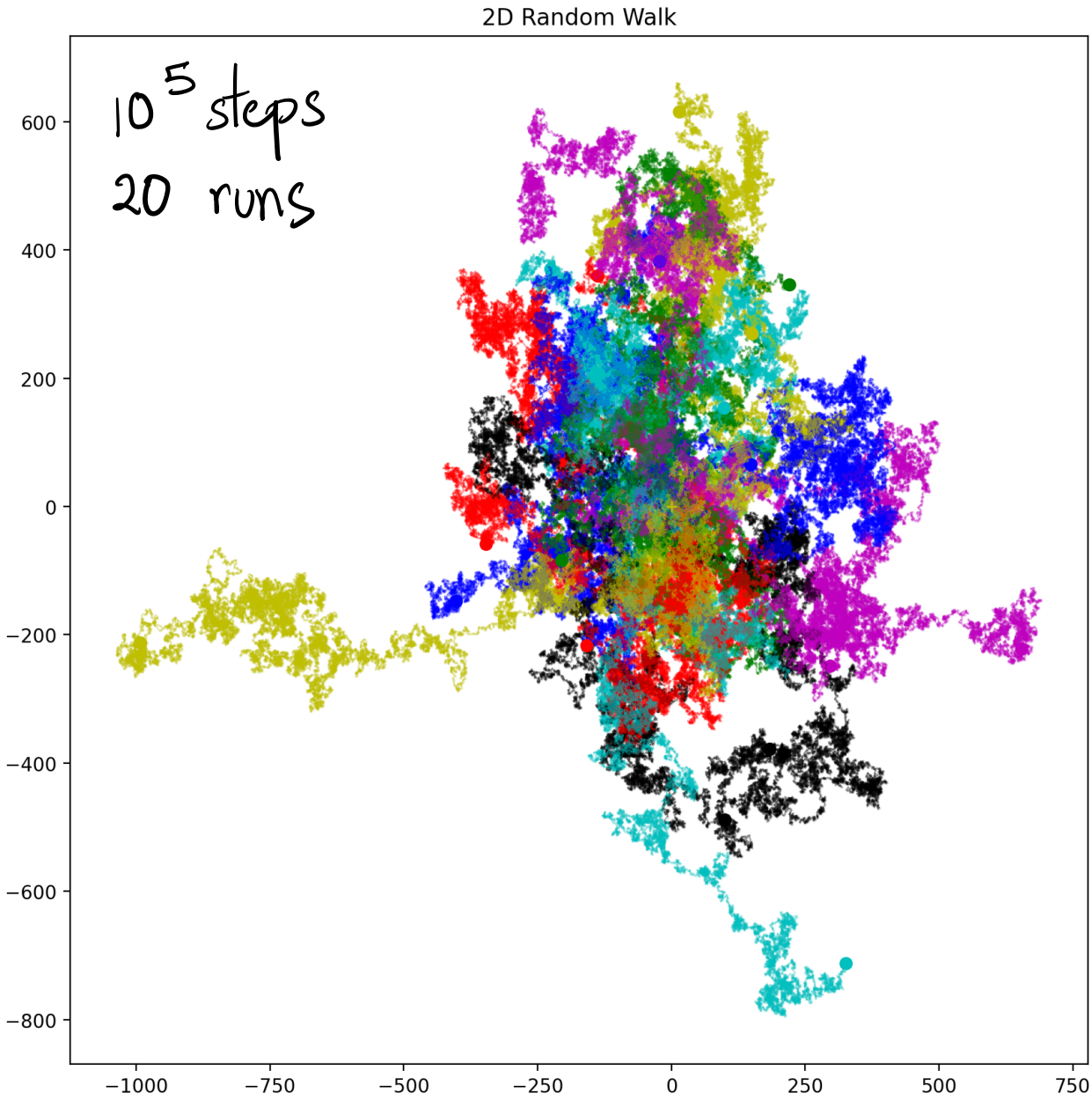
Semester: Winter 2023
Instructor: Paquette, Elliot

Stochastic Processes



Course Content. Markov chains. Random walks. Martingales.

Summary of Contents

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. 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. 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. 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

[1] G

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.

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 $\{X \in \{3,4\}\}$.

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

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 $\mathcal{G} \subset \mathcal{F}$. This will be a direct generalization of the conditioning considered above by taking $\mathcal{G} = \sigma(X)$.

Definition 7. 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 [2][Chapter 4]):

Theorem 6. 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:

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

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 $\mathcal{G} \subseteq \mathcal{F}$

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

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 $\mathcal{H} \subseteq \mathcal{G}\mathcal{F}$,

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

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

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

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

1. For each $B \in \mathcal{A}$, $\mathcal{P}(B, \cdot)$ is a version of the conditional expectation $\mathbb{E}(\mathbf{1}_B(Y) \mid \mathcal{G})$.
2. There is a $\mathcal{G}$–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. 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. 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. 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.

[3]

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

Theorem 12. 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*:

[3] 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.

Definition 11. 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 kerernel 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 distribted, 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

Auppose 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 just induction and the Markov property. In the other direction, from Theorem 12, the finite-dimensional marginal determines the law and moreover it should just be checked that the Markov property follows from the claim.

Exercise 1. 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 $(X_k : k > j)$ conditionally on the past $(X_k : k \leq j)$ is the same as the law of the future conditionally on the present X_j.
2. For any $j \geq 0$, conditionally 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 [a]).
You can model this with a Markov chain having tpm:

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

[a] [

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 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 $\mu \boldsymbol{P}^n$, meaning the vector matrix multiplication.

Proof. For the first claim, by Baye's law

$$\Pr(X_j = a_j, \quad \forall 0 \le j \le n) = \Pr(X_n = a_n \mid X_j = a_j, \quad \forall 0 \le j \le n-1) \times \Pr(X_j = a_j, \quad \forall 0 \le j \le n-1).$$

By the Markov property

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

Finally by the definition of the tpm, we conclude

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

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. □

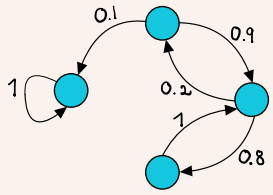
We can also give a graphical representation of THFS Markov chains, which can be helpful in understanding the behavior of small chains.

Definition 18. 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 $\boldsymbol{P}_{a,b}$.

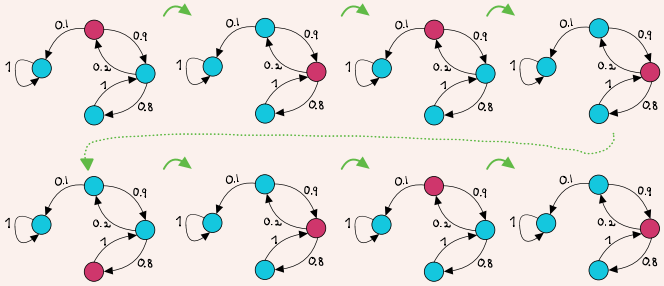
Example 9: Simple transition graph



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

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.** 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. 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,

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

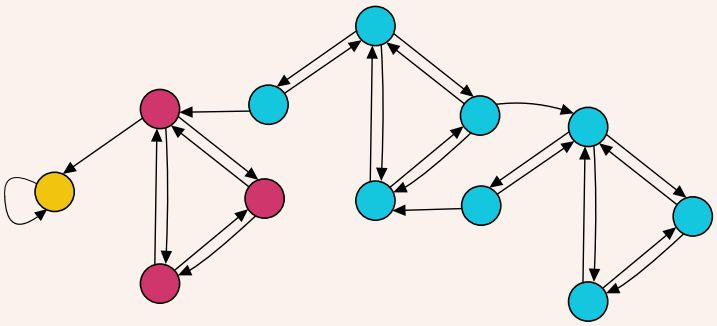
□

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. 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. Show that positive recurrence is a class property.

Exercise 4. 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 $\Pr_y(\tau_x) < \infty$.

Exercise 5. 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. 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. The *period* of a state x is the greatest common divisor of the set

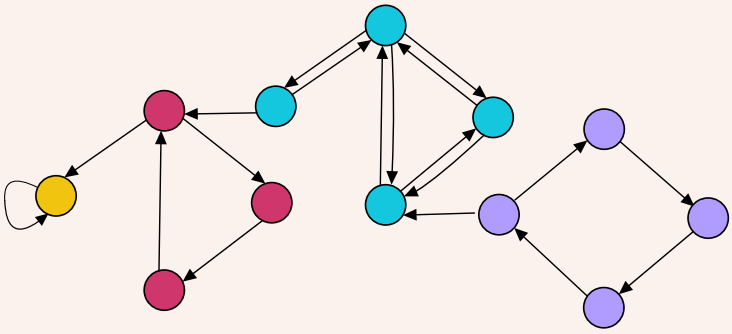
$$\{n \in \mathbb{N} : \boldsymbol{P}^n_{x,x} > 0\}.$$

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

Theorem 19. 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.

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. 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. An irreducible THCS Markov chain is positive recurrent if and only if it has a stationary distribution π.

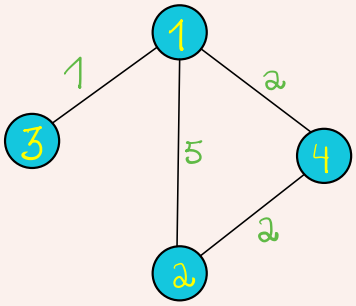
Example 12: Edge weighted graphs

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 $x \in S$

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

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 7. 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,

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

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

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

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 7. 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. 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. 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. 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 $(X_j : j \geq 0)$ and $(Y_j : j \geq 0)$ 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

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

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. □

> **Exercise 8.** (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.*

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 **??**. 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. [4]

Definition 29. 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 $S' = \{x \in S : \pi(x) > 0\}$.

[4] 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 9. 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*.[5]

[5] 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 $\boldsymbol{P} = \boldsymbol{Q}$.[6] [7]

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

[7] 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). [8]

[8] 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.

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,

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

Theorem 25. 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 17: Statisical decipher

This is a decoding

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. See [9]. 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.

[9]

Example 18: IsingModel

The Ising Model

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. 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.

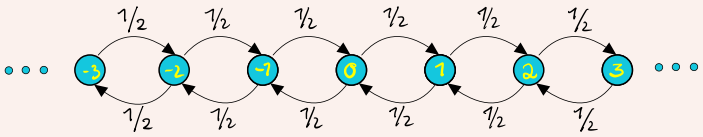
The canonical example of a martingale is simple symmetric random walk:

Example 19: 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. One of the key examples is the following:

Example 20: 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 21: hitting

Suppose that $(X_n : n \geq 0)$ is an irreducible finite state martingale 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:

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

It is also helpful to extend the definition of martingale to processes in which the equality in Definition 33 is rather an inequality. [10]

Definition 33. 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 $n \geq 0$.
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 10. 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 11. (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

[10] 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!

an $n \in \mathbb{N}$. Define $h(k,x) := \mathbb{E}(\phi(X_n)|X_k = k)$ 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 set $x \in S$ with $p = \Pr(X_n = x) > 0$. If we take $\phi(x) = \mathbf{1}_{x\in S}\frac{1}{p}$, then the law $\mathbb{Q}(\cdot) := \Pr(\cdot\phi(X_n))$ 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. *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. 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)[11]

Any adapted process can be made into a martingale:

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 12. Check that a process is a submartingale if and only if its compensator is almost surely non-decreasing.

The bracket process is an important is an important special case. [12] Define

[11] 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.

[12] 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.

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.

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

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*. [13]

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

[13] 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. 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 9. Suppose $(M_n : n \geq 0)$ is a $(\mathscr{F}_n : n \geq 0)$–adapted processn 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*. [14]

[14] 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).

Definition 37. 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. 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! [15]

[15] 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

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

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

(Note)[16]

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

Proof. The process M^τ is again a submartingale, and hence $\mathbb{E}M_n^\tau$ is increasing in n and so is always larger than $\mathbb{E}M_0^\tau = \mathbb{E}M_0$. As τ is almost surely bounded, there is some M deterministic so that $\Pr(\tau > M) = 0$. Hence taking $n = M$, we have $M_n^\tau = M_\tau$ almost surely.

The latter two are applications of dominated convergence. □

Example 22: 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 $(1 - p)^k \dots$). 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 13. 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.

Exercise 14. 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 $M_n := e^{\beta X_n - n a(\beta)}$ is a martingale for any $\beta \in \mathbb{R}$. *This is called the exponential martingale.*

2. Compute $\mathbb{E}_1 s^{\tau_0^+}$ for $s \in (0,1)$.
3. Conclude that $\Pr_1(s^{\tau_0^+} = k) \asymp \frac{1}{\sqrt{k}}$ in the case of SSRW.
4. Bonus: you can find from this the probability of never returning for the biased case (for $p > \frac{1}{2}$) and you can show it has an exponential tail in the case that $p < \frac{1}{2}$.

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. [17] The prototypical constant speed martingale is 1-d SSRW, which is null-recurrent. [18] 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

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

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 $k \ge 2$ 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*:

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

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

Lemma 10. Suppose that $(M_n : n \ge 0)$ is a submartingale. Then

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

[17] 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).

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

Proof. Let $Y_n := a + (M_n - a)_+$. By Exercise 10, if M is a submartingale, then so is Y. Set

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

Then by how Y is chosen,

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

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 28. (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) [19]

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

Corollary. A nonnegative supermartingale $(M_n : n \geq 0)$ converges almost surely.

[19] 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.

Example 23: 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 15. Using optional stopping, and martingale convergence, bound the probability that 1-d SSRW started 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 24: 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 16. 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)[20]

As promised, another structure under which the discrete martingale structure is useful is when the increments of a random walk are bounded.

Theorem 29. 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 fi-

[20] 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 $p_{=}\theta$. 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).

nite 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 28), 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

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

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 $Y^\vartheta_\infty = Y_\vartheta$ almost surely on the event $[Y]_\infty = \infty$ By optional stopping, [21]

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

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 ∞.[22] □

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

[22] 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 .

The Pólya theorem & harmonic functions

In Example 22, 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 30. (Pólya Theorem) SSRW on $\mathbb{Z}^d$ is recurrent if and only if $d \leq 2$.

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 [23]

$$\sum_{n=1}^{\infty} \boldsymbol{P}^{2n}_{o,o} = \infty \quad \text{iff} \quad d \leq 2.$$

In dimension 1, the return probability is explicit, since

$$\boldsymbol{P}^{2n}_{o,o} = \Pr(\text{Binom}(2n, \tfrac{1}{2}) = n) = \binom{2n}{n} 2^{-2n} \asymp \frac{1}{\sqrt{n}}.$$

[23] We also use that all the SSRWs on $\mathbb{Z}^d$ are 2-periodic.

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 17.** 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 38.** 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.

> **Theorem 31.** 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

$$h(y) := \mathbb{E}_y N_x = \sum_{n=0}^{\infty} \boldsymbol{P}_{y,x}^n,$$

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 [24]

$$\begin{aligned}\mathbb{E}_y(h(X_1)) &= \sum_{n=0}^{\infty} \mathbb{E}_y(\boldsymbol{P}^n_{X_1,x}) \\ &= \sum_{n=0}^{\infty} \boldsymbol{P}^{n+1}_{y,x} \\ &\leq h(y).\end{aligned}$$

□

[24] 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 11. 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 31, 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 [25]

$$\Pr_y(X_n = x) \approx e^{-\|y-x\|_2^2/cn} n^{-d/2}.$$

[25] 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:

$$\begin{aligned}(\|y\|^2 + \pm 2y_j + 1)^{-\alpha} &= (\|y\|^2)^{-\alpha} \\ &\quad \mp 2\alpha y_j (\|y\|^2)^{-1-\alpha} \\ &\quad + (4(1+\alpha)\alpha y_j^2 - 2\alpha\|y\|^2)(\|y\|^2)^{-2-\alpha} \\ &\quad + \mathcal{O}((\|y\|^2)^{-3/2-\alpha}).\end{aligned}$$

Using this, we conclude that

$$|\mathbb{E}_y(h_0(X_1) - h_0(y))| \leq C\|y\|^{-d-1},$$

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 $h_\epsilon(y) := (\|y\|_2^2)^{-(d/2-1-\epsilon)}$

$$\mathbb{E}_y(h_0(X_1) - h_0(y)) \leq 0$$

for all $\|y\|_2 \geq C(\epsilon)$. Thus if we take $h(y) = \min\{h_\epsilon, \delta\}$ for some sufficiently small $\delta > 0$, we conclude

$$\mathbb{E}_y(h(X_1) - h(y)) \begin{cases} = 0 & \|y\| \text{ if } \|y\| \text{ is sufficiently small} \\ \leq 0 & \text{otherwise .} \end{cases}$$

Hence h is a positive, nonconstant martingale, and so SSRW is transient. □