Stochastic processes

Nathanaël Berestycki

January 13, 2023

Contents

Introduction

Lecture 1; Thursday 06.10.2022

This course is an introduction to **stochastic processes**, a notion which is central to both classical and modern probability theory. A stochastic process simply describes the evolution of a system which is subject to randomness. Thus formally, a stochastic process is nothing else but a sequence

$$X = (X_0, X_1, \ldots)$$

of random variables taking values in some space S (called the **state space**) which may be the real line or some more exotic space. Such a sequence will be denoted thourghout the notes indifferently by $(X_n, n \geq 0)$ or $(X_n)_{n\geq 0}$. Plainly, this is a very broad notion, and we will need to focus our attention on *how* the underlying randomness affects the evolution of the system in order to be able to say something interesting. In this course and in these notes we will encounter two main types of stochastic processes: on the one hand, **Markov chains** and on the other, **martingales**.

Markov chains are the ones which will occupy us the most in this course, and therefore we only discuss these in this introduction. Informally, a Markov chain is simply a stochastic process in which the future evolution of the system depends only on its current state, and is otherwise independent of its past: to put it another way, at any time $n \geq 0$, if we want to guess something about some future state of the system (say at time $m \geq n$) then the only relevant information is the current state X_n, whereas additional information about $X_0, \ldots, X_{n-1}$ would not impact our guesses. A formal definition will follow, but at this stage it is already sufficient and useful to see that a huge variety of systems intuitively fit this definition.

Example 0.1. The evolution of the genome as it undergoes successive mutations. If we think of the genome of an individual as initially given by some sequence of letters, say $X_0 = \texttt{ATTTCATG}\ldots$, and suppose that at each time step a random mutation occurs (e.g. substitution, deletion, reversals/transpositions, etc.). Clearly, for any time $n \geq 0$, only the current state of the genome is relevant to make predictions about what the genome might look like in the future, whereas *how* it got to that state is irrelevant. This therefore forms an example of a Markov chain.

Example 0.2. A deck of card is being repeatedly shuffled at random. We can describe this evolution by a stochastic process $(X_n, n \geq 0)$ where X_n denotes the state of the deck at time n and the state space is the premutation group $S = S_{52}$ of 52 elements (cards). Clearly, only the current state of the deck is relevant to describe its future, not the previous shuffles leading to the current deck. This too forms a Markov chain.

Example 0.3. The motion of a particle in a turbulent fluid. Here the random variable X_n might be the position and velocity of the particle at time n, so the state space is $S = \mathbb{R}^3 \times \mathbb{R}^3$. This too forms a Markov chain: the past velocities and positions of the particle have no impact on the future trajectory of the particle beyond its current values.

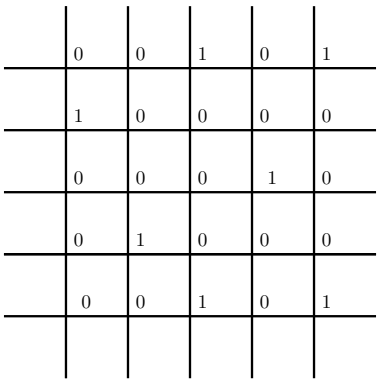


Figure 1: How will the infection spread?

Example 0.4. The spread of an infection through a population. We are given a graph $G = (V, E)$, say the square lattice. Here X_n could denote the assignment of a value 0 or 1 to each vertex of the graph, where a 0 denotes a healthy individual, and a 1 denotes an infected individual (so the state space is $S = \{0, 1\}^V$). See Figure 1. Let us suppose that, at each successive time step, an infected individual infects a random proportion of its neighbours. Then X_n forms a Markov chain: only the current state of the infection is relevant to make predictions about its future evolution, whereas its past would not give us additional information.

Example 0.5. The *Markov Chain Monte Carlo* (MCMC) algorithm – one of the most used algorithms throughout industry, is based on Markov chains. It is a little too sophisticated to explain at this stage, but let us say for now that it involves simulating a sequence of random variables $(X_n, n \geq 0)$ which forms a Markov chain, in order to approximate some desired distribution.

As these examples already show, Markov chains are ubiquitous and arise in a truly bewildering variety of contexts. Perhaps surprisingly, it is possible to build an elegant theory which encompasses all these examples at once. The results are nontrivial mathematically, and interesting for applications. It is a uniquely successful theory in that regard!

In all cases, the questions which will be of interest to us will be of the following type: will the Markov chain ever reach a particular configuration (or sets of configurations) we are interested in? If, so how long can we expect it to take? And what about the *long-term behaviour of the Markov chain*? What will the infection look like in the long run? Will it die out or survive forever? If so, how will the infected individuals be distributed in the population? What about the distribution of the deck of cards – can we guarantee it will be well shuffled in some suitable sense? What about the genome evolution? And in the case of the MCMC algorithm, can we guarantee that the approximation procedure works?

Surprisingly, all these questions can to some extent be answered by a unified theory, which we are about to develop. A remarkable fact is that part of the answer to these questions

involves beautiful connections to harmonic analysis and more specifically to sets of equations that are best viewed as discrete versions of Partial Differential Equations (PDEs). On the one hand, such connections illuminate our understanding of PDEs by giving a concrete ("microscopic") description of the systems they represent. Conversely, the physical intuition one gains from formulating these problems as discrete PDEs is far-reaching. This connection is a recurring theme of these notes, and, I believe, of probability theory in general.

1 Markov chains: definition, basic properties

1.1 Basic definition

The theory we will develop is restricted for convenience to countable state spaces (some considerable analytical subtleties arise in the case where the state space isn't countable). Let $S = \{x_1, x_2, \ldots\}$ denote a *countable* set. We call S the **state space** of the Markov chain, and we call its element states. We say that $\lambda = (\lambda_x)_{x \in S}$ is a **measure** on S if

$$0 \leq \lambda_x < \infty, \qquad x \in S.$$

If furthermore $\sum_{x \in S} \lambda_x = 1$, we say that λ is a **distribution** on S.

Let X denote a random variable with values in S. (Technically, this means we are given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a measurable function $X : \Omega \to S$ but we will never specify this in the future). Then if we define $\lambda_x = \mathbb{P}(X = x)$, $\lambda = (\lambda_x)_{x \in S}$ is a distribution on S, which we call the **law** of X.

Example 1.1. An unbiased die is rolled, let X be the outcome. Then the corresponding state space is $S = \{1, \ldots, 6\}$ and the law of X is uniform: $\lambda_x = 1/6$ for any $x \in S$.

We are now ready to give the definition of stochastic processes and Markov chains.

Definition 1.2. *A **stochastic process** $(X_n, n \geq 0)$ with values in S is simply a sequence of random variables taking values in S.*

Definition 1.3. *A stochastic process $(X_n, n \geq 0)$ with values in S is called a **Markov chain** if there exists functions $P_n : S \times S \to [0, 1]$, called the **transition matrices** of the chain, such that for every $n \geq 0$, for every $x_0, \ldots, x_n, x_{n+1} \in S$,*

$$\mathbb{P}(X_{n+1} = x_{n+1} | X_0 = x_0, \ldots, X_n = x_n) = P_n(x_n, x_{n+1}). \tag{1.1}$$

It is necessary to make a few comments on this definition.

(1) Given some arbitrary stochastic process $X = (X_n, n \geq 0)$, we would normally expect the left hand side of (1.1) to depend on $x_0, \ldots, x_n$ and x_{n+1}. The defining property of a Markov chain is to say that it depends only on x_n (current state), x_{n+1} (the next state), and n.

(2) It is not hard to check (exercise!) using the law of total probability that if (1.1) holds, then the right hand side is necessarily equal to $\mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n)$ (exercise!). Thus in words, $P_n(x, y)$ gives us the probability that, if the chain was in state x at time n, it would jump to state y at time $n + 1$.

(3) While P_n is really a function from $S \times S \to [0, 1]$, we think of it as a matrix

$$P_n = (P_n(x, y))_{x,y \in S}$$

indexed by the elements of S. When S is finite, then the dimension of that matrix equals the size of S and we are in the realm of linear algebra (as we will see later, this is extremely

useful). We will also think of P_n as a matrix even when S is infinite. In that case, P_n is (doubly) infinite. Nevertheless, the entry of that matrix for the row x and column y gives us the probability to jump from x to y at time n.

Definition 1.4. *We call a Markov chain* ***time-homogeneous*** *if its transition matrix P_n does not depend on n: that is, $P_n = P_0$ for all $n \geq 0$. In that case, $P = P_0$ is called* the *transition matrix of the chain.*

Example 1.5. We throw n dice, and let X_n be the sum of the dice modulo 2. Then X is a Markov chain. Its transition matrix is

$$P = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$$

and so X is time-homogeneous.

It is not hard to see that if X is a (possibly time-inhomogeneous) Markov chain, then $Y_n = (n, X_n)$ is a time-homogeneous Markov chain (exercise!). There is therefore no loss of generality in considering such chains. In the rest of this course we focus only on time-homogeneous chains, and do so without further mention.

If X is a Markov chain, the transition matrix P is an example of what in linear algebra is called a **stochastic matrix**: that is, $P(x, y) \geq 0$ and

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

i.e., the sum of each row is equal to 1 (exercise: prove it!). In other words, for each fixed $x \in S$, $(\lambda_y)_{y \in S} = (P(x, y))_{y \in S}$ defines a distribution on S. This is none other than the law of the Markov chain after one step, if it starts from x.

1.2 Specifying a Markov chain; diagram

To determine a Markov chain on a countable state space S, one must specify:

- A stochastic matrix P (which will be the transition matrix of the chain)
- A distribution λ which will be the initial distribution of the chain.

Given such a λ and P, we say that the stochastic process $(X_n, n \geq 0)$ is **Markov** (λ, P) if X is a Markov chain with transition matrix P, and for every $x \in S$, $\mathbb{P}(X_0 = x) = \lambda_x$.(Together, λ and P determine a unique law on S-valued stochastic processes – showing this would require a bit of measure theory, but this will not be needed in the following).

Often, but not always, the starting point of a Markov chain is a fixed state, call it x_0. The associated starting distribution λ is simply the **Dirac mass** at x_0: $\lambda_x = 1_{x=x_0}$. We write $\lambda = \delta_x$ in the following.

Example 1.6. Let $\lambda = (1/2, 1/2)$ and for $0 \leq \alpha, \beta \leq 1$ let

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}.$$

What is the Markov chain described by λ and P? First of all, the state space is $S = \{0, 1\}$ say – we could of course choose the label the states S differently, but S has no matter what two states. Since $\lambda = (1/2, 1/2)$ the chain is equally likely to start in 0 or 1. Reading the rows of P, we see that:

- if at $x = 0$, the chain stays at x with probability $1 - \alpha$, and jumps to $y = 1$ otherwise.
- if at $x = 1$, the chain stays at x with probability $1 - \beta$, and jumps to $y = 0$ otherwise.

It is sometimes convenient to represent a transition matrix P by a **diagram**, where each possible transition is represented by an arrow, and we label the arrow with the probability to make this transition. In the above example, the diagram is as in Figure 2.

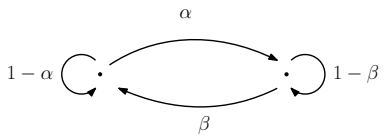


Figure 2: The diagram associated to the transition matrix of Example 1.6

Since the sum of all transitions from a point x is always equal to 1, we can also choose not to include any self-loops. This is a little easier to read if the chain is a bit bigger.

1.3 Examples: random walks on graphs

Lecture 2; Friday 7.10.2022 The following example is a fundamental example of a Markov chain, which will come back on many occasions in this course. Let $G = (V, E)$ be a graph. That is, V is a (countable) collection of vertices, and $E \subset V \times V$. We assume that the graph G is locally finite, i.e., for every $x \in V$, $\deg(x) < \infty$.

Associated to this graph, we can define a transition matrix P as follows:

$$P(x, y) = \begin{cases} \frac{1}{\deg(x)} & \text{if } (x, y) \in E \\ 0 & \text{else.} \end{cases}$$

In words, P describes the transitions of an ant walking on G, which at each time step jumps to a uniformly chosen neighbour of its current position. See Figure 3 for an example.

Example 1.7. Random walk on $\mathbb{Z}^d$. A very important graph is the cubic lattice $\mathbb{Z}^d$. Here $d \geq 1$ is the dimension, and $x \in \mathbb{Z}^d$ is called a neighbour of $y \in \mathbb{Z}^d$ if $\sum_{i=1}^{d} |x_i - y_i| = 1$. With an abuse of notation we call $\mathbb{Z}^d$ both the set of vertices and the corresponding graph.

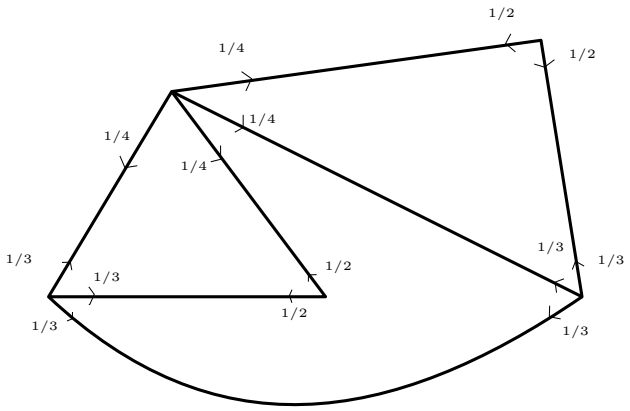


Figure 3: Graph (here undirected for simplicity) and associated transition matrix and diagram

For instance, you might think of a tourist walking at random in Manhattan ($d = 2$). A natural question, whose answer will occupy us for a significant portion of this course, would be: will she ever return to her hotel? Or in $d = 3$, you could think of a bird flying at random. Will it ever return to its nest? We will later see that the answer to these questions depends very much on the dimension d. In this case, we will prove that tourists are more lucky than birds...

Many more examples can be described as random walks on graphs, even if that is not always immediately apparent.

Example 1.8. Card shuffling. Consider the following card shuffling method. We start with an ordered deck. Then at each step, we select a pair of cards uniformly at random in the deck, and exchange them. (Thus shuffling method is very inefficient, but is a canonical example from the mathematical point of view.) This is a Markov chain which can be realised as a random walk on a suitable graph. Let V denote the permutation group on n elements (with $n = 52$ if it is a real deck of cards). Define a graph on V as follows: say that two permutations σ, σ' are nieghbours if

$$\sigma' \cdot \sigma^{-1} \in T$$

where $T = \{(i, j) : 1 \leq i < j \leq n\}$ is the set of transpositions. (In group theory, this graph is called the **Cayley graph** of S_n generated by the set of transpositions.) Then the above card shuffling is simply the random walk on this graph, started from the identity permutation.

1.4 (Weak) Markov property

While the definition of a Markov chain via conditioning (see (1.1)) is somewhat intuitive, it is not very convenient to work with in practice. That is because conditional probabilities

tend to be awkward to manipulate and cumbersome. The following equivalent condition is somewhat a little easier.

Proposition 1.9. *Let S be a countable state space, P a transition on matrix and λ a distribution on S. A stochastic process $(X_n, n \geq 0)$ with values in S is Markov (λ, P), if and only if, for every $n \geq 0$ and every $x_0, \ldots, x_n \in S$,*

$$\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) = \lambda_{x_0} P(x_0, x_1) \ldots P(x_{n-1}, x_n). \tag{1.2}$$

Proof. Suppose X is Markov (λ, P). We proceed by induction. The formula (1.2) is clear for $n = 0$. Furthermore, for $n \geq 1$,

$$\begin{aligned}\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) &= \mathbb{P}(X_0 = x_0, \ldots, X_{n-1} = x_{n-1}) \times \mathbb{P}(X_n = x_n | X_0 = x_0, \ldots, X_{n-1} = x_{n-1}) \\ &= \lambda_{x_0} P(x_0, x_1) \ldots P(x_{n-2}, x_{n-1}) \times P(x_{n-1}, x_n)\end{aligned}$$

by the induction hypothesis and the definition of a Markov chain. This proves (1.2).

Conversely, suppose (1.2) holds. Then clearly (applying it with $n = 0$) $\mathbb{P}(X_0 = x) = \lambda_x$ so λ is the initial distribution of X. Now let us show it is a Markov chain with transition matrix P: we have, if $n \geq 1$,

$$\begin{aligned}\mathbb{P}(X_n = x_n | X_0 = x_0, \ldots, X_{n-1} = x_{n-1}) &= \frac{\mathbb{P}(X_n = x_n; X_0 = x_0, \ldots, X_{n-1} = x_{n-1})}{\mathbb{P}(X_0 = x_0, \ldots, X_{n-1} = x_{n-1})} \\ &= \frac{\lambda_{x_0} P(x_0, x_1) \ldots P(x_{n-2}, x_{n-1}) \times P(x_{n-1}, x_n)}{\lambda_{x_0} P(x_0, x_1) \ldots P(x_{n-2}, x_{n-1}) \times P(x_{n-2}, x_{n-1})} \\ &= P(x_{n-1}, x_n)\end{aligned}$$

as desired. □

With this characterisation we can formulate an important property of Markov chains, called the Markov property. In some sense it is just another way of expressing the definition of a Markov chain. Later on this property will be superseded by a more powerful version which will be called the strong Markov property.

Proposition 1.10. *Let X be Markov (λ, P) and let $m \geq 0$ and $x \in S$. Conditionally given $\{X_m = x\}$, the sequence $(X_m, X_{m+1}, \ldots)$ is Markov (δ_x, P) and is independent of $(X_0, \ldots, X_m)$.*

To explain the statement, recall that δ_x is a Dirac mass at x (equal to 1 at x and 0 elsewhere). Note also that the statement would be unchanged if the conclusion was "... and is independent of $(X_0, \ldots, X_{m-1})$". This is because we are conditioning on X_m. Under this conditional probability, X_m is not random (indeed it is equal to x) and constants are independent of everything: if c is a constant random variable, then X is independent of Y if and only if it is independent of (Y, c).

Proof. We first reformulate the statement in a more concrete way which can be verified by computations. Let $Y_0 = X_m, Y_1 = X_{m+1}, \ldots$. Let $\mathbb{P}^* = \mathbb{P}(\cdot | X_m = x)$ denote the conditional probability given $X_m = x$. Then to prove the proposition, we claim it suffices to show the following: for any $x_0, \ldots, x_m = x$ and any $x = y_0, y_1, \ldots, y_n$,

$$\begin{aligned} &\mathbb{P}^*(X_0 = x_0, \ldots, X_m = x_m, Y_0 = y_0, \ldots, Y_n = y_n) \\ &= \mathbb{P}^*(X_0 = x_0, \ldots, X_m = x_m) \times \delta_x(y_0) P(y_0, y_1) \ldots P(y_{n-1}, y_n). \end{aligned} \tag{1.3}$$

Let us explain why (1.3) implies Proposition 1.10. Indeed if (1.3) holds, let us denote by A the event $A = \{X_0 = x_0, \ldots, X_m = x_m\}$ and by B the event $B = \{Y_0 = y_0, \ldots, Y_n = y_n\}$. We first deduce from (1.3) by summing over $x_0, \ldots, x_{m-1} \in S$ that the second term in the right hand side is $\mathbb{P}^*(B)$. By Proposition 1.9, this implies that Y has the desired law. Furthermore, we deduce that $\mathbb{P}^*(A \cap B) = \mathbb{P}^*(A)\mathbb{P}^*(B)$ and thus A and B are independent under $\mathbb{P}^*$. Since A and B are arbitrary events describing completely the sequence $(X_0, \ldots, X_m)$ and Y respectively, we deduce that these two sequences are independent under $\mathbb{P}^*$.

Thus it suffices to verify (1.3). As this is a somewhat tedious computation using the definition of conditional probability and Markov chains, the proof is left as an exercise. □

1.5 Chapman–Kolmogorov equations and diagonalisation

Lecture 3; Thursday 13.10.2022

Given a Markov chain, we might ask what is the probability to find the chain in some given state y starting from x, not just after one step but after n steps. As we will see, this can be reduced to computing the matrix P^n.

It is worth reviewing the definition of P^n (also because this was only defined in linear algebra for finite matrices). Let $P^0 = I$ denote the identity matrix (that is, $I(x, y) = 1_{y=x}$). Set $P^0 = I$, and define by induction

$$P^n(x, y) = \sum_{z \in S} P^{n-1}(x, z) P(z, y).$$

When S is finite, the above definition corresponds to the matrix multiplication $P^n = P^{n-1} \times P$. When S is infinite, a few comments are needed to justify this definition: it is easy to see that this defines $P^n(x, y)$ as a nonnegative number, meaning that the sum which serves to define $P^n(x, y)$ is actually well-defined.

Given a measure λ, we may define λP in the same manner:

$$(\lambda P)(y) = \sum_x \lambda_x P(x, y)$$

which also corresponds to matrix multiplication if we think of λ as a row vector.

Theorem 1.11 (Chapman–Kolmogorov equations)**.** *Let X be Markov (λ, P). Then for all $n \geq 0$, $x, y \in S$, we have:*

(i) $\mathbb{P}(X_n = y | X_0 = x) = P^n(x, y)$

(ii) $\mathbb{P}(X_n = y) = (\lambda P^n)(y)$.

Proof. Note that conditionally given $X_0 = x$, X is simply Markov (δ_x, P) (by the Markov property at time $m = 0$). Let us write $\mathbb{P}_x$ for $\mathbb{P}(\cdot|X_0 = x)$. Then by the law of total probability,

$$\begin{aligned}\mathbb{P}_x(X_n = y) &= \sum_z \mathbb{P}_x(X_{n-1} = z, X_n = y) \\ &= \sum_z \mathbb{P}_x(X_{n-1} = z)\mathbb{P}_x(X_n = y|X_{n-1} = z) \\ &= \sum_z \mathbb{P}_x(X_{n-1} = z)P(z, y)\end{aligned}$$

by the Markov property at time $n-1$. The result (i) therefore follows by induction and the definition of P^n.

Point (ii) follows by conditioning on the value of X_0 and the law of total probability. □

To put it another way, P^n can by induction be seen to satisfy

$$P^n(x, y) = \sum_{x_1,\ldots,x_{n-1}\in S} P(x, x_1)\ldots P(x_{n-1}, y)$$

and the sum specifies all the ways that there are to reach y from x in n steps (and the summand indicates the probability of that particular path).

The Chapman–Kolmogorov equations can be used to compute $\mathbb{P}(X_n = y|X_0 = x)$ explicitly in some cases, usually when the state space is finite (and in fact typically quite small, as otherwise the computations become too unwieldly). In that case, suppose we can **diagonalise** the matrix P, i.e., we can write

$$P = U^{-1}DP$$

where

$$D = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ & & \ddots & \\ 0 & & \ldots & \lambda_k \end{pmatrix}$$

is a diagonal matrix of size $k = |S|$, and $(\lambda_i)_{1\le i\le k}$ are the **eigenvalues** of P. Then

$$P^n = U^{-1}D^nU = U^{-1}\begin{pmatrix} \lambda_1^n & 0 & \ldots & 0 \\ 0 & \lambda_2^n & \ldots & 0 \\ & & \ddots & \\ 0 & & \ldots & \lambda_k^n \end{pmatrix} U.$$

As a consequence, $P^n(x, y)$ is given as a fixed linear combination of the eigenvalues to the nth power, as summarised by the following result.

Theorem 1.12. *Suppose that the transition matrix P on the finite space S of size $k = |S|$ is diagonalisable. Then for each $x, y \in S$, there exist $a_1(x,y), \ldots, a_k(x,y)$ such that for any $n \geq 0$,*

$$P^n(x,y) = a_1(x,y)\lambda_1^n + a_2(x,y)\lambda_2^n + \ldots + a_k(x,y)\lambda_k^n.$$

The coefficients of this linear combination are often not so hard to compute in practice, by computing by hand the first few values of $P^n(x,y)$. We illustrate this method with an example.

Example 1.13. Consider the example of Example 1.6 and Figure 2, i.e.,

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}.$$

How can we compute $P^n(x,y)$?

We first compute the characteristic polynomial of P,

$$\begin{aligned} \chi(\lambda) &= \begin{vmatrix} 1-\alpha-\lambda & \alpha \\ \beta & 1-\beta-\lambda \end{vmatrix} \\ &= (1-\alpha-\lambda)(1-\beta-\lambda) - \alpha\beta \\ &= (\lambda-1)(\lambda-q) \end{aligned}$$

for some $q \in \mathbb{R}$ to be determined. (We know a priori that the expression of the second line may be simplified into an expression of the type in the third line, since it is a polynomial of the second degree with $\lambda = 1$ obviously a root and the coefficient of λ^2 is obviously equal to one). To compute q, we note that $\chi(0) = q$ so

$$q = (1-\alpha)(1-\beta) - \alpha\beta = 1 - \alpha - \beta.$$

Hence P is diagonalisable with two distinct eigenvalues, namely

$$\lambda_1 = 1, \lambda_2 = 1 - \alpha - \beta.$$

We deduce that

$$P^n(x,y) = A(x,y) + B(x,y)(1-\alpha-\beta)^n.$$

For instance say $x = y = 0$ is the state corresponding to the first row of P and let $A = A(0,0)$ and $B = B(0,0)$. Then

$$P^n(0,0) = A + B(1-\alpha-\beta)^n.$$

To compute A and B, note that $P^0(x,x) = 1$ so $A + B = 1$, and $P^1(x,x) = 1 - \alpha$, so

$$A + B(1-\alpha-\beta) = 1 - \alpha.$$

We deduce that $B(\alpha+\beta) = \alpha$, so

$$B = \frac{\alpha}{\alpha+\beta}; A = \frac{\beta}{\alpha+\beta}.$$

Thus, the final answer is

$$P^n(0,0) = \frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}(1-\alpha-\beta)^n.$$

As an **exercise**, compute $P^n(x,y)$ in all remaining three cases.

Remark 1.14. Note that the theorem assumes that P is diagonalisable, which can be a little difficult to check in practice. When we compute the characteristic polynomial χ of a transition matrix P and find that the roots are all distinct, then this implies by a theorem of linear algebra that P is indeed diagonalisable and thus Theorem 1.12 applies.

In some examples however, we find the correct number of roots but some coincide, and so we cannot immediately deduce that P is diagonalisable and apply Theorem 1.12. However, it can be shown that even in these cases we get a formula for $P^n(x,y)$: suppose λ is a root of χ of multiplicity equal to d. Then the "coefficient" in front of λ is, instead of a constant independent of n, a *polynomial* in n of degree $d-1$. For instance, if λ is a double root of χ, then the term λ^n in the expression for $P^n(x,y)$ comes with a "coefficient" of the form $An+B$. Thus this eigenvalue contributes $(An+B)\lambda^n$ to $P^n(x,y)$, and the expression of $P^n(x,y)$ is a sum over all eigenvalues of expressions of this type.

Remark 1.15. Keep in mind that in order to diagonalise P you are allowed to diagonalise it over the complex numbers, i.e. the eigenvalues are allowed to be complex. In that case, they will necessarily come in pairs of complex-conjugates, since $P^n(x,y)$ is a real quantity. This can sometimes be useful in reducing the number of unknowns that we need to solve for.

2 Hitting probabilities and recurrence/transience

2.1 Class structure

Our first task in our analysis of Markov chains will be to break the state space into pieces where the state "communicate" with one another: within each such piece, it is possible to eventually visit each state. To formulate this idea, we need the following definition.

Definition 2.1. *Let P be the transition matrix of a Markov chain on some state space S. Let $x, y \in S$. We say that x **leads to** y, and we write $x \to y$, if*

$$\mathbb{P}_x(X_n = y \text{ for some } n \geq 0) > 0.$$

Here and in the rest of these notes, $\mathbb{P}_x$ denote the law of the Markov chain conditioned to start in the state x, i.e., $\mathbb{P}_x(A) = \mathbb{P}(A|X_0 = x)$.

Remark 2.2. Observe two things. First, the property that x leads to y does *not* depend on the starting distribution of the chain. It depends only on what the chain can (or cannot) do *if* it starts from x.

Second, the choice of n in the definition above could be random (i.e., depend on the actual realisation of the Markov chain starting from x). That is, we are not asking that there is some fixed $n \geq 0$ such that the Markov chain has positive probability to be at y at time n, simply that there is some time, which can be random, at which the chain visits y.

Definition 2.3. *We say that x **communicates with** y, and we write $x \leftrightarrow y$, if $x \to y$ and $y \to x$.*

There is a useful characterisation of the notion of $x \to y$ in terms of the transition probabilities of the chain.

Proposition 2.4. *Fix two states $x, y \in S$. The following are equivalent.*

(i) $x \to y$.

(ii) there exists $n \geq 0$, and a sequence of states $x_0, \ldots, x_n$ with $x_0 = x$ and $x_n = y$, such that
$$P(x_0, x_1) \ldots P(x_{n-1}, x_n) > 0.$$

(iii) there exists $n \geq 0$ such that $P^n(x, y) > 0$.

This proposition is perhaps a bit surprising in view of Remark 2.2. In point (iii), we are precisely requiring that there is some fixed n (which, in this statement, is not allowed to be random) such that the chain has positive probability to be at y at this time n which was fixed in advance. It is surprising that this is equivalent to (i), since in this definition, the requirement seemed at first sight less strict.

Proof. We observe that (iii) $\Rightarrow$ (i) is immediate. For (i) $\Rightarrow$ (iii), observe that by σ-additivity (more specifically, Boole's inequality) we have

$$\begin{aligned}\mathbb{P}_x(X_n = y \text{ for some } n \geq 0) &= \mathbb{P}_x(\bigcup_{n=0}^{\infty} \{X_n = y\}) \\ &\leq \sum_{n \geq 0} \mathbb{P}_x(X_n = y) \\ &= \sum_{n \geq 0} P^n(x, y).\end{aligned}$$

Thus if the right hand side is 0, so is the left hand side and x cannot lead to y. This show (i) $\Rightarrow$ (iii).

The equivalence (ii) $\iff$ (iii) is even simpler: we have, by Chapman–Kolmogorov,

$$P^n(x, y) = \sum_{x_1, \ldots, x_{n-1} \in S} P(x, x_1) \ldots P(x_{n-1}, y).$$

So the left hand side is positive if and only if the right hand side is positive, which is equivalent to at least one term being positive. □

From (ii) it is immediate that if $x \to y$ and $y \to z$ then $x \to z$. Furthermore $x \to x$ always. Thus

$$x \leftrightarrow y \text{ defines an equivalence relation.}$$

Definition 2.5. *The partition of S defined by $\leftrightarrow$ are called the* ***communicating classes*** *of P. A chain in which there is a single communicating class is called* ***irreducible****.*

Example 2.6. What are the communicating classes in this example?

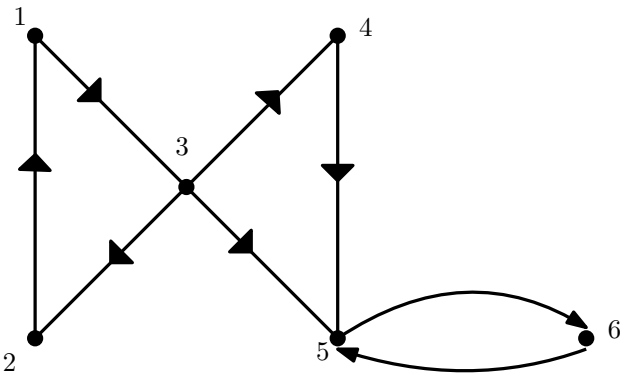


Figure 4: Only the transitions with positive probability are represented

Answer: $\{1, 2, 3\}, \{4\}, \{5, 6\}$.

Exercise 2.7. True or false? Let C be a communicating class. Starting from a state in C, the Markov chain remains in C forever.

2.2 Hitting times and probabilities

We begin with an important definition.

Definition 2.8. *Let $A \subset S$ be a collection of states. We call* ***hitting time of*** *A, and denote by T_A, the random variable*

$$T_A = \inf\{n \geq 0 : X_n \in A\}$$

with the convention $\inf \emptyset = +\infty$*. If $A = \{x\}$ where $x \in S$, we simply write T_x instead of $T_{\{x\}}$.*

In words, T_A is the first (smallest) time at which the chain enters in A and tells us how long we have to wait until the chain visits A. This is why we take $\inf \emptyset = +\infty$.

Example 2.9. A supermarket gives away stickers (also known as coupons) from a set of N possible coupons. At each visit to the supermarket, you receive a randomly chosen coupon. How long must you wait until you collect the complete set?

This can be formulated as the hitting of a Markov chain. Set X_k to be the number of coupons that remain to be collected after k visits to the supermarket (so initially $X_0 = N$). Then the time of interest to us is the hitting time of zero, namely T_0.

Lecture 4; Friday 17.10.2022

Usually we are interested in whether the chain will ever visit A (with what probability?), and how long this will take on average. We state the theorem regarding the probability to hit A; this cannot be computed using algebraic methods such as Theorem 1.12. Instead, this result tells us that the probability to hit A, viewed as a function of the starting point, satisfies a set of equations. This set of equations can be viewed as a discrete analogue to the **Dirichlet problem** from (harmonic) analysis, as we will discuss below.

Theorem 2.10. *Let A be as above and T_A the hitting time of A. Set $h_A(x) = \mathbb{P}_x(T_A < \infty)$ (recall that $\mathbb{P}_x$ denotes the probability for the chain starting from x). Then $h_A(x)$, viewed as a function of $x \in S$, is the **minimal nonnegative** solution to*

$$\begin{cases} h_A(x) = \sum_y P(x,y) h_A(y) & (\text{if } x \notin A) \\ h_A(x) = 1 & (\text{if } x \in A). \end{cases} \tag{2.1}$$

To explain the meaning of the word *minimal* in the above statement, this means if g is any other nonnegative solution to the equation, then $h_A(x) \leq g(x)$ for any $x \in S$.

Remark 2.11. A result of similar flavour can be established regarding the expectation of T_A and this will be part of the exercise sheets.

We will first prove this theorem and then discuss its significance.

Proof. Let us first check that h_A solves the "Dirichlet problem" (2.1). If $x \in A$ it is straightforward that $h_A(x) = 1$ since $T_A = 0 < \infty$ in that case. Let us suppose $x \notin A$. Then T_A cannot be equal to zero so $T_A \geq 1$. Let us decompose over the position of the chain at the first step: using the law of total probability, we get

$$\begin{aligned} h_A(x) = \mathbb{P}_x(T_A < \infty) &= \sum_{y \in S} \mathbb{P}_x(T_A < \infty, X_1 = y) \\ &= \sum_{y \in S} P(x,y) \mathbb{P}_x(T_A < \infty | X_1 = y) \end{aligned}$$

The second term in the sum depends only on the future since $T_A \geq 1$. Using the (weak) Markov property at time 1, we deduce that

$$\begin{aligned} h_A(x) &= \sum_{y \in S} P(x,y) \mathbb{P}_y(T_A < \infty) \\ &= \sum_{y \in S} P(x,y) h_A(y). \end{aligned}$$

So h_A is a solution of the Dirichlet problem (2.1).

Now let us check minimality. Let $g(x)$ denote another nonnegative solution and let us show $g(x) \geq h_A(x)$ for any $x \in S$ Obviously $g = h_A$ on A so let us consider the case $x \notin A$.

Then writing the equation that g satisfies and considering separately the cases where $y \in A$ and the cases where $y \notin A$, we get:

$$\begin{aligned} g(x) &= \sum_{y \in S} P(x,y) g(y) \\ &= \sum_{y \in A} P(x,y) \times 1 + \sum_{y \notin A} P(x,y) g(y) \\ &= \mathbb{P}_x(X_1 \in A) + \sum_{y \notin A} P(x,y) g(y). \end{aligned}$$

The first term has an interpretation but the second one is less obvious. The only thing we can do is use the equation satisfied by g but at the point y now (which is allowed, since $y \notin A$). Thus

$$\begin{aligned} g(x) &= \mathbb{P}_x(X_1 \in A) + \sum_{y \notin A} P(x,y) \sum_{z \in S} P(y,z) g(z) \\ &= \mathbb{P}_x(X_1 \in A) + \sum_{y \notin A} P(x,y) \left(\sum_{z \in A} P(y,z) \times 1 + \sum_{z \notin A} P(y,z) g(z) \right) \\ &= \mathbb{P}_x(X_1 \in A) + \mathbb{P}_x(X_1 \notin A, X_2 \in A) + \sum_{y \notin A} \sum_{z \notin A} P(y,z) g(z) \\ &= \mathbb{P}_x(T_A = 1) + \mathbb{P}_x(T_A = 2) + \sum_{y \notin A} \sum_{z \notin A} P(y,z) g(z) \end{aligned}$$

after once again considering separately the cases $z \in A$ and the cases $z \notin A$. A pattern is beginning to emerge. Reasoning by induction, we get for all $n \geq 0$:

$$g(x) = \mathbb{P}_x(T_A = 1) + \mathbb{P}_x(T_A = 2) + \ldots + \mathbb{P}_x(T_A = n) + \sum_{x_1, \ldots, x_n \notin A} P(x, x_1) \ldots P(x_{n-1}, x_n) g(x_n).$$

Since g is assumed nonnegative, we deduce for all $n \geq 0$.

$$g(x) \geq \mathbb{P}_x(T_A = 1) + \mathbb{P}_x(T_A = 2) + \ldots + \mathbb{P}_x(T_A = n)$$

Letting $n \to \infty$, this implies

$$g(x) \geq \sum_{n=1}^{\infty} P_x(T_A = n) = \mathbb{P}_x(T_A < \infty) = h_A(x),$$

as desired (where we used σ-additivity of $\mathbb{P}_x$ above). □

2.3 Physical interpretation and example

Theorem 2.10 calls for numerous comments, which we try to gather here. The equation

$$h(x) = \sum_{y} P(x,y) g(y) \tag{2.2}$$

is saying, in words, that the value of h is equal to the **average value** of h at its neighbours (where the averaging is done according to the transition probabilities $P(x, \cdot)$). This is perhaps clearest of in the case of a random walk on a graph, but remains true on any Markov chain. In other words, the function h_A satisfies the **mean-value property**. If h was a function on $\mathbb{R}^d$ instead of a function defined on S, the mean-value property is the property that $h(x)$ is equal to the average of its values on any sphere centered at x and contained on the domain in which h is defined. We could deduce from such a property that h must be a **harmonic function**:

$$\Delta h(x) = 0. \tag{2.3}$$

The equation (2.2) must be thought of as a discretised form of the Laplace equation (2.3). For the same reason, the equation (2.1) must be thought of as a discretised version of the continuous **Dirichlet problem**

$$\begin{cases} \Delta h(x) = 0 & \text{on } A^c \\ h(x) = 1 & \text{if } x \in A. \end{cases} \tag{2.4}$$

The second of these equations must be viewed as a boundary condition. This partial differential equation, or PDE for short, has a physical interpretation which is useful to bear in mind as it helps us mentally guess the hitting probabilities of A by a chain. Namely, the Dirichlet problem is precisely the equation satisfied by the **temperature** (viewed as a function of the space variable x) if we impose a temperature of 1 (in normalised units) on A. So if you view A as some kind of oven in intersideral space in which you maintain a constant temperature (say 200C if you are baking), this doesn't just heat your oven but also the space around it. In the stationary regime you would expect the temperature to be pretty close to 200C close to the oven, and to decay to zero "at infinity". The equation satisfied by this temperature would obey a non-normalised version of (2.4). This remarkable fact was discovered by Laplace (possibly in conjunction with the chemist Lavoisier) around 1782-1783.

Furthermore, as we will discuss in some examples below, the fact that h_A is not just *a* solution to the Dirichlet problem but *the* minimal nonnegative solution may be viewed as a way of specifying an additional boundary condition "at infinity". This is best understood and appreciated on examples; in the physical example of the oven above, one can see that harmonic functions equal to 1 are not necesarily unique: one can imagine setting a positive temperature "at infinity", which could modify the temperature away from A and from ∞, but cannot affect the fact that the temperature is a harmonic function of the space variable (so the resulting temperature would still be a solution of the Dirichlet problem). The fact that we consider the minimal nonnegative solution to the Dirichlet problem can be viewed as setting a boundary condition at infinity to be the lowest permissible value. (This is often, but not always, zero). This requirement clearly makes solutions of the Dirichlet problem unique.

We now switch to an example (which has nothing to do with these physical considerations) but where the Dirichlet problem (i.e., Theorem 2.10) is useful for a concrete problem.

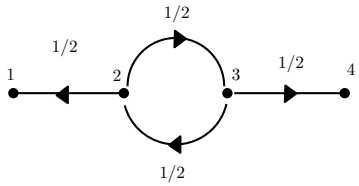


Figure 5: Markov chain representation of the two-player game in Example 2.12.

Example 2.12. Two players, A and B, toss a fair coin, until one of them gets a head. If the first player is A, what is the chance that B wins this game?

This can be formulated as a Markov chain with $S = \{1, 2, 3, 4\}$. States 2 or 3 indicate that A or B toss the coin. 1 indicates that A has obtained a head. 4 indicates that B has obtained a head. The corresponding Markov chain can be represented as in Figure 5.

We are then interested in $\mathbb{P}_x(T_A < \infty)$ where $x = 2$ and $A = \{4\}$. It is notationally more convenient to write $h_x = h_A(x)$ here. The equations we get from the Dirichlet problem are: $h_1 = 0, h_4 = 1$ (obvious). Moreover,

$$h_2 = \frac{1}{2}(h_1 + h_3) = \frac{h_3}{2}.$$
$$h_3 = \frac{1}{2}(h_4 + h_2) = \frac{1}{2}(1 + h_2).$$

Thus

$$h_2 = \frac{1}{4}(1 + h_2), \text{ i.e. }, h_2 = 1/3, h_3 = 2/3.$$

So, starting from A, B has only a probability of $1/3$ to win this game!

Exercise 2.13. What if the coin is biased? Can you compute the winning probabilities in this case?

2.4 Biased random walk on $\mathbb{Z}$

Lecture 5; Thursday 20.10.2022

We are now going to apply what we discovered about hitting probabilities to some fundamental examples. The first one we consider is the *biased random walk* on $\mathbb{Z}$. This is the Markov chain on $\mathbb{Z}$ whose transition matrix P verifies

$$P(x, y) = \begin{cases} p & \text{if } y = x + 1 \\ q & \text{if } y = x - 1 \\ 0 & \text{else.} \end{cases}$$

where $0 \leq p \leq 1$ and $p + q = 1$. See Figure 6. In other words this is a random walk on a *directed* graph (for every $x \in \mathbb{Z}$, the weight of the directed edge $(x, x+1)$ is p and that of its reverse $(x+1, x)$ is q).

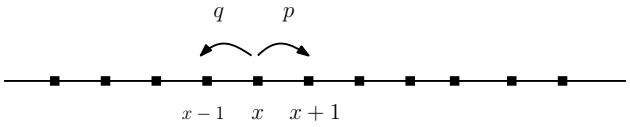


Figure 6: Biased random walk on $\mathbb{Z}$.

We will be interested in knowing, starting from some point $x \in \mathbb{Z}$ (possibly far away from 0), how likely it is that the biased random walk ever touches 0. Let us set $h_x = \mathbb{P}_x(T_0 < \infty)$, where, as before, T_0 is the hitting time of zero. Note that by symmetry (possibly exchanging the roles of p and q we may assume without loss of generality that $x \geq 0$. Also, since we are interested in whether the chain hits $y = 0$, we could modify the chain so that $y = 0$ is an absorbing state (i.e., once the chain hits 0, it stays there forever). There is therefore no need to consider negative states.

Applying Theorem 2.10, we see that $h_0 = 1$ and h is the minimial nonnegative solution to

$$h_i = ph_{i+1} + qh_{i-1}; \quad i \geq 1. \tag{2.5}$$

This is a recurrence of order two, which can be solved explicitly. Such equations arise often in Markov chains, so we explain carefully how to solve them. (The method is similar to solving an Ordinary Differential Equation of the form $ay'' + by' + cy = 0$, and for a good reason: a recursion of order two is in fact nothing but a discretised form of this ODE). To solve (2.5), we consider the associated characteristic equation:

$$x = px^2 + q. \tag{2.6}$$

We note that $x = 1$ is a solution, thus we can compute the other root by considering the coefficient of the highest degree and the value of the polynomial at $x = 0$, which is q: that

$$px^2 - x + q = p(x - 1)(x - q/p)$$

so that the other root of (2.6) is necessarily $x = q/p$.

Consider first the case where $p \neq q$ so the two roots are distinct. Then it can be shown that the general solution of (2.5) is obtained by considering a linear combination of the form

$$h_i = A1^i + B\left(\frac{q}{p}\right)^i = A + B\left(\frac{q}{p}\right)^i, \tag{2.7}$$

where $A, B \in \mathbb{R}$ are unknown parameters to be determined. This solution should remind you of the calculations in Theorem 1.12 and Example 1.13. This is not a coincidence: in fact, the recursion (2.5) can be written in matrix form in terms of the unknown vector $Y_i = (h_{i+1}, h_i)$

in such a way that we are looking to compute the power of a certain 2×2 matrix and hence need to diagonalise it; hence (2.6) is nothing but the characteristic polynomial for this matrix.

Either way, we can compute the unknown parameters A and B from the boundary conditions. We obtain one equation from the knowledge $h_0 = 1$ so $A + B = 1$. To obtain the second equation we exploit the fact that h is the minimal nonnegative solution of (2.5) (as mentioned informally, we can think of the minimality condition as a boundary condition at infinity, which therefore gives us a second equation). We have assumed that $p \neq q$ but we distinguish further $p < q$ or $q > p$.

Case 1. Suppose $p > q$. Then since $A + B = 1$ we can write

$$h_i = A + B\left(\frac{q}{p}\right)^i = A\left(1 - \left(\frac{q}{p}\right)^i\right) + \left(\frac{q}{p}\right)^i.$$

When $q < p$, the term in the bracket on the right hand side is ≥ 0 for every $i \geq 0$. Hence the *minimal nonnegative* solution of (2.5) is attained for $A = 0$ and we deduce in this case

$$h_i = \left(\frac{q}{p}\right)^i; \quad i \geq 0. \tag{2.8}$$

Note that, in this case, when the starting point i is far away from zero, the probability to ever reach zero becomes vanishingly small. This is somewhat intuitive since, when $p > q$, the walk is more likely to go to the right than to the left, and starts already quite far from zero.

Case 2. Now suppose instead $p < q$. Then $q/p > 1$ and the term $(q/p)^i$ diverges to ∞ as $i \to \infty$. Since h_i is in fact bounded by 1 it follows necessarily that $B = 0$. Hence $A = 1$ and

$$h_i = 1; \quad i \geq 0. \tag{2.9}$$

Thus in this case, no matter how far away from zero the walk starts, it is always *guaranteed* to return to zero!

Case 3. To some extent we could use our intuition to guess the qualitative nature of the results when $p \neq q$, but when $p = q$, the symmetric situation in which the walk doe not favour either direction, it becomes harder to make use of our intuitition and guess the answer: starting from far away, does the walk eventually return to zero, or does it get lost at infinity? We return to the recursion (2.5) satisfied by h_i and recall that when $p = q$ the two roots to the characteristic equation (2.6) coincide and are both equal to 1. As explained in Remark 1.14, it is possible to give the general form of the recursion: namely, in that case, any solution to (2.5) is of the form

$$h_i = A + Bi, i \quad i \geq 0, \tag{2.10}$$

(where A, B are parameters to be determined) and h remains the minimal nonnegative solution to (2.10) satisfying $h_0 = 1$. The latter gives us $A = 1$ so h_i is the minimal nonnegative solution to $h_i = 1 + Bi, i \geq 0$. This minimum is clearly attained for $B = 0$. We deduce,

$$h_i = 1; \quad i \geq 0. \tag{2.11}$$

Thus, as in Case 2, no matter how far away from zero the walk starts, it is always *guaranteed* to return to zero!

We summarise these answers through the following theorem.

Theorem 2.14. *Consider the biased random walk on $\mathbb{Z}$ described above, and set $h_i = \mathbb{P}_i(T_0 < \infty)$ for $i \geq 0$. Then*

- *if $p \leq q$, $h_i = 1$ for any $i \geq 0$.*
- *if $p > q$ then $h_i = (\frac{q}{p})^i$ for any $i \geq 0$. In particular $h_i \to 0$ as $i \to \infty$.*

To think of what this might mean in practice, consider playing in a casino in a game of chance in which at stage, you either win or lose 1 euro. The game stops when (if) you become ruined (i.e., when you fortune reaches zero). In a casino, the game is always biased ever so slightly against you, so $p < q$. What will happen to your fortune in the long run? The answer is, no matter how rich you are to begin with, you will end up with certain ruin!

2.5 Branching processes

We now introduce another favourite example (which we will keep returning to throughout the course) called branching processes. This Markov chain can be thought of as a crude (but remarkably useful) model for the spread of an epidemic in a population, ignoring any spatial, demographic effect – in fact pretty much everything except for the growth of the epidemic itself. Traditionally branching processes are often described in terms of the growth of a population rather than an epidemic within a population, but this makes no difference in terms of mathematics, and we find the epidemic interpretation more relevant these days...! In the first interpretation, the branching process will count the number of infected individuals at time n, while in the second interpretation, the branching process counts the size of the population at generation n.

The stochastic process is defined as follows. Suppose we are given a distribution $(p_k)_{k\geq 0}$ a distribution on $\mathbb{N}$. We think of this distribution as describing the law of the number of individuals I will infect if I am infected, and is called the *offspring distribution*.

Definition 2.15. *The branching process with offspring distribution $(p_k)_{k\geq 0}$, is the stochastic process $(Z_k)_{k\geq 0}$ defined by induction as follows: initially $Z_0 = 1$. Furthermore, for $n \geq 0$, given $Z_n = k$,*

$$Z_{n+1} = \sum_{i=1}^{k} \xi_{n+1,i} \tag{2.12}$$

where $\xi_{n,i}$ are i.i.d. and have common law $(p_k)_{k\geq 0}$.

In words, each of the k individuals alive at generation n gives rise to independent and identically distributed offsprings with law $(p_k)_{k\geq 0}$, and Z_{n+1} counts the total number of offsprings in generation $n+1$. Note that, although we think of Z_n as counting the number of infected individuals at time n (where n could be the number of days since the start of the pandemic), in practice it is more convenient to use the language associated with the interpretation of Z_n as a population count in generation n (and so we speak of generations and individuals *alive* rather than infected at generation n).

It is not hard to check that a branching process defines a Markov chain on $\mathbb{N} = \{0, 1, \ldots\}$.

Exercise 2.16. Write down a formula for the transition probabilities of a branching process with offspring distribution $(p_k)_{k\geq 0}$. You may find useful to recall that the law of the sum of k random variables with distribution λ is given by the k-fold convolution of λ with itself, λ^{*k}, where the convolution of two distributions λ and μ is given by $\lambda * \mu(m) = \sum_i \lambda_i \mu_{m-i}$.

Both in the pandemic and in the population interpretations, we are interested in knowing whether the process will ever become extinct, i.e., whether $T_0 < \infty$. (Note that once the population reaches zero, it remains extinct for ever after that time). Applying Theorem 2.10, the following fundamental result can be shown:

Theorem 2.17. *Let us set $\theta = \mathbb{P}_1(T_0 < \infty)$. Then θ is the smallest nonnegative solution to the equation $\theta = F(\theta)$, where*

$$F(\theta) = \sum_{n=0}^{\infty} \theta^n p_n.$$

An exercise in the next example sheet will guide you towards a proof of this important theorem. In practice, we care above all whether $\theta < 1$: i.e., is there a positive probability for the infection (or the population) to survive forever?

Using Theorem 2.17 and a bit of analysis, the following fundamental dichotomy is not very hard to show (for this we need to exclude the trivial and unrealistic case where $(p_k)_{k\geq 0}$ is the Dirac mass at 1, i.e., we assume $p_1 \neq 1$):

Theorem 2.18. *We have $\theta < 1$ if and only if the mean number of offsprings, $m = \sum_{k=0}^{\infty} k p_k$, satisfies $m > 1$.*

We will not need this result in the rest of the course so skip its proof, which is in any case not difficult. In the applied literature and in the media, the mean number of offsprings is usually called R instead of m (and represents, once again, the mean number of secondary infections created by a single infected individual). This theorem shows that what governs the long term behaviour of the pandemic depends only on this single number R and not on the details of the offspring distribution $(p_k)_{k\geq 0}$: when $R > 1$ the epidemics survives forever with positive probability (and that probability is itself very high if the initial number of infected individuals is very high), whereas if $R \leq 1$ then pandemics is *guaranteed* to die out. Theorem 2.18, although based on a very crude model, explains the almost obsessive focus on this quantity in the context of a pandemic. More sophisticated epidemiological models often share a similar structure (towards the end of the course will see how spatiality changes the result).

2.6 Strong Markov property

We have already discussed the (weak) or simple Markov property, which says that the law of a Markov chain, starting from some fixed time onwards, is that of a Markov chain given the state of the chain at that time, and is furthermore independent of the past of the chain (still conditionally on the current state of the chain).

The strong Markov property extends this fact to a certain class of very important *random* times, known as stopping times. We begin with the definition.

Definition 2.19. *Let $(X_n, n \geq 0)$ be a stochastic process with values in a state space S. We say that the random variable T, with values in $\{0, 1, \ldots\} \cup \{\infty\}$, is a **stopping time** if, for every $n \geq 0$, the event $\{T = n\}$ is determined by the values of $X_0, \ldots, X_n$ only. That is, there exists a function $F_n : S^{n+1} \to \{0, 1\}$ such that, almost surely, the two following random variables are equal:*

$$1_{\{T=n\}} = F_n(X_0, \ldots, X_n).$$

The formal definition is at first quite hard to parse, but is in fact very intuitive. It says we can determine whether $T = n$ simply by considering the stochastic process up to time n. In other words, whether or not $T = n$ depends only on the past (including the present) of the process, not its future. In practice, it is very easy to see if a random time is a stopping time.

Example 2.20. Let X be any stochastic process with values in a state space S. If $A \subset S$ and $T_A = \inf\{n \geq 0 : X_n \in A\}$ is the hitting time of A, then T_A is a stopping time. Indeed, $T_A = n$ if and only if

$$\{X_0 \notin A, \ldots, X_{n-1} \notin A, X_n \in A\}.$$

As a consequence, the event $\{T_A = n\}$ depends only on $X_0, \ldots, X_n$ and not on the future of X after time n.

Example 2.21. Let $T_A^{(2)}$ be the time of the second visit to A: that is,

$$T_A^{(2)} = \inf\{n > T_A : X_n \in A\}.$$

Is this a stopping time?

Answer: yes! Indeed, we can write

$$\{T_A^{(2)} = n\} = \bigcup_{0 \leq j \leq n-1} \Big(\{T_A = j\} \cap \{X_{j+1} \notin A, \ldots, X_{n-1} \notin A, X_n \in A\} \Big).$$

which depends only on $X_0, \ldots, X_n$.

Exercise 2.22. Consider the coupon (or sticker) collector problem. Let T denote the first time at which there is a repetition (i.e., a sticker is collected twice). Is T a stopping time?

Exercise 2.23. Let X be a stochastic process on S and let $A \subset S$. Let $T = T_A - 1$. Is T a stopping time?

Exercise 2.24. Let X be a stochastic process on S and let $A \subset S$. Let T denote the time of the *last* visit by X to A. Is this a stopping time?

It should be apparent that by nature, a stopping time does not anticipate the behaviour of the Markov chain after the stopping time. This property lies at the root of the next theorem, which extends the weak Markov property (Proposition 1.10) to stopping times.

Theorem 2.25. *Let X be Markov (λ, P), let $x \in S$ and let T be a stopping time. Conditionally given $T < \infty$ and $\{X_T = x\}$, the sequence $(X_T, X_{T+1}, \ldots)$ is Markov (δ_x, P) and is independent of $(X_0, \ldots, X_T)$.*

Proof. The proof proceeds by applying the law of total probability (summing over all the possible values of T, say $T = m$) and applying the weak Markov property, since $T = m$ depends only on the past by definition of a stopping time. The proof is not particularly informative and is therefore skipped. □

Lecture 6. Friday, 21.10.2022.

Using the strong Markov property in a concrete example is far more informative than checking its somewhat boring proof...

Example 2.26. Consider the biased random walk on $\mathbb{Z}$ of Figure 6, and let $h_i = \mathbb{P}_i(T_0 < \infty)$. (We computed h_i exactly in Theorem 2.14 based on the Dirichlet problem, but ignore this for a moment). We make the following claim:

$$h_i = \mathbb{P}_i(T_{i-1} < \infty) h_{i-1}. \tag{2.13}$$

To see (2.13), simply apply the strong Markov property at the stopping time T_{i-1}. Indeed, if $T_{i-1} < \infty$, Theorem 2.25 implies that, after time T_{i-1}, the future of the Markov chain is simply that of a biased walk starting from $i - 1$, and so the conditional probability to hit zero is just h_{i-1}.

An interesting consequence of (2.13) is that h_i is necessarily of the form z^i for some z. Indeed, note that, by translation invariance, $\mathbb{P}_i(T_{i-1} < \infty) = \mathbb{P}_1(T_0 < \infty) = z$ does not depend on i, so that (2.13) becomes

$$h_i = z h_{i-1}$$

an by induction $h_i = z^i h_0 = z^i$. In Theorem 2.14 we identified z explicitly: if $q \geq p$ then $z = 1$, while if $p > q$ then $z = q/p$. Thus $z = \min(1, q/p)$.

2.7 Recurrence, transience

Let $(X_n, n \geq 0)$ denote a Markov chain on some state space S.

Definition 2.27. *Let $x \in S$. We say that x is **recurrent** if*

$$\mathbb{P}_x(X_n = x \text{ for infinitely many values of } n) = 1.$$

In other words, let $V_x = \sum_{n=0}^{\infty} 1_{\{X_n = x\}}$ denote the total number of visits to x. Then x is recurrent if and only if $\mathbb{P}_x(V_x = \infty) = 1$.

Definition 2.28. *Let us say x is transient if $\mathbb{P}_x(V_x = \infty) = 0$.*

Note that, *a priori*, a state x could be neither transient nor recurrent: the probability to visit x infinitely often could be strictly between 0 and 1. However, we will soon see *a posteriori*, as a consequence of Theorem 2.31, that this is not possible: this probability is, in fact, either 0 or 1.

Note also that the notion of recurrence for a state x does not depend on the starting distribution λ but only on the transition matrix P: the condition states that, if we *were* to start the chain in x, then we *would* visit x infinitely often.

The following lemma is a fundamental observation. Recall that a random variable N taking values in $\{1, 2, \ldots\}$ is said to have a **geometric distribution** with parameter $p \in [0, 1]$ if $\mathbb{P}(N = n) = (1-p)^{n-1}p; n = 1, 2 \ldots$. Equivalently, $\mathbb{P}(N > n) = (1-p)^n$ for $n = 0, 1, \ldots$. In words, we toss a biased coin (where the probability to get a heads at each toss is p) and wait until the first toss N where we get a heads. The parameter p can therefore be thought of as a success probability for independent and identically distributed trials. We also recall that $\mathbb{E}(N) = 1/p$: in other words, we must wait on average for $1/p$ trials for a success. This can be verified by differentiation of the geometric series (and is also very intuitive – consider for instance the case where p is very small!).

Lemma 2.29. *For $x \in S$, set $p_x = \mathbb{P}_x(T_x^+ < \infty)$, where $T_x^+ = \inf\{n \geq 1 : X_n = x\}$ is the first return time to x. Then V_x is a geometric random variable with parameter $1 - p_x$.*

Remark 2.30. In the above definition of the return time, note that this differs from the hitting time defined earlier only in that the inf is taken over $n \geq 1$ rather over $n \geq 0$. p_x is therefore the probability that, starting from x, the chain ever returns to x.

Proof. Define the successive return times to x as $T_x^{(0)} = T_x = 0$, $T_x^{(1)} = T_x^+$, and inductively:

$$T_x^{(i)} = \inf\{n > T_x^{(i-1)} : X_n = x\}.$$

We note that for each $i \geq 1$, $T_x^{(i)}$ is a stopping time (exercise!). Note also that V_x, the number of visits to x, is $> n$ if and only if $T_x^{(n)} < \infty$:

$$V_x > n \iff T_x^{(n)} < \infty.$$

(The inequality $V_x > n$ above is strict because the first visit occurs at $T_x^{(0)}$ with our conventions.) Now, to prove the lemma, it suffices to show that

$$\mathbb{P}(V_x \geq n) = p_x^{n-1}, \quad n = 1, 2, \ldots \tag{2.14}$$

We prove this by induction and using the strong Markov property. For $n = 1$ this is trivial. Now suppose $n \geq 2$. We have,

$$\begin{aligned}\mathbb{P}_x(V_x \geq n) &= \mathbb{P}_x(T_x^{(n)} < \infty) \\ &= \mathbb{P}_x(T_x^{(n)} < \infty, T_x^{(n-1)} < \infty) \\ &= \mathbb{P}_x(T_x^{(n-1)} < \infty)\mathbb{P}_x(T_x^{(n)} < \infty | T_x^{(n-1)} < \infty) \\ &= \mathbb{P}_x(T_x^{(n-1)} < \infty)p_x\end{aligned}$$

where in the last line, we have used that $\mathbb{P}_x(T_x^{(n)} < \infty | T_x^{(n-1)} < \infty)$: this follows from the strong Markov property at time $T_x^{(n-1)}$ (which is indeed a stopping time). By the induction hypothesis, we deduce that $\mathbb{P}_x(V_x \geq n) = p_x^{n-1}$, as desired. This completes the proof of the lemma. □

In words, the lemma says the following. At each subsequent visit to x, the Markov chain tosses an independent coin: with probability p_x it comes back, with probability $1 - p_x$ it does not come back. (The independence of the coin and the fact that they always have the same probability is precisely the strong Markov property). If we interpret a success as "not coming back", we see that V_x counts the number of trials until we have a success, and that is why V_x has a geometric distribution with parameter $1 - p_x$.

From this lemma we deduce the following important characterisation of recurrence and transience.

Theorem 2.31. *We have the following dichotomy.*

(i) Suppose $p_x = 1$. Then x is recurrent, and $\sum_{n=0}^{\infty} P^n(x, x) = \infty$.

(ii) Suppose $p_x < 1$. Then x is transient, and $\sum_{n=0}^{\infty} P^n(x, x) < \infty$.

In particular, a state is either recurrent or transient. It is recurrent if and only if $\sum_{n=0}^{\infty} P^n(x, x) = \infty$.

We recall that $P^n(x, x)$ is the (x, x) entry of the matrix P^n, and so is equal to $\mathbb{P}_x(X_n = x)$ by the Chapman–Kolmogorov equation.

Proof. Suppose $p_x = 1$. By Lemma 2.29, we know that V_x is a geometric random variable with success probability $1 - p_x = 0$. Thus $\mathbb{P}_x(V_x \geq n) = 1$. Since n is arbitrary, we deduce $V_x = \infty$ and hence x is recurrent.

Suppose instead $p_x < 1$. Then $\mathbb{P}_x(V_x \geq n) = p_x^n - 1 \to 0$ as $n \to \infty$. Hence $V_x < \infty$, with $\mathbb{P}_x$-probability equal to one ($\mathbb{P}_x$-a.s.).

Concerning the series, note that

$$\begin{aligned}\sum_{n=0}^{\infty} P^n(x,x) &= \sum_{n=0}^{\infty} \mathbb{P}_x(X_n = x) \\ &= \sum_{n=0}^{\infty} \mathbb{E}_x(1_{\{X_n=x\}}) \\ &= \mathbb{E}_x(\sum_{n=0}^{\infty} 1_{\{X_n=x\}}) \\ &= \mathbb{E}_x(V_x) = 1/(1-p_x).\end{aligned}$$

(In the third line, we exchanged summation and expectation. This requires a bit of measure theory for a proper justification: for instance via Fubini's theorem, or by considering the truncated sum, using linearity of expectation and taking a limit via the monotone convergence theorem since all terms are positive. This can also be proved by hand rather elementarily). The right hand side is finite if and only if $p_x < 1$, and this completes the proof of the theorem. □

Lecture 7: Thursday 27.10.2022 Before seeing concrete examples of the theorem, we explain a few consequences of the above dichotomy.

Corollary 2.32. *Let C be a communicating class. Then one of the two alternative holds:*

(i) For every $x \in C$, x is recurrent.

(ii) For every $x \in C$, x is transient.

In other words, recurrence/transience is a class property.

Proof. Fix $x, y \in C$. Suppose x is transient, and let us show y is transient too. Since $x \leftrightarrow y$ we know that $P^k(x,y) > 0$ for some $k \geq 0$ and $P^j(y,x) > 0$ for some $j \geq 0$. Furthermore, using the (simple) Markov property, for any $n \geq 0$,

$$P^{k+n+j}(x,x) \geq P^k(x,y)P^n(y,y)P^j(y,x).$$

Now let us sum over $n \geq 0$. The left hand side is finite by Theorem 2.31, and thus so is the right hand side. It follows (again by Theorem 2.31) that y is transient, as desired. □

Let us call a collection C of states **closed** if $x \in C$ and $x \to y$ imply $y \in C$. That is, starting from somewhere in C, the chain remains in C forever. (This notion is distinct from that of communicating class introduced earlier in Definition 2.5: indeed, we have already noticed in Exercise 2.7 that a communicating class is not necessarily closed). In the same vein as above, we now show that every finite closed communicating class is recurrent.

Corollary 2.33. *Suppose C is a finite, closed, communicating class. Then C is recurrent (i.e., all states $x \in C$ are recurrent).*

Proof. Fix $x \in C$. Since C is finite and the Markov chain is defined forever, there is necessarily some $y \in C$ which is visited infinitely many times by the chain. This follows from the *pigeonhole principle* or, more simply, from the observation that $\infty = \sum_{y \in C} V_y$, so there is at least one $y \in C$ such that $V_y = \infty$. Note however that this $y \in C$ may be random: that is, we have established that

$$\mathbb{P}_x(\bigcup_{y \in C} \{V_y = \infty\}) = 1.$$

By Boole's inequality (also known as a union bound),

$$1 = \mathbb{P}_x(\bigcup_{y \in C} \{V_y = \infty\}) \leq \sum_{y \in C} \mathbb{P}_x(V_y = \infty)$$

and hence for at least one $y \in C$ (this one is deterministic), it must be the case that

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

On the other hand, by the strong Markov property,

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

so we deduce in particular that $\mathbb{P}_y(V_y = \infty) > 0$. Hence y cannot be transient, and so must be recurrent by Theorem 2.31. By the previous corollary, this implies all the states of C are recurrent. □

2.8 Pólya's theorem

In this section we will apply Theorem 2.31 to state and prove Pólya's remarkable theorem concerning the recurrence and transience of simple random walk in $\mathbb{Z}^d$.

Let G be a locally finite, connected graph. Then the associated random walk is clearly irreducible (recall that this means that there is a single communicating class). Hence there are two possibilities: either all vertices of G are recurrent or all vertices of G are transient. Call the graph G recurrent in the first case, transient in the other.

Given any graph, a fundamental question is whether this graph is recurrent or transient. When the graph is finite, by Corollary 2.33, the graph is necessarily recurrent. For instance, the random transpositions of Example 1.8 are necessarily recurrent: if we shuffle the deck of cards sufficiently many times, it is certain that at some point the deck will return to its original ordering (it might take a long time in practice though! We will actually be able to compute how long it takes on average in the next Chapter, see Example 3.10 – the answer is $n!$).

The question of recurrence/transience therefore only arises when the graph is infinite. Then it might be possible for random walk to get lost in the graph, and wander off to infinity before returning to its starting point, as the next amazing result shows. Recall the

graph $G = \mathbb{Z}^d$, with vertex set $V = \{x = (x_1, \dots, x_d) : x_i \in \mathbb{Z}\}$ formed by points of $\mathbb{R}^d$ such that each coordinate is a (relative) integer, and with edge set determined $x \sim y$ if and only if $\sum_{i=1}^d |x_i - y_i| = 1$. With an abuse of notation we use $\mathbb{Z}^d$ both to denote the vertex set and the graph itself.

Theorem 2.34 (Pólya's theorem)**.** *$\mathbb{Z}^d$ is recurrent if $d = 1$ or $d = 2$, but is transient for $d \geq 3$.*

Apart from the sheer beauty of the theorem and its surprising conclusion, this result also has enormous implications e.g. in condensed matter physics, from Bose–Einstein condensation to superconductivity; it can be seen as the main reason why matter behaves differently in two and three dimensions.

To prove this theorem we will focus mostly on the case $d = 1$. This might seem surprising initially, because in reality we already know the answer in that case: indeed, the random walk on $\mathbb{Z}$ simply corresponds to the case $p = q = 1/2$ of the biased random walk, for which we have shown in Theorem 2.14 that $\mathbb{P}_x(T_0 < \infty) = 1$ for any $x \in \mathbb{Z}$. Hence by the simple Markov property $\mathbb{P}_0(T_0^+ < \infty) = 1$ and the walk is recurrent.

However, it is not straightforward to generalise this approach to higher dimensions. This is because the analysis of difference equations, which lies at the heart of our proof of Theorem 2.14, is much more subtle in higher dimensions: it is hard to get explicit formulae.

Instead, we will provide a different proof of recurrence in dimension $d = 1$, which can be generalised to higher dimensions with relatively little effort. This proof will be based on the characterisation in Theorem 2.31: namely, we will study the asymptotics of $P^n(0,0)$ as $n \to \infty$, and show that the series is not summable. The study of $P^n(0,0)$ is based on the following combinatorial observation.

Lemma 2.35. *For any $n \geq 0$, $\mathbb{P}_0(X_n = 0) = 0$ if n is odd, whereas for even integers,*

$$\mathbb{P}_0(X_{2n} = 0) = \binom{2n}{n} 2^{-2n}.$$

Proof of Lemma 2.35. We first observe that the walk alternates between even and odd positions, so it is impossible to return to zero at an odd time. So we will only consider even times in the following.

Fix any path $(x_0, \dots, x_n)$ with $x_0 = 0$, $x_i \in \mathbb{Z}$ and $|x_i - x_{i-1}| = 1$, for $1 \leq i \leq n$. Then

$$\mathbb{P}_0(X_0 = x_0, \dots, X_n = x_n) = (\tfrac{1}{2})^n \tag{2.15}$$

since each transition has probability $1/2$. It follows that (changing n into $2n$ since we want to consider even times),

$$\mathbb{P}_0(X_{2n} = 0) = (\tfrac{1}{2})^{2n} \# \mathcal{L}_{2n}, \tag{2.16}$$

where $\mathcal{L}_{2n}$ is the set of lattice paths of length $2n$ starting at 0 and ending at 0 at time $2n$:

$$\mathcal{L}_{2n} = \{(x_0, \dots, x_{2n}) : x_0 = x_{2n} = 0, x_i \in \mathbb{Z} \text{ for } 0 \leq i \leq 2n, |x_i - x_{i-1}| = 1 \text{ for } 1 \leq i \leq 2n\}.$$

Indeed we obtain (2.16) by summing (2.15) over all paths in $\mathcal{L}_{2n}$ and each such path contributes the same probability, namely $(1/2)^{2n}$. The lemma is therefore proved once we show

$$\#\mathcal{L}_{2n} = \binom{2n}{n}.$$

However, this is immediate from the following observation: any path in $\mathcal{L}_{2n}$ satisfies the property that there are exactly n times for which the corresponding increment of the path is $+1$, and n times for which the increment is -1: that is, for any $(x_0, \dots, x_{2n}) \in \mathcal{L}_{2n}$,

$$\#\{1 \leq i \leq 2n : x_i - x_{i-1} = +1\} = n.$$

Furthermore, any such path in $\mathcal{L}_{2n}$ is entirely determined by the data of the set above (the set of times for which the increment is $+1$). There are clearly $\binom{2n}{n}$ ways of choosing such a set. Hence $\#\mathcal{L}_{2n} = \binom{2n}{n}$ and the lemma is proved. □

Using Lemma 2.35, we can now determine the asymptotic behaviour of $P^{2n}(0,0)$. We make use of the well known approximation of $n!$ by Stirling:

Lemma 2.36. *We have $n! \sim (n/e)^n \sqrt{2\pi n}$, in the sense that the ratio of the two sides converges to 1 as $n \to \infty$.*

Proof of Theorem 2.34, case $d = 1$. We just need to compute an asymptotic equivalent for $P^{2n}(0,0)$: combining Lemma 2.35 and 2.36, we have:

$$\begin{aligned} \mathbb{P}_0(X_{2n} = 0) &= \frac{(2n)!}{n!n!} 2^{-2n} \\ &\sim \frac{(2n/e)^{2n}\sqrt{2\pi 2n}}{(n/e)^{2n} 2\pi n} 2^{-2n} \\ &\sim \frac{1}{\sqrt{\pi n}}. \end{aligned} \tag{2.17}$$

Now, the series $1/\sqrt{n}$ is not summable and hence $P^{2n}(0,0)$ is equivalent to a non-summable series. We deduce that $\sum_{n=0}^{\infty} P^n(0,0) = \infty$. This completes the proof in the case $d = 1$. □

Now consider the case $d \geq 2$. Before starting the proof properly we explain the intuition. If $(X_n, n \geq 0)$ is a simple random walk on $\mathbb{Z}^d$, observe that for each $1 \leq i \leq d$, the coordinate process $(X_n^i, n \geq 0)$ is a stochastic process taking values in $\mathbb{Z}$. At each step, it stays put with probability $1 - 1/d$, while with the remaining probability it moves by ± 1 with equal probability. In other words it is a simple random walk with random "delays". From that and (2.17) it is natural to expect that if $\vec{0} = (0, \dots, 0)$, then $\mathbb{P}_{\vec{0}}(X_n^i = 0) \approx n^{-1/2}$. (As this is an informal explanation providing heuristics, we do not try to make precise what we mean by $\approx$ here). Furthermore, as i varies between 1 and d, the coordinates are essentially independent of one another. The only dependence between them comes from the fact that only one coordinate changes at a time. Thus it is reasonable to expect, using independence,

$$\mathbb{P}_{\vec{0}}(X_n = \vec{0}) = \mathbb{P}_{\vec{0}}(X_n^1 = 0, \dots, X_n^d = 0) \approx (n^{-1/2})^d = n^{-d/2}.$$

Note that the sum of the series on the right hand side is infinite if $d = 1, 2$ but finite for $d \geq 3$. Combining with Theorem 2.31 gives the result of Theorem 2.34 we are looking for.

Turning this into a proof requires a number of rather delicate arguments, **which go beyond the scope of the course**.

Proof of Theorem 2.34, case $d = 2$. We briefly sketch the proof for the case $d = 2$, for which we adapt an argument due to D. Chafaï. This relies on the Vandermonde convolution identity for binomial coefficients:

$$\binom{n+m}{r} = \sum_{k=0}^{r} \binom{n}{k}\binom{m}{r-k},$$

which is obtained by considering the expansion of $(1+x)^{m+n} = (1+x)^m(1+x)^n$. Taking $m = n = r$ this gives

$$\binom{2n}{n} = \sum_{k=0}^{n} \binom{n}{k}\binom{n}{n-k} = \sum_{k=0}^{n} \binom{n}{k}^2. \tag{2.18}$$

Now note that

$$P^{2n}(\vec{0}, \vec{0}) = \sum_{i,j \geq 0 : i+j=n} \frac{(2n)!}{(i!j!)^2} \left(\frac{1}{4}\right)^{2n}. \tag{2.19}$$

To see this, note that each fixed lattice path of length $2n$ from 0 to 0 has probability exactly $(1/4)^{2n}$. Furthermore, once $i, j \geq 0$ have been chosen such that $i + j = n$, there are exactly $\frac{(2n)!}{(i!j!)^2}$ ways to choose the lattice path in such a way that the first coordinate moves exactly $2i$ times, and the second $2j$ times. This proves (2.19). Rewriting this we get

$$P^{2n}(\vec{0}, \vec{0}) = \binom{2n}{n}(1/4)^{2n} \sum_{k=0}^{n} \binom{n}{k}^2 = \binom{2n}{n}^2 (1/4)^{2n} \tag{2.20}$$

by (2.18). Using Stirling's formula (Lemma 2.36) we deduce that

$$P^{2n}(\vec{0}, \vec{0}) \sim \frac{1}{\pi n}.$$

Since this is not summable we conclude using Theorem 2.31. □

Proof of Theorem 2.34, case $d \geq 3$. We briefly sketch the proof for the case $d \geq 3$. While the Vandermonde identity has an obvious generalisation (think of expanding $(1+x)^n(1+x)^n(1+x)^n = (1+x)^{3n}$) the resulting identity is of little help: which coefficient should be of interest to us? Instead we follow the argument of Norris [Nor98]. First of all, we observe that it suffices to consider the case $d = 3$. To see this, note that $\mathbb{Z}^3 \subset \mathbb{Z}^d$ and that if $Y_n = (X_n^1, X_n^2, X_n^3)$ records the first three coordinates of X, then Y is a delayed (i.e., "lazy", in the terminology of an exercise in the first example sheet) simple random walk in $\mathbb{Z}^3$: more precisely, at each step, Y stays put with probability $1 - 3/d$ and otherwise evolves like a simple random walk on $\mathbb{Z}^3$ with probability $3/d$.

From this observation we claim that it suffices to prove transience in the case $d = 3$. Indeed, if we know transience for $d = 3$, this immediately implies transience for the lazy version of the random walk on $\mathbb{Z}^3$, as the expected number of visits is then also necessarily finite. In turn, this implies that the number of visits to the origin of $\mathbb{Z}^d$ by random walk on $\mathbb{Z}^d$ also has finite expectation: this is because whenever $X_n = \vec{0}$, then necessarily $Y_n = (X_n^1, X_n^2, X_n^3) = (0, 0, 0)$ so each visit to $\vec{0}$ coincides with a visit of Y to $(0, 0, 0)$. Given that Y is transient, it therefore follows that X is transient.

Thus suppose $d = 3$, and note

$$P^{2n}(\vec{0}, \vec{0}) = \sum_{i,j,k \geq 0 : i+j+k=n} \frac{(2n)!}{(i!j!k!)^2} \left(\frac{1}{6}\right)^{2n}, \tag{2.21}$$

which is the direct analogue of (2.19), valid for the same reasons. This can also be rewritten in the following way:

$$P^{2n}(\vec{0}, \vec{0}) = \binom{2n}{n} \sum_{i,j,k \geq 0 : i+j+k=n} \left(\frac{n!}{i!j!k!}\right)^2 \left(\frac{1}{6}\right)^{2n}. \tag{2.22}$$

We combine this with two observations:

$$\sum_{i,j,k \geq 0 : i+j+k=n} \frac{n!}{i!j!k!} \left(\frac{1}{3}\right)^n = 1 \tag{2.23}$$

since the left hand side is the sum over all assignments of n balls into 3 urns of the probability to make that assignment; and secondly, the multinomial coefficient is maximised when all parts are approximately equal: that is,

$$\frac{n!}{i!j!k!} = \binom{n}{i\ j\ k} \leq \binom{n}{m\ m\ n-2m} \tag{2.24}$$

with $m = \lfloor n/3 \rfloor$. These two observations together with (2.22) give us

$$P^{2n}(\vec{0}, \vec{0}) \leq \binom{2n}{n} \left(\frac{1}{2}\right)^{2n} \binom{n}{m\ m\ n-2m} \left(\frac{1}{3}\right)^n \sim cn^{-3/2}$$

for some constant $c > 0$, by Stirling's formula. The result follows in the case $d \geq 3$. □

3 Long-term behaviour

Lecture 8: Friday 28.10.2022

Consider the following silly example. A frog oscillates between jumping and falling asleep. At each minute, if asleep, it wakes up with some given probability α, and otherwise stays asleep. If awake and jumping, it falls asleep with probability β, and otherwise continues jumping. In other words, this is nothing but the silly example of Example 1.6.

Suppose we wait for a long time. What can we say about the probability of finding the frog asleep? And does it matter for this probability if we are told initially the frog was awake or asleep? Intuition suggests a limit exists for the probability to find asleep at time n, and this limit does not depend on the initial state of the frog. Actually, in this case we were able to diagonalise the chain explicitly (see Example 1.13) and check this intuition: the probability to find the frog asleep at a large time is approximately $\beta/(\alpha+\beta)$, independently of the starting state.

What about other chains? Consider for instance random walk on a finite connected graph. In this case too our intuition suggests that there is a limit for the probability to find the walk in some given state after a long time, and that this limit should not depend on the starting state. How to formalise this intuition? How to prove it? This will occupy us throughout this chapter. The key concept is that of invariant distribution which we define below.

3.1 Invariant measures and invariant distributions

Definition 3.1. *Let P denote the transition matrix of a Markov chain on a state space S. A measure μ on S is called **invariant** if for every $y \in S$,*

$$\mu_y = \sum_{x \in S} \mu_x P(x, y). \tag{3.1}$$

*If furthermore μ is a distribution, we say that μ is an **invariant distribution** (ID).*

Sometimes invariant distributions are also referred to as equilibrium distributions or stationary distributions. All these words mean the same thing. In matrix notations, the measure μ is invariant if and only if

$$\mu P = \mu. \tag{3.2}$$

Both (3.1) and (3.2) have a transparent probabilistic interpretation if μ is a distribution. We start the chain in the distribution μ, and let the chain evolve for one step. Then the distribution after one step is still μ (recall for instance Theorem 1.11).

We write this fact as the following more general proposition.

Proposition 3.2. *Suppose λ is an invariant distribution. Let $(X_0, X_1, \ldots)$ be Markov (λ, P), and let $m \geq 0$ be arbitrary and fixed. Then the distribution of $(X_m, X_{m+1}, \ldots)$ is also Markov (λ, P).*

Proof. By the simple Markov property, we only need to prove that the law of X_m is λ. But

$$\mathbb{P}(X_m = y) = \sum_{x \in S} \lambda_x P^m(x, y) = (\lambda P^m)_y = \lambda_y$$

since $\lambda P^m = \lambda P P^{m-1} = \lambda P^{m-1} = \ldots = \lambda P = \lambda$. □

Example 3.3. Recall the silly example of Example 1.6. What are the invariant measures? We have two equations:

$$\begin{cases} \mu_0 &= (1-\alpha)\mu_0 + \beta\mu_1 \\ \mu_1 &= (1-\beta)\mu_1 + \alpha\mu_0. \end{cases}$$

While these are two equations for two unknowns, there is some redundancy: indeed, simplifying, the equations become

$$\begin{cases} \alpha\mu_0 &= \beta\mu_1 \\ \beta\mu_1 &= \alpha\mu_0 \end{cases}$$

so in fact we have only one unknown. This is to be expected: invariant measures are not unique: indeed, if μ is an invariant measure and $r \geq 0$ is arbitrary then $r\mu$ is also an invariant measure (the defining condition (3.2) is linear).

Here $\mu_0 = \beta, \mu_1 = \alpha$ gives a solution, and any invariant measure is of the form $\mu_0 = r\beta, \mu_1 = r\alpha$. In particular, there is a unique invariant distribution: $\mu_0 = \beta/(\alpha+\beta), \mu_1 = \alpha/(\alpha+\beta)$.

Note that this is precisely the limiting probabilities we computed for the frog to be asleep or awake in Example 1.13 by diagonalisation!

The computation in the last example suggests there is a close connection between invariant measures and invariant distributions, and the above is no fluke. We can back this up with the following proposition, which for convenience we only state in the case of a finite state space (see below why).

Proposition 3.4. *Suppose S is finite, and X is Markov (λ, P) for some arbitrary starting distribution λ. Suppose*

$$\pi_y = \lim_{n \to \infty} \mathbb{P}(X_n = y)$$

exists. Then $(\pi_y)_{y \in S}$ is an invariant distribution.

We tend to reserve the letters λ for the starting distribution of a chain, π for an invariant distribution, and μ for an invariant measure. This is purely a matter of convention, of course.

Proof. The proof is simply an application of the Markov property (or the definition of Markov chains). Indeed, fix $y \in S$. Since $\mathbb{P}(X_n = y) \to \pi_y$ as $n \to \infty$ we can also say that $\mathbb{P}(X_{n+1} = y) \to \pi_y$. But decomposing over the possible values of the chain at time n,

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

Now take the limit as $n \to \infty$. Since S is finite there is no difficulty in taking the limit for each term and deducing the value of the limit for the sum: thus

$$\pi_y = \sum_{x \in S} \pi_x P(x, y).$$

In other words, since $y \in S$ was arbitrary, π is invariant. □

In practice, Proposition 3.4 is of little help. What we would really like is a kind converse: if given an invariant distribution π, can we guarantee that $\mathbb{P}(X_n = y)$ actually converges to π_y? From that point of view, Proposition 3.4 only tells us there is no point in looking beyond invariant distributions/measures, and simply serves as a motivation to study those. For instance, do invariant distributions always exist? Are they unique? If so, is the desired convergence actually true?

Before we answer these questions, we give more examples.

Example 3.5. Let $G = (V, E)$ denote a locally finite (undirected) graph. For $x \in V$, let $\mu(x) = \deg(x)$. Then μ is invariant for the simple random walk on G.

To see this, we fix $y \in S$. We want to show

$$\mu(y) = \sum_{x \in V} \mu(x) P(x, y)$$

i.e.

$$\deg(y) = \sum_{x \in V : x \sim y} \deg(x) \frac{1}{\deg(x)}$$

which is plainly true.

From Example 3.5 we deduce for instance that if $G = \mathbb{Z}^d$ then $\mu(x) = 1$ is an invariant measure. Note that this cannot be scaled to give an invariant distribution! (In fact, we will soon see that *no* invariant distribution exists in this case). However, for the card shuffle of random transpositions discussed in Example 1.8, all vertices have equal degree and the graph is finite (the total size is $n!$) so an invariant distribution is given by $\pi_x = (1/n!)$ for every $x \in S_n$, the permutation group of order n. In fact, we will soon see this is the *unique* invariant distribution.

Lecture 9: Thursday 3.11.2022

Exercise 3.6. Renewal theory. Here is another family of examples we have not encountered before (this can be skipped on a first reading). We give ourselves a probability distribution $(p_n)_{n \geq 0}$ on $\{1, 2, \ldots\}$ and we think of p_n as the probability that a lightbulb will last n months before needing to be replaced. Once the lightbulb breaks, we immediately replace it with a new lightbulb, whose lifetime is an independent random variable with same distribution. Alternatively, $(p_n)_{n \geq 1}$ could represent the law of the time (in minutes) between two successive buses at the bus stop, thus p_n is the probability that two successive buses are

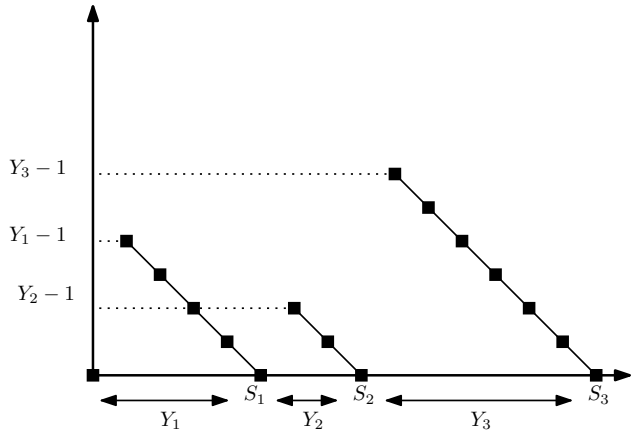


Figure 7: "Simulation" of the renewal chain R_n.

separated by n minutes. Either way we are interested in the stochastic process $(R_n, n \geq 0)$ which denotes the number of remaining months until the next replacement/next bus and its long term behaviour: what can we say about the probability that we have to wait n minutes when we arrive? How long do we have to wait on average until the next replacement/bus? And why do we tend to be *unlucky?!*

Formally, this is defined as follows. Let $Y_1, Y_2, \ldots$ denote i.i.d. random variables with common distribution $(p_n)_{n\geq 1}$. Let $S_n = \sum_{i=1}^{n} Y_i$ (this is the time of the nth replacement), and for $n = 0, 1, 2, \ldots$, let

$$R_n = \inf\{S_m : S_m \geq n\} - n. \tag{3.3}$$

$(R_n, n \geq 0)$ is called a **renewal chain**. Its transition probabilities are as follows:

$$\begin{cases} P(i, i-1) = 1 & \text{if } i \geq 1. \\ P(0, i) = p_{i+1} & \text{for any } i \geq 0, \end{cases}$$

and 0 otherwise. See Figure 7 for an illustration.

From the point of view of the above questions it is natural to ask the following question: what are the invariant measures and invariant distributions (if any) of this Markov chain? A problem on the example sheet will guide you towards a solution. We will also further develop this example to illustrate the theory we will present below.

3.2 Existence, uniqueness for recurrent chains

Given a Markov chain with transition matrix P on a state space S, our first task is to ask if we can find an invariant measure. When S is finite, it is not hard to do this using linear algebra (see one of the exercises in the upcoming sheet). We now present a general construction, which shows existence and uniqueness (up to scaling) of invariant measures, for recurrent and irreducible chains.

Definition 3.7. *Fix a state $x \in S$, and define a measure μ^x on S as follows:*

$$\mu^x(y) = \mathbb{E}_x\Big(\sum_{n=0}^{T_x^+-1} 1_{\{X_n=y\}}\Big),$$

*where, as before, $T_x^+ = \inf\{n \geq 1 : X_n = x\}$ is the return time to x. The measure μ^x is called the **return measure** based at x.*

In words, $\mu^x(y)$ counts the expected number of visits to y between two successive visits to x. Note that there is no guarantee *a priori* that $\mu^x(y) < \infty$. The return measure is useful because of the following result.

Theorem 3.8. *Suppose that P defines a Markov chain. Then μ^x is* the *minimal nonnegative solution of the equations:*

$$\begin{cases} \mu(x) = 1 \\ \sum_{w \in S} \mu(w)P(w,y) \leq \mu(y) \quad \text{for all } y \in S. \end{cases} \tag{3.4}$$

*If furthermore x is recurrent then μ^x solves (3.4) with equality, i.e., μ^x is an invariant measure and satisfies $\mu^x(x) = 1$ (we say that μ is **normalised** at x), and is therefore the minimal invariant measure normalised at x. Finally, if P is also irreducible, then $\mu(y) \in (0,\infty)$ for all $y \in S$. In that case, the nonnegative solutions to (3.4) are unique, so μ^x is the unique invariant measure normalised to be equal to 1 at x.*

A measure which satisfies (3.4) is sometimes called a **super-invariant** measure, or a **super-solution** to the equation $\mu = \mu P$. Thus μ^x is the minimal super-invariant measure normalised at x, and when x is recurrent, μ^x is the minimal invariant measure normalised at x.

Before starting the proof (which is quite long) we state an important corollary to this result.

Corollary 3.9. *Suppose P is irreducible and recurrent. Then P possesses a unique invariant measure up to scaling: if μ, μ' are two invariant measures, there exists $r \in [0,\infty]$ such that $\mu = r\mu'$.*

Proof. Suppose that we can find x such that $0 < \mu(x) < \infty$ and x' such that $0 < \mu'(x') < \infty$ (otherwise the result trivially holds with $r = \infty$ or $r = 0$). Then $\mu/\mu(x)$ is an invariant measure normalised at x. Hence by uniqueness, $\mu/\mu(x)$ must coincide with the return measure based at x. Hence all the entries of μ are positive and finite, in particular $\mu(x') \in (0,\infty)$ also. Hence dividing by $\mu(x')$, we obtain an invariant measure which is now normalised at x' instead of x. By uniqueness this must coincide with the return measure based at x'.

But by the same reasoning, the latter is a multiple of μ': indeed, $\mu'/\mu'(x') = \mu^{x'}$ must coincide with the return measure based at x'. Thus both $\mu/\mu(x')$ and $\mu'/\mu'(x')$ are equal to one another (and equal to $\mu^{x'}$). Consequently,

$$\mu = r\mu'$$

with $r = \mu(x)/\mu(x')$.

□

Let us now begin the proof of the theorem.

Proof of Theorem 3.8. **(Super-Invariance).** There are many things to check. The first is to prove that μ^x indeed solves (3.4). Let us start with checking the normalisation, i.e. $\mu^x(x) = 1$. This holds because, starting from x, by definition of T_x^+, the only visit by the chain to x occurs at time 0 (even on the event $\{T_x^+ = \infty\}$).

The proof that μ^x is invariant is more tricky, but the idea is simple to explain in words. It exploits the so-called **cycle trick**. As we know, $\mu^x(y)$ measures the expected number of visits to y by the chain during $\{0, 1, \dots, T_x^+ - 1\}$. On the other hand, $\mu^x P(y) = \sum_z \mu^x(z)P(z, y)$ measures the total expected number of visits to other sites z, weighted by the probability to make a transition to y immediately after. A moment of thought shows that this counts the total expected number of visits to y, but during the time interval $\{1, \dots, T_x^+\}$ instead of $\{0, \dots, T_x^+ - 1\}$. However when x is recurrent the two quantities are obviously equal (even for $y = x$: the lone visit to x is counted at the end rather than the beginning). When x is not recurrent (i.e., T_x^+ could be infinite), we still have an inequality.

More formally, we notice that no matter what,

$$\mu^x(y) = \mathbb{E}_x(\sum_{n=0}^{T_x^+ - 1} 1_{\{X_n = y\}}) \geq \mathbb{E}_x(\sum_{n=1}^{T_x^+} 1_{\{X_n = y\}}).$$

In fact, if the chain was recurrent this would always be an equality, whereas if the chain is not assumed recurrent, the inequality comes from the case $y = x$ and $T_x^+ = \infty$. Thus, exchanging sums and expectations liberally (the actual justification is provided by measure theory as usual):

$$\begin{aligned}
\mu^x(y) &\geq \mathbb{E}_x(\sum_{n=1}^{T_x^+} 1_{\{X_n = y\}}) \\
&= \mathbb{E}_x(\sum_{n=1}^{\infty} 1_{\{X_n = y, n \leq T_x^+\}}) \\
&= \sum_{n=1}^{\infty} \mathbb{P}_x(X_n = y, n \leq T_x^+) \qquad (3.5) \\
&= \sum_{n=1}^{\infty} \sum_{z \in S} \mathbb{P}_x(X_{n-1} = z, T_x^+ > n - 1, X_n = y).
\end{aligned}$$

Since T_x^+ is a stopping time, the event $\{T_x^+ > n-1\}$ is entirely determined by the $(X_0, \dots, X_{n-1})$.

We can therefore apply the simple Markov property at time $n-1$, to obtain:

$$\begin{aligned}
\mu^x(y) &\geq \sum_{n=1}^{\infty} \sum_{z\in S} \mathbb{P}_x(X_{n-1}=z, T_x^+ > n-1) P(z,y) \\
&= \sum_{z\in S} P(z,y) \sum_{n=1}^{\infty} \mathbb{E}_x(1_{\{X_{n-1}=z, T_x^+>n-1\}}) \\
&= \sum_{z\in S} P(z,y) \sum_{m=0}^{\infty} \mathbb{E}_x(1_{\{X_m=z, T_x^+>m\}}) \\
&= \sum_{z\in S} P(z,y) \mathbb{E}_x(\sum_{m=0}^{\infty} 1_{\{X_m=z, T_x^+>m\}}) \\
&= \sum_{z\in S} P(z,y) \mu^x(z),
\end{aligned}$$

where, to go from the second to third line we changed the index of summation $n \geq 1$ to $m = n-1 \geq 0$, and we recognise in the last line the number of visits to z between 0 and $T_x^+ - 1$, as noted in the heuristic description. This shows that μ^x is super-invariant. If x is recurrent, then all these inequalities are equalities, so μ^x is invariant.

Minimality. Let μ be another super-invariant measure normalised at x, i.e., a solution to (3.4). We aim to show that for every $y \in S$, $\mu(x) \geq \mathbb{E}_x(\sum_{n=0}^{T_x^+-1} 1_{\{X_n=y\}})$. The super-invariance of μ gives us

$$\begin{aligned}
\mu(y) &\geq \sum_{w\in S} \mu(w) P(w,y) \\
&= P(x,y) + \sum_{w\in S\setminus\{x\}} \mu(w) P(w,y)
\end{aligned}$$

by separating the contribution of the term $w = x$ from the rest of the sum. Calling $w = w_1$ above and exploiting again (iteratively) the invariance of μ we get

$$\begin{aligned}
\mu(y) &\geq P(x,y) + \sum_{w_1\in S\setminus\{x\}} \sum_{w_2\in S} \mu(w_2) P(w_2,w_1) P(w_1,y) \\
&= P(x,y) + \sum_{w_1\in S\setminus\{x\}} P(x,w_1) P(w_1,y) + \sum_{w_1,w_2\in S\setminus\{x\}} \mu(w_2) P(w_2,w_1) P(w_1,y) \\
&= \mathbb{P}_x(X_1=y, T_x^+\geq 1) + \mathbb{P}_x(X_2=y, T_x^+\geq 2) + \sum_{w_1,w_2\in S\setminus\{x\}} \mu(w_2) P(w_2,w_1) P(w_1,y).
\end{aligned}$$

Continuing inductively, we find

$$\begin{aligned}
\mu(y) \geq \mathbb{P}_x(X_1=y, T_x^+\geq 1) + \ldots + \mathbb{P}_x(X_n=y, T_x^+\geq n) + \\
+ \sum_{w_n,\ldots,w_1\in S\setminus\{x\}} \mu(w_n) P(w_n,w_{n-1}) \ldots P(w_2,w_1) P(w_1,y).
\end{aligned}$$

Ignoring the sum at the end of the right hand side we get

$$\mu(y) \geq \sum_{i=1}^{n} \mathbb{P}_x(X_i = y, T_x^+ \geq i). \tag{3.6}$$

Letting $n \to \infty$, and recalling (3.5), we recognise the right hand side as $\mu^x(y)$, so $\mu(y) \geq \mu^x(y)$, as desired.

Nondegeneracy. Now suppose P irreducible (and recurrent), and let us show $\mu^x(y) \in (0, \infty)$. Let us first check positivity. Fix $y \in S$. Since P is irreducible, let $n \geq 0$ such that $P^n(x, y) > 0$. Since μ^x is invariant,

$$\begin{aligned}\mu^x(y) &= (\mu^x P^n)(y) \\ &= \sum_z \mu^x(z) P^n(z, y) \\ &\geq \mu^x(x) P^n(x, y) > 0\end{aligned}$$

so $\mu^x(y) > 0$. To check finiteness, let $m \geq 0$ such that $P^m(y, x) > 0$. Then

$$\mu^x(x) = \sum_z \mu^x(z) P^m(z, x) \geq \mu^x(y) P^m(y, x).$$

Since the left hand side is equal to 1 and $P^m(y, x) > 0$, we deduce that $\mu_x(y) < \infty$, as desired.

Uniqueness. The last thing to check is that there is a unique invariant measure normalised at x. Let μ be another such normalised invariant measure, and let $\rho = \mu - \mu^x$. We already know that $\rho \geq 0$ by minimality of μ^x, and furthermore by linearity, ρ is invariant. Hence ρ is an invariant measure, and satisfies $\rho(x) = 0$. Let us show that $\rho(y) = 0$ for any other $y \in S$. By invariance,

$$\rho(x) = \sum_z \rho(z) P^m(z, x) \geq \rho(y) P^m(y, x)$$

where, as before, $m \geq 0$ is chosen so that $P^m(y, x) > 0$. Since the left hand side is zero and the right hand side is nonnegative, it must be zero. But as $P^m(y, x) > 0$ the only possibility is that $\rho(y) = 0$, as desired. This completes the proof of the theorem. □

Example 3.10. Consider $G = \mathbb{Z}^d$ with $d = 1, 2$. Then G is recurrent (by Pólya's theorem) and irreducible. Hence invariant measures are unique up to scaling. Since $\mu(x) = 1$ is an invariant measure, all invariant measures must be constant. In particular there can be no invariant distribution.

Example 3.11. Note that if P is the transition matrix of an irreducible, recurrent chain, and $x, y \in S$ then necessarily μ^x and μ^y are proportional to one another. This is absolutely not obvious *a priori*. For instance, suppose that, between two successive visits to x, the number of visits to y is $\alpha > 0$ on average. Then it follows from Theorem 3.8 (or more precisely Corollary 3.9) that, between two successive visits to y, the number of visits to x is on average $1/\alpha$.

Lecture 10: Friday 4.11.2022

3.3 Positive recurrence and invariant distributions

At this point we know by Proposition 3.4 that, to compute the long-term distribution of a Markov chain over its state space, we must look for an invariant distribution. We also know, by Theorem 3.8, that invariant measures always exist (at least for recurrent chains, which is the only case that is sensible to consider – for transient chains, the chain goes "off to infinity" so there is no point looking for limiting distributions; this will be formalised later). We also know these invariant measures are unique up to scaling. But the question remains: does there exist an invariant distribution? (If one exists, it is clearly necessarily unique by Corollary 3.9.)

This question is clearly more subtle than just recurrence/transience. For instance, in the case of the frog of Example 3.3 or any random walk on a finite graph, (see Example 3.5), we could find an invariant distribution. But on $\mathbb{Z}$ and $\mathbb{Z}^2$, both of which are recurrent, there is no invariant distribution. As we will see, the key notion for the existence of an invariant distribution is the following notion of *positive recurrence*.

Definition 3.12. *Let P be the transition matrix of a Markov chain on a state space S. The state $x \in S$ is called **positive recurrent** if*

$$\mathbb{E}_x(T_x^+) < \infty,$$

*where as before $T_x^+ = \inf\{n \geq 1 : X_n = x\}$ is the return time to x. If however, $T_x^+ < \infty$ with probability one starting from x, but the above expectation is infinite, x is said to be **null recurrent**.*

The condition of positive recurrence is a strong form of recurrence. Not only is the return time finite, but the expectation is finite, so the chain returns to x *quickly*. By contrast, a null recurrent state is one such that the return time is finite with probability one, but in fact one must typically wait a very long time before seeing the chain return to x. As we will see, the long term behaviour of a Markov chain is very different in these two cases. The theorem below provides the basis of this fundamental distinction.

Theorem 3.13. *Let P be an irreducible transition matrix. Then the following are equivalent:*

(i) Every state $x \in S$ is positive recurrent.

(ii) Some state $x \in S$ is positive recurrent.

(iii) There exists an invariant distribution.

Furthermore, if either (i)-(iii) occur, the unique distribution π is necessarily given by the formula

$$\pi_x = \frac{1}{\mathbb{E}_x(T_x^+)}. \tag{3.7}$$

Before seeing a proof of this theorem, we make a few remarks.

Remark 3.14. At first the formula (3.7) appears miraculous. Let $m_x = \mathbb{E}_x(T_x^+)$ (sometimes called the *mean recurrence time*). It's not even clear at first why $\sum_x (1/m_x)$ should be equal to 1: this sum is either zero (all the terms are zero), or equal to 1. The theorem guarantees there are no other possibilities, since $1/m_x$ defines the unique invariant distribution. (It's even less clear why $1/m_x$ should be invariant!)

On the other hand, some basic considerations can help us understand the formula. Suppose the chain is recurrent and consider the successive returns to x. Then the time between the successive returns forms i.i.d. random variable (by the strong Markov property), of mean m_x, by definition. Thus, using the **law of large numbers**, the **fraction** of times spent at x in the long run should be $1/m_x$. But, at least intuitively, if the chain spends 3% of its time in a state x, the probability to find it at x after a very long time should also be 3%. That is, the fraction of times spent at x should coincide with the probability to find it at x, hence it is reasonable to expect that $\pi_x = 1/m_x$.

Let us now begin a proof of that theorem, and defer examples until after the proof.

Proof. (i) $\Rightarrow$ (ii) is trivial. Let us assume (ii) and show (iii). Suppose x is a positive recurrent state. In particular x is recurrent so, by Theorem 3.8, the return measure μ^x based at x is invariant. Let us check that μ^x has finite total mass (so that scaling by that mass, we would get an invariant distribution). Recalling the definition of $\mu^x(y)$ as the expected number of visits to y prior to the return to x,

$$\begin{aligned}\sum_{y\in S}\mu^x(y) &= \sum_{y\in S}\mathbb{E}_x\Big(\sum_{n=0}^{T_x^+-1} 1_{\{X_n=y\}}\Big)\\ &= \mathbb{E}_x\Big(\sum_{n=0}^{T_x^+-1}\sum_{y\in S} 1_{\{X_n=y\}}\Big)\\ &= \mathbb{E}_x(T_x^+)\end{aligned} \tag{3.8}$$

since $\sum_{y\in S} 1_{\{X_n=y\}} = 1$: there is always a unique y where the chain is located at time n. Since x is positive recurrent, the right hand side is finite, and thus

$$\pi = \mu^x/m_x \tag{3.9}$$

(where $m_x = \mathbb{E}_x(T_x^+)$) is an invariant distribution. This proves (iii).

Now let us show (iii) $\Rightarrow$ (i). Let π be an invariant distribution. Let us first show that $\pi_x > 0$ for every $x \in S$. Since $\sum_x \pi_x = 1$, there is at least one x_0 such that π_{x_0} such that $\pi_{x_0} > 0$. Now fix $x \in S$. By irreduciblity, choose $n \geq 0$ such that $P^n(x_0, x) > 0$. Then, since π is invariant,

$$\pi_x = \sum_z \pi_z P^n(z,x) \geq \pi_{x_0} P^n(x_0,x) > 0,$$

as desired. Let us show that x is positive recurrent. Since $\pi_x > 0$ we can consider the measure π/π_x, which is invariant (as a scaling of π) and is normalised at x. Since μ^x is the

minimal such invariant, we deduce that $\mu^x(y) \leq \pi_y/\pi_x$, for any $y \in S$. (Note here that we do not yet know that x is recurrent, but even without this assumption, it holds that $\mu^x(y) \leq \pi_y/\pi_x$: this is because μ^x is the minimal super-invariant, and π, being invariant, is also super-invariant). Hence

$$\sum_{y \in S} \mu^x(y) \leq \sum_{y \in S} \pi_y/\pi_x = 1/\pi_x < \infty. \tag{3.10}$$

We have already noted in (3.8) that $\sum_y \mu^x(y) = \mathbb{E}_x(T_x^+)$. Combining with (3.10), we deduce that x is positive recurrent, as desired. This proves (i).

Hence (i), (ii) and (iii) are equivalent. To prove (3.7), we recall (3.9), valid for any positive recurrent state $x \in S$ (hence for any $x \in S$ by (i)). We evaluate both sides at the state x. Since $\mu^x(x) = 1$, we deduce that

$$\pi_x = 1/m_x,$$

as desired. □

To get a feeling for this fundamental theorem we now discuss a few examples.

Example 3.15. Let $G = \mathbb{Z}^d$. Does G admit an invariant distribution? If so, then the graph is positive recurrent and so in particular recurrent. This clearly rules out $d \geq 3$. What about $d = 1, 2$? Even in that case no invariant distribution can exist, as already mentioned in Example 3.10 (there is uniqueness of invariant measures up to scaling, and $\mu(x) = 1$ is invariant).

Lecture 11: Thursday 10.11.2023

Example 3.16. Let P be an irreducible transition matrix over a *finite* state space. Then the corresponding chain is necessarily positive recurrent. Indeed, the chain is recurrent, so μ^x is an invariant measure by Theorem 3.8. Its total mass is necessarily finite (each entry is finite, as shown in Theorem 3.8, and there are only a finite number of states). Hence the chain possesses an invariant distribution, and so is positive recurrent.

Sometimes we can use the invariant distribution to compute the mean recurrence times, as in this example.

Example 3.17. Suppose we shuffle cards (say with random transpositions). Then this is a random walk on the permutation group S_k (a finite state space), and hence is positive recurrent. As already mentioned, an invariant distribution is given by the uniform distribution $\pi_x = 1/k!$, because on this graph every vertex has a constant degree (more precisely, $\binom{k}{2}$ in the case of random transpositions). By uniqueness (say, Corollary 3.9) the invariant distribution is necessarily unique. Suppose we start from the identity permutation, denoted by e, and let m denote the mean return time to e: i.e., $m_e = \mathbb{E}_e(T_e^+)$. Then

$$\pi_e = 1/m_e$$

and therefore

$$m_e = k!$$

In other words, on average it takes $k!$ for the deck to return to its original configuration. This is finite but very long when k is large!

Example 3.18. Consider the renewal chain of Example 3.6. Does the chain have an invariant distribution? Clearly, the chain is irreducible on $S = \{0, \dots, M-1\}$ where M is the essential supremum of the renewal distribution $(p_n)_{n\geq 1}$, i.e., $M = \sup\{n \geq 1 : p_n > 0\}$ (note that M could be infinite). By Theorem 3.8 there is an invariant distribution if and only if 0 is positive recurrent. Clearly, starting from 0, the return time to 0 has the same law as the inter-renewal distribution $(p_n)_{n\geq 1}$. Thus the renewal chain has an invariant distribution if and only if $\sum_{n=1}^{\infty} np_n < \infty$. This is consistent with the invariant measure you will describe in an exercise on Problem Sheet 5.

3.4 Convergence to equilibrium

We now come to one of the crucial theorems concerning Markov chains. Suppose X is a Markov chain on a state space S and π is an invariant distribution. Then, as suggested by Proposition 3.4, π_y is a good candidate for $\lim_{n\to\infty} \mathbb{P}(X_n = y)$. But does the limit actually exist? And can we prove this?

The answer is basically yes, except for a possible obstruction of a combinatorial nature, called **periodicity**.

Definition 3.19. *We say that a state x is **aperiodic** if $P^n(x,x) > 0$ for* all *sufficiently large n: i.e., there exists n_0 (possibly depending on x) such that $P^n(x,x) > 0$ for all $n \geq n_0$.*

Example 3.20. Consider $G = \mathbb{Z}$ and let $x = 0$. Is x aperiodic? The answer is no, since the walk alternates between even and odd positions so $P^n(0,0) = 0$ for every n odd. (More generally this will be the case for every **bipartite** graph – a graph for which vertices can be coloured black and white in such a way that black vertices only have white neighbours and vice-versa. Thus, on $\mathbb{Z}^d$ which is bipartite thanks to the usual checkerboard colouring, the walk is periodic).

Example 3.21. Consider the n-cycle, $G = \mathbb{Z}/(n\mathbb{Z})$ equipped with edges between x and $x \pm 1$ mod n for every $0 \leq x \leq n-1$, and consider a given state, say $x = 0$. Is x aperiodic?

Actually this depends on the parity of n. If n is even then the walk alternates between even and odd positions, just as in the case of $\mathbb{Z}$. However if n is odd then the walk could return to an even position in an odd number of steps (and vice-versa) by wrapping around the cycle. Hence in that case x is aperiodic.

The kind of restrictions in the above two examples (the walk can only return to its starting point on certain integer multiples) is in fact the most general obstruction to aperiodicity, as shown in the following exercise.

Exercise 3.22. The state x is aperiodic if and only if the set of possible return times

$$\mathcal{R}_x := \{n \geq 1 : P^n(x,x) > 0\}$$

has a greatest common divisor (gcd) equal to 1. In all other cases, $\mathcal{R}_x$ is of the form $\mathcal{R}_x = \{d, 2d, 3d, 4d, \ldots\}$ where $d = d_x$ is called the period of x. Furthermore, if the chain is irreducible, then all the states are either simultaneously aperiodic or simultaneously periodic. In the latter case, the period of a state x does not depend on x and is hence constant across the state space: that is, $d_x = d_y$ for $x, y \in S$.

We will not use this exercise and so do not include its proof, though it is good to keep in mind to understand the notion of periodicity/aperiodicity. The following lemma (which actually forms a small part of the proof of the exercise) will however be needed and so we include it separately and prove it.

Lemma 3.23. *Suppose P is irreducible and let $z \in S$. Suppose z is aperiodic. Then for every $x, y \in S$, there exists $n_0 = n_0(x, y, z)$ such that for all $n \geq n_0$, $P^n(x, y) > 0$.*

Proof. Fix $k, j \geq 0$ such that $P^k(x, z) > 0$ and $P^j(z, y) > 0$, and let m_0 be such that $P^m(z, z) > 0$ if $m \geq m_0$. Then if $m \geq m_0$,

$$P^{k+j+m}(x, y) \geq P^k(x, z)P^m(z, z)P^j(z, y) > 0$$

so the lemma holds with $n_0 = m_0 + j + k$. □

In particular, if P is irreducible, some state is aperiodic if and only if all states are aperiodic. In that case we call the entire Markov chain **aperiodic**. We can now state one of the main theorems of this class.

Theorem 3.24. *Suppose P is an irreducible, aperiodic transition matrix on S with invariant distribution $(\pi_y)_{y \in S}$. Let $X = (X_n, n \geq 0)$ be Markov (λ, P) for some arbitrary starting distribution λ. Then for every $y \in S$,*

$$\mathbb{P}(X_n = y) \to \pi_y$$

as $n \to \infty$. In particular, for every $x, y \in S$, $\lim_{n\to\infty} P^n(x, y)$ exists and equals π_y.

Proof. The proof of this theorem is quite beautiful and uses an idea called **coupling** which is very important in the study of Markov processes and more generally in probability theory. The idea is to introduce another copy of the chain, $(Y_n)_{n \geq 0}$, which is started from the invariant distribution π (hence by Proposition 3.2 its distribution at any given time is also equal to π), and force X_n to be equal to Y_n by modifying the trajectory of X.

Thus let Y be Markov (π, P) and be independent from X. We consider the following stochastic process

$$W_n = (X_n, Y_n),$$

taking values in $S' = S \times S$.

Step 1. We show that W is an irreducible Markov chain on S' (called the **product chain**). Since X and Y are independent and are both Markov with transition matrix P, it is clear that W is a Markov chain on S', with transition matrix Q defined by

$$Q\Big((x,y);(z,w)\Big) = P(x,z)P(y,w)$$

for any $(x,y) \in S'$ and $(z,w) \in S'$ (i.e. for any $x,y,z,w \in S$). The starting distribution μ of W is

$$\mu((x,y)) = \mathbb{P}(X_0 = x, Y_0 = y) = \lambda_x \pi_y$$

by independence between X_0 and Y_0.

Now let us check that Q is irreducible (and in fact aperiodic, although we will not need this). Fix (x,y) and (z,w) arbitrary in S'. Since P is aperiodic and irreducible, we know by Lemma 3.23 that $P^n(x,z) > 0$ for all n large enough, and $P^n(y,w) > 0$ for all n large enough. Therefore,

$$Q^n\Big((x,y);(z,w)\Big) = P^n(x,z)P^n(y,w) > 0$$

for all n large enough. Thus Q is irreducible (and in fact aperiodic).

Step 2: Q is positive recurrent. Since Q is irreducible, in order to show that the product chain is positive recurrent it suffices to check it has an invariant distribution by Theorem 3.13. On the other hand, it is easy to guess what the invariant distribution should be: since X and Y are independent, we guess $\nu_{(x,y)} = \pi_x \pi_y$ (for $x,y \in S$) defines an invariant distribution. Let us check this. First of all, $\nu_{(x,y)} \geq 0$ for all $x,y \in S$, and

$$\sum_{(x,y)\in S'} \nu_{(x,y)} = \sum_{x,y\in S} \pi_x \pi_y = (\sum_{x\in S} \pi_x)(\sum_{y\in S} \pi_y) = 1,$$

so ν is a distribution on $S' = S \times S$. Second of all, if $(z,w) \in S'$ then

$$\begin{aligned}\sum_{(x,y)\in S'} \nu_{(x,y)} Q\Big((x,y);(z,w)\Big) &= \sum_{x\in S, y\in S} \pi_x \pi_y P(x,z)P(y,w) \\ &= \Big(\sum_{x\in S} \pi_x P(x,y)\Big)\Big(\sum_{y\in S} \pi_y P(y,w)\Big) \\ &= \pi_z \pi_w = \nu_{(z,w)}\end{aligned}$$

so ν is invariant. Hence Q is positive recurrent. In particular Q is recurrent!

Let us fix $u \in S$ a reference state, and set

$$T = T_{(u,u)} = \inf\{n \geq 0 : X_n = Y_n = u\}.$$

T is the first time both chains are equal to u. Since Q is recurrent, note that $T < \infty$ almost surely.

Step 3. Coupling and conclusion. We modify the trajectory of X by taking it equal to Y after time T (for this reason T is called the **coupling time**). That is, define

$$X'_n = \begin{cases} X_n & \text{if } n \leq T \\ Y_n & \text{if } n > T. \end{cases}$$

Since T is a stopping time, it is not hard to see by the strong Markov property that X' is also Markov (λ, P) and so has the same law as X. (Indeed, after time T, the stochastic process Y is nothing but Markov (δ_u, P), which is exactly the law that X needs to have after time T.) Thus $\mathbb{P}(X_n = y) = \mathbb{P}(X'_n = y)$ for any $n \geq 0$ and for any $y \in S$. And since $X'_n = Y_n$ for $n \geq T$, we see that

$$\mathbb{P}(X_n = y, T \leq n) = \mathbb{P}(Y_n = y, T \leq n).$$

Consequently,

$$\begin{aligned} \mathbb{P}(X_n = y) &= \mathbb{P}(X_n = y, T \leq n) + \mathbb{P}(X_n = y, T > n) \\ &= \mathbb{P}(Y_n = y, T \leq n) + \mathbb{P}(X_n = y, T > n). \end{aligned}$$

Hence

$$\begin{aligned} |\mathbb{P}(X_n = y) - \mathbb{P}(Y_n = y)| &\leq \mathbb{P}(Y_n = y, T > n) + \mathbb{P}(X_n = y, T > n) \\ &\leq 2\mathbb{P}(T > n) \end{aligned}$$

The right hand side tends to zero since $T < \infty$ with probability one by Step 2. On the other hand, since Y is Markov (π, P) and π is invariant, $\mathbb{P}(Y_n = y) = \pi_y$, by Proposition 3.2. In summary, we have proved

$$|\mathbb{P}(X_n = y) - \pi_y| \to 0$$

as $n \to \infty$, as desired. (In fact, the argument above shows that the convergence is uniform in $y \in S$). □

Example 3.25. To visualise the argument given above in a concrete example, consider for instance a card shuffling process (say we shuffle cards using random transpositions, but this could of course be more general). Let X_n denote the deck of cards after n shuffles, which is an element of the permutation group S_k with $k = 52$ for a real deck of cards. Note that the invariant distribution π is uniform on S_k. Along with our deck of cards X, we consider an imaginary or virtual deck of cards Y_n, which already at the beginning of time is assumed to be perfectly shuffled, i.e. initially Y_0 has law π. Initially the two decks are shuffled independently of one another. We wait until the first time that both decks are perfectly ordered, say both equal to the identity permutation. It will take a very long time for this to happen (roughly $(k!)^2$ on average, in fact) but Step 2 of the proof shows this will eventually happen at some time $T < \infty$. We then let the two decks evolve in parallel after time T with *identical* updates. When we consider the decks individually, at each time we are updating the deck by using the correct shuffling procedure, hence altogether this evolution defines a valid way of constructing a process with the same law as X. On the other hand, Y is already

in equilibrium at the beginning of time, and remains so by invariance of π and Proposition 3.2. Since the two decks are *identical* after time T which is finite, the law of X_n and the law of Y_n are necessarily close to one another, the only difference coming from the possibility that time T (the "coupling time") has not yet occurred, even though n is very large.

We now see a few basic examples of application of the theorem.

Example 3.26. Let $G = (V, E)$ be a finite connected graph, and suppose that the random walk $(X_n, n \geq 0)$ on G is aperiodic. Then for any vertices u and v in V, we have, as $n \to \infty$,

$$\mathbb{P}_u(X_n = v) \to \frac{\deg(v)}{2|E|}. \tag{3.11}$$

To see this, note that the walk is irreducible since G is connected, and aperiodic by assumption. It is furthermore positive recurrent and in fact an invariant distribution is given by scaling the invariant measure $\mu(x) = \deg(x)$ so its total mass becomes one. To conclude it remains to observe that $\sum_x \deg(x) = 2|E|$ (since each edge is counted twice in the sum) and apply Theorem 3.24.

In fact, the same argument applies to undirected graphs, even though in that case we do not know the invariant distribution explicitly (note that this is in general not simply given by the in-degree of a vertex, nor its outdegree). This is of great practical importance in connection with search engines. To explain, the world wide web can be viewed as a graph, with vertices given by webpages and (directed) edges between vertices if there is a hyperlink between the two pages. It is essential in this application that we view the graph as directed rather than undirected: you might have a link to the New York Times on your personal webpage, but typically the converse is not true! In 1998, two PhD students at Stanford named Larry Page and Sergey Brin, together with two professors (Rajeev Motwani and Terry Winograd) came up with an interesting proposal to evaluate the relative importance of webpages (the paper is publicly available, see [PBMW99]). One could simply consider a random walk on the directed graph of the world wide web; its invariant distribution gives the desired ranking. The idea is that if a website is important, then many webpages will point to it; this will be reflected in the invariant distribution of the random walk on this directed graph (as suggested by the formula (3.11) for the undirected case). This simple algorithm was called PageRank by its authors. A year later they quit their PhD to start a company that would implement the algorithm. The company became known as... Google. (In reality, the Markov chain on which PageRank is based has an extra randomisation step compared to the random walk, where every so often the walk restarts on the graph with some given distribution (say proportionally to the in- or out-degree of a vertex.)

There are two reasons why this is a tractable algorithmic problem. One is that the random walk on the world wide web reaches its equilibrium very rapidly. There are ways to assign a rigourous meaning to this last statement, which are in fact related to the "coupling time" of the proof of Theorem 3.24. It can then be shown that if the graph has order n vertices, typically the walk will reach its equilibrium in time of order $O(\log n)$. This type of

results is closely connected with a beautiful area of modern probability theory called **mixing times and cutoff phenomenon for Markov chains**, see [LP17] for more on this topic and [BLPS18] for the above results. Secondly, the invariant distribution can be computed as the leading eigenvector of a matrix related to the transition matrix (i.e. closely related to the adjacency matrix of the graph). Since the matrix in question is typically quite sparse, the numerical computation of its leading eigenvector is not as difficult as what one might fear.

Example 3.27. Consider the renewal chain $(R_n, n \geq 0)$ of Example 3.6. Let $(p_n)_{n\geq 1}$ denote the renewal distribution and let $m = \sum_{n\geq 1} np_n$ denote the average waiting time between two successive renewals. Suppose to simplify that $p_n > 0$ for all $n = 1, \dots, M$ where $M = \sup\{n \geq 1 : p_n > 0\}$ is the essential supremum of $(p_n)_{n\geq 1}$. Then the renewal chain is aperiodic on $\{0, \dots, M\}$. If $m < \infty$, we have already checked in Example 3.18 that there is an invariant distribution. Therefore

$$\mathbb{P}(R_n = 0) \to \pi_0$$

as $n \to \infty$. The event $\{R_n = 0\}$ is precisely the event that there is a renewal at time n (i.e., in the bus example, it is the event that you get the bus as soon as you arrive). The computation of π implies that $\pi_0 = 1/m$.

Lecture 12: Friday 11.11.22

3.5 Time-reversal and detailed balance equations

In the definition of a Markov chain, the past and the present are conditionally independent given the present. It is therefore natural to ask ourselves, whether there is complete time-reversal symmetry?

Intuitively this is typically not possible for "entropic" reasons: intuitively, the entropy of a Markov chain can only increase as time goes on (this can in fact be proved rigourously without too much effort). This clearly prevents a complete time-reversal symmetry. However if the Markov chain is started from the invariant distribution the entropy stays constant (since the distribution is constant). Can there be complete symmetry at equilibrium? The answer, given in the theorem below, is quite close to yes.

Theorem 3.28. *Let P be an irreducible transition matrix, and let π be an invariant distribution for P. Suppose X is Markov (π, P) and let $N \geq 1$ be fixed. Set*

$$\hat{X}_n = X_{N-n}, \quad 0 \leq n \leq N.$$

Then $\hat{X}$ is Markov $(\pi, \hat{P})$, where

$$\hat{P}(x, y) = \frac{\pi_y}{\pi_x} P(y, x). \tag{3.12}$$

Remark 3.29. Before the proof, we note that the formula for $\hat{P}$ in (3.12) is reminiscent of Bayes' elementary formula for conditional probability:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A)}{\mathbb{P}(B)}\mathbb{P}(B|A).$$

In fact, (3.12) is Bayes' formula with $B = \{X_n = x\}, A = \{X_{n-1} = y\}$. As in Bayes' formula, (3.12) shows that in order to observe a transition from y to x in the reverse time-direction (or rather, for that transition to be likely), it is not sufficient for the transition x to y to be likely in the usual direction of time: in addition, the state x itself must be overall likely compared to y, unconditionally on any information.

Proof of Theorem 3.28. We first note that the formula (3.12) is meaningful since $\pi_x > 0$ for any $x \in S$ as the chain is irreducible. The first thing to do is to check that $\hat{P}$ is indeed a transition matrix (i.e., a stochastic matrix). It is obvious that $\hat{P}(x, y) \geq 0$ so let us check that the row sum is one:

$$\sum_{y\in S} \hat{P}(x, y) = \frac{1}{\pi_x} \sum_{y\in S} \pi_y P(y, x) = \frac{1}{\pi_x}\pi_x = 1$$

by invariance of π. So $\hat{P}$ is indeed a transition matrix.

Now let us check that $\hat{X}$ is Markov $(\pi, \hat{P})$ by direct computation (more specifically Proposition 1.9). Then for any sequence of states $x_0, \ldots, x_N$, noting that $\pi_u P(u, v) = \hat{P}(v, u)\pi_v$,

$$\begin{aligned}
\mathbb{P}(\hat{X}_0 = x_0, \ldots, \hat{X}_N = x_N) &= \mathbb{P}(X_0 = x_N, \ldots, X_N = x_0) \\
&= \underbrace{\pi_{x_N} P(x_N, x_{N-1})}\, P(x_{N-1}, x_{N-2}) \ldots P(x_1, x_0) \\
&= \hat{P}(x_{N-1}, x_N)\, \underbrace{\pi_{x_{N-1}} P(x_{N-1}, x_{N-2})} \ldots P(x_1, x_0) \\
&= \ldots \\
&= \hat{P}(x_{N-1}, x_N)\hat{P}(x_{N-2}, x_{N-1}) \ldots \hat{P}(x_0, x_1)\pi_{x_0},
\end{aligned}$$

by induction. Rearranging the terms in the product in the right hand side, we see that this is nothing else than the product of the $\hat{P}$-transition probabilities, so we conclude by Proposition 1.9. □

There is one particular case which is especially interesting, and which is when $\hat{P} = P$. This corresponds to the following definition

Definition 3.30. *Let μ be a measure on a state space S and let P be a transition matrix. We say that μ solves the **Detailed Balance Equations** (DBE) (equivalently: μ is **reversible**) if for all $x, y \in S$:*

$$\mu_x P(x, y) = \mu_y P(y, x).$$

To understand this definition, we make the following observation.

Remark 3.31. Let π be the invariant distribution of an irreducible, aperiodic Markov chain. Then $\pi(x)P(x,y)$ is the probability to observe a transition from x to y at equilibrium (or the frequency of such transitions). So π solves the detailed balance equations if and only if the frequency of transitions from x to y equals the frequency of transitions from y to x, for every $x, y \in S$.

Example 3.32. If G is a locally finite graph and $\mu(x) = \deg(x)$, then μ is reversible. Indeed, if x, y are two neighbouring vertices (there is nothing to prove otherwise), then $\mu(x)P(x,y) = \deg(x) \times (1/\deg(x)) = 1$ which is also $\mu(y)P(y,x)$, so μ solves the detailed balance equations.

It is easy to check that reversibility implies invariance (but the converse is far from true, as we will discuss below):

Lemma 3.33. *Suppose μ is a reversible measure for P. Then μ is invariant.*

Proof. Let $y \in S$. We just check the identity defining invariance:

$$\begin{aligned}\sum_{x\in S} \mu(x)P(x,y) &= \sum_{x\in S} \mu(y)P(y,x) \\ &= \mu(y) \sum_{x\in S} P(y,x) \\ &= \mu(y),\end{aligned}$$

where we used the reversibility of μ in the first line, and the fact that P is a stochastic matrix in the last one. This gives the result. □

For instance, combining Lemma 3.33 and Example 3.32, we recover the fact (already established in Example 3.5) that $\mu(x) = \deg(x)$ defines an invariant measure on a locally finite graph, since it is reversible.

The notion of reversible measure leads us to that of reversible Markov chains. Let X be Markov (λ, P). Let us say that X is a **reversible chain** if for all $N \geq 1$, the distribution of $\hat{X}$ is the same as X, where $\hat{X}_n = X_{N-n}$ for $0 \leq n \leq N$. Then we deduce from Theorem 3.28 the following result.

Corollary 3.34. *Let λ be a distribution and P an irreducible transition matrix on a state space S. The following are equivalent:*

(i) X is reversible

(ii) λ satisfies the detailed balance equations (in particular λ is an invariant distribution).

Remark 3.35. In words, if π is an invariant distribution solving the detailed balance equations, then the movie looks the same forwards *and* backwards (at equilibrium, of course)! Surprisingly, it is possible to have walks on *directed* graphs which nevertheless have a reversible equilibrium. See the exercises on the upcoming example sheet.

Proof. Let us start with the proof of (ii) $\Rightarrow$ (i). If λ is reversible then λ is an invariant distribution by Lemma 3.33. Let $N \geq 1$. Then we can apply Theorem 3.28, and $\hat{X}$ is Markov $(\lambda, \hat{P})$, but then clearly $\hat{P} = P$. So $\hat{X}$ has the same law as X and X is reversible.

In the converse direction suppose X is reversible and let us check that λ is also invariant. Take $N = 1$. Then reversibility shows that (X_1, X_0) has the same law as (X_0, X_1). In particular X_1 has the same law as X_0 so λ is indeed invariant. So Theorem 3.28 applies. Since X is reversible, we deduce that $\hat{P} = P$ and hence λ solves the detailed balance equations. □

In the exercises you will see more examples of reversible Markov chains and examples in which reversibility does not hold. (In fact, it can be shown that any Markov chain on $\mathbb{N}$ in which only transitions to nearest neighbours are allowed, has a reversible measure). We will go through a more complicated example in the next section.

Lecture 13: Thursday 17.11.22

3.6 Example: particle system

In many stochastic processes of interest (be it in statistical mechanics or in more applied fields such as biology) the state space and the process is often too complex to solve directly for an invariant distribution. But solving detailed balance equations is far easier, and if you are given a Markov chain it should be the first thing you try. To illustrate this point we will describe a Markov chain which is a little more complex than other examples treated in class so far, and for which we can solve the detailed balance equations.

The example we will describe is the most basic example of a **particle system**, in which we consider multiple random walks on a finite graph without interaction. Let us suppose that $G = (V, E)$ is a finite, connected graph. We are given a number of particles on this graph (we identify each particle with the position of a random walker on this graph). Suppose this number is N. At each step we pick a particle uniformly at random, and move it to a randomly chosen neighbour. Let us describe the state space and transition matrix a little more formally. The Markov chain will keep track of the number of particles at each site. Thus the state space will be

$$S = \mathbb{N}^V; \text{ with } \mathbb{N} = \{0, 1, \ldots\}$$

and an element of S (a *configuration* of the system) will be denoted by $\vec{n} = (n_x)_{x \in V}$ with $n_x \in \mathbb{N}$ for every vertex $x \in V$. Let $\vec{e}_x = (0, 0, \ldots, 0, 1, 0, \ldots 0)$ with a unique 1 at position $x \in V$. Then we define a transition matrix on S by setting

$$P(\vec{n}, \vec{n} - \vec{e}_x + \vec{e}_y) = \begin{cases} \frac{n_x}{N} \frac{1}{\deg(x)} & \text{if } n_x > 0 \text{ and } y \sim x \\ 0 & \text{else.} \end{cases} \tag{3.13}$$

In words, the new state $\vec{n} - \vec{e}_x + \vec{e}_y$ is the one in which we have taken one particle from x and moved it to y. The probability to do so is assumed to be n_x/N (this is the probability

to choose a particle at x among the n_x present in $\vec{n}$) times $1/\deg(x)$, which is the same as the probability that a random walk at x would move to y.

In this Markov chain, the total number of particles remains constant equal to N. If we restrict S to the configurations in which there are N particles (let us call S_N the corresponding subspace), it is not hard to see we have an irreducible Markov chain on a finite state space (hence positive recurrent). What is the unique invariant distribution of this chain? In other words, as time goes to ∞, how will the particles be distributed across the graph?

Theorem 3.36. *Fix any $\lambda > 0$. Then the measure*

$$\mu(\vec{n}) = \prod_{x \in V} \frac{(\lambda \deg(x))^{n_x}}{(n_x)!}$$

is a reversible (hence invariant) measure.

Let us prove this theorem first and discuss later what it means concretely.

Proof. Let $\vec{m} = \vec{n} - \vec{e}_x + \vec{e}_y$. We need to check that

$$\mu(\vec{n})P(\vec{n}, \vec{m}) = \mu(\vec{m})P(\vec{m}, \vec{n}).$$

Because of the product form of μ, and since nothing changes except at x and y, we need to check that

$$\frac{(\lambda \deg(x))^{n_x}}{(n_x)!} \frac{(\lambda \deg(y))^{n_y}}{(n_y)!} \times \frac{n_x}{N} \frac{1}{\deg(x)} = \frac{(\lambda \deg(x))^{n_x - 1}}{(n_x - 1)!} \frac{(\lambda \deg(y))^{n_y + 1}}{(n_y + 1)!} \times \frac{n_y + 1}{N} \frac{1}{\deg(y)}.$$

But making the cancellations, both sides are equal to

$$\frac{\lambda^{n_x + n_y} (\deg(x))^{n_x - 1} (\deg(y))^{n_y}}{(n_x - 1)! n_y!}$$

so equality holds and the proof of Theorem 3.36 is complete. □

The formula for the reversible measure μ above has an interpretation in terms of Poisson random variables. Recall that a random variable X with values in $\mathbb{N}$ is said to have the Poisson distribution with parameter α if

$$\mathbb{P}(X = k) = \exp(-\alpha) \frac{\alpha^k}{k!}.$$

Corollary 3.37. *Let π denote the law of $(X_v)_{v \in V}$ where X_v are independent Poisson random variables with parameter $\lambda \deg(v)$. Then π is an invariant distribution on S.*

Proof. π is clearly a distribution (since it is the joint law of some set of random variables), and is easily seen to be fixed multiple of the measure μ above. Indeed,

$$\begin{aligned}\pi(\vec{n}) &= \prod_{x\in V} e^{-\lambda \deg(x)} \frac{(\lambda \deg(x))^{n_x}}{(n_x)!} \\ &= \left(\prod_{x\in V} e^{-\lambda \deg(x)}\right) \mu(\vec{n})\end{aligned}$$

so π is indeed proportional to μ with factor of proportionality given by $\prod_{x\in V} e^{-\lambda \deg(x)} = e^{-2\lambda|E|}$. Since μ is reversible, π is also reversible hence a reversible distribution. □

The distribution is not unique because here the number of particles is unconstrained so the Markov chain is not irreducible. When we condition on the number of particles however we obtain the following result:

Corollary 3.38. *For any $\lambda > 0$, the distribution π conditioned on having N particles in total is the unique invariant distribution on S_N.*

It is a priori not obvious that this law does not depend on λ, but since each such choice of parameter λ results in an invariant distribution, and S_N is finite so the chain is irreducible recurrent, the conditioned distribution cannot depend on λ. When $N \to \infty$, it can be shown that choosing λ of order $N/(2|E|)$ would make the conditioning on the total size of the system essentially harmless. Hence in that case the unique invariant distribution is very close to independent Poisson random variables with parameter $\lambda \deg(x)$ at each vertex x of the graph.

3.7 Complements

We briefly mention a few additional results without proof, which won't be needed for the rest of the course, but which you may find interesting.

The first one is an *ergodic theorem*, which says that for an irreducible, aperiodic positive recurrent, not only does the long term distribution of the chain converges to the invariant distribution π, but also the fraction of times spent at any given state becomes approximately equal to π in the long run. Thus *spatial averages* (π) are equal to *time averages*.

Theorem 3.39. *Let P be an irreducible, aperiodic positive recurrent Markov chain and let π be the invariant measure. Let X be Markov (λ, P) for some starting distribution λ. For $y \in S$, let $V_n(y) = \sum_{i=0}^{n-1} 1_{\{X_n=y\}}$ denote the number of visits to y by time n. Then as $n \to \infty$,*

$$\frac{V_n(y)}{n} \to \pi_y \tag{3.14}$$

almost surely. More generally if $f : S \to \mathbb{R}$ is a bounded function on S, then as $n \to \infty$,

$$\frac{1}{n}\sum_{i=0}^{n-1} f(X_i) \to \bar{f}, \tag{3.15}$$

where $\bar{f} = \sum_{x \in S} \pi_y f(y)$ *is the expectation of* f *under the invariant distribution* π*.*

This ergodic theorem (part (3.14)) can be deduced from the law of large numbers by considering the successive visits to y, as the interval of times between successive visits form i.i.d. random variables by the strong Markov property, with mean $\mathbb{E}_y(T_y^+) = 1/\pi_y$. The second part (3.15) can then be deduced using some relatively standard measure theoretic arguments.

Finally, we have focused in this section on positive recurrent chains. But what if the chain is null recurrent, for instance on $\mathbb{Z}$ or $\mathbb{Z}^2$? What can be said about the limiting probability to find the chain in a given state? What about the fraction of times spent at a given state?

Theorem 3.40. *Let* P *be an irreducible, aperiodic* null *recurrent Markov chain. Let* X *be Markov* (λ, P) *for some starting distribution* λ*. For* $y \in S$*, let* $V_n(y) = \sum_{i=0}^{n-1} 1_{\{X_n = y\}}$ *denote the number of visits to* y *by time* n*. Then as* $n \to \infty$*,*

$$\mathbb{P}(X_n = y) \to 0 \tag{3.16}$$

as $n \to \infty$*, and*

$$\frac{V_n(y)}{n} \to 0, \tag{3.17}$$

almost surely.

Formally this is in fact the same conclusion as in the positive recurrent case, but with $\pi_y = 1/\mathbb{E}_y(T_y^+) = 0$.

4 Martingales

We come to one of the most fundamental concepts in probability theory, which is that of martingales. They play a role which is analogous to that of **constants of motion** in physics (i.e., energy, momentum, etc.): as we will see, they identify certain quantities which are conserved throughout the evolution of a stochastic process. Often, finding a martingale associated to a stochastic process is the key which unlocks the understanding of its behaviour. In order to appreciate what a powerful tool martingales are on some concrete examples, we must however first develop a fair amount of theory.

4.1 Definitions

Let $(X_n, n \geq 0)$ be a stochastic process taking values in $\mathbb{R}$ (in this chapter we depart somewhat from the study of Markov chains, and we will not restrict ourselves to a countable state space; in fact, for the notion of martingale, it will be important that X_n takes values in a linear space such as $\mathbb{R}$ in order to make sense of its expectation. It would also be possible to discuss martingales with values in $\mathbb{R}^d$, but we will not do so for simplicity).

For $n \geq 0$, let $\mathcal{F}_n$ denote the collection of events depending only on $(X_0, \ldots, X_n)$: formally, an event A is in $\mathcal{F}_n$ if and only if there is a (measurable) function $\varphi : \mathbb{R}^{n+1} \to \{0, 1\}$ such that $1_A = \varphi(X_0, \ldots, X_n)$. As an example, with these notations, the random variable T is a stopping time if and only if the event $\{T \leq n\}$ is in the collection $\mathcal{F}_n$.

Definition 4.1. *We call $\mathcal{F} = (\mathcal{F}_n)_{n\geq 0}$ the **filtration** generated by the stochastic process X.*

We also write $\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n)$ and say that $\mathcal{F}_n$ is generated by $(X_0, \ldots, X_n)$. In some examples the randomness only kicks in at time $n = 1$. In those cases it will be convenient to let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$; for $n = 0$ the collection on the right is then empty. In this uninteresting case we take $\mathcal{F}_0$ consists only of trivial events, namely $\emptyset$ and the full probability space Ω. (This is a technicality which can safely be ignored for most of the course).

When we observe X progressively from time 0 to time n, we know exactly the status of each event in $\mathcal{F}_n$, i.e., given an event in $\mathcal{F}_n$ we can tell with complete certainty whether it occurs or not. In this sense, $\mathcal{F}_n$ corresponds to and should be thought of as the **information** gained from observing $(X_0, \ldots, X_n)$.

Definition 4.2. *A stochastic process $(Y_n)_{n\geq 0}$ is called **adapted** to the filtration $\mathcal{F}$ if for every $n \geq 0$, the random variable Y_n depends only on $(X_0, \ldots, X_n)$.*

Once again, formally, this means there exists a (measurable) function $\varphi_n : \mathbb{R}^{n+1} \to \mathbb{R}$ such that $Y_n = \varphi_n(X_0, \ldots, X_n)$. Another way to say it in a measure-theoretic way (so we will not do that in the rest of the course) is that Y_n is $\mathcal{F}_n$-measurable.

Example 4.3. If $X = (X_n)_{n\geq 0}$ is a stochastic process, then $Y_n = X_n^2 \sin(X_n)$ is an adapted process. $Z_n = X_n - X_{n-2}$ is also adapted. However, $W_n = X_n + X_{n+1}$ is not: to know the value of W_n would require more information than is contained in $\mathcal{F}_n$.

If X is a stochastic process with filtration $\mathcal{F}$, and Y is another stochastic process adapted to $\mathcal{F}$, this implies that the filtration $\mathcal{G}$ generated by Y satisfies

$$\mathcal{G}_n \subset \mathcal{F}_n; \quad n \geq 0.$$

Indeed any event of $\mathcal{G}_n$ can be written as a function of $(Y_0, \ldots, Y_n)$ and so as a function of $(X_0, \ldots, X_n)$.

Example 4.4. Let $X_n = \pm 1$ be a sequence of i.i.d. fair coin tosses. Let $\mathcal{F} = (\mathcal{F}_n)_{n\geq 0}$ denote the filtration generated by X. Let also $S_n = \sum_{i=1}^n X_i$ if $n \geq 1$ (and $S_0 = 0$). In other words, $(S_n)_{n\geq 0}$ is the unbiased random walk on $\mathbb{Z}$. Let $\mathcal{G} = (\mathcal{G}_n)_{n\geq 0}$ be the filtration generated by S. Then we claim that $\mathcal{F}$ and $\mathcal{G}$ are identical.

Indeed, $(S_n)_{n\geq 0}$ is clearly adapted to $\mathcal{F}$: given the increments up to time n, we obtain the walk by summing these increments. This shows that $\mathcal{G}_n \subset \mathcal{F}_n$ for every $n \geq 0$.

Conversely, $(X_n)_{n\geq 0}$ is also adapted to $\mathcal{G}$. Indeed, we obtain the increment X_n as $S_n - S_{n-1}$ which depends on $\mathcal{F}_n$. This shows that $\mathcal{F}_n \subset \mathcal{G}_n$. Thus $\mathcal{F}_n = \mathcal{G}_n$ for every $n \geq 0$.

We now formulate the crucial definition for this chapter.

Definition 4.5. *Let $X = (X_n)_{n\geq 0}$ be a stochastic process and $\mathcal{F} = (\mathcal{F}_n)_{n\geq 0}$ be its filtration. The stochastic process $(M_n)_{n\geq 0}$ is called a **martingale** if:*

(i) M is adapted to $\mathcal{F}$.

(ii) M is integrable: i.e., for every $n \geq 0$, $\mathbb{E}(|M_n|) < \infty$.

(iii) Finally, M satisfies the martingale condition: for every $n \geq 0$,

$$\mathbb{E}(M_{n+1}|\mathcal{F}_n) = M_n. \tag{4.1}$$

Let us make some comments on the defining condition (4.1) of this definition. In words, this property says that "given all the information up to today, the best prediction we can make about tomorrow's value, is equal to today's value". As an example which should help fix ideas, if M_n denotes the temperature as a function of time (where n is measured in days) then M_n can roughly be thought of as a martingale: in the absence of any other information, the best guess we can make about tomorrow's temperature is roughly the same as what it is today. The temperature might go up or down, but its conditional expectation will be the same as today's temperature. This assumes that the expected increase in temperature balances the expected decrease – an assumption which, at best, ignores seasonal effects, and, at worst, is blatantly untrue.

Formally, if X takes values in a countable state space S, then the property (4.1) can be written more concretely as follows: for every $n \geq 0$, for every $x_0, \ldots, x_n \in S$,

$$\mathbb{E}(M_{n+1}|X_0 = x_0, \ldots, X_n = x_n) = M_n.$$

Another way to rewrite the same thing is to write:

$$\mathbb{E}(M_{n+1} - M_n | X_0 = x_0, \ldots, X_n = x_n) = 0.$$

This is equivalent to the above, because M is adapted, so that, given $X_0 = x_0, \ldots, X_n = x_n$, M_n is known (i.e., it is a constant – so its conditional expectation is simply M_n, the same way that the expectation of the random variable equal to a constant c is c itself).

Remark 4.6. In a measure-theoretical framework, yet another way to rewrite (4.1) would be to write $\mathbb{E}(M_{n+1}1_A) = \mathbb{E}(M_n 1_A)$ for every $n \geq 0$ and every $\mathcal{F}_n-$measurable event A. We will not write things in this manner since we do not assume measure theory.

4.2 Properties of conditional expectation and examples

The defining property (4.1) of a martingale may feel somewhat abstract at this stage, so as usual it is a good idea to turn to examples. These examples are fundamental and will keep coming up in later parts of the theory. To treat any examples it is important to first understand the following fundamental rules for manipulating conditional expectations:

- Conditional expectation is linear: $\mathbb{E}(aX + bY|\mathcal{F}) = a\mathbb{E}(X|\mathcal{F}) + b\mathbb{E}(Y|\mathcal{F})$
- If Y is determined by $\mathcal{F}$ then $\mathbb{E}(Y|\mathcal{F}) = Y$: since Y is known given $\mathcal{F}$, its conditional expectation is simply itself. More generally Y should be regarded as a constant, thus $\mathbb{E}(XY|\mathcal{F}) = Y\mathbb{E}(X|\mathcal{F})$.
- By contrast if Y is independent of $\mathcal{F}$ then $\mathbb{E}(Y|\mathcal{F}) = \mathbb{E}(Y)$: indeed in that case, conditioning on $\mathcal{F}$ has given no relevant information on Y, so its conditional expectation is simply its *un*conditional expectation.

Formally, these properties hold whenever the random variables are nonnegative (whether the conditional expectations are finite or infinite), and whenever the random variables are integrable (i.e., $\mathbb{E}(|X|) < \infty$). It is important to note that $\mathbb{E}(X|\mathcal{F})$ is a random variable; **intuitively**, $\mathbb{E}(X|\mathcal{F})$ describes how our estimate of X changes when we get to observe the information contained in $\mathcal{F}$. Hence if we are given $\mathcal{F}$, the quantity $\mathbb{E}(X|\mathcal{F})$ is completely fixed: in other words, it is a random variable, but depends only on the randomness used to generate $\mathcal{F}$.

It is desirable that, at the end of this course, you will be so familiar with conditional expectation that you will be able to use these properties without thinking about it. Practice with lots of examples is essential for this. Let us see how this works in practice.

Lecture 14: Friday 18.11.22 (zoom)

Example 4.7. Let $(X_i)_{i\geq 1}$ be a sequence of i.i.d. random variables (with values in $\mathbb{R}$). Suppose $\mathbb{E}(|X_i|) < \infty$ and let $m = \mathbb{E}(X_i)$. Let $S_n = \sum_{i=1}^n X_i$ (we will always use the convention in these sums that if $n = 0$, $\sum_{i=1}^0 X_i = 0$, so $S_0 = 0$). Let

$$M_n = S_n - mn.$$

Then M is a martingale in the filtration $(\mathcal{F}_n)_{n\geq 0}$ generated by $(X_i)_{i\geq 1}$. That is, for $n \geq 1$, $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ consists of the events depending only on $(X_1, \ldots, X_n)$; as per our convention, for $n = 0$, $\mathcal{F}_0$ contains only the two trivial events $\emptyset$ and the full probability space Ω. (Once again this technicality can safely be ignored.)

Let us prove this. First, it is clear that $(M_n)_{n\geq 0}$ is adapted: given $(X_1, \ldots, X_n)$ we can say with complete certainty what is the value of $S_n = \sum_{i=1}^n X_i$, and thus the value of $M_n = S_n - mn$.

Second, $\mathbb{E}(|M_n|) < \infty$ for any $n \geq 0$, because M_n is the sum of n random variables which are by assumption integrable.

Finally, let us prove (4.1). We have:

$$\begin{aligned}
\mathbb{E}(M_{n+1}\big|\mathcal{F}_n) &= \mathbb{E}(S_{n+1} - m(n+1)\big|\mathcal{F}_n) \\
&= \mathbb{E}(S_n + X_{n+1}|\mathcal{F}_n) - m(n+1) && \text{(by linearity)} \\
&= S_n - m(n+1) + \mathbb{E}(X_{n+1}\big|\mathcal{F}_n) && \text{(by linearity and since } S_n \text{ is determined by } \mathcal{F}_n) \\
&= S_n - m(n+1) + \mathbb{E}(X_n) && \text{(by independence of } X_{n+1} \text{ with respect to } \mathcal{F}_n) \\
&= S_n - mn = M_n,
\end{aligned}$$

as desired.

In the same vein, let us find another martingale associated with S_n. This will be denoted by Q_n which stands for "quadratic".

Example 4.8. In the same setup as above, let us suppose $m = \mathbb{E}(X_i) = 0$ and that $\sigma^2 = \mathbb{E}(X_i^2) < \infty$ (since $\mathbb{E}(X_i) = 0$, σ^2 is also the variance of each X_i). For $n \geq 0$, let

$$Q_n = S_n^2 - n\sigma^2.$$

Then $Q = (Q_n)_{n\geq 0}$ is a martingale in the filtration $\mathcal{F}$ generated by $(X_1, X_2, \ldots)$ in the same sense as before (that is, $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ for $n \geq 0$, with the same convention as before when $n = 0$ – from now on we will not mention this again).

Let us check this. First of all, it is clear that Q is adapted: indeed, given $\mathcal{F}_n$ we can determine S_n with certainty and hence also S_n^2, and therefore also Q_n. Second, for every $n \geq 0$,

$$\mathbb{E}(|Q_n|) < \infty$$

by the triangle inequality and the fact that we have assumed that $\mathbb{E}(X_i^2) < \infty$. It remains to prove (4.1). Then

$$\begin{aligned}
\mathbb{E}\big(Q_{n+1}\big|\mathcal{F}_n\big) &= \mathbb{E}\big(S_{n+1}^2 - (n+1)\sigma^2\big|\mathcal{F}_n\big) \\
&= \mathbb{E}\big((S_n + X_{n+1})^2|\mathcal{F}_n\big) - (n+1)\sigma^2 \\
&= \mathbb{E}\big(S_n^2 + 2S_nX_{n+1} + X_{n+1}^2\big|\mathcal{F}_n\big) - (n+1)\sigma^2 \\
&= S_n^2 + 2S_n\mathbb{E}\big(X_{n+1}\big|\mathcal{F}_n\big) + \mathbb{E}\big(X_{n+1}^2\big|\mathcal{F}_n\big) - (n+1)\sigma^2.
\end{aligned}$$

So far we have only used linearity and the fact that S_n is known (given $\mathcal{F}_n$) and hence is like a constant – it can be taken out of the conditional expectation. Now recall that

X_{n+1} is independent of $\mathcal{F}_n$ (i.e., of $(X_1, \ldots, X_n)$) and hence $\mathbb{E}(X_{n+1}|\mathcal{F}_n) = \mathbb{E}(X_n) = 0$ (the conditional expectation equals the unconditional expectation). Likewise, $\mathbb{E}\big(X_{n+1}^2\big|\mathcal{F}_n\big) = \mathbb{E}(X_{n+1}^2) = \sigma^2$. We deduce:

$$\begin{aligned}\mathbb{E}\big(Q_{n+1}\big|\mathcal{F}_n\big) &= S_n^2 + 2S_n \times 0 + \sigma^2 - (n+1)\sigma^2 \\ &= S_n^2 - n\sigma^2 \\ &= Q_n.\end{aligned}$$

Thus Q is a martingale, as desired.

Our next example concerns branching processes, already discussed in Definition 2.15 and in Theorem 2.17. For this we will need our last fundamental property of conditional expectation, which is the so-called **tower property**.

Theorem 4.9. *If X is a random variable and $\mathcal{F}$ a collection of events, then $\mathbb{E}(X) = \mathbb{E}[\mathbb{E}(X|\mathcal{F})]$.*

(Here we use $\mathcal{F}$ to denote a single collection of events – like $\mathcal{F}_n$ for instance – rather than a whole filtration $\mathcal{F} = (\mathcal{F}_n)_{n\geq 0}$).To explain the meaning of Theorem 4.9, recall that $\mathbb{E}(X|\mathcal{F})$ is a random variable which describes how our estimate of X changes when we get to observe the information contained in $\mathcal{F}$. This proposition says that the average of our updated guess will be equal to $\mathbb{E}(X)$ (when we average over the randomness in $\mathcal{F}$). It is for instance impossible to systematically underestimate X when we get to observe the information contained in $\mathcal{F}$.

Let us now return to the example of branching processes. Recall that a **branching process** $(Z_n)_{n\geq 0}$ is determined by an offspring distribution $(p_k)_{k\geq 0}$ on $\mathbb{N} = \{0, 1, \ldots\}$ where p_k is the probability for any given individual to have k offsprings in the next generation, and offspring numbers are independent random variables. Thus, $Z_0 = 1$ (initially there is one individual in this population) and given $(Z_1, \ldots, Z_n)$, we obtain Z_{n+1} as

$$Z_{n+1} = \sum_{i=1}^{Z_n} \xi_i,$$

where $(\xi_i)_{i\geq 1}$ are independent of $(Z_1, \ldots, Z_n)$ and are i.i.d. with common law $(p_k)_{k\geq 0}$. The following theorem gives a fundamental martingale attached to the branching process.

Theorem 4.10. *Let $(Z_n)_{n\geq 0}$ be the above branching process and set $m = \sum_{k=0}^{\infty} kp_k$ denote the mean number of offsprings per individual. For $n \geq 0$, set*

$$M_n = \frac{Z_n}{m^n}.$$

Then $M = (M_n)_{n\geq 0}$ is a martingale in the filtration $\mathcal{F} = (\mathcal{F}_n)_{n\geq 0}$ generated by $(Z_0, Z_1, Z_2, \ldots)$.

Lecture 15: Thursday 24.11.22

Proof. We start with the obvious observation that M is adapted to $\mathcal{F}$. Indeed, if we are given $\mathcal{F}_n$ we can determine Z_n with certainty and hence also $M_n = Z_n/m^n$. So M is adapted.

Normally we would then turn to a proof that M_n is integrable, without which the conditional expectation might not necessarily be well defined. However here, we will first establish (4.1) (the definition of conditional expectation and the manipulations we will do are all justified by the nonnegativity of the random variables involved) and deduce that M_n is integrable using the tower property of conditional expectations. Let $n \geq 0$. Then

$$\begin{aligned}\mathbb{E}(M_{n+1}|\mathcal{F}_n) &= \mathbb{E}(\frac{Z_{n+1}}{m^{n+1}}|\mathcal{F}_n) \\ &= \frac{1}{m^{n+1}}\mathbb{E}(Z_{n+1}|\mathcal{F}_n) \text{ (by linearity)} \\ &= \frac{1}{m^{n+1}}\mathbb{E}(\sum_{i=1}^{Z_n}\xi_i|\mathcal{F}_n).\end{aligned}$$

Now we remember our rules for manipulating conditional expectations. When we condition on $\mathcal{F}_n$ (i.e., we pretend we know all the information in $\mathcal{F}_n$), Z_n is no longer a random variable but actually a fixed (constant) quantity. We can thus use the linearity of conditional expectation to get

$$\begin{aligned}\mathbb{E}(M_{n+1}|\mathcal{F}_n) &= \frac{1}{m^{n+1}}\sum_{i=1}^{Z_n}\mathbb{E}(\xi_i|\mathcal{F}_n) \\ &= \frac{1}{m^{n+1}}\sum_{i=1}^{Z_n}\mathbb{E}(\xi_i) \text{ (since } \xi_i \text{ is independent from } \mathcal{F}_n) \\ &= \frac{1}{m^{n+1}}mZ_n \text{ (since each } \xi_i \text{ has mean } m) \\ &= \frac{Z_n}{m^n} \\ &= M_n.\end{aligned}$$

Therefore (4.1) holds in this case. Let us now use this to deduce that M_n is in fact integrable: taking expectations, we have

$$\mathbb{E}(\mathbb{E}(M_{n+1}|\mathcal{F}_n)) = \mathbb{E}(M_n)$$

so that by the tower property (i.e., Theorem 4.9) we have

$$\mathbb{E}(M_{n+1}) = \mathbb{E}(M_n).$$

By induction we have $\mathbb{E}(M_n) = \mathbb{E}(M_0) = 1$ so it is clear that $\mathbb{E}(|M_n|) < \infty$ (as $M_n \geq 0$). This concludes the proof of Theorem 4.10. □

Since we showed in the proof that $\mathbb{E}(M_n) = 1$, we obtain as an important corollary the following result:

Corollary 4.11. *In setup of Theorem 4.10, we have for every $n \geq 0$, $\mathbb{E}(Z_n) = m^n$.*

This should remind you of Theorem 2.18. Indeed when $m < 1$ we have $\mathbb{E}(Z_n) = m^n \to 0$ (as $n \to \infty$) so it is natural to expect that the process will become extinct eventually. On the other hand if $m > 1$ then $\mathbb{E}(Z_n) = m^n \to \infty$ so it is natural to expect that the branching process can survive at least with positive probability. (Of course this heuristic does not immediately suggest what should be happening in the borderline case $m = 1$.) We will soon see martingale proofs of this dichotomy.

4.3 Fair games and martingale transform

In the course of the proof of Theorem 4.10 we observed that the expectation remains constant. This is in fact true of every martingale – the first hint that they are in a certain sense "constants of motion":

Proposition 4.12. *Let $M = (M_n)_{n\geq 0}$ be a martingale in some filtration $\mathcal{F} = (\mathcal{F}_n)_{n\geq 0}$. Then $\mathbb{E}(M_n) = \mathbb{E}(M_0)$ for every $n \geq 0$. Furthermore, for all $n \geq k \geq 0$,*

$$\mathbb{E}(M_n|\mathcal{F}_k) = M_k. \tag{4.2}$$

Proof. This is easy to establish with the tower property: since $\mathbb{E}(M_n|\mathcal{F}_{n-1}) = M_{n-1}$, we take expectation and by the tower property deduce that $\mathbb{E}(M_n) = \mathbb{E}(M_0)$. The proof of (4.2) is similar and uses a generalisation of the tower property of Theorem 4.9 for conditional expectations of conditional expectations. □

We will now discuss a very important interpretation of martingales which are related to the notion of **fair game**. To discuss this (and to prepare for later developments) it is useful to generalise slightly the notion of martingale.

Definition 4.13. *Let $(X_n, n \geq 0)$ be a stochastic process generating a filtration $\mathcal{F} = (\mathcal{F}_n)_{n\geq 0}$. Let $M = (M_n)_{n\geq 0}$ be a stochastic process with values in $\mathbb{R}$. We say that M is a **submartingale** (resp. **supemartingale**) if:*

(i) M is adapted to $\mathcal{F}$

(ii) For every $n \geq 0$, M_n is integrable, i.e., $\mathbb{E}(|M_n|) < \infty$.

(iii) Finally,

$$\mathbb{E}(M_{n+1}|\mathcal{F}_n) \geq M_n, \tag{4.3}$$

(resp. $\mathbb{E}(M_{n+1}|\mathcal{F}_n) \leq M_n$).

The definition therefore only differs from that of a martingale in the last point. The difference is that in a submartingale, M_n underestimates the true conditional expectation. The terminology comes from harmonic analysis: as we will soon see, martingales are connected to harmonic functions, whereas submartingales are related to subharmonic functions (and supermartingales to superharmonic functions).

The notion of martingale is intrinsically related to that of fair game. Let us imagine that a player takes part in a game involving randomness and $(M_n)_{n\geq 0}$ is the fortune of the player at time n (so M_0 is the initial fortune of the player). We will introduce, for $n \geq 1$,

$$\Delta_n = M_n - M_{n-1},$$

the increment (change) of the fortune after the nth game. Then using Δ instead of M, we have that M is a martingale if and only if:

1. $(\Delta_n)_{n\geq 1}$ is adapted,
2. Δ_n is integrable for every $n \geq 1$,
3. and for all $n \geq 1$,
$$\mathbb{E}(\Delta_n|\mathcal{F}_{n-1}) = 0 \tag{4.4}$$

Naturally, the right hand side should be ≥ 0 for a submartingale, and ≤ 0 for a supermartingale. Equation (4.4) says that the game in which the player takes part is *fair* at each step, the expected net change in the fortune is zero. On the other hand in a supermartingale, the game is biased against the player (like playing against a casino) and in a submartingale it is on the contrary biased in the player's favour.

Example 4.14. Consider the biased walk $(X_n)_{n\geq 0}$ on the integers with $P(i, i+1) = p$; $P(i, i-1) = q$ and $p + q = 1$. If $p \geq q$ then X is a submartingale, while if $p \leq q$ then X is a supermartingale.

This interpretation leads us to the following notion of **martingale transform**. This is the martingale that one gets when a player in a fair game *varies the stakes*. To explain what this means, suppose a player bets 1€ at each time step in a fair game. Her resulting fortune M_n (assuming she sticks to this betting strategy) at the end of the nth game will, as already mentioned, form a martingale, since we assume the game to be fair. But it is also possible for her to change betting strategy and to play with different stakes at each game: for instance she might choose to bet 2€ on some games, or 100€ on some others, if she is feeling lucky. If she bets 100€ instead of 1€ her fortune will increase by $100\Delta_n$ instead of by Δ_n. The fortune in this scenario (with the new betting strategy) is what we call a *martingale transform* of the original martingale describing the evolution of the fortune with some fixed (old) betting strategy. (A formal definition will follow).

Let us point out that after changing strategy, the game remains a fair one: no matter what, she cannot bias the game in her favour or against her simply by deciding how much she is going to bet! This intuitive fact explains the fundamental fact that a martingale transform remains a martingale, as we will now see formally.

Definition 4.15. *Let $(\mathcal{F}_n)_{n\geq 0}$ be a filtration generated by a stochastic process. We say that the stochastic process $W = (W_n)_{n\geq 1}$ is **predictable** if for every $n \geq 1$, W_n depends only $\mathcal{F}_{n-1}$.*

You can think of W_n as representing the stakes for the nth game: in general, the player is allowed to decide how much she is willing to bet at the nth game depending on the results of the previous $(n-1)$ games, but not (unfortunately) on what will happen in the nth game itself.

Definition 4.16. *Let $(M_n)_{n\geq 0}$ be a (sub, super-)martingale with respect to a filtration $\mathcal{F} = (\mathcal{F}_n)_{n\geq 0}$, and let $W = (W_n)_{n\geq 1}$ be a predictable process. The **martingale transform** of M by W is the process $M' = (M'_n)_{n\geq 0}$ defined by*

$$M'_n = M_0 + W_1\Delta_1 + \ldots + W_n\Delta_n.$$

(Thus $M'_0 = M_0$.)

It is common to denote the martingale transform M' as $M' = W \cdot M$. Note that each increment of M has been amplified by W. If you study stochastic analysis later on, you will encounter stochastic integrals of the form

$$M'_t = M_0 + \int_0^t W_s \, \mathrm{d}M_s,$$

where M is a (continuous) martingale and W is a process which is predictable in a suitable sense. In fact, martingale transforms are nothing else but discretised versions of these stochastic integrals, in the same way that integrals can be approximated (discretised) by Riemann sums. We now state the theorem alluded to above, which states the martingale transform of a martingale remains a martingale. Let us call a process $(X_n)_{n\geq 0}$ **uniformly bounded** if there exists $C < \infty$ (nonrandom!) such that

$$|X_n| \leq C, \qquad \text{for all } n \geq 0.$$

This inequality is required to almost surely, that is, on an event of probability one – but pay attention to the fact that your upper bound on X_n is not allowed to be random. For instance, if X is an exponential random variable, and $X_n = X$, then $(X_n)_{n\geq 0}$ is *not* bounded (because there is no absolute bound on the exponential distribution).

Theorem 4.17. *Let M be a martingale and let W be a bounded, predictable process. Then $W \cdot M$ is a martingale. If instead M is a submartingale (resp. supermartingale), and we also assume $W_n \geq 0$ for all $n \geq 0$, then $W \cdot M$ remains a submartingale (resp. supermartingale).*

Proof. Suppose first that M is a martingale. Let us consider the increments Δ'_n of the martingale transform $M' = W \cdot M$: we have for $n \geq 1$,

$$\Delta'_n = M'_n - M'_{n-1} = W_n\Delta_n.$$

To prove that M' is a martingale, we need to check three properties. First, it is clear that Δ'_n is adapted: indeed, Δ_n is adapted and W_n is predictable (hence clearly adapted). So the product $W_n\Delta_n$ is adapted. Second, since W_n is bounded,

$$\mathbb{E}(|\Delta'_n|) = \mathbb{E}(|W_n\Delta_n|) \leq C\mathbb{E}(|\Delta_n|) < \infty,$$

so Δ'_n is integrable. Finally, for $n \geq 1$,

$$\begin{aligned}\mathbb{E}(\Delta'_n|\mathcal{F}_{n-1}) &= \mathbb{E}\Big(W_n\Delta_n\Big|\mathcal{F}_{n-1}\Big)\\ &= W_n\mathbb{E}\Big(\Delta_n\Big|\mathcal{F}_{n-1}\Big) \quad \text{because } W \text{ is predictable, so } W_n \text{ is known given } \mathcal{F}_{n-1}\\ &= 0\end{aligned}$$

because M is a martingale. This verifies (4.4), hence M' is a martingale. □

More generally the above result holds if we only assume that W_n is bounded but not uniformly: i.e., $|W_n| \leq C_n$ for some nonrandom $C_n < \infty$. The following example is a well known betting strategy known as St Petersburg's.

Example 4.18. Let $(M_n)_{n\geq 0}$ denote (unbiased) random walk on the integers. Let τ be the first time there is a positive increment: $\tau = \inf\{n \geq 1 : \Delta_n \geq 0\}$.

Let $W_1 = 1$ and define inductively $W_n = 2W_{n-1}$ so long as $n \leq \tau$. Let $W_n = 0$ if $n > \tau$. Then $W \cdot M$ is a martingale. Indeed, W_n is predictable: the decision to switch to zero can be made based solely on the information in the games before time n. Furthermore W_n is bounded (although not uniformly in n) since $|W_n| \leq 2^n$ which is nonrandom. Thus Theorem 4.17 applies and M' is a martingale.

St. Petersburg's strategy is well known since at the stopping time τ, the fortune of the player is guaranteed to have increased: indeed, say $\tau = n$, then

$$\begin{aligned}M'_\tau &= M_0 - (1 + 2 + \ldots + 2^{n-1}) + 2^n\\ &= M_0 - (2^n - 1) + 2^n\\ &= M_0 + 1.\end{aligned}$$

Since the game is fair, this may seem paradoxical! As we are about to see, this is can only happen if M' has very peculiar properties which make this strategy not feasible in practice: indeed, in the next section we will see that for "reasonably behaved" martingales, it is impossible to make money from fair games.

4.4 Optional stopping theorem

We come to a couple of fundamental results about martingales. The first one will be the optional stopping theorem (discussed in this section) and the second will be the martingale convergence theorem (discussed in the next). After these two results we will switch to examples and see how martingale theory can help us analyse stochastic processes.

Our first observation is that quitting a fair game at a stopping time, no matter how well chosen, cannot tilt the game into an unfair game. Recall that if $\mathcal{F}$ is a filtration generated by some stochastic process, then the random variable τ with values in $\{0, 1, \ldots\} \cup \{\infty\}$ is called a stopping time if for every $n \geq 0$, $\{\tau \leq n\} \in \mathcal{F}_n$. Also, given $a, b \in \mathbb{R}$, let $a \wedge b = \min(a, b)$.

Proposition 4.19. *Let M be a (sub)martingale and let τ be a stopping time. For $n \geq 0$, let*

$$M'_n = M_{n\wedge\tau} = \begin{cases} M_n & \text{if } n \leq \tau \\ M_\tau & \text{if } n \geq \tau. \end{cases}$$

Then $(M'_n)_{n\geq\tau}$ is a (sub)martingale.

The process M' is known as the martingale *stopped* at time τ, because it ceases to evolve after τ. For instance, the martingale in St Petersburg's strategy (Example 4.18) is a stopped martingale.

Proof. This is actually a simple consequence of the martingale transform Theorem 4.17. Indeed, we take

$$W_n = 1_{\{\tau\geq n\}}$$

which is bounded and nonnegative. It is furthermore predictable (despite appearances): indeed,

$$\{\tau \geq n\} = \{\tau \leq n-1\}^c,$$

and written in this manner, the event $\{\tau \geq n\}$ depends indeed only on $\mathcal{F}_{n-1}$ (since τ is a stopping time). Furthermore, we claim that $W \cdot M$ is simply the stopped martingale M'. Indeed,

$$(W \cdot M)_n = M_0 + 1_{\{\tau\geq 1\}}(M_1 - M_0) + \ldots + 1_{\{\tau\geq n\}}(M_n - M_{n-1})$$

which is a telescopic sum, summing to M_n if $\tau \geq n$ or M_τ if $\tau \leq n$. □

The consequence of this, the *optional stopping theorem*, is one of the cornerstones of probability. It says that the expectation of a martingale is conserved even at a stopping time (intuitively, one cannot make money from fair games). As mentioned concerning the St Petersburg stratgy, this theorem needs assumptions.

Theorem 4.20. *Let M be a (sub)martingale and τ a stopping time. Suppose at least one of the following two conditions hold:*

(i) τ is bounded, i.e., there exists $n \geq 0$ (nonrandom) such that $\tau \leq n$ almost surely.

(ii) $\tau < \infty$ almost surely and $M_{n\wedge\tau}$ is uniformly bounded (i.e., there exists $C < \infty$ nonrandom such that $|M_n| \leq C$ for all $n \leq \tau$).

Then $\mathbb{E}(M_\tau) = \mathbb{E}(M_0)$ (and if $\mathbb{E}(M_\tau) \geq \mathbb{E}(M_0)$ if M is a submartingale).

Before we give a proof of this theorem, it is useful to compare the conclusion of this theorem with the St Petersburg example (Example 4.18). In that case we have a stopping time τ and a martingale such that $M_\tau = M_0 + 1$. This seems to violate the conclusion of the Optional Stopping Theorem since in such a situation it is impossible to have $\mathbb{E}(M_\tau) = \mathbb{E}(M_0)$. But there is no contradiction: simply, on the one hand, τ is not bounded (even if it is a geometric random variable and so has a tail that decays exponentially fast), and on the other

hand, the stopped martingale is certainly far from bounded – it has the potential to become very negative before turning positive. This feature makes impractical in real life: casinos in particular prevent players from borrowing too much money – this forces players' fortunes to stay bounded and hence makes the conclusion of the optional stopping theorem valid.

Proof. Suppose first (i), i.e., there exists a nonrandom $n \geq 0$ such that $\tau \leq n$ with probability one. Let us consider for simplicity the martingale case. Let $M'_k = M_{\tau \wedge k}$ for $k \geq 0$. As proved in Proposition 4.19, M' is a (sub)martingale. Consequently, by Proposition 4.12,

$$\mathbb{E}(M'_k) = \mathbb{E}(M'_0)$$

for any $k \geq 0$. In particular, take $k = n$. Then $M'_k = M_{n \wedge \tau} = M_\tau$ since $\tau \leq n$ by assumption. Furthermore $M'_0 = M_0$. Hence we learn

$$\mathbb{E}(M_\tau) = \mathbb{E}(M_0),$$

as desired.

Now suppose (ii). We apply the conclusion of (i) to the random time $\tau_n = \tau \wedge n$: note that this is a stopping time (check it!) which is furthermore bounded by n. Hence by (i),

$$\mathbb{E}(M_{\tau \wedge n}) = \mathbb{E}(M_0).$$

At this point we want to let $n \to \infty$ and claim that $\mathbb{E}(M_{\tau \wedge n}) \to \mathbb{E}(M_\tau)$. To see this, you may either use the dominated convergence theorem (if you know measure theory) using the boundedness assumption and the fact that $\tau < \infty$ almost surely, or you may simply observe that

$$\begin{aligned} |\mathbb{E}(M_{\tau \wedge n}) - \mathbb{E}(M_\tau)| &\leq \mathbb{E}(|M_\tau - M_{\tau \wedge n}|) \\ &\leq 2C\mathbb{P}(\tau > n) \end{aligned}$$

since the only contribution to this expectation comes from the event $\{\tau > n\}$ (otherwise, the random variables M_τ and $M_{\tau \wedge n}$ are equal). The right hand side tends to 0 as $n \to \infty$, because $\tau < \infty$ with probability one. This concludes the proof. □

Lecture 16: Friday 25.11.22

We now begin some examples of applications of the Optional Stopping theorem.

Example 4.21. Let $(X_n, n \geq 0)$ denote an (unbiased) random walk on the integers starting from 0. Fix $a, b \in \mathbb{N} = \{0, 1, \ldots\}$. Then

$$\mathbb{P}_0(T_{-a} < T_b) = \frac{b}{a+b}, \tag{4.5}$$

where, as usual, T_x denote the hitting time of x. In fact, we have already seen this by solving an associated Dirichlet problem (see for instance (2.10), from which (2.10) can be deduced without too much difficulties (in fact, this was one of the exercises on an example sheet). Let

us give a new proof of (4.5) based on martingales and in particular the Optional Stopping theorem. Consider $T = T_{-a} \wedge T_b$. This is the first hitting time of the set $A = \{-a, b\}$, and so is a stopping time. Furthermore, X is a martingale (as already proved in Example 4.7) and is uniformly bounded before T (by $C = \max(a, b)$ which is indeed non-random). We may therefore apply the optional stopping theorem. We deduce:

$$\mathbb{E}(X_T) = \mathbb{E}(X_0).$$

The right hand side is simply 0. To evaluate the left hand side, we note that X_T is a random variable that only takes two values, namely $-a$ (with the desired probability $p = \mathbb{P}_0(T_{-a} < T_b)$) or b (with the complementary probability $1 - p$). We deduce

$$\mathbb{E}(X_T) = p(-a) + (1 - p)b.$$

Since $\mathbb{E}(X_T) = 0$, this implies

$$b = p(a + b)$$

and hence

$$p = \frac{b}{a + b},$$

as desired in (4.5).

Of course, this can be translated: for instance, for $0 \leq k \leq n$, so equivalently to (4.5) we may write $\mathbb{P}_k(T_0 < T_n) = (n - k)/n = 1 - k/n$.

Example 4.22. In the same setup as above, let $T = T_{-a} \wedge T_b$. Then we claim

$$\mathbb{E}_0(T) = ab. \tag{4.6}$$

We will also derive this from the optional stopping theorem. In order to do this, it is useful to know a martingale which involves time. We already found such a martingale in Example 4.8, namely

$$M_n = X_n^2 - n\sigma^2,$$

where σ^2 is the variance of the increments Δ_n of the random walk. Since $\Delta_n = \pm 1$, $\mathbb{E}((\Delta_n)^2) = 1$, and since $\mathbb{E}(\Delta_n) = 0$, we see that $\sigma^2 = \mathrm{Var}(\Delta_n) = 1$, hence

$$M_n = X_n^2 - n.$$

We would like to apply the optional stopping theorem to M and the stopping time T. Unfortunately, this is not uniformly bounded (because T is not uniformly bounded: although T is finite almost surely by the recurrence of random walk in dimension 1, T cannot be bounded by a nonrandom constant). We thus apply the optional stopping theorem first to $T \wedge n$, which is a bounded stopping time: we get

$$\mathbb{E}(M_{T \wedge n}) = \mathbb{E}(M_0) = 0.$$

Hence

$$\mathbb{E}(X_{T\wedge n}^2) = \mathbb{E}(T \wedge n).$$

We now let $n \to \infty$ and deduce (using a bit of measure theory: the dominated convergence on the left hand side, since X is uniformly bounded before time T, and the monotone convergence theorem on the right hand side),

$$\mathbb{E}(X_T^2) = \mathbb{E}(T).$$

But we already know that X_T is a random variable that only takes two values ($-a$ or b), and we know with which probability thanks to Example 4.21. Hence we can compute $\mathbb{E}(X_T^2)$: namely, if $p = \mathbb{P}_0(T_{-a} < T_b) = b/(a+b)$,

$$\begin{aligned}\mathbb{E}(X_T^2) &= pa^2 + (1-p)b^2 \\ &= \frac{b}{a+b}a^2 + \frac{a}{a+b}b^2 \\ &= \frac{ba^2+ab^2}{a+b} = \frac{ab(a+b)}{a+b} \\ &= ab.\end{aligned}$$

This completes the proof of (4.6).

Another comment (which may help to remember (4.6)) is that if we take $a = b = n$ then we have

$$\mathbb{E}(T) = n^2$$

i.e. it takes roughly n^2 units of time before the walk leaves the interval $[-n, n]$. This order of magnitude (if not the exact constant in front) could have been guessed from the central limit theorem. Indeed, the position X_m at time m of the random walk is the sum of m i.i.d. random variables (the increments of the walk) which are centered and have variance 1. By the central limit theorem, when m is large, X_m is therefore close in distribution to $\sqrt{m}\mathcal{N}(0,1)$, where $\mathcal{N}(0,1)$ is a standard Gaussian random variable. We must therefore take $m \approx n^2$ before there is a good chance for the walk to exit $[-n, n]$.

4.5 Hitting time of patterns

We will illustrate the power of the Optional Stopping theorem through another very basic and natural question. Suppose we play a game of heads a tails: i.e., we toss a fair coin independently and repeatedly. We do so until we obtain a specified pattern w, say

$$w = THT.$$

More formally, let $(X_n)_{n\geq 1}$ be i.i.d. random variables with $\mathbb{P}(X_n = H) = \mathbb{P}(X_n = T) = 1/2$, and consider the pattern time

$$\tau = \inf\{n \geq 3 : (X_{n-2}, X_{n-1}, X_n) = (T, H, T)\},$$

which is a stopping time. How long should we wait on average for τ? And does this depend on the chosen pattern w? We will show two approaches to this question. The first one will use tools from Markov chains. It is more natural, but does not generalise. The second will use the tools from martingale theory, and will be very elegant – as well as immediate to generalise.

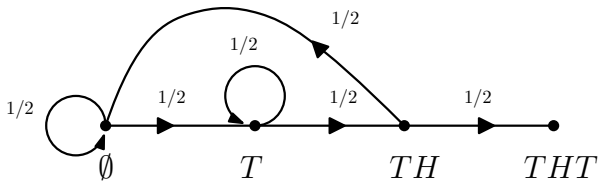
Markov chain solution. We start with the tools from Markov chains. Let us associate to the sequence $(X_n, n \geq 1)$ another Markov chain (call it W_n) which corresponds to the part of the last letters of $(X_0, \dots, X_n)$ which match with w. Thus, if $n = 6$, and

$$(X_1, \dots, X_6) = (H, T, T, H, T, H)$$

then

$$W_n = TH,$$

which is the part of $(X_1, \dots, X_n)$ which may be used to get the pattern $w = THT$. We claim that $(W_n)_{n\geq 0}$ is a Markov chain, on the state space $S = \{\emptyset, T, TH, THT\}$, and this Markov chain may be represented by the following diagram:



In words, the chain W_n starts at time 0 from the state $W_0 = \emptyset$, since at the start no piece of the desired pattern already exists. Then at each step we have probability $1/2$ to start creating the pattern with drawing the letter T. If so we have probability $1/2$ to draw the letter H (which will help us build w and thus advance), or to draw T in which case T is the only pattern that is usable. From TH, either we draw a T and we have w, or we draw a H and must start from scratch.

With these notations, we are interested in $\mathbb{E}_\emptyset(T_w)$. For a state $x \in S$, let $F(x) = \mathbb{E}_x(T_w)$. Using the Markov property, we may write a system of equations for $F(x)$ (essentially the analogue of the Dirichlet problem of Theorem 2.10, but for expected hitting times):

$$\begin{aligned} F(\emptyset) &= \frac{1}{2}F(\emptyset) + \frac{1}{2}F(T) + 1 \\ F(T) &= \frac{1}{2}F(T) + \frac{1}{2}F(TH) + 1 \\ F(TH) &= \frac{1}{2}F(\emptyset) + \frac{1}{2}F(w) + 1 \end{aligned}$$

Since $F(w) = 0$, we get

$$F(TH) = \frac{1}{2}F(\emptyset) + 1$$

and thus plugging back into the equation for $F(T)$, we obtain

$$F(T) = \frac{1}{2}F(\emptyset) + 3$$

and so plugging back into the equation for $F(\emptyset)$ we get

$$F(\emptyset) = \frac{1}{2}F(\emptyset) + 5$$

so that

$$F(\emptyset) = 10.$$

Hence

$$\mathbb{E}(\tau) = 10. \tag{4.7}$$

As you can see, this is hard to generalise. What if we change w to a different pattern, say $w = HHH$? What if the pattern is somewhat longer? each time the Markov chain must be computed, the system of equations must be set up and solved. As you can see, this is not a very efficient method!

Lecture 17, Thursday 1.12.22

Martingale solution. We now explain how to answer this question using the tools from martingales. We first consider a restricted game using three independent fair coins X, Y, Z. The player hopes to get the pattern THT. At first she bets 1€ on T. If she loses, she quits the game. Otherwise, she invests her fortune (now 2€) on the next letter of that pattern, H. If she loses, she quits the game. Otherwise, she invests her fortune (now 4€) on the next letter of that pattern, T.

In total the player either loses her initial 1€, or wins $8 - 1 = 7$€ in case she gets the pattern she desires (which occurs with probability $1/8$). Naturally, since the game is fair, her expected win is $7 \times (1/8) - 1 \times (7/8) = 0$ €.

Now let us get back to the question at hand, the computation of $\mathbb{E}(\tau)$. Suppose that at each $n = 1, 2, \ldots$, a new player comes in, and starts playing the above restricted game with coins (X_n, X_{n+1}, X_{n+2}). For $n \geq 0$, let M_n denote the combined net fortune (sum of all fortunes) at time n of all the players that have come in up to and including time n (i.e., after the nth coin has been tossed). Since each player plays at a fair game, their fortune evolves like a martingale. As a consequence $(M_n)_{n\geq 0}$ is itself a martingale.

We will apply the optional stopping theorem at time τ. Let us consider the random variable M_τ. Suppose $\tau = n$. Then we have had n players who have each invested 1€ to participate in the game. Of those, since τ is the first time that the pattern w occurs, only two players have gained something positive: the player arriving at time $n-2$ (playing with coins X_{n-2}, X_{n-1}, X_n – she has earned 8€) and the player arriving at time n who plays with coins X_n, X_{n+1}, X_{n+2} – she has earned 2€ at time n. (She may well lose it eventually, but at time n, that is the state of her fortune!) This can be summarised by the following table:

player	1	2	...	$n-2$	$n-1$	n
losses by time n	-1	-1		-1	-1	-1
wins by time n	0	0	...	$+8$	0	$+2$

We conclude that

$$M_\tau = -\tau + 10,$$

almost surely. By the Optional Stopping theorem, we must have

$$\mathbb{E}(M_\tau) = 0$$

and hence

$$\mathbb{E}(\tau) = 10,$$

which confirms (4.7).

Remark 4.23. In order to justify the use of the optional stopping theorem, it would be better to first apply it at time $\tau \wedge n$, noting that if W_n denote the wins by time n and L_n the losses by time n, $\mathbb{E}(M_{\tau\wedge n}) = \mathbb{E}[W_{\tau\wedge n}] - \mathbb{E}[L_{\tau\wedge n}]$. Thus

$$\mathbb{E}[W_{\tau\wedge n}] = \mathbb{E}[L_{\tau\wedge n}] = \mathbb{E}(\tau \wedge n).$$

We let $n \to \infty$, and the right converges to $\mathbb{E}(\tau)$ by monotone convergence, while the left hand side converges to $\mathbb{E}(W_\tau) = 10$ by the dominated convergence theorem, since $|W_{\tau\wedge n}| \leq 10$.

This approach is much easier to generalise. For instance, the time τ' to hit the pattern $w' = HHH$ has expectation

$$\mathbb{E}(\tau') = 2 + 4 + 8 = 14,$$

since the gain comes from the last three gamblers (which win respectively 2,4 and 8 at time τ).

Remark 4.24. In particular, the expected time it takes to obtain a given pattern w of a fixed length (here the length of the pattern is $n = 3$) depends on the pattern w, even though, in any sequence of n given coin tosses, the probability to obtain any pattern w is always $1/2^n$, independently of the pattern w.

Intuitively, this comes from the dependence between the sequences. When we fail to build the pattern, in particular, sometimes the failure itself can be used as the start of a new attempt to build the pattern. In the pattern HHH however, when we fail it will always be through the letter T and so this can not be used; we must start from scratch. From that point of view it seems intuitive that HHH (or equivalently TTT) is the worst possible pattern in the sense that the corresponding expected hitting time is maximal among patterns of length three. This is relatively easy to prove with the martingale approach.

Exercise 4.25. Prove that the pattern $w = HH\ldots H$ (or equivalently TTT) maximises the expected hitting time among all patterns of length n.

Obviously this argument is not limited to coin-tossing:

Exercise 4.26. A monkey types at random on a keyboard. How long will it take on average until it will have typed the word ABRACADABRA?

4.6 Discrete Stroock–Varadhan theorem

By now it should be plausible that finding martingales is very helpful in order to describe or understand a stochastic process. But is there a systematic way to find martingales, or is it an art? The following result, which we call a discrete form of the Stroock–Varadhan theorem, shows that at least for Markov chains there is a systematic method. In fact, there are so many martingales attached to a Markov chain that these can be used to completely characterise the Markov property. (Usually this theorem is stated for continuous diffusions, where it is highly nontrivial – in this simple discrete setting, its proof will be rather simple).

First, we need some notations. Fix a Markov chain $(X_n, n \geq 0)$ with transition matrix P on a state space S. Let $f : S \to \mathbb{R}$ be a function on S, with values in $\mathbb{R}$. Define a new function Pf by setting, for $x \in S$,

$$(Pf)(x) = \sum_y P(x,y)f(y). \tag{4.8}$$

Therefore, if the function f is viewed as a column vector, then the function Pf coincides with the vector Pf. However the identity (4.8) signals a change in point of view which may be useful to adopt (and which has been already implicit in some of what we have already discussed): we view here P no longer as a matrix but as an operator on function, in the same way that in analysis a function $K(x,y)$ from $S \times S \to \mathbb{R}$ is identified with an integral operator $f \mapsto Kf$ with $Kf(x) = \int K(x,y)f(y)dy$. In fact, except for the fact that the sum has been replaced by an integral, this analogy can be taken literally. When we wish to emphasise this point of view we will call P an **operator** rather than a matrix.

Remark 4.27. Note that $Pf(x) = \mathbb{E}_x(f(X_1))$. Thus $Pf(x)$, in words, is the following quantity: start from x, and apply one step of the Markov chain, then compute the average new value of f.

The definition (4.8) makes sense for more general matrices. In particular, if $D = P - I$ (where I is the identity matrix), we can define

$$\begin{aligned} Df(x) &= \sum_y D(x,y)f(y) \\ &= \sum_y [P(x,y) - I(x,y)]f(y) \\ &= [\sum_y P(x,y)f(y)] - f(x) \\ &= Pf(x) - f(x) \\ &= \sum_y P(x,y)[f(y) - f(x)], \end{aligned} \tag{4.9}$$

where in the last line we used the fact that $\sum_y P(x,y) = 1$. Either way, as can be seen from any of the last two above equivalent expressions, $Df(x)$ measures the **expected change** in f after applying one step of the Markov chain.

Definition 4.28. *The operator (matrix) D on functions $f : S \to \mathbb{R}$ is called the* ***discrete Laplace operator*** *associated to the transition matrix P.*

Example 4.29. The following examples explains more clearly why D bears the name of a discrete Laplace operator. Fix a small number $\varepsilon > 0$ and consider the simple random walk on $S_\varepsilon = \varepsilon\mathbb{Z}^d$ (which is the scaled lattice, where the mesh size is ε instead of one). Let P_ε denote the associated transition matrix on S_ε and let D_ε denote the associated discrete Laplace operator. Fix a sequence $x_\varepsilon \in S_\varepsilon$ with $x_\varepsilon \to x \in \mathbb{R}^d$ as $\varepsilon \to 0$. Fix a function $f : \mathbb{R}^d \to \mathbb{R}$ to be a twice continuously differentiable. Then, as $\varepsilon \to 0$,

$$D_\varepsilon f(x_\varepsilon) \sim \frac{\varepsilon^2}{2d}\Delta f(x), \tag{4.10}$$

in the sense that the ratio of both sides converges to 1 as $\varepsilon \to 0$. To see (4.10), we simply note that (denoting $(e_i)_{1\le i\le d}$ the canonical basis of $\mathbb{R}^d$, so that from $x_\varepsilon \in S_\varepsilon$, the walk can jump to $x_\varepsilon \pm \varepsilon e_i$ with probability $1/(2d)$ each),

$$\begin{aligned} D_\varepsilon f(x_\varepsilon) &= \sum_{i=1}^{d} \frac{f(x_\varepsilon + \varepsilon e_i) + f(x_\varepsilon - \varepsilon e_i)}{2d} - f(x_\varepsilon) \\ &= \sum_{i=1}^{d} \frac{f(x_\varepsilon + \varepsilon e_i) + f(x_\varepsilon - \varepsilon e_i) - 2f(x_\varepsilon)}{2d} \\ &= \sum_{i=1}^{d} \frac{\varepsilon^2}{2d}\frac{\partial^2 f}{\partial x_i^2}(x_\varepsilon) + o(\varepsilon^2) \\ &\sim \frac{\varepsilon^2}{2d}\Delta f(x) \end{aligned}$$

after Taylor expansion and by continuity of the second derivatives.

Remark 4.30. If you study probability further, you will learn that, speeding time by a factor ε^{-2}, the random walk on $\varepsilon\mathbb{Z}^d$ converges to a random continuous trajectory called the **Brownian motion**. In a suitable sense, its "infinitesimal transition probabilities" are related to the (continuous) Laplacian.

Let us now state the discrete Stroock–Varadhan theorem. In one direction, this result gives us a way to construct many martingales associated with a Markov chain. In other direction, this result says that these martingales completely characterise the law of the Markov chain.

Theorem 4.31 (Discrete Stroock–Varadhan theorem)**.** *Let $X = (X_n)_{n\ge 0}$ be a stochastic process on a state space S. Let P be a transition matrix on S and let $D = P - I$ be the associated discrete Laplace operator. Fix $f : S \to \mathbb{R}$, and define an associated stochastic process M^f by setting (for $n \ge 0$):*

$$M_n^f = f(X_n) - \sum_{k=0}^{n-1} Df(X_k), \tag{4.11}$$

(where if $n = 0$ the sum is by convention null). Suppose $f(X_n)$ and $Pf(X_n)$ are integrable random variables for any $n \geq 0$, i.e., $\mathbb{E}[|f(X_n)|] < \infty$ and $\mathbb{E}[|Pf(X_n)|] < \infty$ for all $n \geq 0$. Then the two conditions are equivalent:

(i) X is a Markov chain with transition matrix P.

(ii) M^f is a martingale for any choice of $f : S \to \mathbb{R}$.

Proof. Let us with the direction (i) $\Rightarrow$ (ii). Fix f as in the theorem, and consider M^f. It is clear that M^f is adapted since, given $\mathcal{F}_n$, $f(X_n)$ is known, and so is $Df(X_k)$ for every $0 \leq k \leq n-1$. M^f is further integrable by the triangle inequality and the assumption that $f(X_n)$ and $Pf(X_k)$ are integrable for every $0 \leq k \leq n$. It remains to prove the martingale property:

$$\begin{aligned}\mathbb{E}[M^f_{n+1}|\mathcal{F}_n] &= \mathbb{E}\left[f(X_{n+1}) - \sum_{k=0}^{n} Df(X_k)\middle|\mathcal{F}_n\right] \\ &= \mathbb{E}[f(X_{n+1})|\mathcal{F}_n] - \sum_{k=0}^{n} Df(X_k) \\ &= Pf(X_n) - \sum_{k=0}^{n} Df(X_k)\end{aligned}$$

by the Markov property and the definition of Pf. Thus, isolating the last term in the sum, and using one of the equivalent definitions of the Laplace operator (more precisely, (4.9)),

$$\begin{aligned}\mathbb{E}[M^f_{n+1}|\mathcal{F}_n] &= Pf(X_n) - Df(X_n) - \sum_{k=0}^{n-1} Df(X_k) \\ &= Pf(X_n) - (Pf(X_n) - f(X_n)) - \sum_{k=0}^{n-1} Df(X_k) \\ &= f(X_n) - \sum_{k=0}^{n-1} Df(X_k) \\ &= M^f_n.\end{aligned}$$

Thus M^f is a martingale, as desired.

Now let us turn to the converse (ii) $\Rightarrow$ (i). Fix $y \in S$. We first claim that it suffices to show that

$$\mathbb{P}(X_{n+1} = y|\mathcal{F}_n) = P(X_n, y). \tag{4.12}$$

Indeed from there it follows directly (using the tower property) that for any $x \in S$, $\mathbb{P}(X_{n+1} = y|\mathcal{F}_n, X_n = x) = P(x, y)$, which is the definition of Markov chains with transition matrix P.

To see (4.12), we simply use the fact that M^f is a martingale for the function $f = \delta_y$: indeed,

$$\begin{aligned}
\mathbb{P}(X_{n+1} = y|\mathcal{F}_n) &= \mathbb{E}(f(X_{n+1})|\mathcal{F}_n) \\
&= \mathbb{E}(M^f_{n+1}|\mathcal{F}_n) + \sum_{k=0}^{n} Df(X_k) \\
&= M^f_n + \sum_{k=0}^{n} Df(X_k) \\
&= f(X_n) - \sum_{k=0}^{n-1} Df(X_k) + \sum_{k=0}^{n} Df(X_k) \\
&= f(X_n) + Df(X_n) \\
&= f(X_n) + Pf(X_n) - f(X_n) \\
&= Pf(X_n),
\end{aligned}$$

as desired. □

Example 4.32. Let X be simple random walk on $\mathbb{Z}$ and let $f(x) = x$. Then $Df(x) = 0$, so X_n is a martingale, as we already know. Now consider $f(x) = x^2$. Then

$$Pf(x) = \frac{1}{2}(x+1)^2 + \frac{1}{2}(x-1)^2 = x^2 + 1.$$

Thus $Df(x) = Pf(x) - f(x) = 1$, and we deduce that $M^f_n = X^2_n - n$ is a martingale. We therefore recover the result obtained in Example 4.8.

We have already alluded to the fact that there are connections between martingales and harmonic functions. The Stroock–Varadhan theorem allows us to make this precise. First let us define precisely the notion of harmonic function and of subharmonic functions.

Definition 4.33. *Let P be a transition matrix on a state space S, and let $D = P - I$ be the associated discrete Laplace operator. Let $f : S \to \mathbb{R}$. We say that f is* ***harmonic*** *(resp. subharmonic, superharmonic) on $A \subset S$ if*

$$Df(x) = 0$$

(resp. $Df(x) \geq 0$, $Df(x) \leq 0$) for all $x \in A$.

In other words, f is harmonic on A if $Pf(x) = f(x)$, i.e., $f(x) = \sum_y P(x,y)f(y)$ (that is, f satisfies the mean-value property). For a subharmonic function f, the definition is equivalent to requiring $f(x) \leq \sum_y P(x,y)f(y)$ for all $x \in A$.

Example 4.34. Let $A \subset S$, and let $h(x) = \mathbb{P}_x(T_A < \infty)$. Then h is harmonic on A^c. In fact, by Theorem 2.10, h is the minimmal nonnegative harmonic function which is equal to 1 on A.

Corollary 4.35. *Let $X = (X_n)_{n\geq 0}$ be Markov (λ, P) on a state space S. Let f be a subharmonic function. For $n \geq 0$ let $Z_n = f(X_n)$ and suppose Z_n is integrable (i.e., $\mathbb{E}(|f(X_n)|) < \infty$). Then $Z = (Z_n)_{n\geq 0}$ is a submartingale. In particular if f is harmonic then Z is a martingale.*

Proof. This is straightforward from Theorem 4.31, except for the integrability of M_n which needs some justification: however when f is harmonic then $Df(X_n) = 0$ which is definitely integrable. This makes the theorem applicable. When f is only assumed to be subharmonic, then $Df(X_n) \geq 0$. Even if we do not know the integrability, this is however sufficient to justify the computations of the conditional expectation in the proof of the theorem: $\mathbb{E}[f(X_{n+1})|\mathcal{F}_n] = Pf(X_n) = f(X_n) + Df(X_n) \geq f(X_n)$. The result follows immediately. □

As an example of application, we obtain a probabilistic proof of an analytic result called the **Liouville property**.

Corollary 4.36. *Let $G = \mathbb{Z}^d$ with $d = 1$ or $d = 2$. Then G possesses the Liouville property: any bounded harmonic function is constant.*

Proof. Let f be a bounded harmonic function and let $x, y \in \mathbb{Z}^d$. We want to show that $f(x) = f(y)$. Consider the random walk X starting at x, and let $T = T_y$ be the hitting time of y by X. Since f is bounded, $f(X_n)$ is integrable, and since f is harmonic, we deduce that $M_n = f(X_n)$ is a bounded martingale. Applying the optional stopping theorem at time T (which is allowed since on the one hand, T finite almost surely by Pólya's theorem, and on the other hand, M is uniformly bounded),

$$\mathbb{E}_x(M_T) = \mathbb{E}_x(M_0).$$

The left hand side $f(y)$ and the right hand side is $f(x)$. This completes the proof. □

Remark 4.37. In dimension $d = 2$, this result should be compared to the Liouville's theorem in complex analysis: any function which is bounded and holomorphic on the entire complex plane is constant. The connection between these two facts is that in two dimension, any real-valued harmonic function $u : \mathbb{R}^2 \to \mathbb{R}$ can be viewed as the real part of a holomorphic function $f : \mathbb{C} \to \mathbb{C}$. Furthermore if u is bounded then so is f, which explains the result.

In fact, the result is true not just in dimension $d = 1, 2$ but also in dimensions $d \geq 3$ (even though the graph is transient). However, the Liouville property is easily shown to fail on graphs such as the binary tree: in fact, by a beautiful theorem of Varopoulos [Var85] (see also the important work of Kaimanovich–Vershik [KV83]), on Cayley graphs (and even more generally), the existence of non-constant bounded harmonic functions is equivalent to the random walk escaping to infinity at positive speed, in the sense that $\lim_{n\to\infty} d(o, X_n)/n$ exists and is positive, with $d(x, y)$ being the graph distance and o the starting point of the walk.

4.7 Eigenfunctions and the escape problem

Lecture 18: Friday 2.12.22

We will show another application of the optional stopping theorem to the following escape problem for a random walk on a locally finite, connected graph $G = (V, E)$. Fix a starting point for the walk (call it x) and a finite subset $A \subset V$ containing x. We suppose that A is a strict subset of V and without loss of generality suppose it is connected. The problem we consider (the escape problem) is the following: *how likely is it for the random walk to remain in A for a long time*? Let

$$\tau = \inf\{n \geq 0 : X_n \notin A\}$$

be the first time that the walk leaves A (thus, with our previous notations, $\tau = T_{A^c}$ is the hitting time of A^c). We are asking to obtain asymptotics as $n \to \infty$ for the probability $\mathbb{P}_x(\tau > n)$. Intuition suggests that this probability decays exponentially fast with n: after all, since A is finite and a strict subset of vertices of the graph, every few units of time, there is always a constant probability to escape A if we haven't done so far. By the Markov property we thus expect an exponential decay of $\mathbb{P}_x(\tau > n)$. But can we prove this? And can we determine the exact exponential rate of decay?

We will see how to answer this problem with martingales and through the help of a suitable notion of eigenfunctions.

Definition 4.38. *Let P the transition matrix of an irreducible Markov chain on S. Let $f : S \to \mathbb{R}$. f is called an **eigenfunction** on $A \subset S$ (with eigenvalue $\lambda \geq 0$) if f is not constantly equal to zero and*

$$\begin{cases} Pf(x) = \lambda f(x) & \text{if } x \in A \\ f(x) = 0 & \text{if } x \notin A. \end{cases}$$

*If furthermore $f(x) > 0$ for all $x \in A$ then we say that f is a **principal eigenfunction**.*

Note that a principal eigenvalue λ satisfies $\lambda > 0$ necessarily (since then both $Pf(x)$ and $f(x)$ are strictly positive).

Example 4.39. Let $G = \mathbb{Z}$ and for $L \geq 1$ let $A = \{1, \dots, L-1\}$. Then for $k \geq 1$, the function

$$f_k(x) = \sin\left(\frac{\pi k x}{L}\right); \quad x \in \{0, \dots, L\};$$

and $f_k(x) = 0$ if $x \in \mathbb{Z} \setminus \{0, \dots, L\}$, defines an eigenfunction with eigenvalue $\lambda_k = \cos(k\pi/L)$. In particular, for $k = 1$, $f = f_1$ is a principal eigenfunction with associated eigenvalue $\lambda = \cos(\pi/L)$.

Let us check this carefully. Fix $k \geq 1$. First let us observe that $f_k(x) = 0$ for all $x \notin A$ (including at the boundary $x = 0$ and $x = L$) since $\sin(0) = \sin(\pi k) = 0$. Furthermore, for

any $x \in A$,

$$\begin{aligned} Pf_k(x) &= \frac{1}{2}f_k(x+1) + \frac{1}{2}f_k(x-1) \\ &= \frac{1}{2}\sin\left(\frac{\pi k(x+1)}{L}\right) + \frac{1}{2}\sin\left(\frac{\pi k(x-1)}{L}\right) \\ &= \frac{1}{2}\Im(\exp(i\frac{\pi k(x+1)}{L})) + \frac{1}{2}\Im(\exp(i\frac{\pi k(x-1)}{L})) \\ &= \Im\left(\exp(i\frac{\pi kx}{L})[\frac{1}{2}\exp(i\frac{\pi k}{L}) + \frac{1}{2}\exp(-i\frac{\pi k}{L})]\right) \\ &= \Im\left(\exp(i\frac{\pi kx}{L})\cos(\frac{\pi k}{L})\right) \\ &= \lambda_k f_k(x). \end{aligned}$$

Thus f_k is an eigenfunction. Furthermore, if $k = 1$, for $x \in A$ we have $\pi x/L \in (0, \pi)$ so $f_k(x) > 0$. Thus f_k is a principal eingenfunction, as desired.

Remark 4.40. The occurrence of trigonometric functions here should only be mildly surprising. Informally, our definition of eigenfunction can also be reformulated in terms of the discrete Laplace operator as $Df(x) = (\lambda - 1)f(x)$, so our Definition 4.38 amounts to requiring that f is an eigenfunction for the discrete Laplace operator in A with *Dirichlet boundary condition*. In the continuum, it is easy to check that $f_k(x) = \sin(\pi kx/L)$ then $f_k''(x)$ is indeed a multiple of $f_k(x)$.

More generally, one can show with a bit of linear algebra (going outside the scope of this course – this is the Perron–Frobenius theorem) that if A is nonempty, there always exists a unique principal eigenfunction. Sometimes these eigenfunctions can even be computed explicitly as above (and you will some other examples in the exercises).

The relevance of the eigenfunctions to the escape problem is explained by the following result.

Theorem 4.41. *Let f be a principal eigenfunction and let λ be the associated eigenvalue. There exists constants c, C, depending only on A, such that for all $x \in A$, and for all $n \geq 0$,*

$$c\lambda^n \leq \mathbb{P}_x(\tau > n) \leq C\lambda^n.$$

(In fact, as mentioned above, principal eigenvalues are necessarily unique, but we will not need this uniqueness here). For instance, in dimension $d = 1$, if $A = \{1, \ldots, L-1\}$, $\mathbb{P}_x(\tau > n)$ decays exponentially fast, at rate $\lambda = \cos(\pi/L)$.

Proof. The proof is a nice application of the optional stopping theorem. The key is to find a suitable martingale. We define:

$$M_n = \lambda^{-n} f(X_{n\wedge\tau}); n \geq 0.$$

We claim M is a martingale. Indeed, it is first clear that M is adapted, and M_n is also integrable for every $n \geq 0$, since λ^{-n} is a constant and f is bounded by an absolute constant C (indeed, f is nonzero only on the set A, which is finite). To check the martingale property, we condition on $\mathcal{F}_n$ and distinguish two cases.

(i) $\tau \leq n$. If $\tau \leq n$, then $X_{n\wedge\tau} = X_{(n+1)\wedge\tau} = X_\tau \in A^c$, so $f(X_{(n+1)\wedge\tau}) = f(X_{n\wedge\tau}) = 0$. Hence, on the event $\{\tau \leq n\}$,

$$\mathbb{E}(M_{n+1}|\mathcal{F}_n) = \mathbb{E}(0|\mathcal{F}_n) = 0 = M_n.$$

(ii) Now suppose instead $\tau > n$, in particular $X_{n\wedge\tau} = X_n \in A$ and $(n+1)\wedge\tau = n+1$. Then

$$\begin{aligned}\mathbb{E}[M_{n+1}|\mathcal{F}_n] &= \mathbb{E}[\lambda^{-(n+1)}f(X_{n+1})|\mathcal{F}_n] \\ &= \lambda^{-(n+1)}Pf(X_n) \\ &= \lambda^{-n}f(X_n) = M_n.\end{aligned}$$

since f is an eigenfunction with eigenvalue λ. All in all, M is indeed a martingale.

Let us apply the optional stopping theorem at the bounded stopping time $\tau \wedge n$. Then

$$\mathbb{E}_x(M_{\tau\wedge n}) = \mathbb{E}_x(M_0) = f(x).$$

On the other hand, in $M_{\tau\wedge n}$, the only nonzero contribution comes from the event $\tau > n$: if $\tau \leq n$ then $M_n = 0$ since $f = 0$ outside of A. Thus

$$\begin{aligned}\mathbb{E}_x(M_{\tau\wedge n}) &= \mathbb{E}_x(M_{\tau\wedge n}1_{\{\tau>n\}}) \\ &= \mathbb{E}_x(\lambda^{-n}f(X_n)1_{\{\tau>n\}}).\end{aligned}$$

Let $c = \min_{x\in A} f(x) > 0$, and let $C = \max_{x\in A} f(x)$. Then the above expectation satisfies

$$\begin{aligned}&\mathbb{E}_x(\lambda^{-n}f(X_n)1_{\{\tau>n\}}) \\ &\geq \lambda^{-n}\mathbb{E}_x(c1_{\{\tau>n\}}) \\ &\geq c\lambda^{-n}\mathbb{P}_x(\tau > n).\end{aligned}$$

We deduce

$$\begin{aligned}\mathbb{P}_x(\tau > n) &\leq (1/c)\lambda^n\mathbb{E}_x(M_{\tau\wedge n}) \\ &= (1/c)\lambda^n\mathbb{E}_x(M_0) \\ &= (1/c)\lambda^n f(x) \\ &\leq (C/c)\lambda^n\end{aligned}$$

from which the desired upper bound follows. The lower bound can be proved in exactly the same way. □

4.8 Doob's martingale convergence theorems

Lecture 19: Friday 9.12.22. NB: no class on Thursday 8.12.22

By now it should be clear that martingales are not only a fundamental tool in the study of stochastic processes, but are also ubiquitous. It turns out there is yet another reason why martingales are so useful. This comes from the Doob martingale convergence theorem which we will discuss in this section. In essence, the theorem can be understood as showing a **dichotomy** for martingales, in the sense that there are only *two* ways a stochastic process can be a martingale. For the conditional expectation to be zero all the time (i.e., for the negative fluctuations to balance the positive fluctuations on average), the only two possibilities are the following:

- Either the fluctuations of the martingales become larger and larger and the martingale oscillates between larger and larger positive and negative values
- Or, on the other hand, the fluctuations become smaller and smaller; in that case the martingale will converge to a finite limit.

Hence it is not possible for any martingale to stay bounded and oscillate indefinitely between two values. The theorem can be stated in various forms, which have the following flavour. We can suppose that the martingale is in some sense bounded (either uniformly or in some weaker sense, say in the L^2 sense). Effectively this rules out scenario one in the above dichotomy. According to this dichotomy, scenario two must holds! And indeed the conclusion of the theorem will be that, under this sole assumption of boundedness, the martingale must have an almost sure limit.

Alternatively, we could make a different assumption, by supposing that the martingale is bounded, but only in one direction: say, we assume that the martingale is nonnegative. Although by doing so we did not explicitly rule out large upward fluctuations, this nevertheless prevent scenario one to occur, since no large downward fluctuations can occur. Hence in this case too, according to this dichotomy, convergence must occur!

We will start our preparations for this amazing theorem by defining the crucial notion of downcrossings, which makes sense for any given sequence (random or not).

Definition 4.42. *Let $(x_n)_{n\geq 0}$ be a sequence with values in $\mathbb{R}$. Fix $\alpha < \beta$. A **downcrossing** of the interval $[\alpha, \beta]$ by the sequence x is an interval of time, say $[n_1, n_2]$, such that $x_{n_1} \geq \beta$, $x_{n_2} \leq \alpha$, but $x_k > \alpha$ for $n_1 \leq k < n_2$.*

We will be interested in counting the number of downcrossings completed by time n. To this end, let us introduce the following stopping times:

$$\sigma_1 = \inf\{n \geq 0 : x_n \geq \beta\}; \quad \tau_1 = \inf\{n \geq \sigma_1 : x_n \leq \alpha\}.$$

The interval $[\sigma_1, \tau_1]$ is the first interval at which a downcrossing of $[\alpha, \beta]$ is completed. The inductively, for $n \geq 1$ let

$$\sigma_{n+1} = \inf\{n \geq \tau_n : x_n \geq \beta\}; \quad \tau_{n+1} = \inf\{n \geq \sigma_{n+1} : x_n \leq \alpha\}.$$

Figure 8 shows the first few values of these stopping times on a concrete example.

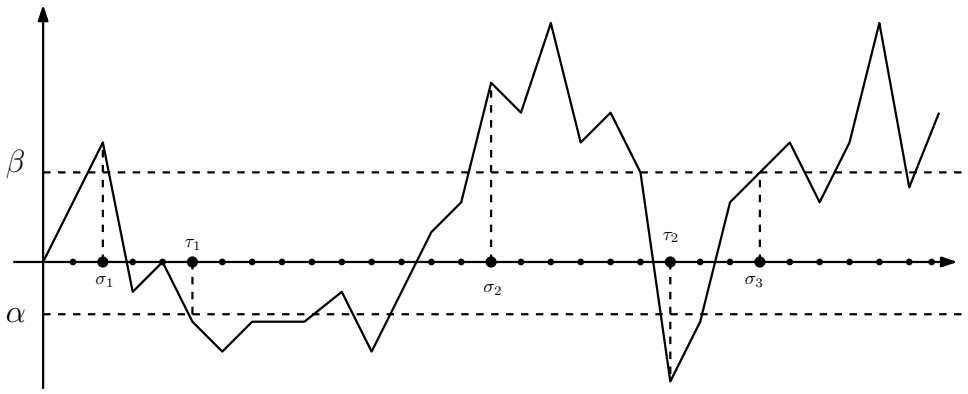


Figure 8: The sequence $(x_k)_{k\geq 0}$ is drawn as a continuous curve (the linear interpolation of its values) for ease. It has completed two downcrossings by time n (during $[\sigma_1, \tau_1]$ and $[\sigma_2, \tau_2]$ respectively). A third one begins at time σ_3 before time n, but is not completed.

Definition 4.43. *For $\alpha < \beta$, the **number of downcrossings** of the sequence $(x_k)_{k\geq 0}$ by time n, $d_n = d_n[\alpha, \beta]$, is the quantity:*

$$d_n = \max\{k : \tau_k \leq n\}$$

with $d_n = 0$ by convention if $\tau_1 > n$ (i.e., when this set is empty).

The reason we care about downcrossings is that they can be used to characterise the convergence of the sequence $(x_n)_{n\geq 0}$:

Lemma 4.44. *The sequence $(x_n)_{n\geq 0}$ is convergent in $\mathbb{R} \cup \{-\infty, +\infty\}$ if and only if for any $\alpha, \beta \in \mathbb{Q}$ with $\alpha < \beta$, there exists $C = C(\alpha, \beta)$ such that $d_n[\alpha, \beta] \leq C$.*

Proof. This is easy to see using the limsup and liminf of the sequence, $\ell^+ = \limsup_{n\to\infty} x_n$ and $\ell^- = \liminf_{n\to\infty} x_n$. We recall that these are always well defined in $\mathbb{R} \cup \{-\infty, \infty\}$, and the sequence is convergent if and only if $\ell^- = \ell^+$. Hence non convergence is equivalent to $\ell^- < \ell^+$ which is equivalent to the existence of rational numbers $\alpha < \beta$ such that $\ell^- < \alpha < \beta < \ell^+$. By definition of ℓ^- and ℓ^+, this is equivalent to the existence of subsequences along which $x_n \leq \alpha$ and $x_n \geq \beta$, which in turn is equivalent to an unbounded number of downcrossings. □

We are therefore interested in bounding from above the number of downcrossings. The crucial observation, due to Doob, is the following inequality called **Doob's downcrossing lemma**.

Lemma 4.45. *Let $(X_n)_{n\geq 0}$ be a submartingale, and for $\alpha < \beta$ let $D_n = D_n[\alpha, \beta]$ denote the number of downcrossings by X of $[\alpha, \beta]$ by time n. Then*

$$\mathbb{E}(D_n) \leq \frac{\mathbb{E}(X_n^+) + |\beta|}{\beta - \alpha}$$

where $X_n^+ = (X_n)_+$, *and* $x_+ = \max(x, 0)$ *denote the nonnegative part of* x_n.

Let us first give the proof of the lemma, and then discuss its significance.

Proof. Let $\theta = \beta - \alpha$ which is the gap between the two sides of the downcrossing. We define a predictable process $(W_n)_{n\geq 0}$ by setting

$$W_n = \begin{cases} 1 & \text{if } \sigma_i < n \leq \tau_i \text{ for some } i \geq 1, \\ 0 & \text{else.} \end{cases}$$

In words, the betting strategy is 1 if a downcrossing has started, and 0 if it has been completed and the next one has not yet started. Clearly, W is predictable, bounded, and nonnegative, from which it follows that the martingale transform $Z_n = (W \cdot X)_n - X_0$ is a submartingale, by Theorem 4.17. Furthermore, with W we only bet a nonzero amount (1€, in fact) during a downcrossing. Let us consider the increments $\Delta'_n = Z_n - Z_{n-1} = W_n(X_n - X_{n-1})$ of the martingale transform. We note the following properties:

- For each completed downcrossing interval $[n_1, n_2]$, the sum of the increments of Z during $(n_1, n_2]$ is at least θ in absolute value, i.e., $\sum_{n=n_1+1}^{n_2} \Delta'_n \leq -\theta$. (This is only an upper-bound, because at the start n_1 of the downcrossing, $X_{n_1} \geq \beta$, and at the end of the downcrossing, $X_{n_2} \leq \alpha$, so overall X has decreased by more than θ in absolute value.)
- Outside of a downcrossing the increments of Z are zero.
- The only way these downcrossings are potentially offset is through a downcrossing which started at time σ before n but not been completed by time n. In that case, during that interval W is nonzero, and X could potentially increase – but in any case, by no more than $(X_n - \beta)_+$:

$$\sum_{k=\sigma+1}^{n} \Delta'_k \leq (X_n - \beta)_+$$

As a consequence,

$$Z_n = (W \cdot X)_n - X_0 \leq (X_n - \beta)_+ - \theta D_n. \tag{4.13}$$

Exploiting the submartingale property of Z, we get

$$\mathbb{E}((X_n - \beta)_+ - \theta D_n) \geq \mathbb{E}(Z_n) \geq \mathbb{E}(Z_0) = 0.$$

In other words,

$$\mathbb{E}(D_n) \leq \frac{1}{\theta}[\mathbb{E}((X_n - \beta)_+)] \leq \frac{1}{\theta}[\mathbb{E}(X_n^+) + |\beta|],$$

as desired. □

Remark 4.46. As already mentioned, Doob's downcrossing lemma is the crucial argument which is used in the proof of the martingale convergence below. It is worth pausing a moment to consider the various ideas used in that proof. At the heart of the proof lies the idea that, even if we bet 1€ during each downcrossing, the resulting fortune remains a submartingale (so the fortune has a tendency to increase overall). These downcrossings can only be offset by a large (positive) value at the end. Hence if $\mathbb{E}(X_n^+)$ is known to be not too large, it follows there cannot be too many downcrossings. In other words, if the martingale remained bounded and oscillated a lot, there would be many downcrossings (and hence a way to lose money systematically) without anything to offset it – an impossibility.

Lecture 20: Thursday 12.01.23. NB: no class on Thursday 15.12.22 and Friday 16.12.22

With this we can now state the first version of Doob's martingale convergence.

Theorem 4.47 (Doob's martingale convergence theorem 1)**.** *Let $(X_n)_{n\geq 0}$ be a submartingale, and suppose that there exists $K < \infty$ such that $\mathbb{E}(X_n^+) \leq K$ for every $n \geq 0$. Then there exists a random variable X such that $X_n \to X$, almost surely. Furthermore, $\mathbb{E}(|X|) < \infty$.*

Remark 4.48. The above notion of convergence is that of almost sure convergence: for any given realisation of the sequence $(X_n)_{n\geq 0}$, e.g. one that you simulate on a computer, you will observe the sequence X_n converging to a limit. That limit could depend on the realisation of the sequence, and is hence a random variable.

Proof. We will use the criterion of Lemma 4.44 together with Doob's downcrossing Lemma 4.45. Fix $\alpha < \beta$ and let $D_n[\alpha, \beta]$ denote the number of downcrossings of $[\alpha, \beta]$ by time n. Since $\mathbb{Q} \times \mathbb{Q}$ is countable, it suffices to show that, almost surely for this fixed choice of α and β in $\mathbb{Q}$, $D_n[\alpha, \beta]$ is a bounded sequence (note: the bound $C = C(\omega, \alpha, \beta)$ is allowed to be random here). Now, $D_n[\alpha, \beta]$ is a nondecreasing sequence and so converges to a limit $D = D[\alpha, \beta]$. It suffices to check that $\mathbb{E}(D) < \infty$ (which implies $D < \infty$ almost surely and hence $D_n \leq D < \infty$ is a sequence bounded by a finite random variable, as desired). By Doob's downcrossing lemma,

$$\begin{aligned}\mathbb{E}(D_n[\alpha, \beta]) &\leq \frac{1}{\beta - \alpha}(\mathbb{E}(X_n^+) + |\beta|) \\ &\leq \frac{K + |\beta|}{\beta - \alpha}\end{aligned}$$

using the assumption. Letting $n \to \infty$, and using the monotonicity of D_n with n, we conclude that

$$\mathbb{E}(D) < \infty,$$

as desired. This proves convergence of the sequence X_n to a limit X. The proof that X is integrable uses some measure-theory arguments: first, $\mathbb{E}(X_n^+) \leq K$ which implies by Fatou's lemma that $\mathbb{E}(X^+) \leq \liminf_{n\to\infty} \mathbb{E}(X_n^+) \leq K$. Secondly, since $\mathbb{E}(X_n^+) \leq K$ and $\mathbb{E}(X_n^+) - \mathbb{E}(X_n^-) = \mathbb{E}(X_n) \geq \mathbb{E}(X_0)$ we also see that $\mathbb{E}(X_n^-) \leq \mathbb{E}(X_n^+) - \mathbb{E}(X_0)$ is also bounded, from which it follows (once again by Fatou's lemma) that $\mathbb{E}(X^-) < \infty$. □

A simple but very important corollary is the following:

Corollary 4.49. *Let $(X_n)_{n\geq 0}$ be a supermartingale, and suppose $X_n \geq 0$ for every $n \geq 0$. Then X_n converges to a limit $X \geq 0$ such that $\mathbb{E}(X) \leq \mathbb{E}(X_0) < \infty$.*

Proof. Let $Y_n = -X_n$. Then Y is a submartingale. Furthermore $Y_n^+ = \max(Y_n, 0) = 0$, so clearly we can apply the previous theorem to Y_n (with $K = 0$ trivially!) Hence Y_n converges to a finite limit, and hence so must $X_n = -Y_n$. The bound on the expectation is a consequence of Fatou's lemma (or the observation already contained in the proof of Theorem 4.47). □

Remark 4.50. A sometimes useful way of remembering Corollary 4.49 is to view it as a stochastic analogue of the property that nondecreasing sequences which are bounded above converge. But it seems to me more useful to keep in mind the dichotomy explained in the introduction to this section in order to understand *why* the result is true.

Note also the fundamental fact that nonnegative martingales converge!

As an example of application, let us give a very short proof of recurrence for the random walk on $\mathbb{Z}$ using martingales.

Example 4.51. Consider the random walk $(X_n)_{n\geq 0}$ on $\mathbb{Z}$, started from $X_0 = 1$. Let $\tau = T_0$ denote the hitting time of zero. Then $\tau < \infty$, with probability one. In particular X is recurrent.

Proof. Recall that X is a martingale. Hence if $Y_n = X_{n\wedge\tau}$ then $(Y_n)_{n\geq 0}$ is also a martingale by Proposition 4.19. It is furthermore nonnegative since we stop at time τ. By Corollary 4.49, Y_n converges almost surely to a limit, call it L. But since Y_n is integer-valued, the only that $(Y_n)_{n\geq 0}$ converges is if Y_n is constant (equal to L) from a certain point onwards; thus $Y_n = L$ for $n \geq n_0$ for some (random) n_0. But Y_n changes by ± 1 at every time n, except if $n \geq \tau$. Thus $\tau \leq n_0$ must be finite (and $L = 0$, but this is not important). This proves the first claim. To prove X is recurrent, if X starts from 0, then after the first step it will be at ± 1. If it jumps to $x = 1$, then it will necessarily return to zero by what we just proved and the simple Markov property. If it jumps to $x = -1$, then the same argument (or symmetry) also shows that the walk will return to zero. Either way, it is guaranteed to return to zero, hence X is recurrent. □

Remark 4.52. We now know multiple proofs of recurrence in one dimension: the first one follows from the computation of hitting probabilities in Theorem 2.14. The second one follows from the non-summability of the series $P^n(0,0)$ which is computed combinatorially in Theorem 2.34. The last proof is the shortest, and follows from basic martingale considerations.

We now state the second version of Doob's martingale convergence theorem, in which both the assumption and the conclusion involve a second moment (so we have both a stronger assumption and a stronger conclusion that Theorem 4.47). For this (in the proof and for

using this theorem), the concept of **conditional variance**, already encountered in Example sheet 8 exercise 1, is very useful: if X is a random variable,

$$\mathrm{Var}(X|\mathcal{F}_n) = \mathbb{E}(X^2|\mathcal{F}_n) - \mathbb{E}(X|\mathcal{F}_n)^2. \tag{4.14}$$

It is easy to check that this can be rewritten as $\mathbb{E}[(X - m_{X,n})^2|\mathcal{F}_n]$ where $m_{X,n} = \mathbb{E}(X|\mathcal{F}_n)$. Conditional variances obey all the rules you know for variances, bearing in mind that – just as in the computation of conditional expectations – all the random variables which depend just on $\mathcal{F}_n$ should be treated as constants. For instance,

$$\mathrm{Var}(\lambda X|\mathcal{F}_n) = \lambda^2 \,\mathrm{Var}(X|\mathcal{F}_n) \quad \text{(scaling)} \tag{4.15}$$

whenever $\lambda \in \mathbb{R}$, and this remains true if λ is a random variable depending solely on $\mathcal{F}_n$. Likewise, if X and Y are independent given $\mathcal{F}_n$, then

$$\mathrm{Var}(X + Y|\mathcal{F}_n) = \mathrm{Var}(X|\mathcal{F}_n) + \mathrm{Var}(Y|\mathcal{F}_n) \quad \text{(additivity)}. \tag{4.16}$$

Finally, if X is independent of $\mathcal{F}_n$ then

$$\mathrm{Var}(X|\mathcal{F}_n) = \mathrm{Var}(X), \tag{4.17}$$

which mirrors what we already know of conditional expectations.

Theorem 4.53. *Let $(X_n)_{n\geq 0}$ be a martingale and suppose that there exists $K < \infty$ such that $\mathbb{E}(X_n^2) \leq K$. Then there exists a random variable X such that $X_n \to X$ almost surely as $n \to \infty$, and $\mathbb{E}(|X_n - X|^2) \to 0$ (i.e., X_n converges to X in $L^2(\mathbb{P})$). In particular, $\mathbb{E}(X) = \mathbb{E}(X_0)$.*

Proof. By Cauchy–Schwarz, $\mathbb{E}(X_n^+) \leq \mathbb{E}(|X_n|) \leq \mathbb{E}(X_n^2)^{1/2} \leq \sqrt{K}$, so the first Doob martingale convergence theorem applies, and hence X_n converges to some limit X almost surely. Hence it suffices to prove that X_n converges to its limit in $L^2(\mathbb{P})$. The proof uses some elements of measure theory, so its understanding is not required for this course. Let $\Delta_n = X_n - X_{n-1}$ be the increment, and let $V_n = \mathbb{E}(\Delta_n^2|\mathcal{F}_{n-1})$ for $n \geq 1$. (Recall that V_n can be interpreted as the conditional variance of X_n, given $\mathcal{F}_{n-1}$.) Recall that, as proved in Example sheet 8, exercise 1, if we define

$$M_n = X_n^2 - \sum_{i=1}^{n} V_i, \qquad n \geq 0,$$

then $(M_n)_{n\geq 0}$ is a martingale. Thus $\mathbb{E}(M_n) = 0$ and $\mathbb{E}(X_n^2) = \sum_{i=1}^n \mathbb{E}(V_i)$. Now, V_i is clearly nonnegative and thus $\mathbb{E}(V_i) \geq 0$ too. On the other hand, we have assumed that $\mathbb{E}(X_n^2) \leq K$ is bounded. Hence the series $\sum_{i=1}^{\infty} \mathbb{E}(V_i)$ must be convergent.

Let us show that $(X_n)_{n\geq 0}$ is a Cauchy sequence in $L^2(\mathbb{P})$. Let $1 \leq m < n$. Then

$$\begin{aligned}(X_n - X_m)^2 &= \left(\sum_{i=m+1}^{n} \Delta_i\right)^2 \\ &= \sum_{i=m+1}^{n} \Delta_i^2 + 2 \sum_{m+1\leq i<j\leq n} \Delta_i \Delta_j.\end{aligned}$$

Taking expectations, we deduce

$$\mathbb{E}((X_n - X_m)^2) = \sum_{i=m+1}^{n} \mathbb{E}(\Delta_i^2) + 2 \sum_{m+1 \leq i < j \leq n} \mathbb{E}(\Delta_i \Delta_j). \tag{4.18}$$

But using the tower property (Theorem 4.9), we note that $\mathbb{E}(\Delta_i^2) = \mathbb{E}(\mathbb{E}(\Delta_i^2 | \mathcal{F}_{i-1})) = \mathbb{E}(V_i)$, while if $i < j$,

$$\begin{aligned} \mathbb{E}(\Delta_i \Delta_j) &= \mathbb{E}(\mathbb{E}(\Delta_i \Delta_j | \mathcal{F}_i)) \\ &= \mathbb{E}(\Delta_i \mathbb{E}(\Delta_j | \mathcal{F}_i)) \text{ (because } \Delta_i \text{ is known given } \mathcal{F}_i) \\ &= \mathbb{E}(\Delta_i \times 0) = 0, \end{aligned}$$

by the martingale property (more precisely, by Proposition 4.12). Plugging back into (4.18), we deduce that

$$\mathbb{E}((X_n - X_m)^2) = \sum_{i=m+1}^{n} \mathbb{E}(V_i).$$

But we have already observed that the series on the right hand side is convergent. The Cauchy property follows immediately, hence X_n converges to X in $L^2(\mathbb{P})$, as desired. □

Example 4.54. Let ε_i be independent random variables, with $\varepsilon_i = \pm 1$ each with probability $1/2$. Then the series

$$\sum_{n=1}^{\infty} \frac{\varepsilon_n}{n}$$

converges almost surely and in $L^2(\mathbb{P})$ (although it does not converge absolutely). Indeed, the partial sum $M_N = \sum_{n=1}^{N} \varepsilon/n$ is a martingale which is bounded in $L^2(\mathbb{P})$ (as $\mathbb{E}(M_N^2) = \sum_{n=1}^{N} 1/n^2$ is bounded).

Remark 4.55. In practice, to use this theorem it is very convenient to recall that $M_n = X_n^2 - \sum_{i=1}^{n} V_i$ is a martingale, as it gives a way of evaluating $\mathbb{E}(X_n^2)$: the assumption of the theorem is satisfied as soon as the conditional variances have summable expectations, i.e.,

$$\sum_{n=1}^{\infty} \mathbb{E}(V_n) < \infty. \tag{4.19}$$

This can be useful to (somewhat) demistify the proof of Theorem 4.53: essentially, to prove convergence in $L^2(\mathbb{P})$, we find a good martingale (M) associated to X_n^2. The fact M is a martingale shows that X is bounded in $L^2(\mathbb{P})$ if and only if the conditional variances of the increments are so small that their expectation is summable, i.e., (4.19) holds. It should not be too surprising that, once the variances of the increments are small, the whole sequence $(X_n)_{n \geq 0}$ is Cauchy and hence converges.

In the applications below it will be essential to check that some martingale is bounded in $L^2(\mathbb{P})$. We will do so by checking (4.19). We **warn the reader** that there is no tower property for conditional variance: while $V_n = \mathrm{Var}(X_n | \mathcal{F}_{n-1})$, it is not true in general that $\mathbb{E}(V_n) = \mathrm{Var}(X_n)$.

4.9 Application 1: exponential growth of branching processes

As an example of application of Doob's martingale convergence theorem, let us now return to the example of branching processes. Recall once again that a branching process $(Z_n)_{n\geq 0}$ is determined by an offspring distribution $(p_k)_{k\geq 0}$ on $\mathbb{N} = \{0, 1, \ldots\}$ where p_k is the probability for any given individual to have k offsprings in the next generation, and offspring numbers are independent random variables. Thus, $Z_0 = 1$ (initially there is one individual in this population) and given $(Z_1, \ldots, Z_n)$, we obtain Z_{n+1} as

$$Z_{n+1} = \sum_{i=1}^{Z_n} \xi_i,$$

where $(\xi_i)_{i\geq 1}$ are independent of $(Z_1, \ldots, Z_n)$ and are i.i.d. with common law $(p_k)_{k\geq 0}$.

We have already stated (and partly proved) in Theorem 2.18 that the survival/extinction dichotomy depends crucially on the mean number of offspring, $m = \sum_{k=0}^{\infty} kp_k$, also known as reproductive number in an epidemiological context. We will give a different proof of this result here based on martingales. But more importantly, we will be able to show that when the process survives, it grows geometrically (i.e., exponentially) fast, at rate m: not only does the branching process survives in this case, but it grows extremely quickly. This is stated in the following theorem, in which we make the additional assumption that the offspring distribution has finite variance.

Theorem 4.56. *Let $(Z_n)_{n\geq 0}$ be the above branching process, and suppose $\sigma^2 = \mathrm{Var}(\xi) < \infty$, where ξ has the distribution $(p_k)_{k\geq 0}$ on $\mathbb{N} = \{0, 1, \ldots\}$. Then the following dichotomy holds:*

- *If $m \leq 1$, with $p_1 \neq 1$, then $(Z_n)_{n\geq 0}$ becomes extinct almost surely: there exists n_0 (random but finite almost surely) such that $Z_n = 0$ for $n \geq n_0$.*
- *However if $m > 1$ then $\mathbb{P}(Z_n > 0 \text{ for all } n \geq 0) > 0$. Furthermore, Z_n/m^n converges almost surely to a limit W which satisfies $\mathbb{E}(W) = 1$, hence $W > 0$ with positive probability.*

Proof. The proof of this theorem relies entirely on the martingale identified in Theorem 4.10. Recall that this theorem states that if

$$M_n = \frac{Z_n}{m^n},$$

then $(M_n)_{n\geq 0}$ is a martingale.

Let us first prove extinction in the case $m \leq 1$ and $p_1 \neq 1$, for which the proof is slick and elegant. Since M is a martingale and $m \leq 1$, note that $(Z_n)_{n\geq 0}$ is a supermartingale (as $\mathbb{E}(Z_{n+1}|\mathcal{F}_n) = mZ_n \leq Z_n$ for any $n \geq 0$). Furthermore, $Z_n \geq 0$ and so we may apply the first Doob convergence theorem (more precisely, Corollary 4.49): therefore, $(Z_n)_{n\geq 0}$ converges almost surely to a limit, call it L. Since $Z_n \in \mathbb{N}$ is integer-valued, we deduce (as in Example 4.51) that Z_n must be constant from some point onwards, equal to its limit L:

that is, there exists n_0 (random but finite almost surely) such that $Z_n = L$ for all $n \geq n_0$. There are various ways to conclude from here: Now, because we have assumed $p_1 \neq 1$, at each step there is always a positive probability that $Z_{n+1} \neq Z_n$, unless $Z_n = 0$. Thus the only way for Z_n to be eternally constant from some point onwards, is if $Z_n = 0$ is extinct. This shows that extinction takes place at a finite time, almost surely.

Now let us assume that $m > 1$ and prove the second point. In fact, it suffices to prove that

$$\mathbb{E}(M_n^2) \leq K \tag{4.20}$$

is bounded: indeed, given (4.20) we may apply the second Doob convergence theorem (Theorem 4.53) to deduce that M_n converges almost surely and in L^2 to a limit W, which satisfies $\mathbb{E}(W) = 1$. As $W = \lim_{n\to\infty} Z_n/m^n$, clearly $W \geq 0$. Since we also have $\mathbb{E}(W) = 1$, it follows that W is not identically zero and hence $W > 0$ with positive probability. The fact that Z_n/m^n converges to a limit which is strictly positive on an event E of positive probability shows that, at least on the event E, Z_n is nonzero for all $n \geq 0$ and hence E implies survival.

Thus, it suffices to prove (4.20). As indicated in Remark 4.55, we do so by bounding from above the conditional variance of M_n. Let $V_n = \mathbb{E}(\Delta_n^2|\mathcal{F}_{n-1})$ and recall that V_n can be interpreted as the conditional variance of M_n given $\mathcal{F}_{n-1}$: that is, we compute the variance of M_n given all the information in $\mathcal{F}_{n-1}$ (using the same rules we might use to compute a conditional expectation). By Remark 4.55, in order to show (4.20) it suffices to show

$$\sum_{i=1}^{\infty} \mathbb{E}(V_i) < \infty. \tag{4.21}$$

Let us compute V_n. We have:

$$\begin{aligned} V_n &= \mathrm{Var}(M_n|\mathcal{F}_{n-1}) = \mathrm{Var}(\frac{Z_n}{m^n}|\mathcal{F}_{n-1}) \\ &= \frac{1}{m^{2n}} \mathrm{Var}(Z_n|\mathcal{F}_{n-1}) \text{ (by scaling)} \\ &= \frac{1}{m^{2n}} \mathrm{Var}(\sum_{i=1}^{Z_{n-1}} \xi_i|\mathcal{F}_{n-1}). \end{aligned}$$

Given $\mathcal{F}_{n-1}$, Z_{n-1} is known and so we need to compute the variance of the sum of a finite number of i.i.d. random variables. Using the elementary fact that $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ when X and Y are independent, we deduce that

$$V_n = \frac{1}{m^{2n}} \sum_{i=1}^{Z_{n-1}} \mathrm{Var}(\xi_i|\mathcal{F}_{n-1}).$$

The ξ_i are independent of $\mathcal{F}_{n-1}$ and so the conditional variance $\mathrm{Var}(\xi_i|\mathcal{F}_{n-1})$ is simply the unconditional variance, i.e., σ^2. Thus

$$V_n = \frac{1}{m^{2n}} Z_{n-1}\sigma^2.$$

Hence

$$\mathbb{E}(V_n) = \frac{\sigma^2}{m^{2n}}\mathbb{E}(Z_{n-1}) = \frac{\sigma^2}{m^{n+1}}.$$

Since $m > 1$, we deduce that $\sum_{n=1}^{\infty} \mathbb{E}(V_n) < \infty$. This proves (4.21) and hence concludes the proof. □

4.10 Application 2: branching random walk

As our next application, we will describe another system of particles known as branching random walks. This can be thought of as a (very interesting) generalisation of the branching processes studied in some details in these notes, but in which we also take into account spatial effects. We suppose given a locally finite graph G and an offspring distribution $(p_k)_{k\geq 0}$ on $\mathbb{N}$. We start the process with a single particle (or individual) at time $n = 0$, located in some pre-determined vertex of the graph. As before, the process is defined inductively: given all the particle locations at generation n, each particle gives rise to a random number of offsprings, distributed according to $(p_k)_{k\geq 0}$. If the parent's location is at $x \in V$, then each of the children's location is given by $y \in V$ with probability $P(x, y)$, where P is the transition probability of the random walk on G. In other words, the children's location is obtained from that of the parent by taking a random walk step. As always, all the displacements and the number of offsprings are assumed to be independent of one another. (A more formal definition will be given below). Thus the total size of the population at generation n is given by a branching process, but the particles now have an interesting, nontrivial distribution across the vertices of the graph.

To understand why such a system of particles is natural, one may think again of the growth of a population. For instance, a plant will usually place its offsprings in nearby locations rather than very far away. Alternatively, in an epidemiological context, an individual tends to infect its friends or contacts, who usually live nearby rather than far away.

In such a system we are not just interested in the total population size (which, as already mentioned, is simply a branching process) but in how the particles are distributed across the graph. How far away from the starting point can you find particles after n generations, when the populations survive? How many particles will be on a given site? Where is the center of mass of the population? Many of these questions, although very basic, have only been understood very recently (say less than ten years ago) and are still the subject of intense research.

Let us define the branching random walk a bit more formally. Again, we fix a locally finite graph $G = (V, E)$ and an offspring distribution $(p_k)_{k\geq 0}$ on $\mathbb{N} = \{0, 1, \ldots\}$. This will require some slightly tedious notations (but hopefully the above intuition will help). This will be defined inductively. We will have for each generation $n \geq 0$ a set of particles, with Z_n many particles in total, and whose locations are given by $\mathcal{Z}_n = (X_{n,1}, \ldots, X_{n,Z_n})$, where for each $1 \leq i \leq Z_n$, $X_{n,i}$ is a vertex of the graph (the particles are labelled from $i = 1$ to $i = Z_n$ in some arbitrary order). The label n will be reserved for the generation in the branching process ("time"), while the label i will be reserved for the index of the ith particle in some given generation n. We will reserve a thid label, j, for the location of the jth child

of the ith particle in some generation n. The branching random walk is defined inductively as follows:

Definition 4.57. *Initially we have $Z_0 = 1$ and the unique particle comprising the population at time $n = 0$ is located at $X_{0,1} = u$, for some given vertex u of the graph. Now suppose given the process $\mathcal{Z}_n = (X_{n,1}, \dots, X_{n,Z_n})$ at time $n \geq 0$. For each $i = 1, \dots, Z_n$, let $\xi_{n,i}$ denote independent random variables having the law $(p_k)_{k\geq 0}$ (thus $\xi_{n,i}$ is the number of offsprings of individual i in generation n). Given $\xi_{n,i}$, let $(Y_{n,i,j})_{1\leq j\leq \xi_{n,i}}$ be independent random variables with*

$$\mathbb{P}(Y_{n,i,j} = w | X_{n,i} = v) = P(v, w),$$

where $P(v, w)$ is the transition matrix of simple random walk on G. The position $Y_{n,i,j} \in V$ is the location of the jth child of the ith particle in generation n. We then set

$$\mathcal{Z}_{n+1} = (Y_{n,i,j})_{1\leq i\leq Z_n, 1\leq j\leq \xi_{n,i}}. \tag{4.22}$$

We emphasise again that in this definition, we take the $(Y_{n,i,j})_{1\leq i\leq n, 1\leq j\leq \xi_{n,i}}$ to be independent of one another and independent of $(\mathcal{Z}_0, \dots, \mathcal{Z}_n)$.

To make things somewhat more interesting, we will add another ingredient to this definition. Namely, we will fix a finite subset A of vertices such that the initial ancestor's location $u \in A$, and declare that every particle falling outside of A is immediately killed. Formally speaking, this means that, compared to (4.22), the index j is only running over $1 \leq j \leq \xi_{n,i}$ such that $Y_{n,i,j} \in A$:

$$\mathcal{Z}'_{n+1} = (Y_{n,i,j})_{1\leq i\leq Z_n, 1\leq j\leq \xi_{n,i}: Y_{n,i,j}\in A}. \tag{4.23}$$

The resulting process $(\mathcal{Z}'_n)_{n\geq 0}$ is called the **branching random walk killed** on A^c. From the point of view of interpretation, you could think that A represents an island and the plant population cannot grow beyond A. Alternatively, in an epidemiological context, you could imagine that A represents a region that has not yet been completely vaccinated against an infection, while A^c represents a region in which everyone is vaccinated. See Figure 9 for an illustration.

Note that the total population size $Z'_n = \#\mathcal{Z}'_n$ is now no longer simply a branching process, because of the fact that some particles are killed when they leave A. Because of this, it is unclear under what conditions the killed branching random walk survives. Clearly, using Theorem 4.56, survival requires $m = \sum_{k=0}^{\infty} kp_k$ to be greater than 1, otherwise the entire branching random walk dies out (even if we don't kill particles when they leave A). So does it suffice that $m > 1$? We might initially be tempted to say yes; indeed if $m > 1$ we know by Theorem 4.56 that the population size grows exponentially. If we start from the center which is far away from the boundary of A, it will take time before even a single particle dies off and by that time we should already have a very large number of particles. So we might expect the killing to play very little role in the survival of the process. In fact, this is not the case, as demonstrated by the following result:

Theorem 4.58. *Let $(\mathcal{Z}'_n)_{n\geq 0}$ be the above killed branching random walk, and suppose $\sigma^2 = \mathrm{Var}(\xi) < \infty$, where ξ has the distribution $(p_k)_{k\geq 0}$ on $\mathbb{N} = \{0, 1, \dots\}$. Let λ be a principal eigenvalue of A, in the sense of Definition 4.38. Then the following dichotomy holds:*

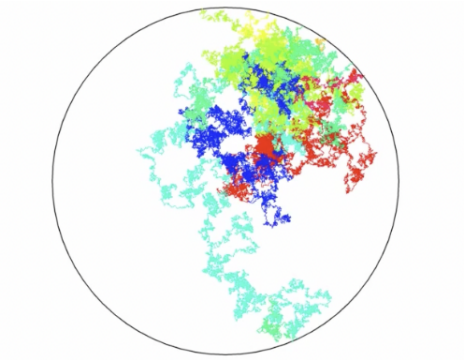


Figure 9: An example of a branching random walk on a graph which is the square lattice with fine mesh size. Different colours represent different particles. Particles are killed when they leave the unit disc.

- *If $m < 1/\lambda$, then $(\mathcal{Z}'_n)_{n\geq 0}$ becomes extinct almost surely: there exists n_0 (random but finite almost surely) such that $Z'_n = \#\mathcal{Z}'_n = 0$ for $n \geq n_0$.*
- *However if $m > 1/\lambda$ then $\mathbb{P}(Z_n > 0 \text{ for all } n \geq 0) > 0$.*

Thus the critical threshold for survival is not $m = 1$ but $m = 1/\lambda$. It is not hard to see (e.g., using Theorem 4.41) that $\lambda < 1$. Thus the threshold for survival is strictly greater than 1. To put this is an epidemiological context, if the world outside of A is fully vaccinated against the infection, then the reproductive number m can be greater than one and yet the epidemics will die out (so long as it is not greater than $1/\lambda$)! The proof we give below does not cover the case $m\lambda = 1$, but pushing the arguments just a little further would show that the killed branching random walk also dies out at the critical value.

Proof. As always, the key is to identify a suitable martingale. Since our result involves the principal eigenvalue λ, it is natural to seek a martingale that involves it and the associated principal eigenfunction f. Recall that f is positive on A and satisfies

$$Pf(x) = \lambda f(x).$$

Taking some inspiration from Theorem 4.41, we propose

$$M_n = \frac{1}{(m\lambda)^n} \sum_{i=1}^{Z'_n} f(X_{n,i}). \tag{4.24}$$

Lemma 4.59. *$M = (M_n)_{n\geq 0}$ is a martingale.*

Proof of Lemma 4.59. Using the fact that f is bounded above (as A is finite) it is not hard to see that M is integrable, and it is clearly adapted. Let us check the martingale property:

we have

$$\mathbb{E}(M_{n+1}|\mathcal{F}_n) = \mathbb{E}(\frac{1}{(m\lambda)^{n+1}} \sum_{i=1}^{Z_n'} \sum_{j=1}^{\xi_{n,i}} f(Y_{n,i,j}) 1_{\{Y_{n,i,j} \in A\}} | \mathcal{F}_n)$$
$$= \frac{1}{(m\lambda)^{n+1}} \sum_{i=1}^{Z_n'} \mathbb{E}(\sum_{j=1}^{\xi_{n,i}} f(Y_{n,i,j}) | \mathcal{F}_n) \tag{4.25}$$

Now to compute the last conditional expectation, $\xi_{n,i}$ is independent of $\mathcal{F}_n$. Thus the conditional expectation is equal to an (unconditional) expectation of the form

$$\mathbb{E}(\sum_{j=1}^{N} f(Y_j)),$$

where the Y_j are i.i.d., having law $P(X_{n,i}, \cdot)$ and independent of N. We compute this expectation by conditioning on N and applying the tower property: find

$$\mathbb{E}(\sum_{j=1}^{N} f(Y_j)|N) = N\mathbb{E}(f(Y)),$$

so

$$\mathbb{E}(\sum_{j=1}^{N} f(Y_j)) = \mathbb{E}(N)\mathbb{E}(f(Y)).$$

Here $\mathbb{E}(N)$ is simply the expected number of offsprings, m, and $\mathbb{E}(f(Y))$ is simply $Pf(X_{n,i})$. Since f is an eigenfunction with eigenvalue λ, we see that

$$\mathbb{E}(\sum_{j=1}^{\xi_{n,i}} f(Y_{n,i,j})|\mathcal{F}_n) = m\lambda f(X_{n,i}).$$

Plugging back into (4.25), we deduce

$$\mathbb{E}(M_{n+1}|\mathcal{F}_n) = \frac{1}{(m\lambda)^{n+1}} \sum_{i=1}^{Z_n'} m\lambda f(X_{n,i}) = M_n,$$

as desired. □

Let us see why the fact that M is a martingale implies the result. Note that M is nonnegative, hence by Doob's convergence theorem (more precisely, Corollary 4.49) converges almost surely to a (finite) limit, call it W.

Suppose first that $m\lambda < 1$. Then

$$\sum_{i=1}^{Z_n'} f(X_{i,n}) = (m\lambda)^n M_n.$$

Let $\varepsilon = \min_{x\in A} f(x)$. Then

$$\varepsilon Z'_n \leq (m\lambda)^n M_n. \tag{4.26}$$

Since M_n converges to a finite limit and $m\lambda < 1$, we see that Z'_n converges to 0. But Z'_n is integer valued, so Z'_n must eventually be equal to zero.

Now suppose that $m\lambda > 1$. To prove survival it suffices to show (just as in Theorem 4.56) that M is bounded in L^2. As indicated in Remark 4.55, the most convenient way is to compute the conditional variance of M_n, take the expectation and show that this is summable. We sketch the argument here: let

$$V_{n+1} = \mathrm{Var}(M_{n+1}|\mathcal{F}_n).$$

Clearly,

$$V_{n+1} = \frac{1}{(m\lambda)^{2n+2}} \sum_{i=1}^{Z'_n} \mathrm{Var}(\sum_{j=1}^{\xi_{n,i}} f(Y_{n,i,j})|\mathcal{F}_n). \tag{4.27}$$

We could compute the conditional variance on the right hand side of (4.27), but in fact all we need is the fact that this conditional variance is uniformly bounded by a constant, call it C. (This follows from the fact that f is uniformly bounded and the offspring distribution has finite variance.)

We deduce that

$$V_{n+1} \leq \frac{1}{(m\lambda)^{2n+2}} C Z'_n.$$

Taking expectations,

$$\mathbb{E}(V_{n+1}) \leq \frac{C\mathbb{E}(Z'_n)}{(m\lambda)^{2n+2}}.$$

Using (4.26),

$$\mathbb{E}(V_{n+1}) \leq \frac{C}{\varepsilon(m\lambda)^{n+2}}.$$

Since $m\lambda > 1$, we see that the right hand side is summable and hence

$$\sum_{n=1}^{\infty} \mathbb{E}(V_n) < \infty.$$

This completes the proof that M is bounded in $L^2(\mathbb{P})$. □

References

[BLPS18] Nathanaël Berestycki, Eyal Lubetzky, Yuval Peres, and Allan Sly. Random walks on the random graph. *The Annals of Probability*, 46(1):456–490, 2018.

[KV83] Vadim A Kaimanovich and Anatoly M Vershik. Random walks on discrete groups: boundary and entropy. *The annals of probability*, 11(3):457–490, 1983.

[LP17] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[Nor98] James R Norris. *Markov chains*. Cambridge university press, 1998.

[PBMW99] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[Var85] Nicholas Varopoulos. Long range estimates for Markov chains. *Bulletin des sciences mathématiques (Paris. 1885)*, 109(3):225–252, 1985.

Index