge University Press. 1936, pp. 477–481 and Robert B Griffiths. "Peierls proof of spontaneous magnetization in a two-dimensional Ising ferromagnet". In: *Physical Review* 136.2A (1964), A437

is selected, then there is a chance the Hamming distance increases. If neither is selected, the Hamming distance stays the same. So if N is the number of vertices,

$$\mathbb{E}[\rho(X_{n+1}^{(j)}, X_{n+1}^{(j+1)}) \mid \mathscr{F}_n] = (1 - \tfrac{D+1}{N}) + \tfrac{2}{N} \sum_{w \sim v^{(j)}} |p_w(X_n^{(j)}) - p_w(X_n^{(j+1)})|$$

Now by calculus, we check

$$|p_w(X_n^{(j)}) - p_w(X_n^{(j+1)})| \leq \max_{s \in \mathbb{R}} \left| \tfrac{e^{\beta(s+2)}}{e^{\beta(s+2)} + e^{-\beta(s+2)}} - \tfrac{e^{\beta s}}{e^{\beta s} + e^{-\beta s}} \right| \leq \tanh(\beta).$$

In all

$$\mathbb{E}[\rho(X_{n+1}^{(j)}, X_{n+1}^{(j+1)}) \mid \mathscr{F}_n] \leq (1 - \tfrac{D+1}{N}) + \tfrac{2D \tanh(\beta)}{N}.$$

Hence, we get that

$$\mathbb{E}[\rho(X_{n+1}, Y_{n+1}) \mid \mathscr{F}_n] \leq \big(1 + \tfrac{2D \tanh(\beta) - D - 1}{N}\big)\rho(X_n, Y_n).$$

Thus if we iterate this inequality, we get

$$\mathbb{E}[\rho(X_n, Y_n)] \leq N\big(1 + \tfrac{2D \tanh(\beta) - D - 1}{N}\big)^n.$$

Hence under the condition $2D \tanh(\beta) < D + 1$, this converges to 0 and moreover, when $n = C \log N$ for large enough C, this is less than $\frac{1}{2}$. □

Martingales

The martingale is a fundamental stochastic process which is essential for basically all modern probability theory, whether it is for stochastic processes or otherwise. Even for the study of Markov chains, we need some of these techniques, and so we pause the theory of Markov chains.

(Discrete time) Martingales are processes satisfying the two following properties:

Definition 32 (Martingale): A Martingale $(M_n : n \geq 0)$ adapated to a filtration $(\mathscr{F}_n : n \geq 0)$ is a real-valued stochastic process satisfying:

1. $\mathbb{E}|M_n| < \infty$ for all $n \geq 0$.
2. $\mathbb{E}(M_{n+1} \mid \mathscr{F}_n) = M_n$.

Thus martingales are processes for which the best guess of their next position is right where they are.

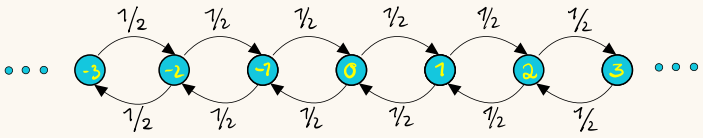
The canonical example of a martingale is simple symmetric random walk:

Example 20: 1-d SSRW

The 1-dimensional simple symmetric random walk on $\mathbb{Z}$ is the markov chain $(X_j : j \geq 0)$ for which

$$X_j = \begin{cases} 1 + X_{j-1}, & \text{with } \Pr(\cdot \mid \mathscr{F}_{j-1}) = \frac{1}{2}. \\ -1 + X_{j-1}, & \text{with } \Pr(\cdot \mid \mathscr{F}_{j-1}) = \frac{1}{2}. \end{cases}$$

This is to say the process has iid increments with distribution Unif$(\{1, -1\}$.



However the power martingales is that they describe many other processes beyond simple random walk. Martingales can be pulled from thin air: all one needs is a filtration (or even built, see Exercise 14). One of the key examples is the following:

Example 21: Doob Martingale

Let Y be any real valued random variable with $\mathbb{E}|Y| < \infty$ and let $(\mathscr{F}_n : n \geq 0)$ be any filtration. Then $M_n := \mathbb{E}(Y \mid \mathscr{F}_n)$

is a martingale (by the tower property of conditional expectations).

This is only an interesting martingale if the filtration and the random variable are connected in some interesting way, such as in the following.

Example 22: hitting

Suppose that $(X_n : n \geq 0)$ is an irreducible finite state markov chain with two absorbing states a and b. Let $Y = \mathbf{1}_{X \text{ absorbed at a}}$, and let $(\mathscr{F}_n : n \geq 0)$ be the filtration generated by $(X_n : n \geq 0)$ Then the associated martingale is:

$$\begin{aligned} M_n &:= \mathbb{E}(Y \mid \mathscr{F}_n) \\ &= \Pr((X_m : m \geq 0) \text{ absorbed at } a \mid \mathscr{F}_n). \end{aligned}$$

Exercise 14 (Binary splitting): Suppose Y is a real-valued random variable with $\mathbb{E}|Y| < \infty$. Define the following sequences: $(X_n : n \geq 0)$, $(Y_n : n \geq 0)$, and $(\mathscr{F}_n : n \geq 0)$. Here (X_n) and (Y_n) are stochastic processes and $(\mathscr{F}_n)$ is a filtration. These are constructed inductively. Set $X_0 = 0$, and $\mathscr{F}_0 = \{0, \Omega\}$ and $Y_0 = \mathbb{E}Y$. Then provided these have been constructed for some n, we define

$$X_{n+1} = \mathbf{1}_{Y > Y_n}, \ \mathscr{F}_{n+1} = \sigma(X_0, \ldots, X_{n+1}), \text{and } Y_{n+1} = \mathbb{E}[Y \mid \mathscr{F}_{n+1}].$$

Hence $(Y_n : n \geq 0)$ is a Doob martingale. Show that for every $n \in \mathbb{N}$, conditioning on $(X_1, X_2, \cdots, X_n)$, the distribution of Y is the same as Y conditioned to lie in some interval $(a, b]$ (one of which may be ∞ and both of which depend on the binary sequence $(X_1, X_2, \cdots, X_n)$), and that these intervals are disjoint for different binary sequences $(X_1, X_2, \cdots, X_n)$. *Hint: induction. Remark: this construction gives a sequence of finitely supported approximations (Y_n) to Y. We will show later that these converge to Y.*

It is also helpful to extend the definition of martingale to processes in which the equality in Definition 33 is rather an inequality. [13]

Definition 33 (Submartingale): A submartingale $(M_n : n \geq 0)$ adapated to a filtration $(\mathscr{F}_n : n \geq 0)$ is a real-valued stochastic process satisfying:

1. $\mathbb{E}|M_n| < \infty$ for all $n \geq 0$.

[13] Remembering the direction of the inequality is really hard. The nomenclature comes from complex function theory, where it mirrors subhmarmonic/superharmonic functions. It may be helpful to think: "sub" means the process is below its predicted value. Or perhaps: martingale betting stratgies all get ruined, supermartingale strategies get ruined even faster!

2. $\mathbb{E}(M_{n+1} \mid \mathscr{F}_n) \geq M_n$.

A process $(M_n : n \geq 0)$ is a *supermartingale* if $(-M_n : n \geq 0)$ is a submartingale.

Exercise 15 (convex): Suppose that $\phi : \mathbb{R} \to \mathbb{R}$ is convex and that $(M_n : n \geq 0)$ is a martingale. Show that if $(\phi(M_n) : n \geq 0)$ has finite expectation, then it is a submartingale. If further ϕ is nondecreasing, then the same holds if $(M_n : n \geq 0)$ is a submartingale

Exercise 16 (The h-transform): Let $(X_n : n \geq 0)$ be a homogeneous time Markov chain with tpm $\boldsymbol{P}$. Let $\phi : S \to \mathbb{R}$ be bounded. Fix an $n \in \mathbb{N}$. Define $h(k, x) := \mathbb{E}(\phi(X_n)|X_k = x)$ for $k \leq n$.

1. Show that
$$M_k = h(k \wedge n, X_{k \wedge n})$$
is a Martingale. A function $h : \mathbb{N} \times S \to \mathbb{R}$ with this property is called a spacetime-harmonic function.

2. Fix a state $x \in S$ with $p = \Pr(X_n = x) > 0$. If we take $\phi(y) := \mathbf{1}_{x=y}\frac{1}{p}$, then the law $\mathbb{Q}(\cdot) := \mathbb{E}(\mathbf{1}_{(\cdot)}\phi(X_n))$ is of a chain conditioned to end at x after n steps. Check that under $\mathbb{Q}(\cdot)$, $(X_n : n \geq 0)$ is an inhomogeneous Markov chain and give its tpms in terms of $\boldsymbol{P}$ and h. *Remark: A Markov chain conditioned to start at x and end at y is called a bridge.*

3. Find the tpms of a 1-d SSRW bridge conditioned to start and end at 0 after $2n$ steps.

Predictable processes, the Doob decomposition, and the bracket

Another method for manufacturing martingales is the *Doob decomposition*.

Definition 34 (Predictable): A stochastic process $(X_n : n \geq 0)$ is *predictable* if X_0 is deterministic and X_n is $\mathscr{F}_{n-1}$–measurable for all $n \in \mathbb{N}$.

(Note)[14]
Any adapted process can be made into a martingale:

[14] This implies adaptedness, but moreover, it means that at the n-th step, you could have determined the process available in the $(n-1)$-st.

Theorem 26: Doob decomposition

Any real-valued process $(X_n : n \geq 0)$ having $\mathbb{E}|X_n| < \infty$ for all n and adapted to a filtration $(\mathscr{F}_n : n \geq 0)$ can be uniquely decomposed as $X_n = M_n + A_n$ where $M_0 = 0$, $(M_n : n \geq 0)$ is a martingale and $(A_n : n \geq 0)$ is predictable. Moreover

$$A_n = \mathbb{E}X_0 + \sum_{j=1}^{n} \mathbb{E}(X_j - X_{j-1} \mid \mathscr{F}_{j-1}).$$

The process $(A_n : n \geq 0)$ is called the *compensator* of $(X_n : n \geq 0)$.

Exercise 17 (Downwards Doob): Check that a process is a sub-martingale if and only if its compensator is almost surely non-decreasing.

The bracket process is an important is an important special case. [15] Define

Definition 35 (Bracket process): For a martingale $(M_n : n \geq 0)$, the *bracket process* $[M_n]$ is the compensator of M_n^2, i.e.

$$\begin{aligned}[M_n] &= \mathbb{E}M_0^2 + \sum_{j=1}^{n} \mathbb{E}(M_j^2 - M_{j-1}^2 \mid \mathscr{F}_{j-1}) \\ &= \mathbb{E}M_0^2 + \sum_{j=1}^{n} \mathbb{E}((M_j - M_{j-1})^2 \mid \mathscr{F}_{j-1}).\end{aligned}$$

[15] This is going to intuitively represent the accumulated ammount of "randomess" of a martingale. This measure can be skeweed to be larger than in some sense it should be if the second moments of increments of the martingale barely exist (or do not exist at all!) in which case this is not really useful. So it is almost always appears paired with the condition that $|M_j - M_{j-1}| \leq 1$ almost surely, which is more helpful.

Predictable processes also play an important role as "betting strategies." One way to conceptualize a martingale $(M_n : n \geq 0)$, or rather the stochastic process of increments $(M_n - M_{n-1} : n \geq 1)$, is as the winnings from playing a *fair game*. [16]

In this case a predictable process can play the role of *wager size*, which is to say the amount the player wishes to bet in the n-th step. In this case

[16] For example: the return for betting 1 dollar on a fair coin flip, where the payout is the amount bet, in which case the payout is 1 or -1 with probability 1/2–i.e. 1-d SSRW.

Definition 36 (Discrete stochastic integral): The (discrete) *stochastic integral* $(M \circ A)$ of two $(\mathscr{F}_n : n \geq 0)$–adapted processes is the stochastic process

$$(M \circ A)_n = M_0 A_0 + \sum_{j=1}^{n} (M_j - M_{j-1}) A_j$$

for $n \geq 0$, which is again adapted.

Moreover, this process remains a (sub)-martingale in some cases:

Lemma 11: Suppose $(M_n : n \geq 0)$ is a $(\mathscr{F}_n : n \geq 0)$–adapted process and $(A_n : n \geq 0)$ is a $(\mathscr{F}_n : n \geq 0)$–predictable process with $|A_n|$ almost surely bounded for each $n \geq 0$:

1. If M is a martingale, then $M \circ A$ is a martingale.
2. If M is a submartingale and $A \geq 0$ then $M \circ A$ is a submartingale.

In the context of the betting strategy interpretation, this means that regardless of how you choose to bet, your winnings remain a martingale.

The application of this that we'll use the most frequently is the *stopped process*. [17]

Definition 37 (Stopped process): If $(X_n : n \geq 0)$ is a stochastic process and τ is a stopping time, then the process X^τ given by $X_n^\tau := X_{\tau \wedge n}$ is called the *stopped process*. Setting $A_n := \mathbf{1}_{\tau \geq n}$, then $X^\tau = X \circ A$

Note that by construction, A_n is predictable, as to check $\tau \geq n$, you just have to check that τ did *not* occur at times $0, \cdots, n-1$.

$$\begin{aligned} X \circ A &= X_0 \mathbf{1}_{\tau \geq 0} + \sum_{j=1}^{n} (X_j - X_{j-1}) \mathbf{1}_{\tau \geq j} \\ &= X_0 + \sum_{j=1}^{n \wedge \tau} (X_j - X_{j-1}). \end{aligned}$$

Corollary 3: If M is a martingale and τ a stopping time, M^τ is again a martingale. The same holds for submartingales and supermartingales.

As betting strategies, this naturally represents stopping criteria: i.e. you have to know when to walk away! [18]

[17] The notation $a \wedge b := \min\{a, b\}$ is handy for working with stochastic processes. There is also $a \vee b := \max\{a, b\}$. The direction of the carat is the same as conjunction (logical and) and disjunction (logical or).

[18] It's also important to know when to hold 'em and when to fold 'em.

Optional stopping

Theorem 27: Optional stopping

Suppose that for a submartingale M and a stopping time τ one of the following three conditions holds:

1. $\tau \leq K$ almost surely for some constant $K > 0$,
2. $\tau < \infty$ almost surely and $|M| \leq 1$ almost surely, or
3. $\mathbb{E}\tau < \infty$ and $|M_n - M_{n-1}| \leq 1$ almost surely.

Then for any stopping time σ with $\sigma \leq \tau$ almost surely,

$$\mathbb{E}M_0 \leq \mathbb{E}M_\sigma \leq \mathbb{E}M_\tau,$$

and if furthermore $(M_n : n \geq 0)$ is a martingale, then the above is an equality.

(Note)[19]

[19] As a matter of formulation, the constant 1 in parts 2 and 3 can be replaced by any other positive real number.

Proof. Note that the constant $\sigma = 0$ is a stopping time, so it suffices to show $\mathbb{E}M_\sigma \leq \mathbb{E}M_\tau$ under the conditions given. The process $M^\tau - M^\sigma$ is again a submartingale, and hence $\mathbb{E}(M_n^\tau - M_n^\sigma)$ is increasing in n and so is always larger than 0. As τ is almost surely bounded, there is some K deterministic so that $\Pr(\tau > K) = 0$. Hence taking $n = K$, we have $M_K^\tau = M_\tau$ almost surely.

The latter two are applications of dominated convergence, applied to the bounded stopping times $\tau \wedge K$ and $\sigma \wedge K$ as $K \to \infty$. □

Example 23: Null–recurrence 1d-SSRW

We can use this to analyze SRW. Let $(X_n : n \geq 0)$ be 1-d SSRW started from 0. Let τ_0^+ be the time of first return, and let τ_a for $a \in \mathbb{N}$ be the hitting time of $\{a, -a\}$. Now $\tau_a < \infty$ almost surely (from any state $x \in [-a, a]$ there is a probability p that in a stepes that the SRW hits one of these. So in ka steps, the probability it still hasn't hit is at most $(1-p)^k \ldots$). Hence $\tau_a \wedge \tau_0^+$ is finite almost surely and so we can apply Theorem 27. Now to get something from it, we should let time advance 1 step. The process $(X_n : n \geq 1)$ is still a martingale, and so:

$$\mathbb{E}[X_{\tau_a \wedge \tau_0^+} | X_1 = 1] = 1.$$

Now

$$\mathbb{E}[X_{\tau_a \wedge \tau_0^+} | X_1 = 1] = a \Pr(\tau_a < \tau_0^+ | X_1 = 1) + 0 \Pr(\tau_a > \tau_0^+ | X_1 = 1).$$

Hence

$$\Pr(\tau_a < \tau_0^+ | X_1 = 1) = \frac{1}{a}.$$

The same claim conclusions holds if $X_1 = -1$, and so we have unconditionally

$$\Pr(\tau_a < \tau_0^+) = \frac{1}{a}.$$

There are two conclusions from this: the first is that 1-d SSRW is recurrent, as each τ_a is finite almost surely, and so

$$\Pr(\tau_0^+ = \infty) < \Pr(\tau_a < \tau_0^+) = \frac{1}{a}.$$

Second, while it is finite, there is a probability of at least $1/a$ that the process takes a steps or longer to return to 0, and hence

$$\sum_{a=1}^{\infty} \Pr(\tau_0^+ \geq a) \geq \sum_{a=1}^{\infty} \frac{1}{a} = \infty.$$

Therefore, this process is null-recurrent.

Exercise 18 (Exit): Use optional stopping to compute the following:

1. The probability of SSRW first exiting the integer interval $[a, b]$ at b if started at some $x \in (a, b)$.
2. The expected time of SSRW to first exit $\{a, b\}$ given it starts at some $x \in (a, b)$. *Hint: look at a martingale made from* X^2.

Example 24: Pattern matching (20°)

Suppose we look for the first occurrence of some pattern in a sequence of random letters. Let $\mathcal{A}$ be a finite alphabet, and let $p = (p_\alpha : \alpha \in \mathcal{A})$ be a probability vector. Suppose we are interested in the first appearance of a word $w = w_1 w_2 \cdots w_k$. We can use optional stopping to find how long it takes to see the word in a sequence $(X_n : n \geq 1)$ of iid samples from $\mathcal{A}$ with probability vector p, which we can assume puts positive probability on every letter of the alphabet.
Let τ be the first hitting time of the word w, i.e.

$$\tau = \min\{n : X_{n-j} = w_{k-j}, \forall\, 0 \leq j < k\}.$$

To define the martingale, we suppose that Rivendell casino offers you a fair game betting on the outcome of the next digit, i.e. if you bet \$1 on a, and a occurs, you win $1/p_a$. Now since w is your lucky word, you just bet on sequences that look like w. So at every step, you bet \$1 on w_1 appearing (it could be the start of your lucky word!). If you win, you then "double-down" and bet your winnings of $1/p_{w_1}$ of $1/p_{w_1}$ on w_2. If your lucky word appears, you just put your winnings in your pile.
Now this is a betting strategy on a martingale (or in fact a combination of many martingales), and so it actually has to be a martingale. These basic martingales have increments:

$$Y_{n+1}^{(a)} - Y_n^{(a)} := \mathbf{1}_{X_{n+1}=a} - \tfrac{1}{p_a}.$$

On the letter w_1, we always bet 1. On the letter w_2, we bet $1/w_1$, but only if the current letter is w_1. Hence, for each $1 \leq j \leq k$ we have a betting strategy:

$$A_n^{(j)} = \prod_{\ell=1}^{j-1} \frac{\mathbf{1}_{X_{n-j+\ell}=w_\ell}}{p_{w_\ell}}.$$

Our data winnings can be represented by

$$M_n = \sum_{j=1}^{k} \left(Y^{(w_j)} \circ A^{(j)}\right)_n,$$

with $M_0 = 0$.
Now the increments of $Y^{(a)}$ are all bounded, and our betting strategies are bounded, so to apply optional stopping, it suffices to show $\mathbb{E}\tau < \infty$. In every consecutive block of k letters, we have a positive probabilility $q = \prod_j p_{w_j}$ of seeing the pattern. This is in every consecutive block, and so

$$\Pr(\tau > mk) \leq (1-q)^m,$$

which implies τ has finite expectation.
Thus

$$0 = \mathbb{E}M_0 = \mathbb{E}M_\tau.$$

The first time that τ occurs, every betting line we had started from before time $\tau - k$ has failed (meaning we have no winnings). We also have winnings coming from all the successful bets that started (inclusively) at times $j > \tau - k$. So our winnings at time τ is (since if τ happens we know precisely what the last k letters are!)

$$\sum_{j=1}^{k} \prod_{\ell=1}^{j} \frac{\mathbf{1}_{w_{k-j+\ell}=w_\ell}}{p_{w_\ell}}$$

We have bet a total of τ dollars, as every betting line cost \$1, and so

$$M_\tau = \sum_{j=1}^{k} \prod_{\ell=1}^{j} \frac{\mathbf{1}_{w_{k-j+\ell}=w_\ell}}{p_{w_\ell}} - \tau.$$

Thus we conclude from optional stopping

$$\mathbb{E}\tau = \sum_{j=1}^{k} \prod_{\ell=1}^{j} \frac{\mathbf{1}_{w_{k-j+\ell}=w_\ell}}{p_{w_\ell}}.$$

For example, the expected time to see the pattern HTH in a string of fair coins is $2^3 + 2 = 10$, while the expected time to see HTT is just $2^3 = 8$.

[20] This construction is due to Shuo-Yen Robert Li. "A Martingale Approach to the Study of Occurrence of Sequence Patterns in Repeated Experiments". In: *The Annals of Probability* 8.6 (1980), pp. 1171 –1176. DOI: 10.1214/aop/1176994578. URL: https://doi.org/10.1214/aop/1176994578.

One lesson to draw from this example is that to find the expected time of a Markov chain S_n to hit a state, it is helpful to construct a martingale M_n of the form $f(S_n) - n$. The Markov chain S_n in this pattern matching example is $(X_n, X_{n-1}, \ldots, X_{n-k})$.

Exercise 19 (SRW): Let $(X_n : n \geq 0)$ be biased 1-d SRW started at 0 (so $X_{n+1} - X_n$ is 1 or -1 with probability p or $1-p$, respectively).

1. Find $a(\beta)$ so that $M_n := e^{\beta X_n - na(\beta)}$ is a martingale for any $\beta \in \mathbb{R}$. *This is called the exponential martingale.*
2. Compute $\mathbb{E}_1 s^{\tau_0^+}$ for $s \in (0,1)$. *Be careful that you are applying optional stopping correctly.* (21[o])
3. Use this to compute the probability of never returning in the case $p > \frac{1}{2}$.

[21] From this expression, in the case that $p = \frac{1}{2}$, it is possible to conclude that $\Pr_1(\tau_0^+ = k) \asymp \frac{1}{k^{3/2}}$ in the case of SSRW (the symbol $a_k \asymp b_k$ means there are constants c, C so that $ca_k \leq b_k \leq Ca_k$ for all $k \geq k_0$). More precisely one shows that the number of simple random walk paths first returning to 0 after $2k$ steps is given by the k-th Catalan number.

One of the big applications of optional stopping are *maximal inequalities*, which give stochastic bounds for the maximum (in time) of a martingale. The basic idea is contained in the following argument:

Proof. Suppose that we have a martingale $(M_n : n \geq 0)$ and suppose that for some $a > 0$, τ is the first time that $|M_n| \geq a$. Then $|M_n|$ is a submartingale (as it is a convex function of a martingale), and by Optional Stopping

$$\mathbb{E}|M_{\tau \wedge n}| \leq \mathbb{E}|M_n|$$

Now on the event that $\tau \leq n$, we have $|M_{\tau \wedge n}| \geq a$, from which it follows that

$$\Pr(\tau \leq n) a \leq \mathbb{E}|M_{\tau \wedge n}| \leq \mathbb{E}|M_n|.$$

The probability on the left is nothing but $\Pr(\max_{0 \leq k \leq n} |M_k| \geq a)$, and so we have proven the following. □

Lemma 12 (Doob Maximal inequality): For any $(M_n : n \geq 0)$ and any $a > 0$ and all $n \geq 1$

$$\Pr(\max_{0 \leq k \leq n} |M_k| \geq a) \leq \frac{\mathbb{E}|M_n|}{a}.$$

[22] This same argument also applies to non-negative submartingales $(X_n : n \geq 0)$ (in place of $|M_n|$) and the positive part of submartingales (which is non-negative a submartingale). We can also replace the absolute value function $|\cdot|$ by any convex function ϕ. It's convenient to refer to all of these versions of this argument as Doob's maximal inequality.

Now this is a useful inequality itself, but the idea of the proof is even more important than the statement. [22]

One of the classic applications of the maximal inequality is the classic strong law of large numbers.

Corollary 4 (Kolmogorov Strong Law): Suppose that $(X_n : n \geq 1)$ are independent and satisfy $\sum_{n=1}^{\infty} \frac{\mathrm{Var}(X_n)}{n^2} < \infty$ then $\frac{1}{n}\sum_{k=1}^{n}(X_k - \mathbb{E}X_k) \xrightarrow[n\to\infty]{\text{a.s.}} 0$.

[23] This can be used to prove the usual strong law of large numbers for iid random variables $(Y_n : n \geq 0)$ with $\mathbb{E}|Y_1| < \infty$ by looking at the truncations $X_n = Y_n \mathbf{1}_{|Y_n| \leq n}$, which was Kolmogorov's original proof.

(Note)[23]

Proof. Note that without loss of generality, we can assume that all $\{X_n\}$ are mean 0. Set $S_n := \frac{1}{n}\sum_{k=1}^n X_k$. By a direct application of Borel–Cantelli, and Chebyshev's inequality, the variance condition implies[24] that along the subsequence $n_k = 2^k$,

$$\Pr(|S_{n_k}| > \delta \quad \text{i.o.}) = 0. \tag{3}$$

[24] If checking this, use the form $\Pr(|X| > a) \le \mathbb{E}\frac{(|X|^2\mathbf{1}_{|X|>a})}{a^2}$, *then* apply Borel Cantelli, and use Fubini to carefully rearrange the sum and finally compare to the summability condition.

Now we need to fill in the gaps between $S_{n_{k+1}}$ and S_{n_k}. Let $M_n = \sum_{\ell=n_k}^n X_\ell$, which is a martingale. By the Doob maximal inequality (applied to $|M_n|^2$)

$$\Pr(\max_{n_k \le n \le n_{k+1}} |M_n|^2 \ge a^2) \le \frac{\mathbb{E}(M_{n_{k+1}}^2)}{a^2} = \frac{1}{a^2}\sum_{\ell=n_k}^{n_{k+1}} \operatorname{Var}(X_\ell).$$

Then

$$\Pr(\max_{n_k \le n \le n_{k+1}} |\tfrac{n}{n_k}S_n - S_{n_k}| \ge \delta) \le \frac{1}{\delta^2 n_k^2}\sum_{\ell=n_k}^{n_{k+1}} \operatorname{Var}(X_\ell) \le \frac{4}{\delta^2}\sum_{\ell=n_k}^{n_{k+1}} \frac{\operatorname{Var}(X_\ell)}{\ell^2}.$$

So by another application of Borel-Cantelli,

$$\Pr(\{\max_{n_k \le n \le n_{k+1}} |\tfrac{n}{n_k}S_n - S_{n_k}| \ge \delta\} \text{ i.o.}) = 0.$$

By combining this with the bound (3), we conclude $|S_n| > 2\delta$ i.o. with probability 0. As this holds for any $\delta > 0$, we have shown the almost sure convergence. □

Lemma 13 (L^p-Maximal inequality): For any martingale $(M_n : n \ge 0)$, any $p > 1$ and all $n \ge 1$

$$\mathbb{E}(\max_{0 \le k \le n} |M_k|^p) \le \left(\frac{p}{p-1}\right)^p \mathbb{E}(|M_n|^p).$$

Proof. Set $M_n^* = \max_{0\le k\le n} |M_k|$ and fix a cutoff $K > 0$. Applying Fubini's theorem and using τ as the first time that $|M_n| \ge \lambda$

$$\begin{aligned}
\mathbb{E}(M_n^* \wedge K)^p &= \int_0^\infty p\lambda^{p-1} \Pr(M_n^* \wedge K \ge \lambda)\, \mathrm{d}\lambda \\
&= \int_0^K p\lambda^{p-2}\mathbb{E}(|M_{\tau\wedge n}|\mathbf{1}_{\tau\le n})\, \mathrm{d}\lambda \\
&= \int_0^K p\lambda^{p-2}\mathbb{E}(|M_n|\mathbf{1}_{\tau\le n})\, \mathrm{d}\lambda \\
&= \int_0^K p\lambda^{p-2}\mathbb{E}(|M_n|\mathbf{1}_{M_n^*\ge\lambda})\, \mathrm{d}\lambda.
\end{aligned}$$

[25] Then interchanging the order of integration, we get

$$\mathbb{E}(M_n^* \wedge K)^p \le \frac{p}{p-1}\mathbb{E}(|M_n|(M_n^* \wedge K)^{p-1}).$$

[25] In the third line, we used that

$$\begin{aligned}
&\mathbb{E}([|M_n| - |M_{\tau\wedge n}|]\mathbf{1}_{\tau\le n}) \\
&= \mathbb{E}([|M_n| - |M_{\tau\wedge n}|]) \ge 0,
\end{aligned}$$

as on the event $\tau > n$ the random variable $|M_n| - |M_{\tau\wedge n}| = 0$.

Applying Hölder's inequality with exponent $p/(p-1)$

$$\mathbb{E}\big(M_n^* \wedge K\big)^p \leq \frac{p}{p-1}\big(\mathbb{E}|M_n|^p\big)^{1/p}\big(\mathbb{E}(M_n^* \wedge K)^p\big)^{1-1/p}.$$

Now we can divide both sides by the appropriate power of $\mathbb{E}\big(M_n^* \wedge K\big)^p < \infty$, and conclude

$$\mathbb{E}\big(|M_n^* \wedge K|^p\big) \leq \left(\frac{p}{p-1}\right)^p \mathbb{E}\big(|M_n|^p\big).$$

Taking $K \to \infty$ and applying monotone convergence, the theorem follows. □

The reflection principle

The cases above give bounds for the maximum of martingales. For the case of SSRW, it is actually possible to go a step further and actually compute the distribution function of the maximu, using what is known as the *reflection principle*.

We can slightly generalize the setup to random walks with symmetric step distributions. A real valued random variable X is symmetric if $X \overset{\text{law}}{=} -X$. Let $S_n = \sum_{j=1}^n X_j$.

Theorem 28: Reflection principle

Suppose that $\{X_n\}$ are independent, symmetric random variables. Let $S_n = \sum_{j=1}^n X_j$. For all $n \geq 1$ and for all $t > 0$,

$$\Pr\big(\max_{1\leq j\leq n} S_j \geq t\big) \leq 2\Pr\big(S_n \geq t\big).$$

Proof. Let τ be the first time j that $S_j \geq t$. Then

$$\Pr\big(\max_{1\leq j\leq n} S_j \geq t\big) = \Pr(\tau \leq n).$$

Now we can define the evil twin random walk E_j by

$$E_n := \sum_{j=1}^n X_j(-1)^{\mathbf{1}\{\tau<j\}}.$$

In other words, the increments of E_n and S_n are the same up to τ, and after τ they are opposite.

Now $(E_n : n \in \mathbb{N})$ and $(S_n : n \in \mathbb{N})$ have the same law, but they are correlated! On the event $\tau \leq n$, we have that at least one of $E_n \geq t$ or $S_n \geq t$, and hence

$$\Pr(\tau \leq n) \leq \Pr(\{E_n \geq t\} \cup \{S_n \geq t\}) \leq 2\Pr(S_n \geq t),$$

using the union bound and equality in law. □

This is extremely sharp.

Martingale convergence

Part of the picture we want to develop for martingales is that in some sense, all martingales look the same: the only thing that changes is the speed at which they run. [26] The prototypical constant speed martingale is 1-d SSRW, which is null-recurrent. [27] Hence it has excursions that travel arbitrarily far from its starting point, but it nonetheless returns to where it starts albeit in infinite expected time. But it can also happen that the accumulated amount of randomness of martingale, over its entire lifetime, is finite. In this case, the martingale must converge.

The key idea to prove this is the "buy–low–sell–high" betting strategy. Let $a < b$ be two real numbers. So given a martingale M, we will design A by wagering whenever the process has decreased below a and is on its way back up! In mathematical terms, we design a sequence of stopping times

$$\alpha_1 \leq \beta_1 < \alpha_2 \leq \beta_2 < \cdots$$

by the following inductive rules. We say that α_1 is the first time the martingale crosses below a. We then let β_1 be the first time after α_1 that the process crosses above b. Then, for $k \geq 2$ we define

$$\alpha_k = \inf\{n > \beta_{k-1} : M_n < a\} \quad \text{and} \quad \beta_k = \inf\{n > \alpha_{k-1} : M_n > b\}.$$

We then let $A_n = \sum_{k=1}^{\infty} \mathbf{1}_{n \in [\alpha_k, \beta_k]}$, which is to say we wager on (M_n) when the process crosses below a and we sell it once it goes back above b.

We define the number of *upcrossings*:

$$N_n := \max\{k \leq n : \beta_k < \infty\},$$

where we formally take $\beta_0 = 0$, so that $N_n \geq 0$ for all n. We show the following:

Lemma 14: Suppose that $(M_n : n \geq 0)$ is a submartingale. Then

$$(b-a)\mathbb{E}N_\infty \leq \sup_n \mathbb{E}((M_n - a)_+ - (M_0 - a)_+).$$

Proof. Let $Y_n := a + (M_n - a)_+$. By Exercise 15, if M is a submartingale, then so is Y. Set

$$(Y \circ A)_n = \sum_{k=1}^{N_n} (Y_{\beta_k} - Y_{\alpha_k}) + (Y_{n \wedge \alpha_{N_n}} - Y_{\alpha_{N_n}}).$$

Then by how Y is chosen,

$$(Y \circ A)_n \geq (N_n)(b-a).$$

[26] There is an important caveat for discrete-time martingales: it can happen that the increments of a martingale are so wild, that in a single step, a near-eternity of randomness has passed. (For example the martingale could have an increment of infinite variance). Otherwise said, for such a martingale, the actual martingale structure is not sufficiently fine-grained to be interesting. So discrete-time martingales often need additional structure to be interesting: two prominent examples are (1) the increments are bounded (a.s.) or (2) the martingale is almost surely positive (this implicitly bounds the increments).

[27] Note that its bracket process $[X_n] = n$–the bracket can be used as a measurement of accumulated randomness.

Taking expectation, we have

$$\mathbb{E}(Y \circ A)_n \geq \mathbb{E}(N_n)(b-a).$$

Now $Y \circ A = Y - Y \circ (1-A)$, and so

$$\mathbb{E}(Y \circ A)_n = \mathbb{E}Y_n - \mathbb{E}Y_0.$$

Taking $n \to \infty$ and applying monotone convergence, the lemma follows. □

Theorem 29: Martingale Convergence

Suppose $(M_n : n \geq 0)$ is a submartingale with $\sup_n \mathbb{E}(M_n)_+ < \infty$ then there is a random variable M_∞ with $\mathbb{E}|M_\infty| < \infty$ so that

$$M_n \xrightarrow[n\to\infty]{\text{a.s.}} M_\infty.$$

Proof. Using

$$(b-a)\mathbb{E}N_\infty \leq \sup_n \mathbb{E}((M_n - a)_+ - (M_0 - a)_+),$$

and using that

$$(M_n - a)_+ \leq (M_n)_+ + |a|,$$

we have that

$$\sup_n \mathbb{E}((M_n - a)_+ - (M_0 - a)_+) \leq 2|a| + \mathbb{E}|M_0| + \sup_n \mathbb{E}(M_n)_+ < \infty$$

We therefore have $\mathbb{E}N_\infty < \infty$ and hence $N_\infty < \infty$ almost surely. Thus for all pairs of rationals $a < b$, we have that the number of upcrossings over (a,b) is finite almost surely.

Now on the event $\liminf_n M_n < \limsup M_n$, there are two rational numbers $a < b$, so that

$$\liminf_n M_n < a < b < \limsup M_n,$$

but then it would follow that the number of (a,b)–upcrossings is infinite. And so we have that almost surely $\liminf_n M_n = \limsup_n M_n$, and so $M_\infty = \limsup_n M_n$ is the almost sure limit of M_n.

To see that it is finite, first observe that, by the submartingale property

$$\mathbb{E}M_0 \leq \mathbb{E}M_n = \mathbb{E}((M_n)_+ - (M_n)_-),$$

and hence

$$\mathbb{E}(M_n)_- \leq -\mathbb{E}M_0 + \mathbb{E}(M_n)_+.$$

It follows that

$$\sup_n \mathbb{E}|M_n| < -\mathbb{E}M_0 + 2\sup_n \mathbb{E}(M_n)_+ < \infty.$$

By Fatou's Lemma

$$\liminf_n \mathbb{E}|M_n| \geq \mathbb{E}|M_\infty|$$

Hence $\mathbb{E}|M_\infty| < \infty$. □

(Note) [28]

One of the most useful special cases is when the the process is just positive:

Corollary 5: A nonnegative supermartingale $(M_n : n \geq 0)$ converges almost surely to a nonnegative random variable M_∞ with $\mathbb{E}M_\infty < \infty$.

Example 25: Gambler's Ruin

Suppose $(M_n : n \geq 0)$ is a non-negative integer valued martingale. Then $M_n \xrightarrow[n\to\infty]{\text{a.s.}} M_\infty$. This means that there is an $N \in \mathbb{N}$ random so that $M_n = M_N$ for all $n > N$, as this is the only way that an integer-valued sequnce can converge. So if $(M_n : n \geq 0)$ represents the winnings of a gambler, then the gambler almost surely stops playing at some point, either because they quit ($M_\infty > 0$) or because they were ruined ($M_\infty = 0$).
(Note that if $M_n = 0$ at some time n, then $M_k = 0$ for all larger times as for a non-negative martingale, $\mathbb{E}[M_{n+1} \mid \mathscr{F}_n] = 0$ implies that M_{n+1} is 0 almost surely).

Exercise 20 (Binary splitting Converges): Suppose Y is a real-valued random variable with $\mathbb{E}|Y| < \infty$. Define the following sequences: $(X_n : n \geq 0)$, $(Y_n : n \geq 0)$, and $(\mathscr{F}_n : n \geq 0)$. Here (X_n) and (Y_n) are stochastic processes and $(\mathscr{F}_n)$ is a filtration. These are constructed inductively. Set $X_0 = 0$, and $\mathscr{F}_0 = \{0, \Omega\}$ and $Y_0 = \mathbb{E}Y$. Then provided these have been constructed for some n, we define

$$X_{n+1} = \mathbf{1}_{Y > Y_n},\ \mathscr{F}_{n+1} = \sigma(X_0, \ldots, X_{n+1}), \text{and } Y_{n+1} = \mathbb{E}[Y \mid \mathscr{F}_{n+1}].$$

Hence $(Y_n : n \geq 0)$ is a Doob martingale. Show that $Y_n \xrightarrow[n\to\infty]{\text{a.s.}} Y$.
Hint: To rule out that $\lim Y_n > Y$, *suppose* Y *is in an interval* $[a, a + \epsilon)$ *which has positive probability* δ *and suppose* $Y_n > a + \epsilon$. *Now estimate how much closer to a* Y_n *will get using* ϵ, δ *and the current location. Use this to rule out that* $Y_n > Y$ *for all n sufficiently large.*

Exercise 21 (Hitting a line): Using optional stopping, and martingale convergence, bound the probability that 1-d SSRW started

[28] This shows that while *a priori*, the assumption in the Theorem is weaker than assuming expected absolute values of M_n are uniformly bounded, for submartingales, it's actually the same.

at 0 ever crosses the line $n \mapsto a + bn$ for $a > 0$ and $b \in (0,1)$ by

$$e^{-\lambda(b)(a+1)} \leq \Pr(\text{crossing}) \leq e^{-\lambda(b)a},$$

where $\lambda(b)$ is appropriately chosen (this can be computed using the exponential martingale).

Example 26: The Pólya Urn

This is one of the most central examples: of Martingale convergence, of time-inhomogeneous Markov chains, of (critically-tuned) self reinforcing behavior, and to boot, it has foundational statistical applications.
The urn can described by running the following Markovian procedure. Suppose after n steps, an urn as (R_n, B_n) balls in it. Now sample a ball from the urn uniformly at random, and add a ball to the urn of the same type as was selected. Then

$$(R_{n+1}, B_{n+1}) = \begin{cases} (R_n + 1, B_n) & \text{with } \Pr(\cdot \mid \mathscr{F}_n) = \frac{R_n}{R_n + B_n}, \\ (R_n, B_n + 1) & \text{with } \Pr(\cdot \mid \mathscr{F}_n) = \frac{B_n}{R_n + B_n}. \end{cases}$$

Noticing that the number of balls always increases by n, one may actually instead just record the red ball probability $p_n := \frac{R_n}{R_n + B_n}$, and note that the denominator can just be expressed as $n + R_0 + B_0$. Now besides being a Markov chain, it turns out that $(p_n : n \geq 0)$ is actually a non-negative martingale. So by martingale convergence, there exists a random variable p_∞ so that $p_n \xrightarrow[n\to\infty]{\text{a.s.}} p_\infty$.

Exercise 22 (Pólya): Suppose that $\theta \overset{\text{law}}{=} \text{Unif}([0,1])$. Let $(X_n : n \geq 1)$ be iid random variables with law Bernoulli(θ).

1. Show that, with $S_n := 1 + \sum_{j=1}^{n} X_j$, $(S_n : n \geq 0)$ has the same law as $(R_n : n \geq 0)$ where $R_0 = B_0 = 1$.
2. Show that if we define $(p_n : n \geq 0)$ in terms of the R_n defined above, then $p_\infty = \theta$.
3. Show that if we condition on $(X_k : 1 \leq k \leq n)$, and that there are a successes and b failures, then the law of p_∞ is Beta$(a+1, b+1)$.

(Note)[29]
As promised, another structure under which the discrete martin-

[29] The Pólya urn process is foundational to Bayesian statistics. Without getting too deep into the meaning, suppose we are in the business of trying to determine the success probability θ of a biased coin coming up heads by repeatedly flipping it. Having never flipped it, we might assert that any θ is equally likely, which is to say that $\theta \overset{\text{law}}{=} \text{Unif}([0,1])$. The $(p_n : n \geq 0)$ of the Pólya urn describe the natural estimator you would make for θ based on the first n coin flips. Now martingale convergence shows that $p_\infty = \theta$. If we want to decide when to stop in a structured way, we might stop when $\mathbb{E}(\ell(p_n - p_\infty) \mid \mathscr{F}_n) < \epsilon$ for some loss function ℓ (this expectation is called the risk). The law of $p_\infty \mid \mathscr{F}_n$ (the "posterior") can be used to compute this risk. A more exotic loss may even be measured in terms of the whole path $(p_k : k \geq n)$, which is then described by the Pólya urn (for example – imagine that your investors lose confidence in you when your estimator oscillates by more than 10% in a short window).

gale structure is useful is when the increments of a random walk are bounded.

Theorem 30: Bracket process & convergence

Suppose that $(M_n : n \geq 0)$ is a martingale. By monotonicity, $[M]_\infty := \lim_{n\to\infty}[M]_n$ exists almost surely (but may be infinite).

1. On the event $[M]_\infty < \infty$, $M_n \xrightarrow[n\to\infty]{\text{a.s.}} M_\infty$, which exists and is finite almost surely.
2. If furthermore $|M_n - M_{n-1}| \leq 1$ almost surely, then on the event $[M]_\infty = \infty$,

$$\limsup M_n = \infty \text{ a. s.} \quad \text{and} \quad \liminf M_n = -\infty \text{ a. s.}.$$

Proof. For the first claim, let τ be the first time $[M]_{n+1} \geq R$, which is still a stopping time, owing to the fact that $[M]$ is predictable. Then setting $Y = M^\tau$, $Y^2 - [Y]$ is a martingale and $[Y]_n < R$ for all n. So by the martingale propery, for all $n \in \mathbb{N}$

$$\mathbb{E}Y_n^2 = \mathbb{E}[Y]_n \leq R.$$

Thus we can apply martingale convergence (Theorem 29), and we conclude that $\lim_{n\to\infty} Y_n$ exists a.s. So, on the event that $[M]_\infty < R$, we conclude that $\lim M_n$ exists and is the limit of Y_n. As the event $[M]_\infty < \infty$ is the union of the events $[M]_\infty < R$ over integer R, we can construct the limit R-by-R and also conclude the convergence.

For the second claim, fix an $a \in \mathbb{R}$. We may assume wlog that $M_0 = 0$, or else we apply the argument to $(M_n - M_0 : n \geq 0)$. We further assume $a > 0$, or else we negate the martingale. Now we alternate between excursions above and below a, letting $\{\tau^{(j)}\}$ be defined inductively by

$$\tau^{(2j)} = \inf\{k \geq \tau^{(2j-1)} : M_k > a\}$$
$$\tau^{(2j+1)} = \inf\{k \geq \tau^{(2j)} : M_k < a\}.$$

Let $\sigma_R^{(j)}$ be the first time n after $\tau^{(j)}$ that $|M_n - M_{\tau^{(j)}}| \geq R$.

Suppose $\tau^{(j)} < \infty$ (taking for $j = 0$, $\tau^{(0)} = 0$) Now $Y_n := M_n - M_{n\wedge\tau^{(j)}}$ is a martingale, and $Y^2 - [Y]$ is a martingale. The bracket $[Y]$ is just

$$[Y] = [M] - [M^{\tau^{(j)}}],$$

which still tends to ∞ on the event $[M]_\infty = \infty$.

Now it cannot be that Y is bounded on the event $[Y]_\infty = \infty$ (which is the same event as $[M]_\infty = \infty$), since stopping at $\sigma = \sigma_R^{(j)}$ (the first

time that $|Y_n| > R$)

$$\mathbb{E}(Y^2 - [Y])_{\sigma\wedge n} = 0,$$

and so

$$(R+1)^2 \geq \mathbb{E}[Y]_{\sigma\wedge n}.$$

But if the event $\{\sigma = \infty\} \cap [Y]_\infty = \infty$ has positive probability, the right hand side goes to ∞ as $n \to \infty$. As R was arbitrary we must have that on the event $[Y]_\infty = \infty$, $\sup_n |Y_n| = \infty$ almost surely.

Let $Y = M^2 - [M]$ and let $\vartheta = \tau^{(j+1)} \wedge \sigma_R^{(j)}$. The stopped process Y^ϑ is a bounded martingale and so converges. As Y is unbounded on the event $[Y]_\infty = \infty$, we therefore have $Y_\infty^\vartheta = Y_\vartheta$ almost surely on the event $[Y]_\infty = \infty$ By optional stopping, [30]

$$0 = \mathbb{E}(Y_\vartheta \mid \mathscr{F}_{\tau^{(j)}}) \geq R\Pr(\{\sigma_R^{(j)} < \tau^{(j+1)}\} \cap \{[Y]_\infty = \infty\} \mid \mathscr{F}_{\tau^{(j)}}) - (1+a).$$

Thus

$$\Pr(\{\sigma_R^{(j)} < \tau^{(j+1)}\} \cap \{[Y]_\infty = \infty\}) \leq \frac{a+1}{R}.$$

By taking $R \to \infty$, it follows that

$$\Pr(\{\tau^{(j+1)} = \infty\} \cap \{[Y]_\infty = \infty\}) = 0.$$

Hence we have shown that on the event $[M]_\infty = \infty$, the process visits a neighborhood of $[a-1, a+1]$ infinitely many times almost surely. It follows that the lim sup and lim inf of the process are both ∞.[31] □

[30] The subtracted term accounts for the case that $\tau^{(j+1)}$ happened first. The first excursion leads to the extra factor of a.

[31] In fact, as the increments of the martingale are all at most 1, visiting intervals $[a-1, a+1]$ infinitely often for all $a \in \mathbb{R}$ is equivalent to the statement on lim sup and lim inf.

One nice corollary of these convergences is a major upgrade to Borel-Cantelli (which is especially helpful in showing that things *do* occur infinitely often). [32]

> **Corollary 6 (Better Borel Cantelli):** Suppose for there is sequence of events $(B_n : n \geq 1)$ adapted to some filtration $(\mathscr{F}_n : n \geq 0)$. Then
>
> $$\sum_{n=1}^\infty \mathbf{1}_{B_n} = \infty \quad \text{a. s.} \Leftrightarrow \sum_{n=1}^\infty \mathbb{E}(\mathbf{1}_{B_n} \mid \mathscr{F}_{n-1}) = \infty \quad \text{a. s.}$$

[32] Both first and second Borel Cantelli lemmas follow from this statement. The first follows from taking expectations of the right hand side (and hence if the probabilities of B_n summable, then finitely many B_n occur almost surely). The second lemma follows as if the events are independent, then taking $\mathscr{F}_n$ to be the natural filtration generated by the sequence, the conditional expectations are expectations.

Proof. Let $X_k = \sum_{n=1}^k \mathbf{1}_{B_n}$, and let M_k and A_k be the martingale and predictable parts of X_k. Then the compensator and bracket process are given by

$$A_k = \sum_{n=1}^k p_n \quad \text{and} \quad [M_k] = \sum_{n=1}^k (p_n - p_n^2), \quad \text{where} \quad p_n := \mathbb{E}(\mathbf{1}_{B_n} \mid \mathscr{F}_{n-1}).$$

Note that these p_n are random variables with $0 \leq p_n \leq 1$, and note that in this notation we are trying to show that

$$\sup_n X_n = \infty \Leftrightarrow \sup_n A_n = \infty.$$

Now suppose we look at the event $[M_\infty] < \infty$, which could happen if $p_n \to 0$ or $p_n \to 1$ sufficiently quickly. In this case, the martingale M_k converges, in which case

$$\sup_n X_n = \infty \Leftrightarrow \sup_n A_n = \infty,$$

as the martingale part is bounded.

We turn to working on the event $\{[M_\infty] = \infty\}$. Observe that $[M_k] \leq A_k$ almost surely, and so $\sup_n A_n = \infty$. It follows that $\limsup_n M_n = \infty$ (from Theorem 30) and so

$$\limsup_n X_n = \limsup_n (M_n + A_n) = \limsup_n M_n + \lim_{n\to\infty} A_n = \infty \text{ a. s.}$$

As X_n is increasing, it follows that $\sup_n X_n = \infty$. □

Example 27: Uniform division

Suppose that we take $(X_n : n \geq 1)$ iid uniform on $[0,1]$. These divide the interval $[0,1]$ into subintervals, by letting $(I_k^{(n)} : 0 \leq k \leq n)$ be the $(n+1)$ intervals with endpoints given by the set $(X_k : 1 \leq k \leq n)$.
Let A_n be the length of the largest interval amonst $(I_k^{(n)} : 0 \leq k \leq n)$. While $(A_n : n \geq 0)$ do *not* form a Markov chain (since when the largest interval is divided, we may need to know the length of the second-largest interval at time n to decide A_{n+1} after division), it is nonetheless true that at every step, the largest interval is divided with probability A_n.
We claim that the largest interval gets divided infinitely often, almost surely. By better Borel-Cantelli, it suffices to check

$$\sum_{n=1}^{\infty} \Pr(A_{n+1} < A_n \mid \mathscr{F}_n) = \sum_{n=1}^{\infty} A_n = \infty \quad \text{a. s.}$$

By the pigeonhole principle, $A_n \geq 1/n$, and so this sum diverges.

Exercise 23 (Discrete OU): Recall that the discrete OU process

$$X_{n+1} = \sqrt{1-\alpha} X_n + \sqrt{\alpha} Z_n,$$

for an iid sequence of standard Gaussians $(Z_n : n \geq 0)$ and a fixed $\alpha \in (0,1)$. Show that with probability 1,

$$\limsup_{n\to\infty} |X_n| = \infty.$$

Hint: show that for any $R > 0$, the process goes above R infinitely often.

Pathologies and Uniform integrability

While all martingales in some sense look like simple random walk, we can have situations in which the martingale can be made to look more and more like a constant, at least in the sense of in-probability convergence.

Example 28: Slow restarts

Consider the inhomogeneous time Markov chain (X_n) on $\mathbb{Z}$ which at step k has the following rule. If at 0, it transitions to $\{1, -1\}$ with probability $1/(2k)$ respectively, or otherwise stays at 0. If at $x \neq 0$, it transitions to $2x$ with probability $1/2$ and to 0 with probability $1/2$. The resulting Markov chain is a martingale. When it leaves 0, it spends a geometrically distributed amount of time before returning. When at 0, it tends to spend longer and longer times, and hence we actually have that $X_n \xrightarrow[n\to\infty]{\Pr} 0$. It on the other hand does not converge almost surely.

Exercise 24 (Slow restarts): In this exercise, you will prove the claims hold in the example above.

1. Suppose that $n, m \in \mathbb{N}$, show that

$$\lim_{m\to\infty} \liminf_{n\to\infty} \Pr(X_{n+m} = 0 \mid X_n) = 1 \quad \text{a.s.}$$

   Hint: Break this into cases, according to value of X_n, and then give lower bounds for the probability in terms of m and n. The memoryless property of the geometric random variable is helpful.

2. Using the bound above, show that $\lim_{n\to\infty} \Pr(X_n = 0) = 1$.

3. Show by Better Borel-Cantelli that for some $m \in \mathbb{N}$ sufficiently large, $X_{km} \neq 0$ for infinitely many k, almost surely (and hence $\neg(X_n \xrightarrow{\text{a.s.}} 0)$).

In the previous sections we developed some convergence properties under control on the increments (either one-sided, as in Theorem 29 or increment control, as in 30).[33] If the increments are too large, (or equivalently, the filtration is too discontinuous), we can create martingales that "look" wrong. In these cases, the martingale property is basically not meaningful.

[33] The finiteness of the bracket implies the second moments have finite increments

Exercise 25 (Off to ∞): Construct a martingale $(M_n : n \geq 0)$ with independent (non iid), mean 0 increments such that

> $M_n \xrightarrow{\text{a.s.}} \infty$.*Hint: necessarily, the increments must become more and more spread out.*

Both of these examples have limits that in some sense "do not look" like the martingale itself. There is a related question of *right closability* of a martingale, which is to say when it is possible to find an $M_\infty \in L^1(\Pr)$ so that not only does $M_n \xrightarrow[n\to\infty]{\text{a.s.}} M_\infty$, but in fact $M_n = \mathbb{E}(M_\infty \mid \mathscr{F}_n)$.

It turns out this has a complete answer, and it is connected to the idea of uniform integrability.

> **Definition 38 (Uniform integrability):** A collection of random variables $(X_\alpha : \alpha \in \mathcal{A})$ is uniformly integrable if and only if for every $\epsilon > 0$ there is an $M > 0$ so that
>
> $$\sup_\alpha \mathbb{E}(|X_\alpha| \mathbf{1}_{|X_\alpha|>M}) \leq \epsilon.$$

Uniform integrability is strictly stronger than $\sup_\alpha \mathbb{E}|X_\alpha| < \infty$. Conversely, by Markov's inequality, it is implied by $\sup_\alpha \mathbb{E}\psi(|X_\alpha|) < \infty$ for any increasing function $\psi : (0,\infty) \to (0,\infty)$ with $\psi(x)/x \xrightarrow[x\to\infty]{\text{a.s.}} \infty$. In particular, families of random variables that are bounded in $L^2(\Pr)$ are uniformly integrable.

The key value of uniform integrability is that it characterizes when the expectations of random variables converge:

> **Lemma 15 (UI to L1):** If $(X_n : n \geq 0)$ is a sequence of real valued random variables having expectations and converging almost surely to X_∞ having finite expectation, then $\lim_{n\to\infty} \mathbb{E}|X_n - X_\infty| = 0$ if and only if the family $(X_n : n \geq 0)$ is uniformly integrable.

Proof. Suppose the family is uniformly integrable. Then for any $\epsilon > 0$, there is an $M > 0$ so that

$$\sup_n \mathbb{E}(|X_n| \mathbf{1}_{|X_n|>M}) \leq \epsilon.$$

By Fatou's lemma,

$$\epsilon > \liminf_{n\to\infty} \mathbb{E}(|X_n| \mathbf{1}_{|X_n|>M}) \geq \mathbb{E}(|X_\infty| \mathbf{1}_{|X_\infty|>M}).$$

Thus setting $Y_n = X_n \mathbf{1}_{|X_n|\leq M}$ for $n \geq 0$ and $n = \infty$, by dominated convergence

$$\lim_{n\to\infty} \mathbb{E}(|Y_n - Y_\infty|) = 0.$$

On the other hand

$$\begin{aligned}\mathbb{E}(|X_n - X_\infty|) &\leq \mathbb{E}(|Y_n - Y_\infty|) + \mathbb{E}|X_n \mathbf{1}_{|X_n|>M}| + \mathbb{E}|X_\infty \mathbf{1}_{|X_\infty|>M}| \\ &\leq \mathbb{E}(|Y_n - Y_\infty|)2\epsilon.\end{aligned}$$

Hence we conclude that

$$\limsup_{n\to\infty} \mathbb{E}(|X_n - X_\infty|) \leq 2\epsilon,$$

but the ϵ is arbitrary, and so we can now take $\epsilon \to 0$.

For the reverse implication, fix an $\epsilon > 0$. There is an N sufficiently large that for all $n > N$,

$$\mathbb{E}(|X_n - X_\infty|) \leq \epsilon/3 \quad \text{and} \quad \mathbb{E}(|X_\infty|\mathbf{1}_{|X_n - X_\infty|\geq 1}) \leq \epsilon/3.$$

By dominated convergence,

$$\lim_{M\to\infty} \mathbb{E}(\sum_{n=1}^{N}(|X_n|\mathbf{1}_{|X_n|>M})) = 0,$$

and similarly,

$$\lim_{M\to\infty} \mathbb{E}(|X_\infty|\mathbf{1}_{|X_\infty|>M-1}) = 0,$$

Thus we can pick an M large enough that both of these are less than $\epsilon/3$. It now follows that with this M, for $n \leq N$ we have

$$\mathbb{E}(|X_n|\mathbf{1}_{|X_n|>M}) \leq \epsilon/3 \leq \epsilon$$

and for $n > N$

$$\begin{aligned}\mathbb{E}(|X_n|\mathbf{1}_{|X_n|>M}) \leq &\mathbb{E}(|X_n - X_\infty|) \\ &+\mathbb{E}(|X_\infty|\mathbf{1}_{|X_n - X_\infty|\geq 1}) \\ &+\mathbb{E}(|X_\infty|\mathbf{1}_{|X_\infty|>M-1}) \leq \epsilon.\end{aligned}$$

□

A martingale $(M_n : n \geq 0)$ is uniformly integrable if the family of random variables $(M_n : n \geq 0)$ is uniformly integrable. Uniformly integrable martingales are precisely right closable, or equivalently, they are exactly the Doob martingales corresponding to some random variable M_∞ and some filtration $(\mathscr{F}_n : n \geq 0)$.

Theorem 31: These are the Doobs

A martingale $(M_n : n \geq 0)$ is uniformly integrable if and only if there is an M_∞ of finite expectation so that $M_n \xrightarrow[n\to\infty]{\text{a.s.}} M_\infty$ and for all $n \geq 0$

$$M_n = \mathbb{E}(M_\infty \mid \mathscr{F}_n).$$

Proof. Suppose $(M_n : n \geq 0)$ is uniformly integrable. Then by definition of uniform integrability, with $\epsilon = 1$, there is an $M > 0$ so that

$$\sup_n \mathbb{E}|M_n| \leq \sup_n \mathbb{E}(|M_n|\mathbf{1}_{|M_n|>M}) + \sup_n \mathbb{E}(|M_n|\mathbf{1}_{|M_n|\leq M}) \leq 1 + M.$$

Hence by martingale convergence, there is an M_∞ of finite expectation so that $\mathbb{E}|M_\infty| < \infty$ and so that $M_n \xrightarrow[n\to\infty]{\text{a.s.}} M_\infty$. Morover from uniform integrability and from Lemma 15,

$$\lim_{n\to\infty} \mathbb{E}|M_n - M_\infty| = 0.$$

Now for any n and any $N > n$

$$M_n = \mathbb{E}(M_N \mid \mathscr{F}_n),$$

and

$$\mathbb{E}|\mathbb{E}(M_N \mid \mathscr{F}_n) - \mathbb{E}(M_\infty \mid \mathscr{F}_n)| \leq \mathbb{E}|\mathbb{E}(|M_N - M_\infty| \mid \mathscr{F}_n)| = \mathbb{E}|M_N - M_\infty|.$$

Thus we have, for any $N > n$

$$\mathbb{E}|M_n - \mathbb{E}(M_\infty \mid \mathscr{F}_n)| \leq \mathbb{E}|M_N - M_\infty|.$$

Sending $N \to \infty$, we conclude that $M_n = \mathbb{E}(M_\infty \mid \mathscr{F}_n)$ almost surely.

The converse is a consequence of the following general fact. □

Lemma 16 (Conditional expectations preserve ui): Suppose $(\mathscr{F}_\alpha : \alpha \in \mathcal{A})$ is any family of sub-σ-algebras in a probability space $(\Omega, \mathscr{F}, \Pr)$ and suppose Y is a real-valued random variable with $\mathbb{E}|Y| < \infty$. Then the collection

$$(\mathbb{E}(Y \mid \mathscr{F}_\alpha) : \alpha \in \mathcal{A})$$

is uniformly integrable.

Proof. Let $Y_\alpha := \mathbb{E}(Y \mid \mathscr{F}_\alpha)$. The function $\phi_M(x) = (|x| - M)_+$ is convex, and so by conditional Jensen's inequality

$$\mathbb{E}\phi_M(Y_\alpha) \leq \mathbb{E}\phi_M(Y). \tag{4}$$

Likewise by Markov's inequality and conditional Jensen's inequality

$$\begin{aligned}\Pr(|Y_\alpha| \geq M) &\leq \mathbb{E}(|Y_\alpha|\mathbf{1}_{|Y_\alpha|>M})\tfrac{1}{M} \\ &\leq \mathbb{E}(\mathbb{E}(|Y| \mid \mathscr{F}_\alpha)\mathbf{1}_{|Y_\alpha|>M})\tfrac{1}{M} \\ &\leq \mathbb{E}(|Y|\mathbf{1}_{|Y_\alpha|>M})\tfrac{1}{M}.\end{aligned}$$

In the last line we have applied the definition of conditional expectation.

Now for any $\epsilon > 0$ there is an $K > 0$ sufficiently large that

$$\mathbb{E}(|Y|\mathbf{1}_{|Y|>K}) \leq \epsilon/3.$$

Then using Markov's inequality once more

$$\mathbb{E}(|Y|\mathbf{1}_{|Y_\alpha|>M}) \leq K\Pr(|Y_\alpha| > M) + \epsilon \leq K\mathbb{E}|Y|/M + \epsilon.$$

Thus we conclude

$$\Pr(|Y_\alpha| \geq M) \leq K\mathbb{E}|Y|/M^2 + \epsilon/(3M).$$

Using (4) and the previous line,

$$\begin{aligned}\mathbb{E}(|Y_\alpha|\mathbf{1}_{|Y_\alpha|>M} &\leq \mathbb{E}\phi_M(Y_\alpha) + M\Pr(|Y|_\alpha > M) \\ &\leq \mathbb{E}\phi_M(Y) + K\mathbb{E}|Y|/M + \epsilon/3.\end{aligned}$$

Now picking M sufficiently large that the first and second term are less than $\epsilon/3$, we conclude

$$\mathbb{E}(|Y_\alpha|\mathbf{1}_{|Y_\alpha|>M}) \leq \epsilon.$$

□

Uniform integrability, or the lack thereof, is one explanation for how a non-negative martingale can hit 0.

> **Exercise 26 (UI or Bust):** Suppose that $(M_n : n \geq 0)$ is a martingale which is positive almost surely. Show that $(M_n : n \geq 0)$ can be decomposed as $M_n = Z_n + Y_n$, two non-negative martingales, where $(Z_n : n \geq 0)$ is uniformly integrable and $Y_n \xrightarrow[n\to\infty]{\text{a.s.}} 0$.

The Pólya theorem & harmonic functions

In Example 23, we saw that SSRW on $\mathbb{Z}$ was null–recurrent. For physical reasons, it can be reasonable to look at higher–dimensional analogues of this process. That is, we consider the set $\mathbb{Z}^d$ as a graph, with nearest-neighbor adjacencies: i.e. for all $x, y \in \mathbb{Z}^d$

$$x \sim y \quad \text{iff} \quad \|x - y\|_1 = 1.$$

In other words, they are adjacent if and only if they differ in exactly one coordinate by 1. The SSRW on this graph is to select, at each step, a neighbor uniformly at random and move to it. In dimension-1, this agrees with the definition of SSRW already given.

In higher dimensions, which in some sense has more directions to move away from the origin, the process becomes more transient. Pólya's theorem crystallizes this in the following way:

Theorem 32: Pólya's Theorem

SSRW on $\mathbb{Z}^d$ is recurrent if and only if $d \leq 2$.

(Joke)[34]

This theorem has many proofs, and we will illustrate one which is more analytic in nature. But it is helpful to keep in mind the following heuristic argument, which explains the transition (and which can be turned into a proof). By the fundamental theorem of recurrence Theorem 17, it suffices to show that [35]

$$\sum_{n=1}^{\infty} \boldsymbol{P}_{o,o}^{2n} = \infty \quad \text{iff} \quad d \leq 2.$$

In dimension 1, the return probability is explicit, since

$$\boldsymbol{P}_{o,o}^{2n} = \Pr(\text{Binom}(2n, \tfrac{1}{2}) = n) = \binom{2n}{n} 2^{-2n} \asymp \frac{1}{\sqrt{n}}.$$

If you had independent coordinates, then you could say that the probability of returning to 0 is just the probability of *simultaneously* having d 1-dimensional random walks return to 0 in n steps, from which one would get $\boldsymbol{P}_{o,o}^{2n} \asymp n^{-d/2}$. This is non-summable when $d = \{1, 2\}$.

> **Exercise 27 (Return-Pr):** Directly argue in $d = 3$ that $\boldsymbol{P}_{o,o}^{2n} \asymp n^{-d/2}$ by exhibiting an exact expression for the return probability.

We pursue a different approach, based on harmonic functions. This will lead to a general strategy for showing transience & recurrence, called the method of Lyapunov functions.

> **Definition 39 (Harmonic functions):** A function $h : S \to \mathbb{R}$ is subharmonic for a time-homogeneous Markov chain $(X_n : n \geq 0)$ if for all $x \in S$ $\mathbb{E}_x|h(X_1)| < \infty$ and $\mathbb{E}_x(h(X_1)) = h(x)$. The function is *subharmonic* if $h(x) \leq \mathbb{E}_x(h(X_1))$ and *superharmonic* if $h(x) \geq \mathbb{E}_x(h(X_1))$.

Harmonic functions (resp. sub/super-harmonic functions) give rise to martingales (resp. sub/super-martingales), when applied to the Markov chain. If the harmonic function has some extra structure, then martingale convergence will tell us things about the behavior of the underlying chain.

[34] Famously: "the drunk man will always find his way home, but the drunk bird is lost forever." S Kakutani

[35] We also use that all the SSRWs on $\mathbb{Z}^d$ are 2-periodic.

Theorem 33: Bounded superharmonics

An irreducible THCS Markov chain is transient if and only if there exists a nonconstant, bounded superharmonic function $h : S \to \mathbb{R}$.

Proof. First, if if this condition holds, then we can apply martingale convergence to $M_n := h(X_n)$ and conclude that $M_n \xrightarrow[n\to\infty]{\text{a.s.}} M_\infty$ (for concreteness, let us say we have fixed the starting point of the chain at some $x \in S$). Now suppose x is recurrent, and that X_n visited x infinitely often. As h is nonconstant, there is another state y with $h(y) \neq h(x)$, and there is a positive probability (by irreducibility) for the chain to travel from x to y – we could choose our favorite way it could happen. Then by the Strong markov property, the chain would have to also travel from x to y infinitely often, and hence the value of $h(X_n)$ would not converge.

To show the converse, it suffices to construct for an irreducible THCS Markov chain, a bounded, nonconstant superharmonic function h. Let N_x be the number of visits to state x , and define

$$h(y) := \mathbb{E}_y N_x = \sum_{n=0}^{\infty} \boldsymbol{P}^n_{y,x},$$

Then $h(y) < \infty$ for all $y \in S \setminus \{x\}$, as by the Strong Markov property

$$\mathbb{E}_y N_x = \Pr_y(\tau_x < \infty)\mathbb{E}_x N_x < \infty,$$

which further shows that h is bounded. We just need that the chain is nonconstant. Note that $\mathbb{E}_y N_x \leq \mathbb{E}_x N_x$, and hence if h is constant, we must have $\Pr_y(\tau_x < \infty) = 1$ for all $y \in S \setminus \{x\}$. This would imply that $(X_n : n \geq 0)$ is recurrent, however, as by letting the chain evolve one step started from x, we would also have $\Pr_x(\tau_x < \infty) = 1$.

Finally, we note that h is in fact superharmonic, since [36]

$$\begin{aligned}\mathbb{E}_y(h(X_1)) &= \sum_{n=0}^{\infty} \mathbb{E}_y(\boldsymbol{P}^n_{X_1,x}) \\ &= \sum_{n=0}^{\infty} \boldsymbol{P}^{n+1}_{y,x} \\ &\leq h(y).\end{aligned}$$

□

[36] Note that this is very nearly harmonic: $\mathbb{E}_y(h(X_1)) = h(y)$ everywhere except at $y = x$, where $\mathbb{E}_x(h(X_1)) - h(x) = -1$. Hence we can view this h as a solution of the equation $\boldsymbol{I} - \boldsymbol{P} = \delta_x$, which is called the Laplacian of the Markov chain. The function h is called the Green's function.

This lets us show 1/2 of Pólya's theorem.

Lemma 17: SSRW on $\mathbb{Z}^d$ is transient for $d \geq 3$.

Proof. Most of the proof here is getting the right guess for h: verifying the proof is just a little bit of calculus.

The idea is to try to construct a harmonic function that looks like the example that appeared in the proof of Theorem 33, which is to say we would like an approximate guess for $\mathbb{E}_y N_x$. Furthermore, the behavior of our guess can be altered as we wish on any finite set, and so we really care about x and y separated.

Using the central limit theorem, we can approximate [37]

$$\Pr_y(X_n = x) \approx e^{-\|y-x\|_2^2/cn} n^{-d/2}.$$

[37] Ignoring constants and parity issues and the fact that we are actually computing the probability of a point, which is rather the local limit theorem.

Hence

$$\mathbb{E}_y N_x \approx \sum_{n=1}^{\infty} e^{-\|y-x\|_2^2/cn} n^{-d/2}.$$

For the first $\|y-x\|_2^2$ terms the Gaussian term is small. For larger terms, the Gaussian is irrelevant, and so we have the cartoon that with $m = \|y-x\|_2^2$

$$\mathbb{E}_y N_x \approx \sum_{n=m}^{\infty} n^{-d/2} \approx m^{1-d/2}.$$

So we're led to consider and h which is approximately $h_0(y) = (\|y\|_2^2)^{1-d/2}$ (setting $x = 0$).

Now to turn this into a proof, we do a Taylor approximation for $\|y\|$ large and $\alpha > 0$ fixed:

$$\begin{aligned}(\|y\|^2 + \pm 2y_j + 1)^{-\alpha} &= (\|y\|^2)^{-\alpha} \\ &\mp 2\alpha y_j (\|y\|^2)^{-1-\alpha} \\ &+ (4(1+\alpha)\alpha y_j^2 - 2\alpha\|y\|^2)(\|y\|^2)^{-2-\alpha} \\ &+ \mathcal{O}((\|y\|^2)^{-3/2-\alpha}).\end{aligned}$$

Using this, we conclude that

$$|\mathbb{E}_y(h_0(X_1) - h_0(y))| \le C\|y\|^{-d-1},$$

which is on the order of the error term.

Now that isn't directly useful to the proof, since it doesn't give a supermartingale. However, if we pick $\alpha = \frac{d}{2} - 1 - \epsilon$ for any ϵ, then the same Taylor computation shows that with $h_\epsilon(y) := (\|y\|_2^2)^{-(d/2-1-\epsilon)}$

$$\mathbb{E}_y(h_0(X_1) - h_0(y)) \le 0$$

for all $\|y\|_2 \ge C(\epsilon)$. Thus if we take $h(y) = \min\{h_\epsilon, \delta\}$ for some sufficiently small $\delta > 0$, we conclude

$$\mathbb{E}_y(h(X_1) - h(y)) \begin{cases} = 0 & \|y\| \text{ if } \|y\| \text{ is sufficiently small} \\ \le 0 & \text{otherwise .} \end{cases}$$

Hence h is a positive, nonconstant martingale, and so SSRW is transient. □

Recurrence can be done by an analogous strategy, but instead of looking for a bounded superharmonic function, we look for a subharmonic function whose every level set is finite and which has bounded increments.

Theorem 34: Potentials for recurrence

An irreducible THCS Markov chain is recurrent if there exists a function $h : S \to \mathbb{R}$ with $\mathbb{E}_y|h(X_1)| < \infty$ having finite and finite sub-level sets (which is to say $\{x : h(x) \leq M\}$ is finite for every M) and so that for all but finitely many $y \in S$, $h(y) \geq \mathbb{E}_y h(X_1)$.

Proof. Let F be the set of y which are exceptions to $h(y) \geq \mathbb{E}_y h(X_1)$. Let τ_F be the first hitting time of $(X_n : n \geq 0)$. Then if M_n is $h(X_n)$, the stopped process $M_n^{\tau_F}$ is a supermartingale which is bounded below. By irreducibility, there is no finite set A so that X_n is eventually in A, and hence from the finite-sublevel-set property, either $\tau_F < \infty$ or $\limsup M_n = \infty$. From Martingale convergence, it must therefore be that $\tau_F < \infty$.

□

This lets us complete Pólya's theorem.

Lemma 18: SSRW on $\mathbb{Z}^d$ is transient for $d = 2$.

Proof. Continuing the strategy, from the Proof of Lemma 17, we formally take d to 2 (this is critical) and consider

$$h_0(y) = \log(1 \vee \|y\|_2^2).$$

Then Taylor expanding this,

$$\log(\|y\|^2 \pm 2y_j + 1) = \log(\|y\|^2) \pm \frac{2y_j}{\|y\|^2} + \frac{1}{2}\frac{-4y_j^2 + 2\|y\|^2}{\|y\|^2} + \cdots$$

As before when taking expectation over a step of SSRW, the 1st,2nd,and 3rd order terms cancel (the 0th) and 4th survives. So we have for all y sufficiently large that there is some constant $C > 0$ so that

$$|\mathbb{E}_y h_0(X_1) - h_0(y)| \leq C\|y\|^{-4}.$$

Now this is not necessarily a superharmonic, but we can fix it to be superharmonic by subtracting a multiple of $\|y\|^{-2}$ (following the computation from the Proof of Lemma 17), so get that for some

$M > 0$ sufficiently large, there is a $C > 0$ so that if $\|y\| > C$, $h_0(y) - M\|y\|^{-2}$ is superharmonic.

□

Exercise 28 (Birth-death): A *birth-death* chain on $\mathbb{N}_0$ is a Markov chain with $p_{i,i+1} = p_i$ and $p_{i,i-1} = q_i = 1 - p_i$. To truly model death, we should make 0 an absorbing state, but it is actually helpful if we consider the *reflected* version that just jumps back to 1 (i.e. $p_0 = 1$). Show that if

$$\limsup_j (p_j - \tfrac{1}{2})j < \tfrac{1}{4}$$

then the chain is recurrent, while if

$$\liminf_j (p_j - \tfrac{1}{2})j > \tfrac{1}{4}$$

then the chain is transient. *Hint: functions of the form $\phi(x) = x^\alpha$ should be helpful.*

Stochastic approximation

Martingale convergence is one of the most powerful tools for showing the limit behavior of stochastic processes, such as the success probabilities in the Pólya urn. Now, the Pólya urn has a huge amount of special structure which one cannot hope to appear in more general contexts. However, martingale convergence is still generally the right basic tool, which we will illustrate here.

Suppose that $(X_n : n \geq 0)$ is a real-valued stochastic process adapted to a filtration $(\mathscr{F}_n : n \geq 0)$, and suppose that there are constants $\gamma_n > 0$ satisfying for some continuous F

$$\mathbb{E}(X_{n+1} \mid \mathscr{F}_n) = X_n + \gamma_n F(X_n).$$

This can be thought of as a type of stochastic algorithm, whose updates, in mean are a small multiple of $F(X_n)$.[38] Now provided we have some control on the fluctuations and provided γ_n is taken to 0, then we can actually ensure that this process converges to a 0 of F.

A *stochastic approximation* scheme is a general version of this, in which we can actually ensure that all of this happens. [39]

Theorem 35: Robbins-Monro condition

Suppose that F is continuous and has finitely many zeros $\{z_k\}$. Suppose that $(X_n : n \geq 0)$, $(\gamma_n : n \geq 0)$ are adapted

[38] If F is negative the derivative of some function G, then this is 1-dimensional "stochastic gradient descent," a process that on average decreases the value of G.

[39] This is adapted from (Robin Pemantle. "A survey of random processes with reinforcement". In: *Probability surveys* 4 [2007], pp. 1–79) which has lots of discussion. The original source for this is (Herbert Robbins and Sutton Monro. "A stochastic approximation method". In: *The annals of mathematical statistics* [1951], pp. 400–407).

to $(\mathscr{F}_n : n \geq 0)$ and satisfy

$$\sum_{n=1}^{\infty} |\mathbb{E}(X_{n+1} \mid \mathscr{F}_n) - (X_n + \gamma_n F(X_n))| < \infty \quad \text{a. s.}$$

Suppose that $\sum_n \gamma_n = \infty$ a. s. and suppose that

$$\sum_{n=1}^{\infty} \operatorname{Var}(X_{n+1} \mid \mathscr{F}_n) < \infty \quad \text{a. s.}$$

Then

$$\min_k |X_n - z_k| \xrightarrow[n\to\infty]{\text{a.s.}} 0.$$

This is to say that the process almost surely converges to the zero-set of F.

Proof. Define

$$\begin{aligned} R_n &:= \mathbb{E}(X_{n+1} \mid \mathscr{F}_n) - (X_n + \gamma_n F(X_n)), \\ \xi_{n+1} &:= X_{n+1} - (X_n + \gamma_n F(X_n) + R_n). \end{aligned}$$

Then the hypotheses can be expressed as

$$\sum_{n=1}^{\infty} |R_n| < \infty \quad \text{a. s.} \quad \text{and} \quad \sum_{n=1}^{\infty} \mathbb{E}(\xi_n^2 \mid \mathscr{F}_{n-1}) < \infty \quad \text{a. s.}$$

If we write a Doob decomposition of the stochastic process $(X_n : n \geq 0)$, then we have

$$X_n = M_n + A_n,$$

where M_n is given by the partial sum of ξ_n and A_n is the partial sum of $\gamma_n F(X_n) + R_n$. By Martingale convergence (specifically Theorem 30), the martingale M_n converges. By the assumption on R_n, the partial sum of R_n converges almost surely as well.

Now suppose that $[a, b]$ is an interval on which $F(x) > 0$. Then by continuity, there is $\delta > 0$ and an $\epsilon > 0$ so that on $I = (b, b + \epsilon)$, $F(x) > \delta$. Now for all times sufficiently large, whenever X_n enters the interval it will exit the interval to the right (as the M_n is converging and the R_n is absolutely summable while the $\sum \gamma_n F(X_n)$ would diverge if it stayed in there forever). Hence it actually follows there is almost surely a last time that the process is in the interval $[a, b]$ (as the process will eventually be unable to jump over the interval, it eventually, almost surely exits to the right, and in I the process always exits to the right, eventually).

The same argument shows that for an interval $[a, b]$ on which $F(x) < 0$, the process almost surely visits the interval only finitely many times. Now there are only countably many intervals $[a, b]$ which are preimages of $(-\infty, -q]$ and $[q, \infty)$ for positive rational

q, and so we conclude that $F(X_n) \xrightarrow[n\to\infty]{\text{a.s.}} 0$, which is equivalent to the claim. □

(Note) [40]

[40] We can furthermore ensure that the only zeros to which X_n converges are the "stable zeros", meaning x so that F is positive to the left of x and negative to the right.

Example 29: Two-armed bandits

Suppose we have a slot machine with two levers, one of which is definitely luckier than the other (because the probability of them being equally lucky is clearly 0). Suppose for simplicity we always wager \$1, and one arm wins with probability p and the other with probability q, for $p \neq q$ and both in $(0,1)$.

The problem is that we do not know which is the correct lever to maximize our winnings. So we need to experiment some between pulling levers. On the other hand, we would like to win as much as possible.

Let (L_n, R_n) be the winnings from the left lever and the right lever respectively. A good strategy would have that $(L_n + R_n)/n \xrightarrow[n\to\infty]{\text{a.s.}} \max\{p,q\}$, in that it at least would eventually maximize the rewards. Even better, we would in some sense maximize $n\max\{p,q\} - \mathbb{E}(L_n + R_n)$, which is to say it would minimize the *regret*.

Suppose we take the following strategy: let $X_n := \frac{L_n}{L_n+R_n+1}$ and select the left arm with probability X_n and the right arm with probability $(1 - X_n)$.

Let $(\mathscr{F}_n : n \geq \mathbb{N})$ be the natural filtration generated by L and R. Then

$$\begin{aligned}\mathbb{E}(X_{n+1} \mid \mathscr{F}_n) = (1-X_n)\Big(\frac{L_n+1}{L_n+R_n+2}p + X_n(1-p)\Big) \\ + X_n\Big(\frac{L_n}{L_n+R_n+2}q + X_n(1-q)\Big).\end{aligned}$$

Note that each of these fractions we can expand by

$$\frac{L_n+a}{L_n+R_n+1+b} - \frac{L_n}{L_n+R_n+1} = \frac{a(L_n+R_n+1) - bL_n}{(L_n+R_n+1)(L_n+R_n+1+b)}.$$

Setting $W_n = L_n + R_n + 1$, we can therefore simplify this as

$$\frac{L_n+a}{L_n+R_n+1+b} - \frac{L_n}{L_n+R_n+1} = (a - bX_n)/W_n + \mathscr{O}(1/W_n^2).$$

Applying these simplifications,

$$\begin{aligned}\mathbb{E}(X_{n+1} - X_n \mid \mathscr{F}_n) = (1-X_n)\Big(\frac{X_n}{W_n}p\Big) \\ + X_n\Big(\frac{1-X_n}{W_n}q\Big) + \mathscr{O}(1/W_n^2)\end{aligned}$$

So if we set $F(x) = x(1-x)(p-q)$, and $\gamma_n = \frac{1}{W_n}$, then we can check this satisfies the stochastic approximation theorem. Note that $W_n - 1$ is *stochastically larger* (see the remark that follows this example) than a sum of iid Bernoulli($\min\{p,q\}$) and stochastically larger than a sum of iid Bernoulli($\max\{p,q\}$), and hence we can check

$$\sum_n \frac{1}{W_n} = \infty \quad \text{a.s.} \quad \text{and} \quad \sum_n \frac{1}{W_n^2} < \infty \quad \text{a.s.}$$

The variance of the martingale increment can also be checked to be bounded by $\mathcal{O}(1/W_n^2)$.
Theorem 35 shows that X_n converges to either 1 or 0. The stability analysis shows it converges to the "right" one when $p \neq q$.

(Note)[41]

Remark 3: For two random variables X, Y say X is stochastically larger than Y (or X stochastically dominates Y) if $\Pr(X \leq t) \leq \Pr(Y \leq t)$ for all $t \in \mathbb{R}$. This is equivalent to saying it is possible to realize a coupling of X and Y so that $X \geq Y$ almost surely. For two random walks, $(X_n), (Y_n)$, say that X is stochastically larger if one can realize a coupling of X and Y such that $X_n \geq Y_n$ for all n.
For a random walk, this is implied by saying the conditional distribution of $X_n | \mathscr{F}_n$ is stochastically dominates $Y_n | \mathscr{F}_n$ for all realizations of either history.

[41] Bandits are a huge subjet. See for example (Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020) for a comprehensive discussion.

Convergence of stochastic gradient descent

The prototypical application of this is the convergence of stochastic gradient descent, an optimization algorithm. [42] Suppose that $F : \mathbb{R}^n \to \mathbb{R}$ is a smooth function with bounded first, second and third derivatives. We suppose that we have the following non-degeneracy condition:

$$\langle \nabla F(x), (\nabla^2 F(x)) \nabla F(x) \rangle = 0 \quad \Longrightarrow \quad \nabla F(x) = 0. \tag{5}$$

The simplest example of functions that satisfy this are strictly convex. [43]

Consider a stochastic algorithm defined by

$$X_{n+1} := X_n - \gamma_n (\nabla F(X_n) + \xi_n)$$

for some random vectors ξ_n satisfying $\mathbb{E}(\|\xi_n\|^2 \mid \mathscr{F}_n) \leq K$. This additional randomness can come from a lot of sources: it may be added

[42] See (Léon Bottou, Frank E Curtis, and Jorge Nocedal. "Optimization methods for large-scale machine learning". In: *Siam Review* 60.2 [2018], pp. 223–311) for a comprehensive treatment.

[43] These are functions whose Hessian is positive definite everywhere.

artificially to improve the behavior of the algorithm, but frequently it is due to employing a randomized estimator for the gradient which is computationally simpler than the computing the actual gradient. [44]

Now,

$$\|\nabla F(X_{n+1})\|^2 = \|\nabla F(X_n)\|^2 - 2\langle \gamma_n(\nabla F(X_n) + \xi_n), (\nabla^2 F(X_n))\nabla F(X_n)\rangle + R_n$$

Here R_n carries a factor of γ_n^2, and so will be absolutely summable. Then under Assumption (5) and by continuity, the set of $\{x : \|\nabla F(x)\| > \delta\}$ can be covered by countably many (necessarily disconnected) sets of the form $\{x : |\langle \nabla F(x), (\nabla^2 F(x))\nabla F(x)\rangle| > \epsilon\}$ for some $\epsilon > 0$. By repeating the same argument as in Theorem 35 it follows that

$$\|\nabla F(X_n)\|^2 \xrightarrow[n\to\infty]{\text{a.s.}} 0,$$

which is to say that SGD converges to a stationary point.

[44] The classical Robbins and Monro argument actually uses a vector version of the scalar stochastic approximation argument above. It shows that the path of the algorithm asymptotically almost surely converges to the trajectory of an ODE, gradient flow, $\frac{d}{dt}X_t = -\nabla F(X_t)$, and so if all limit points of gradient flow satisfy some property (such as $\|\nabla F(X_t)\|^2 = 0$), then so must the limit point of the stochastic algorithm.

Martingale concentration

When the increments of a martingale are sufficiently bounded, it is possible to make much stronger estimates of the maximum value of a martingale, and this leads to some of the most important applications of martingales: tail bounds for random variables.

Definition 40 (Subgaussian): centered random variable X is V-subgaussian if

$$\mathbb{E}e^{\lambda X} \le e^{\lambda^2 V/2} \quad \text{for all } \lambda \in \mathbb{R}.$$

[45]

For a martingale, we can define an upgraded bracket process, replaces a sum of conditional variances by the sum of conditional subgaussian increments.

[45] If the random variable is not centered, there are competing definitions of what V-subgaussian should mean. If one requires $|\mathbb{E}X|^2 \le V$ as well, then the square-root of the subgaussian constant is equivalent, up to universal constants, to a norm.

Definition 41 (Subgaussian Bracket): martingale $(M_n : n \ge 0)$ is (V_n)–conditionally subgaussian for an adapted process $(V_n : n \ge 1)$ if for all $n \ge 1$ and all $\lambda \in \mathbb{R}$

$$\mathbb{E}[e^{\lambda(M_n - M_{n-1})} \mid \mathscr{F}_{n-1}] \le e^{\lambda^2 V_n} \text{ a. s.}$$

Define the subgaussian bracket $[\![M_n]\!]$ as the smallest, non-negative, non-decreasing adapted process so that $(M_n : n \ge 0)$ is conditionally subgaussian with process $([\![M_n]\!] - [\![M_{n-1}]\!] : n \ge 1)$.

Say a Martingale is subgaussian if $[\![M_n]\!] < \infty$ a. s.

This leads immediately to a tail bound for a martingale which enjoys this conditional subgaussian property.

Theorem 36: Subgaussian Azuma

Suppose that $(M_n : n \geq 0)$ that is a subgaussian with martingale. Then for any $n, t \geq 0$,

$$\Pr(\{\sup_{0\leq k\leq n}(M_k - M_0) \geq t\} \cap \{[\![M_n]\!] \leq S\}) \leq \exp\left(-\frac{t^2}{2S}\right).$$

Proof. By subtracting M_0 from the martingale, we may assume M_0 is 0. Define a new process, for any $\lambda \in \mathbb{R}$,

$$\mathcal{E}_n := \exp(\lambda M_n - \lambda^2 [\![M_n]\!]/2).$$

Then by the conditional subgaussian assumption $(\mathcal{E}_n : n \geq 0)$ is a supermartingale. Let T be the first time k that $M_k \geq t$ or that $[\![M_k]\!] > S$. Then by optional stopping, for $\lambda \geq 0$

$$1 \geq \mathbb{E}(\mathcal{E}_{T\wedge n}).$$

On the event $\{T \leq n\} \cap \{[\![M_n]\!] \leq S\}$, we have

$$\mathcal{E}_{T\wedge n} \geq \exp(\lambda t - \lambda^2 [\![M_T]\!]/2) \geq \exp(\lambda t - \lambda^2 S/2).$$

Thus

$$1 \geq \Pr(\{T \leq n\} \cap \{[\![M_n]\!] \leq S\}) \exp(\lambda t - \lambda^2 S/2).$$

Rearranging we have shown that for any $\lambda \geq 0$,

$$\Pr(\{\sup_{0\leq k\leq n} M_k \geq t\} \cap \{[\![M_n]\!] \leq S\}) \leq \exp(-\lambda t + \lambda^2 S/2).$$

Optimzing over $\lambda \geq 0$, we select $\lambda = t/S$ which shows the bound. □

A simple special case is for increments that are bounded.

Lemma 19 (Compact implies subgaussian): Suppose that X is mean 0 and $X \in (a, b)$ for $a, b \in \mathbb{R}$. Then

$$\mathbb{E}\exp(\lambda X) \leq \exp((b-a)^2\lambda^2/8).$$

Or simply, X is $(b-a)^2/4$–subgaussian.

Proof. Suppose without loss of generality that $b \leq a$. We can represent X as a convex combination, by

$$X = b\frac{X-a}{b-a} + a\frac{b-X}{b-a}.$$

Then by convexity for all $\lambda \in \mathbb{R}$

$$\mathbb{E}\exp(\lambda X) \le \mathbb{E}\left(\exp(\lambda b)\frac{X-a}{b-a} + \exp(\lambda a)\frac{b-X}{b-a}\right).$$

Using that X has mean 0,

$$\mathbb{E}\exp(\lambda X) \le \exp(\lambda b)\frac{-a}{b-a} + \exp(\lambda a)\frac{b}{b-a} =: f(\lambda).$$

Taking the log-derivative

$$\frac{d}{d\lambda}\log f(\lambda) = \frac{-ab\exp(\lambda b) + ab\exp(\lambda a)}{-a\exp(\lambda b) + b\exp(\lambda a)}.$$

With courage, we take another derivative, and then bound it above by $(b-a)^2/4$, uniformly in $\lambda \in \mathbb{R}$. Then, integrating twice,

$$\log f(\lambda) \le \frac{\lambda^2}{2}\frac{(b-a)^2}{4}.$$

□

As a corollary, we derive the classical Azuma inequalities.

Corollary 7 (Azuma): Suppose that $(M_n : n \ge 0)$ is a martingale and $(A_n : n \ge 1$ is a predictable process such that for all $1 \le k \le n$, $|M_k - M_{k-1}| \le A_k$, then for all $t \ge 0$

$$\Pr\left(\{\max_{0\le k\le n}(M_k - M_0) \ge t\} \cap \{\sum_1^n A_k \le A\}\right) \le \exp\left(-\frac{t^2}{2A}\right).$$

If A_k are in fact deterministic, then we derive the convential Azuma inequality

$$\Pr\left(\max_{0\le k\le n}(M_k - M_0) \ge t\right) \le \exp\left(-\frac{t^2}{2\sum_{k=1}^n A_k^2}\right).$$

Example 30: Vertex exposure martingale

Let $G = (V, E)$ be a graph on vertex set $V = \{1, 2, \cdots, n\}$, where E is a subset of all subsets of V of size 2. For each edge $e \in E$, let $X_e = \text{Bernoulli}(p)$ for some $p \in (0, 1)$. The resulting random graph, where we consider the random edge set $E' = \{e : X_e = 1\}$ is called Bernoulli percolation on G. In the case that G is the complete graph, this is called the binomial random graph, or the Erdős-Rényi random graph.

The *chromatic number* of a graph G is the smallest $\chi \in \mathbb{N}$ so that there is a function $f : V \to \{1, 2, \cdots, \chi\}$ with the property that $\{v, w\} \in E$ implies $f(v) \ne f(w)$. While the chromatic

number of a graph is difficult to compute, for any Bernoulli percolation G', we have that for all $t \geq 0$,

$$\Pr(|\chi(G') - \mathbb{E}\chi(G')| \geq t) \leq 2\exp(-t^2/2n).$$

To see this we form a martingale in which we reveal vertices one at a time, which is to say $\mathscr{F}_k = \sigma(X_{ij} : i, j \leq k)$. We then let $M_k = \mathbb{E}\chi(G') \mid \mathscr{F}_k$. Now for every realization of the (X_{ij}), changing the edges adjacent to vertex k can only change the chromatic number of the graph by at most 1; that is if we have (X'_{ij}) another realization of the percolation that differs from (X_{ij}) only on those edges incident to k, we have

$$|\chi((X_{ij})) - \chi((X'_{ij}))| \leq 1.$$

This is because best case, changing an edge incident to k could allow us to remove the color class of k. Worst case, changing an edge incident to k might force us to add a new color class for k. Hence $|M_k - M_{k-1}| \leq 1$ almost surely.

To evaluate if this is useful, it is always possible to bound the chromatic number by $2D + 1$ where D is the maximum degree in the graph. So if G is the complete graph on n vertices, and $p \gg 1/\sqrt{n}$, the degrees are much bigger than $\sqrt{n}$ (see the Exercise below), and the concentration is much better than the trivial bound on the chromatic number.

The chromatic number of a random graph has a long, rich story. When $p \leq n^{-1/2-\epsilon}$, it is known from Alon Krivelevich '97 that the chromatic number actually concentrates on one of 2 numbers with probability going to 1. It is not known if there is a sharper concentration bound then the one that comes from Azuma.

Exercise 29 (Maximum Degrees in $G(n,p)$): Show that when $p\sqrt{n} \to \infty$, the minimum degree δ and the maximum degree of the graph satisfy

$$\frac{\delta}{np} \xrightarrow[n\to\infty]{\Pr} 1 \quad \text{and} \quad \frac{\Delta}{np} \xrightarrow[n\to\infty]{\Pr} 1.$$

Azuma + Union bound.

Example 31: Thermodynamic limit of continuum percolation

Suppose that $(X_1, \cdots, X_N)$ are uniformly distributed on $B = [0,1]^d$ for some dimension $d \geq 1$. Let $A = A(X_1, \cdots, X_N)$ be the volume of the subset of B which is within distance $(R/N)^{1/d}$ of a point X_n for some $1 \leq n \leq N$. Let $(\mathscr{F}_n)$ be the natural filtration associated to this sequence, and let (M_n) be the Doob martingale $M_n = \mathbb{E}(A \mid \mathscr{F}_n)$ for $1 \leq n \leq N$. Then setting c_d to be the volume of the d-dimensional unit

ball,

$$|M_n - M_{n-1}| \leq c_d R/N,$$

using that the volume in the ball around the point X_n is $c_d R/N$, and hence the conditional expectation can change by at most π/N.(46°) Thus by Azuma's inequality, for all $t \geq 0$

$$\Pr(|A - \mathbb{E}A| \geq t) \leq 2\exp(-Nt^2/2c_d^2R^2).$$

Hence the area (for fixed d, R and large N) is practically deterministic.

With a slight modification of the setup, we can easily compute the expected volume. If we consider the torus, where opposite boundaries of the cube wrap around, then the expected volume E_n after adding n points satisfies

$$E_{n+1} - E_n = \mathbb{E}\int \mathbf{1}_{\|x-X_{n+1}\|^2 \leq \frac{R}{N}} \mathbf{1}_{x\in B\setminus V_n}\, \lambda(dx).$$

Applying Fubini,

$$E_{n+1} - E_n = c_d \frac{R}{N}\int \mathbf{1}_{x\in B\setminus V_n}\, \lambda(dx) = \frac{c_d R}{N}(1 - E_n).$$

Thus the area that remains satisfies

$$(1 - E_N) = (1 - \tfrac{c_d R}{N})^N \approx e^{-c_d R}.$$

Note that if R is large then the area left uncovered should be small, while our concentration inequality degrades. So for larger R, a different strategy is needed.

Example 32: Crazy Percolation

This is like above, but more crazy. Suppose that once more $(X_1, \cdots, X_N)$ are uniformly distributed on $B = [0,1]^d$ for some dimension $d \geq 1$. Now suppose that these are arriving in order, and the k-th point to arrive creates a ball around it of volume which depends on the existing configuration of points. It chooses one of the radii $\{r_k\}_1^N$ without replacement, possibly randomly. The radii are given by $r_k = (1/k\log N)^{1/d}$. Let A be the resulting volume.

Letting $M_n = \mathbb{E}(A \mid \mathscr{F}_n)$, we have

$$[\![M_N]\!] \leq \frac{c_d \pi^2}{6(\log N)^2}.$$

We just didn't know in which order that occurred.
If the r_k are sampled without replacement, uniformly, then we still have that on the torus,

$$E_N = \prod_{n=1}^{N} (1 - \tfrac{c_d}{n \log N}) \to e^{-c_d},$$

as $N \to \infty$.

One classic minimization problem is the travelling salesman problem. The stochastic travelling salesman problem.

Example 33: Stochastic Travelling Salesman Problem

Suppose that $\{X_j\}$ are iid uniform points in $[0,1]^2$. The length of the shortest tour is given by

$$L_n := \min_{\pi \in \mathfrak{S}_n} \sum_{j=1}^{n} \|X_{\pi(j+1)} - X_{\pi(j)}\|,$$

where we take $\pi(n+1) = \pi(1)$. Once more we have that, by using that each random variable can only change L_n by at most 2, we just get that there is a $c > 0$ so that for all $t > 0$,

$$\Pr(|L_n - \mathbb{E}L_n| \geq t) \leq 2\exp(-ct^2/n).$$

The mean $\mathbb{E}L_n \asymp \sqrt{n}$. One can always construct a strategy by dividing the square into blocks of side length $\sqrt{n}$, and then snaking up and down the columns of blocks in supermarket ordering, stopping in each block to cover all points in each block in any arbitrary ordering. This is not hard to check that it has at most order $\sqrt{n}$ expected length (covering all $\sqrt{n}$ aisles of height 1).

On the other hand, one can show that with probability $1/2$ at least half the points have a closest point which is at least $1/(100\sqrt{n})$ away from it. To get to these $n/2$ points, one must travel at least this distance, and so

$$\mathbb{E}L_n \geq \frac{1}{2}\frac{n}{4}\frac{1}{100\sqrt{n}}.$$

In this situation, Azuma's inequality is actually better than just bounded the increments. Given a collection of points $\{X_j\}_1^n$, let L denote the shortest tour, and let $\widehat{L}$ denote the shortest tour not using X_k. Then by simply deleting the point X_k (skipping it in order and going to the next point), we conclude

$$\widehat{L} \leq L.$$

[46] Writing the area function $A(x_1, x_2, \ldots, x_N)$ as a function of its centers, we use here that

$$|A(x_1, \ldots, x_{n-1}, x_n, \ldots, x_N) - A(x_1, \ldots, x_{n-1}, y, \ldots, x_N)| \leq \frac{R\pi}{N},$$

for all choices $\{x_n\}$ and y in B. Because of this, the difference of martingales (which are expectations of this function) also satisfy this bound.

On the other hand, if we take the shortest tour through $\widehat{L}$, we can always extend it to include X_k by simply adding a detour through X_k to the closest point in the tour. With foresight, let $d_k = \min_{p>k} \|X_p - X_k\|$. Then

$$\widehat{L} \le L \le \widehat{L} + 2d_k.$$

Hence taking conditional expectation of this, given $\mathscr{F}_k$, we have

$$|\mathbb{E}(L \mid \mathscr{F}_{k-1}) - \mathbb{E}(L \mid \mathscr{F}_k)| \le \mathbb{E}(2d_k \mid \mathscr{F}_k) \le \frac{C}{\sqrt{n-k}}.$$

Applying Azuma's inequality, we conclude that there is a $c > 0$ so that for all $t > 0$,

$$\Pr(|L_n - \mathbb{E}L_n| \ge t) \le 2\exp(-ct^2/\log n).$$

Another classic optimization problem is the following.

Example 34: Longest Increasing subsequence

An increasing subsequence in a random permutation $\pi \in \mathfrak{S}_n$ is a collection $\{i_1 < i_2 < \cdots < i_k\}$ so that $\pi(i_1) < \pi(i_2) < \cdots \pi(i_k)$. The longest increasing subsequence problem, or Ulam's problem, is to determine the statistical behavior of the longest increasing subsequence in a uniformly random permutation. This was one of the first examples of computing, and ties to the origins of Markov chain Monte Carlo.

To generate a random permutation, one way is to throw n uniform random points $(x_j, y_j)_1^n$ on $[0,1]^2$. This defines a permutation π by doing the following sorting procedure: reorder the points so that $x_1 < x_2 < \cdots < x_n$ and then let $\pi(j)$ be the relative ordering of y_j amongst $\{y_i\}_1^n$ with 1 being the smallest and n being the largest.

Now this has the virtue that increasing subsequences can be described graphically. An increasing subsequence in this representation is just a sequence of points so that $\{i_1 < i_2 < \ldots < i_k\}$ so that x_{i_j} is increasing in j and y_{i_k} is increasing in j. In short, if you draw a path between the points, it always goes up and to the right.
Now the length (counted by cardinality) of the longest path L_n can only change by at most 1 when adding or moving a

point, and so we have for all $t \geq 0$

$$\Pr(|L_n - \mathbb{E}L_n| \geq t) \leq 2\exp(-t^2/2n).$$

It can be shown that $\mathbb{E}L_n \asymp \sqrt{n}$, and so this bound is on the same order as the mean. This is notably an example where a straight martingale concentration does not give something clearly useful.

See the textbook of Dan Romik. *The surprising mathematics of longest increasing subsequences*. 4. Cambridge University Press, 2015 for a complete discussion of this problem, which has many interesting surprising connections.

Branching processes

Branching processes provide a final important example of stochastic processes. The basic version of a branching process describes the size of population that evolves in time. Each (asexual) amoeba in this population can give birth to some other number of descendants, which then continue to divide ad infinitum. In the standard Bienaymé-Galton-Watson branching process, the offspring distribution of every amoeba is fixed and indepdently distributed.

These can be described by a genealogical tree, showing which amoeba in a population are descended from which other. The easiest notion of *time* in this case is the number of generations in this genealogical tree from the single common ancestor. There are also more sophisticated "wall-clock" time versions of these processes, which the population is described a continuous time, and the length of time to reproduction is random, but in many nice cases, these processes can be related.

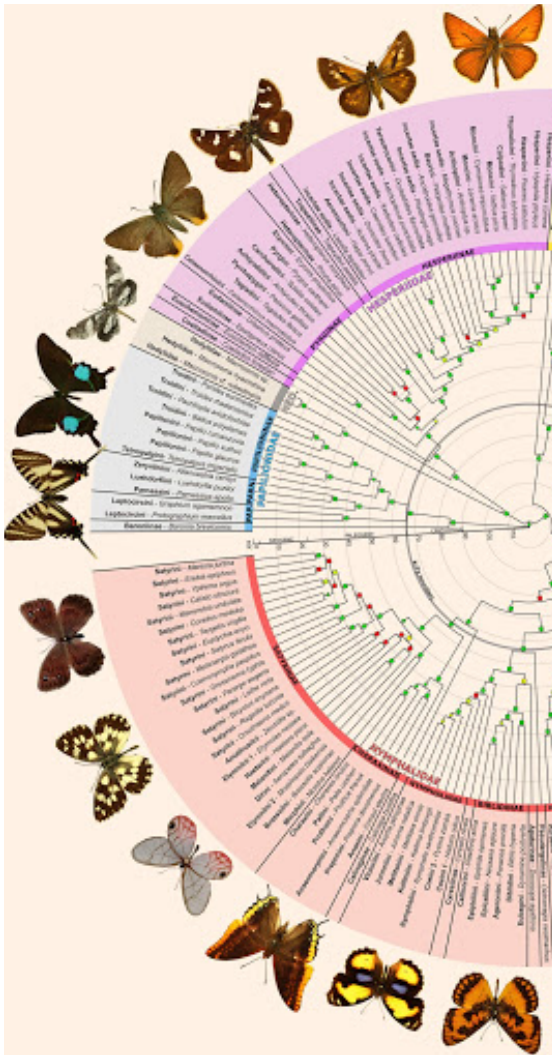
In genealogical time, we can simply record the number of amoebas that are alive in generation n, in which case the population size of the Bienaymé-Galton-Watson process becomes a homogeneous-time Markov chain on the non-negative integers $\mathbb{N}_0$.

> **Definition 42 (Branching Process):** Let Z_n be a random variable that denotes the size of the population of amoebae. The Markov chain $(Z_n)_{n\geq 1}$ with values in $\mathbb{N}_0$ is a *branching process* if,
>
> $$Z_{n+1} = \sum_{j=1}^{Z_n} X_{n,j}$$
>
> where $X_{n,j}$ denotes the number of children born to the jth person in the nth generation. The family $(X_{n,j} : n, j \geq 0)$ is an iid family on $\mathbb{N}_0$, the distribution of which is called the *offspring distribution* of the branching process. We will assume that $\mathbb{E}Z_0 < \infty$ and that $Z_0 > 0$ almost surely. If not otherwise specified, we take $Z_0 = 1$.

In this definition, if Z_n is empty, we consider the sum to be empty, and so 0 is an absorbing state.

Extinction and survival

The first motivating question is the extinction time question: does the population go extinct?



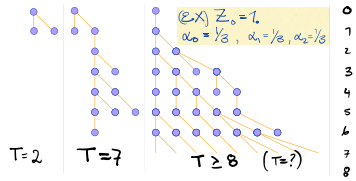


Figure 2: Extinction times and geneological trees.

Definition 43 (Extinction Time): The *extinction time* T_0 of a branching process is the hitting time of Z_n to zero, that is, $T_0 := \tau_0$.

Note that since 0 is absorbing, if 0 is accessible – meaning if the offspring distribution puts positive probability on 0 offspring – then the chain is transient. Moreover, it will always be possible that the population goes extinct in 1 step (although it might be extremely rare). So in summary, if 0 offspring is possible $\Pr(T_0 < \infty) > 0$ (and conversely, if 0 offspring is impossible, then $\Pr(T_0 < \infty) = 0$, unless $Z_0 = 0$). The real question is if $\Pr(T_0 = \infty) > 0$, which is to say does the population have positive probability of surviving forever.

There is once case where we can quickly answer this question.

Definition 44 (Mean Offspring Size): The *mean offspring size* μ is the mean of the offspring distribution, that is, $\mu := \mathbb{E}(X)$ where X follows the offspring distribution.

Theorem 37: Additive martingale

Let $(\mathscr{F}_n : n \geq 0)$ be the natural filtration associated to the process $(Z_n : n \geq 0)$.

1. If $\mu < 1$ then $(Z_n : n \geq 0)$ is a supermartingale.
2. If $\mu = 1$ then $(Z_n : n \geq 0)$ is a martingale,
3. and if $\mu \in (1, \infty)$, then $(Z_n : n \geq 0)$ is a submartingale.
4. Further, for $\mu \in (0, \infty)$, Z_n / μ^n is a martingale.

Proof. All of these statements follow from the computation of the conditional expectation of Z_{n+1} given $\mathscr{F}_n$. For all these cases,

$$\mathbb{E}(Z_{n+1} \mid \mathscr{F}_n) = \sum_{j=1}^{Z_n} \mathbb{E}(X_{n+1,j} \mid \mathscr{F}_n) = \sum_{j=1}^{Z_n} \mu = Z_n \mu.$$

□

This gives rise to an immediate corollary:

Corollary 8 (Extinction): If $\mu \leq 1$, then $\Pr(T_0 < \infty) = 1$, unless the offspring distribution puts all its mass on 1.

Proof. By positive supermartingale convergence, $Z_n \xrightarrow[n \to \infty]{\text{a.s.}} Z_\infty$. As the sequence $(Z_n : n \geq 0)$ is integer valued, it must be that $Z_n = Z_\infty$ for all $n \geq n_0$ for some random n_0. This is impossible if $Z_\infty > 0$ and the offspring distribution puts positive probability on a number besides 1, and so we must have $Z_\infty = 0$. □

So the population always dies out if $\mu \geq 1$. The case where $\mu < 1$ does have some different phenomenology, however, and so we distinguish three cases.

Definition 45 (Criticality): A branching process is *subcritical* if $\mu < 1$, *critical* if $\mu = 1$, and *supercritical* if $\mu \in (1, \infty)$. Moreover, as Z_n/μ^n is a martingale,

$$\lim_{n\to\infty} \mathbb{E}[Z_n] = \lim_{n\to\infty} \mu^n = \begin{cases} 0, & \text{if } \mu < 1 \\ \mathbb{E}[Z_0], & \text{if } \mu = 1 \\ \infty, & \text{if } \mu > 1 \end{cases}$$

Generating Functions

To decide what happens with $\mu > 1$, we need to introduce another tool, the probability generating function.

Definition 46 (Generating Function): Let X be a discrete random variable with values in $\mathbb{N}_0$. The *probability generating function* or *pgf* of X is,

$$G(s) = \mathbb{E}[s^X] = \sum_{k=0}^{\infty} s^k \cdot \Pr(X = k).$$

for all $|s| \leq 1$, for which values the series converges absolutely

The probability generating function allows probabilistic questions to be encoded as analytic questions. Probabilities can be extracted from the generating function by making analytic operations.

Theorem 38: Differentiating the pgf

Probabilities for X can be obtained from the generating function by successive differentiation. If $G^{(j)}$ is the jth derivative of G,

$$G^{(j)}(s) = \sum_{k=j}^{\infty} k(k-1)\cdots(k-j+1)s^{k-j} \Pr(X = j).$$

In particular $G^{(j)}(0) = j! \cdot \Pr(X = j)$

This can be seen by just successively differentiating the power series term by term.

Example 35: A few common generating functions

1. Let $X \sim \text{Unif}(\{0,1,2\})$. Then,
$$G(s) = \frac{1}{3} + s\left(\frac{1}{3}\right) + s^2\left(\frac{1}{3}\right) = \frac{1}{3}\left(1+s+s^2\right).$$
2. Let $X \sim \text{Geom}(p)$. For $|s| < 1$,
$$G(s) = \sum_{k=1}^{\infty} s^k p(1-p)^{k-1} = sp\sum_{k=1}^{\infty}(s(1-p))^{k-1} = \frac{sp}{1-s(1-p)}.$$
3. Let $X \sim \text{Poisson}(\mu)$. For $\mu > 0$,
$$G(s) = \sum_{k=0}^{\infty} \frac{e^{-\mu}\mu^k}{k!}\cdot s^k = e^{-\mu}\cdot\sum_{k=0}^{\infty}\frac{(\mu s)^k}{k!} = e^{-\mu}e^{\mu s} = e^{\mu(s-1)}.$$

Theorem 39: Properties of Generating Functions

We mention in passing a few properties of probability generating functions

1. If X and Y are random varibles on $\mathbb{N}_0$ that satisfy,
$$G_X(s) = G_Y(s) \quad \forall s \in (0,1)$$
then $X \overset{\text{law}}{=} Y$.
2. If X and Y are independent, then,
$$G_{X+Y}(s) = G_X(s)\cdot G_Y(s)$$

Proof. For the first point, a pgf of a random variable has an absolutely convergence power series expansion in $|s| \leq 1$, and so they are equal to their Taylor series expansions at 0. So if $G_X(s) = G_Y(s)$ for all $s \in (0,1)$ (and hence in $[0,1]$), all their derivatives at 0 are equal, and so all they have the same Taylor series. But this means they have same law, as the pgf is the generating function of the probability vector of a random variable.

For the second point, we compute for any $s \in (0,1)$
$$G_{X+Y}(s) = \mathbb{E}[s^{X+Y}] = \mathbb{E}[s^X s^Y] = \mathbb{E}[s^X]\mathbb{E}[s^Y] = G_X(s)G_Y(s).$$
□

The main reason we have introduced the pgf is that it is especially well suited to branching processes, for the following reason:

Theorem 40: Branching processes and the pgf

The generating function of the nth generation size Z_n is the n-fold composition of the offspring distribution generating function,

$$G_n(s) = \mathbb{E}[s^{Z_n}] = \mathbb{E}(\underbrace{G \circ G \circ \cdots \circ G}_{n \text{ times}}(s)^{Z_0}).$$

Proof. The generating function of the nth generation size Z_n is,

$$G_n(s) = \mathbb{E}[s^{Z_n}] = \mathbb{E}\left[s^{\sum_{k=1}^{Z_{n-1}} X_k}\right] = \mathbb{E}\left[\mathbb{E}\left[s^{\sum_{k=1}^{Z_{n-1}} X_k} \mid Z_{n-1}\right]\right]$$

where the last inequality is by the Total Law of Expectation.

$$\begin{aligned}
\mathbb{E}\left[s^{\sum_{k=1}^{Z_{n-1}} X_k} \mid Z_{n-1} = z\right] &= \mathbb{E}\left[s^{\sum_{k=1}^{z} X_k} \mid Z_{n-1} = z\right] \text{ by conditioning} \\
&= \mathbb{E}\left[s^{\sum_{k=1}^{z} X_k}\right] \text{ by independence} \\
&= \mathbb{E}\left[\prod_{k=1}^{z} s^{X_k}\right] \\
&= \prod_{k=1}^{z} \mathbb{E}[s^{X_k}] \text{ by independence} \\
&= [G(s)]^z \text{ for all } z.
\end{aligned}$$

So we have shown that

$$\mathbb{E}[s^{\sum_{k=1}^{Z_{n-1}} X_k} \mid Z_{n-1}] = [G(s)]^{Z_{n-1}}$$

Taking expectations,

$$G_n(s) = \mathbb{E}[G(s)^{Z_{n-1}}] = G_{n-1}(G(s))$$

The result follows by induction on n. □

Corollary 9 (Extinction by time n): For any $n \in \mathbb{N}$,

$$\Pr(T_0 \leq n) = H(\underbrace{G \circ G \circ \cdots \circ G}_{n \text{ times}}(0))$$

where $H(s) = \mathbb{E}s^{Z_0}$.

Proof. The generating function for the n-th generation size Z_n is,

$$G_n(s) = \sum_{k=0}^{\infty} s^k \Pr(Z_n = k)$$

Because 0 is an absorbing state,

$$\Pr(T_0 \leq n) = \Pr(Z_n = 0) = G_n(0).$$

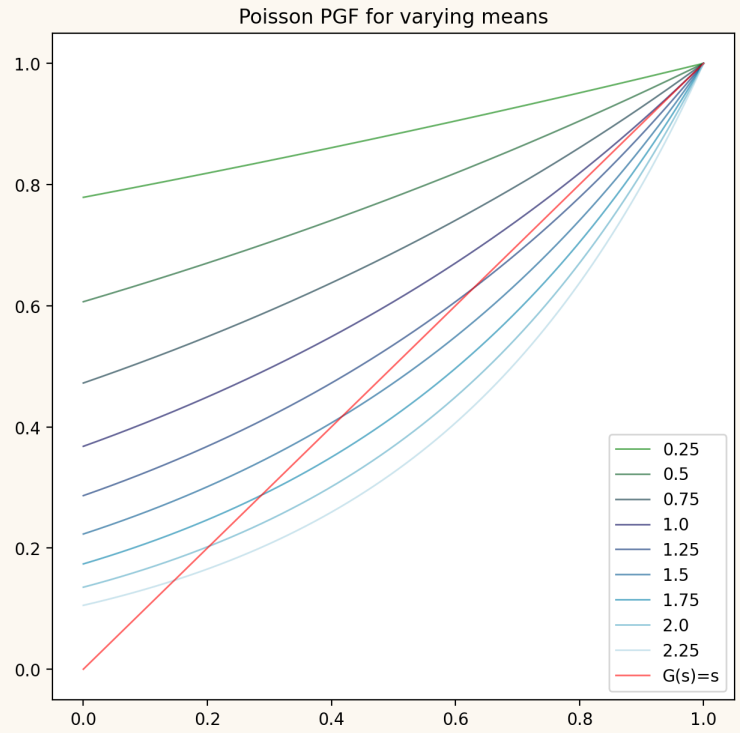
Applying Theorem 40

$$G_n(s) = H(\underbrace{G \circ G \circ \cdots \circ G}_{n \text{ times}}(s)).$$

□

So we can compute the probability of extinction by time n by understanding the behavior of iterated compositions of the pgf of the offspring distribution. This motivates the following collection of analytic facts about pgfs.

Example 36: Poisson PGFs visualized



Note that the pgfs dip below the straight line $s \mapsto s$ when the means of the Poisson are strictly greater than 1.

Lemma 20 (Analytic properties of pgfs): The probability generating function $G(s)$ of a random variable X on $\mathbb{N}_0$ is convex and non-decreasing on $[0,1]$ with $G(1) = 1$. Furthermore, it is strictly increasing, unless $\Pr(X = 0) = 1$, and it is strictly convex, unless $X \leq 1$ almost surely.

Proof. From the definition, for $s \in [0,1]$.

$$G(s) = \sum_{n=0}^{\infty} s^n \Pr(X = n).$$

So $G(1) = 1$, as $\Pr(X \in \mathbb{N}_0) = 1$. Further G is non-decreasing and convex as all terms in its Taylor expansion are non-negative. It is further strictly increasing if $\Pr(X > 0) > 0$ and strictly convex if $\Pr(X > 1) > 0$. □

We can also relate the mean of the distribution to the slope of the pgf at 1

> **Lemma 21 (The slope at 1 is the mean):** $G'(1) = \mathbb{E}X$.

Proof. $G'(1) = \sum_{k=1}^{\infty} k \cdot \Pr(X = k) = \mathbb{E}[X]$. □

All these facts have the following consequence.

> **Lemma 22 (Intersecting the line):** If X is a random variable on $\mathbb{N}_0$ and $\mathbb{E}X > 1$, there are exactly two solutions of $G_X(s) = s$ on $[0,1]$, one at 1 and one at $s_* \in [0,1)$. Moreover, for $s < s_*$, $G_X(s) > s$ and for $s \in (s_*, 1)$, $G_X(s) < s$.

Proof. If $\mathbb{E}X > 1$, then $\Pr(X > 1) > 0$, and so G_X is a strictly convex increasing function on $[0,1]$. It therefore intersects the straight-line $s \mapsto s$ either $\{0,1,2\}$ times. One intersection is at $s = 1$, and the slope of G_X' is greater than 1, so G_X descends below the line $s \mapsto s$ from 1. Further $G_X(0) \geq 0$, and so there must be a second intersection somewhere on $[0,1)$. As there cannot be more than 2 intersections, we conclude the statements in the lemma. □

Theorem 41: The root of extinction

Let $G(s)$ be the probability generating function of a random variable X on $\mathbb{N}_0$ with $\Pr(X > 1) > 0$. For any initial distribution with pgf H the smallest positive root of the equation $G(s) = s$, s_* is the probability of eventual extinction, that is, $P(T_0 < \infty) = H(s_*)$.

Proof. The iterates

$$p_k := (G \circ G \circ G \cdots G)(0)$$

are nondecreasing in k, and by definition of s_*, they are always below $s_* < 1$. Hence they have a limit p_∞. The iterates satisfy

$$p_{k+1} = G(p_k),$$

and so by taking limits on both sides and using continuity of G, we conclude that $p_\infty = G(p_\infty)$. Hence we must have that in fact on taking $k \to \infty$, $p_\infty = s_*$. From Corollary 9 and continuity of H, we conclude the proof. □

Thus we have proven the Bienaymé–Galton–Watson theorem, which states that there is a positive probability of surviving forever when the process is supercritical, while the process goes extinct almost surely if the process is critical or subcritical:

Corollary 10 (Galton-Watson): If the branching process is supercritical, i.e. $\mathbb{E}X > 1$, then $\Pr(T_0 = \infty) > 0$.

Example 37: Computing Extinction Probabilities

Consider a branching process with,

$$Z_0 = 1 \quad \text{and} \quad \vec{\pi} = \begin{pmatrix} 1/6 & 1/3 & 1/2 \end{pmatrix}$$

where $\vec{\pi}$ is the offspring distribution. The curves,

$$y = s \quad \text{and} \quad y = G(s) = \frac{1}{6}(1 + 2s + 3s^2)$$

intersect at $s = 1$. In this case the smallest nonnegative solution of the quadratic equation

$$0 = 1 - 4s + 3s^2 = (1 - 3s)(1 - s)$$

is the extinction probability, and so the tree goes extinct with probability $1/3$.

Example 38: Computing Extinction Probabilities

Consider a branching process with,

$$Z_0 = 1 \quad \text{and} \quad \vec{\pi} \sim \text{Po}(\mu)$$

where $\vec{\pi}$ is the offspring distribution. Recall that,

$$G(s) = e^{\mu(s-1)}$$

Solving $s = e^{\mu(s-1)}$ numerically by iteration,

$G_1(0)$	$G_2(0)$	$G_3(0)$	$G_4(0)$	$G_10(0)$	$G_{15}(0)$
0.135335	0.177403	0.192975	0.199079	0.203169	0.203187

Almost sure growth rates, and the Kesten–Stigum theorem

Suppose that we were unsatisfied by simply knowing that a population survived, in the supercritical case $\mu > 1$, and we furthermore wished to know the size of the population. In Theorem 37, we showed that $M_n := Z_n/(\mathbb{E}Z_0\mu^n)$ is a martingale, which strongly suggests that the correct order of magnitude is μ^n. Indeed, martingale convergence implies that M_n has an almost sure, almost surely finite limit M_∞. So we know, in a sense that the growth rate is *at most* μ^n in some sense, and we know it basically for free. We do not know however, that $M_\infty > 0$. If we let Δ be the extinction probability of the supercritical GW tree, then we know for example that there is *at least* a probability Δ that $M_\infty = 0$, on account of the fact that if $T_0 < \infty$, then we have $M_{T_0} = 0$, and hence $M_\infty = 0$. But is this the only way that the GW tree can die out?

This is answered by the Kesten–Stigum theorem:

Theorem 42: Kesten-Stigum

The event $\{M_\infty = 0\}$ has probability Δ if and only if the offspring random variable X satisfies $\mathbb{E}X\log X < \infty$ (at $x = 0$, we take $x\log x = 0$).

Otherwise said, under this $\mathbb{E}X\log X < \infty$ condition, almost surely either the population dies out in finite time or there is an almost surely positive random variable W so that $M_n = Z_n/(\mathbb{E}Z_0\mu^n) \xrightarrow[n\to\infty]{\text{a.s.}} W$.

This proof introduces a couple of beautiful ideas (some of which were introduced in Russell Lyons, Robin Pemantle, and Yuval Peres. "Conceptual proofs of L log L criteria for mean behavior of branching processes". In: *The Annals of Probability* (1995), pp. 1125–1138). While the theorem is an if and only if, we show just the sufficiency condition.

The basic strategy here is to consider a *change of measure*, much like the Doob h-transform, in which we look at a new probability measure $\mathbb{Q}$ which biases the probability space by W, i.e.

$$\mathbb{Q}(A) = \mathbb{E}(\mathbf{1}_A W).$$

If W is 0 on some positive probability event besides the obvious one (where the population is extinct at a finite time), then that entire event will be a measure 0 set in $\mathbb{Q}$. Now this idea is only useful if we can work with this probability measure.

Now it turns out it is possible to give an explicit probabilistic interpretation to the measures

$$\mathbb{Q}_n(A) = \mathbb{E}(\mathbf{1}_A M_n),$$

but these will only be related to $\mathbb{Q}$ if $\mathbb{E}(W \mid \mathscr{F}_n) = M_n$,[47] in which case for events $A \in \mathscr{F}_n$, we will have by definition of conditional expectation

$$\mathbb{Q}_n(A) = \mathbb{E}(\mathbf{1}_A M_n) = \mathbb{E}(\mathbf{1}_A W) = \mathbb{Q}(A).$$

The idea here is that the brnaching process $(Z_n : n \geq 0)$ will have a new explicit, Markovian description, at least for $0 \leq k \leq n$ under the measure $\mathbb{Q}_n$. In fact, it will turn out to be time-homogeneous.

To describe this, we need the idea of size-biasing.[48]

Definition 47 (Size bias): For a law μ on $[0,\infty)$ having finite expectation $\mathbb{E}X = \int_{\mathbb{R}} x\mu(dx) < \infty$, the size bias of μ is given by

$$\mu^s(dx) = \frac{x\mu(dx)}{\mathbb{E}X},$$

which is again a law on $[0,\infty)$. For a non-negative real valued random variable X with $\mathbb{E}X < \infty$, we say that a random variable X^s is a realization of the size-bias of X if its law is the size bias of the law of X.

It is worth noting that the size bias of a random variable always puts 0 probability on 0.

Exercise 30 (Size-biased Poisson): Show that $X \overset{\text{law}}{=} \text{Poisson}(\mu)$ for some μ if and only if $X^s \overset{\text{law}}{=} 1 + X$.

One of the main tools for working with size biasing is the following.

Lemma 23 (Size biased sums): Suppose that $X = \sum_{j=1}^n Y_j$ is a sum of iid non-negative random variables with finite expectation. Then we can realize the size bias of X by defining

$$X^s = Y^s + \sum_{j=2}^n Y_j$$

where Y^s follows the size biased distribution of Y_1 but is otherwise independent of (Y_j) and I.

Proof. Because the summands are iid, we can also realize the size bias of X by choosing $I = \text{Unif}(\{1,2,\cdots,n\})$ uniformly at random and independently of all (Y_j), since

$$X^s \overset{\text{law}}{=} \sum_{j=1}^n (Y_j \mathbf{1}_{j\neq I} + Y^s \mathbf{1}_{j=I}).$$

[47] So we will need to show that $(M_n : n \geq 0)$ is uniformly integrable. We will circle back to this.

[48] Size biasing appears naturally in life. The most famous tangible example is waiting for the bus. If the interarrival distribution between busses is X, and these have densities and are iid, then it can be proven that the length of the interarrival time when YOU arrive follows the size-bias of the normal interarrival distribution. In particular, probability is systematically disadvantaging you by making the bus arrival times longer when you arrive. This is not a joke. For what it's worth, it affects everyone equally.

We need to check the distribution of X^s. Let ϕ be a bounded Borel function from $\mathbb{R} \to \mathbb{R}$. Then

$$\mathbb{E}(\phi(X^s) \mid I) = \int \phi(\sum_{j=1}^{n} y_j) \frac{y_I}{\mathbb{E}Y_1} \mu^{\otimes n}(dy).$$

Hence taking expectation over I,

$$\mathbb{E}(\phi(X^s)) = \int \phi(\sum_{j=1}^{n} y_j) \frac{\sum_{j=1}^{n} y_j}{n\mathbb{E}Y_1} \mu^{\otimes n}(dy) = \mathbb{E}(\phi(X) \frac{X}{\mathbb{E}X}).$$

□

Remark 4 (Distribution of I): The idea in the proof allows the lemma to generalize to the non-iid setting, where the additional randomization is important. In general, for non-iid (Y_j), one must choose I so that $\Pr(I = j)$ is proportional to $\mathbb{E}Y_j$. With this choice, the size bias can always be realized by choosing

$$X^s \overset{\text{law}}{=} \sum_{j=1}^{n} (Y_j \mathbf{1}_{j \neq I} + Y^s \mathbf{1}_{j=I}).$$

This leads us to the definition of the size-biased Galton-Watson tree.

Definition 48 (Size-biased Galton Watson tree): Say that $(Z_n : n \geq 0)$ is the population alive at generation n in a size-biased Galton-Watson tree if conditionally on $\mathscr{F}_n$,

$$Z_{n+1} \overset{\text{law}}{=} X^s + \sum_{j=2}^{Z_n} X_j$$

where (X_j) and X^s are independent of each other and $\mathscr{F}_n$, and they follow the offspring distribution and size bias of the offspring distribution, respectively.

If viewing this is a genealogical tree, then there is a single lineage within this tree which is special, in that it always follows the size bias distribution. As the size bias of an integer random variable is always at least 1 almost surely, this lineage is never ending. One can also represent this tree as a single distinguished infinite path, *the spine*, to which one attaches Galton-Watson trees with offspring distribution X. This leads to an explicit representation of the size of the branching process:

Lemma 24 (Spinal representation): Suppose Z_n^s follows the distribution of a size-biased Galton Watson tree with initial pop-

ulation Z_0 and offspring random variable X. Let $((W_n^{(k)} : n \geq 0) : k \geq 1)$ be an infinite family of iid Galton Watson trees with initial distributions given by $X^s - 1$ and offspring distributions given by X. Let $W^{(0)}$ be a Galton Watson tree with initial population $Z_0 - 1$. Then for any $n \geq 1$

$$Z_n = 1 + \sum_{k=0}^{n} W_{n-k}^{(k)}$$

Proof. At step 0, this representation is

$$Z_0 = 1 + W_0^{(0)},$$

in other words, 1 population is the "chosen 1", which will be on the spine, and the rest are not. In step 1, this means

$$Z_1 = 1 + W_1^{(0)} + W_0^{(1)},$$

where we have divided the descendants into those not on the spine in the first generation, $(W_1^{(0)}) \stackrel{\text{law}}{=} Z_0 - 1$, all those non-chosen descendants of the chosen 1, $W_0^{(1)} \stackrel{\text{law}}{=} X^s - 1$, and the new chosen 1. Proving the representation holds can be checked by induction. □

Lemma 25 ($\mathbb{Q}_n$ is the size bias): Under $\mathbb{Q}_n$ the random variable $(Z_k : 0 \leq k \leq n)$ has the law of a size-biased Galton-Watson tree.

Proof. This is similar to the exercise 16 in that we are biasing a measure by a space-time harmonic function. Without repeating the details, if P is the tpm of the chain $(Z_k : 0 \leq k \leq n)$ under Pr, then the tpm at step k under $\mathbb{Q}_n$ is given by

$$Q_k(x, y) = P(x, y)\frac{h(k+1, y)}{h(k, x)},$$

where $h(k, x) = x/\mu^k$. Thus this ratio is

$$\frac{h(k+1, y)}{h(k, x)} = \frac{y}{\mu x},$$

which means that Z_{k+1} conditioned on $\mathscr{F}_k$ follows precisely the size-bias of the conditional law of Z_{k+1} given Z_k. By Lemma 23, this is exactly the transition probability matrix in the definition of the size biased Galton Watson tree. □

So under $\mathbb{Q}_n$ we actually have a probabilistic description of the population size. The main estimate that we need to make is a bound for Z_n^s/μ^n.

Lemma 26 (Boundedness of the size bias GW): Suppose that $\mathbb{E}X \log X < \infty$ and $\mathbb{E}X > 1$. Then $(M_n : n \geq 0)$ is a uniformly integrable martingale.

This already implies $M_n = \mathbb{E}(M_\infty \mid \mathscr{F}_n)$, and moreover $\mathbb{E}M_\infty = 1$. Hence, the growth rate Z_n/μ^n is asymptotically nontrivial, which is to say it is not 0 almost surely.

Proof. We need to show that for every $\epsilon > 0$ there is an K sufficiently large that

$$\epsilon > \mathbb{E}(M_n \mathbf{1}_{|M_n|>K}) = \mathbb{Q}_n(Z_n/\mu^n > K\mathbb{E}Z_0).$$

Now we have already seen that under $\mathbb{Q}_n$, the law of Z_n has the law of a size-bias Galton Watson process, i.e.

$$\mathbb{Q}_n(Z_n/\mu^n > K\mathbb{E}Z_0) = \Pr(Z_n^s/\mu^n > K\mathbb{E}Z_0).$$

Using the representation in Lemma 24, if we condition on $\mathcal{G} := \sigma(W_0^{(k)} : k \geq 1)$, we have

$$\mathbb{E}(Z_n^s/\mu^n \mid \mathcal{G}) \leq \sum_{k=0}^{n} \frac{W_0^{(k)}}{\mu^k} \leq \sum_{k=0}^{\infty} \frac{W_0^{(k)}}{\mu^k}.$$

To lighten the notation let $Y_k = W_0^{(k)}$. It suffices to show that this sum is finite almost surely.

These are now iid random variables, which have the law of the size biases of X. By assumption $\mathbb{E}X \log X < \infty$, which implies $\mathbb{E} \log Y_k < \infty$.

Now, for any $\alpha > 1$,

$$\Pr(Y_k > \alpha^k) \leq \Pr(\log Y_k > k \log \alpha).$$

Using that these are iid, for some constant C_α,

$$\sum_{k=1}^{\infty} \Pr(\log Y_k > k \log \alpha) \leq \sum_{k=1}^{\infty} \Pr(\log Y_1 > k \log \alpha) \leq C_\alpha \mathbb{E} \log Y_1,$$

and so by Borel Cantelli, there are at most finitely many k so that $Y_k > \mu^{k/2}$. It follows that

$$\sum_{k=1}^{\infty} \frac{Y_k}{\mu^k} < \infty \text{ a. s.},$$

which implies the uniform integrability. □

Theorem 43: Half of Kesten-Stigum

If $\mathbb{E}X \log X < \infty$, and $\mathbb{E}X > 1$ then $\{M_\infty = 0\}$ has probability Δ (and so with probability 1, if the tree does not go extinct, its population grows like $M_\infty \mu^k$).

Proof. Suppose that the initial distribution has 1 element almost surely. By performing a first step analysis, each descendant of the root must die out almost surely for $M_\infty = 0$. However all of these have the same probability of dying out. So the probability q that $M_\infty = 0$ can be seen to satisfy

$$q = f(q),$$

where f is the pgf of the offspring distribution. The only roots of this are Δ and 1. As we know it is not 1, it must be Δ.

□