Introductory lecture notes on

MARKOV CHAINS AND RANDOM WALKS

Takis Konstantopoulos[*]

Autumn 2009

Contents

Preface

These notes were written based on a number of courses I taught over the years in the U.S., Greece and the U.K. They form the core material for an undergraduate course on Markov chains in discrete time. There are, of course, dozens of good books on the topic. The only new thing here is that I give emphasis to probabilistic methods as soon as possible. Also, I introduce stationarity before even talking about state classification. I tried to make everything as rigorous as possible while maintaining each step as accessible as possible. The notes should be readable by someone who has taken a course in introductory (non-measure-theoretic) probability.

The first part is about Markov chains and some applications. The second one is specifically for simple random walks. Of course, one can argue that random walk calculations should be done before the student is exposed to the Markov chain theory. I have tried both and prefer the current ordering. At the end, I have a little mathematical appendix.

There notes are still incomplete. I plan to add a few more sections:

– On algorithms and simulation

– On criteria for positive recurrence

– On doing stuff with matrices

– On finance applications

– On a few more delicate computations for simple random walks

– Reshape the appendix

– Add more examples

These are things to come...

A few starred sections should be considered "advanced" and can be omitted at first reading. I tried to put all terms in blue small capitals whenever they are first encountered. Also, "important" formulae are placed inside a coloured box.

PART I: MARKOV CHAINS

1 Introduction

A Markov chain is a mathematical model of a random phenomenon evolving with time in a way that the past affects the future only through the present. The “time” can be discrete (i.e. the integers), continuous (i.e. the real numbers), or, more generally, a totally ordered set. We are herein constrained by the Syllabus to discuss only discrete-time Markov chains. In the module following this one you will study continuous-time chains.

In Mathematics, a phenomenon which evolves with time in a way that only the present affects the future is called a DYNAMICAL SYSTEM.

An example from Arithmetic of a dynamical system in discrete time is the one which finds the greatest common divisor $\gcd(a, b)$ between two positive integers a and b. Recall (from elementary school maths), that $\gcd(a, b) = \gcd(r, b)$, where $r = \text{rem}(a, b)$ is the remainder of the division of a by b. By repeating the procedure, we end up with two numbers, one of which is 0, and the other the greatest common divisor. Formally, we let our “state” be a pair of integers (x_n, y_n), where $x_n \geq y_n$, initialised by

$$x_0 = \max(a, b), \quad y_0 = \min(a, b),$$

and evolving as

$$\begin{aligned} x_{n+1} &= y_n \\ y_{n+1} &= \text{rem}(x_n, y_n), \quad n = 0, 1, 2, \ldots \end{aligned}$$

In other words, there is a function F that takes the pair $X_n = (x_n, y_n)$ into $X_{n+1} = (x_{n+1}, y_{n+1})$. The sequence of pairs thus produced, $X_0, X_1, X_2, \ldots$ has the property that, for any n, the future pairs $X_{n+1}, X_{n+2}, \ldots$ depend only on the pair X_n.

An example from Physics is provided by Newton’s laws of motion. Suppose that a particle of mass m is moving under the action of a force F in a straight line. For instance, the particle is suspended at the end of a spring and force is produced by extending the spring and letting the system oscillate (Hooke’s law). Newton’s second law says that the acceleration is proportional to the force:

$$m\ddot{x} = F.$$

Here, $x = x(t)$ denotes the position of the particle at time t. Its velocity is $\dot{x} = \frac{dx}{dt}$, and its acceleration is $\ddot{x} = \frac{d^2x}{dt^2}$. Assume that the force F depends only on the particle’s position and velocity, i.e. $F = f(x, \dot{x})$. The state of the particle at time t is described by the pair $X(t) = (x(t), \dot{x}(t))$. It is not hard to see that, for any t, the future trajectory $(X(s), s \geq t)$

can be completely specified by the current state $X(t)$. In other words, past states are not needed.

Markov chains generalise this concept of dependence of the future only on the present. The generalisation takes us into the realm of Randomness. We will be dealing with random variables, instead of deterministic objects.

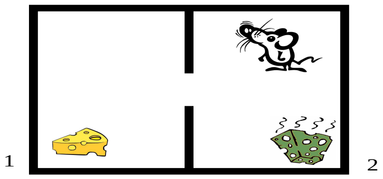
Other examples of dynamical systems are the algorithms run, say, by the software in your computer. Some of these algorithms are deterministic, but some are stochastic.[1] They are stochastic either because the problem they are solving is stochastic or because the problem is deterministic but "very large" such as finding the determinant of a matrix with 10,000 rows and 10,000 columns or computing the integral of a complicated function of a large number of variables. Indeed, an effective way for dealing with large problems is via the use of randomness, i.e. via the use of the tools of Probability.

We will develop a theory that tells us how to describe, analyse, and use those mathematical models which are called Markov chains. We will also see why they are useful and discuss how they are applied. In addition, we will see what kind of questions we can ask and what kind of answers we can hope for.

2 Examples of Markov chains

2.1 A mouse in a cage

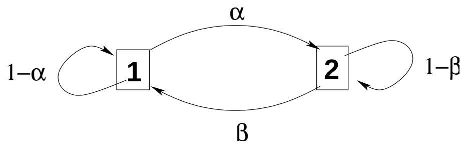
A mouse is in a cage with two cells, 1 and 2, containing fresh and stinky cheese, respectively. A mouse lives in the cage. A scientist's job is to record the position of the mouse every minute. When the mouse is in cell 1 at time n (minutes) then, at time $n+1$ it is either still in 1 or has moved to 2.



Statistical observations lead the scientist to believe that the mouse moves from cell 1 to cell 2 with probability $\alpha = 0.05$; it does so, regardless of where it was at earlier times. Similarly, it moves from 2 to 1 with probability $\beta = 0.99$.

We can summarise this information by the TRANSITION DIAGRAM:

[1] Stochastic means random. An example of a random algorithm is the MONTE CARLO method for the approximate computation of the integral $\int_a^b f(x)dx$ of a complicated function f: suppose that f is positive and bounded below B. Choose a pair (X_0, Y_0) of random numbers, $a \le X_0 \le b$, $0 \le Y_0 \le B$, uniformly. Repeat with (X_n, Y_n), for steps $n = 1, 2, \ldots$. Each time count 1 if $Y_n < f(X_n)$ and 0 otherwise. Perform this a large number N of times. Count the number of 1's and divide by N. The resulting ratio is an approximation of $\int_a^b f(x)dx$. Try this on the computer!



Another way to summarise the information is by the 2×2 TRANSITION PROBABILITY MATRIX

$$\mathsf{P} = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix} = \begin{pmatrix} 0.95 & 0.05 \\ 0.99 & 0.01 \end{pmatrix}$$

Questions of interest:
1. How long does it take for the mouse, on the average, to move from cell 1 to cell 2?
2. How often is the mouse in room 1 ?
Question 1 has an easy, intuitive, answer: Since the mouse really tosses a coin to decide whether to move or stay in a cell, the first time that the mouse will move from 1 to 2 will have mean $1/\alpha = 1/0.05 \approx 20$ minutes. (This is the mean of the binomial distribution with parameter α.)

2.2 Bank account

The amount of money in Mr Baxter's bank account evolves, from month to month, as follows:

$$X_{n+1} = \max\{X_n + D_n - S_n, 0\}.$$

Here, X_n is the money (in pounds) at the beginning of the month, D_n is the money he deposits, and S_n is the money he wants to spend. So if he wishes to spend more than what is available, he can't.

Assume that D_n, S_n, are random variables with distributions F_D, F_S, respectively. Assume also that $X_0, D_0, S_0, D_1, S_1, D_2, S_2, \ldots$ are independent random variables.

The information described above by the evolution of the account together with the distributions for D_n and S_n leads to the (one-step) transition probability which is defined by:

$$p_{x,y} := P(X_{n+1} = y | X_n = x), \quad x, y = 0, 1, 2, \ldots$$

We easily see that if $y > 0$ then

$$\begin{aligned} p_{x,y} = P(D_n - S_n = y - x) &= \sum_{z=0}^{\infty} P(S_n = z, D_n - S_n = y - x) \\ &= \sum_{z=0}^{\infty} P(S_n = z, D_n = z + y - x) \\ &= \sum_{z=0}^{\infty} P(S_n = z) P(D_n = z + y - x) \\ &= \sum_{z=0}^{\infty} F_S(z) F_D(z + y - x), \end{aligned}$$

something that, in principle, can be computed from F_S and F_D. Of course, if $y = 0$, we have $p_{x,0} = 1 - \sum_{y=1}^{\infty} p_{x,y}$. The transition probability matrix is

$$\mathsf{P} = \begin{pmatrix} p_{0,0} & p_{0,1} & p_{0,2} & \cdots \\ p_{1,0} & p_{1,1} & p_{1,2} & \cdots \\ p_{2,0} & p_{2,1} & p_{2,2} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

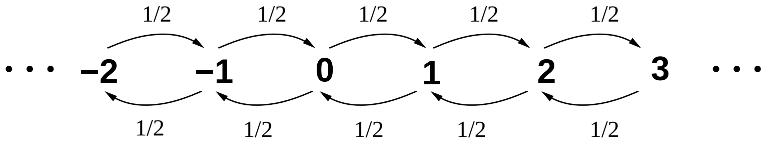
and is an infinite matrix.

Questions of interest:
1. Will Mr Baxter's account grow, and if so, how fast?
2. How long will it take (if ever) for the bank account to empty?

We shall develop methods for answering these and other questions.

2.3 Simple random walk (drunkard's walk)

Consider a completely drunk person who walks along a street. being drunk, he has no sense of direction. So he may move forwards with equal probability that he moves backwards.

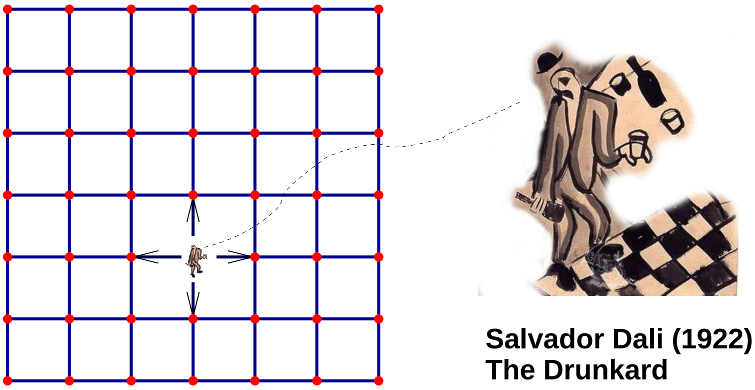


Questions:
1. If he starts from position 0, how often will he be visiting 0?
2. Are there any places which he will never visit?
3. Are there any places which he will visit infinitely many times?
4. If the pub is located in position 0 and his home in position 100, how long will it take him, on the average, to go from the pub to home?

You might object to the assumption that the person is totally drunk and has no sense of direction. We may want to model such a case by assigning probability p for the drunkard to go one step to the right (and probability $1 - p$ for him to go one step to the left).

2.4 Simple random walk (drunkard's walk) in a city

Suppose that the drunkard is allowed to move in a city whose streets are laid down in a square pattern:



Salvador Dali (1922)
The Drunkard

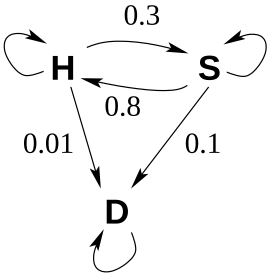
Suppose that the drunkard moves from corner to corner. So he can move east, west, north or south, and let's say he's completely drunk, so we assign probability 1/4 for each move.

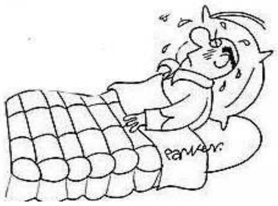
We may again want to ask similar questions. But, observe, that since there are more degrees of freedom now, there is clearly a higher chance that our man will be lost.

These things are, for now, mathematically imprecise, but they will become more precise in the sequel.

2.5 Actuarial chains

A life insurance company wants to find out how much money to charge its clients. Clearly, the company must have an idea on how long the clients will live. It proposes the following model summarising the state of health of an individual on a monthly basis:





Thus, there is probability $p_{H,S} = 0.3$ that the person will become sick, given that he is currently healthy, etc. Note that the diagram omits $p_{H,H}$, and $p_{S,S}$ because $p_{H,H} = 1 - p_{H,S} - p_{H,D}$, and $p_{S,S} = 1 - p_{S,H} - p_{S,D}$. Also, $p_{D,D}$ is omitted, the reason being that the company does not believe that its clients are subject to resurrection; therefore, $p_{D,D} = 1$.

Question: What is the distribution of the lifetime (in months) of a currently healthy individual?

Clearly, the answer to this is crucial for the policy determination.

3 The Markov property

3.1 Definition of the Markov property

Consider a (finite or infinite) sequence of random variables $\{X_n, n \in \mathsf{T}\}$, where T is a subset of the integers. We say that this sequence has the MARKOV PROPERTY if,

for any $n \in \mathsf{T}$,
the future process $(X_m, m > n, m \in \mathsf{T})$
is independent of
the past process $(X_m, m < n, m \in \mathsf{T})$,
conditionally on X_n.

Sometimes we say that $\{X_n, n \in \mathsf{T}\}$ is Markov instead of saying that it has the Markov property.

Usually, $\mathsf{T} = \mathbb{Z}_+ = \{0, 1, 2, 3, \ldots\}$ or $\mathbb{N} = \{1, 2, 3, \ldots\}$ or the set $\mathbb{Z}$ of all integers (positive, zero, or negative). We focus almost exclusively to the first choice. So we will omit referring to T explicitly, unless needed. We will also assume, almost exclusively (unless otherwise said explicitly), that the X_n take values in some countable set S, called the STATE SPACE. The elements of S are frequently called STATES.

Since S is countable, it is customary to call (X_n) a MARKOV CHAIN.

Let us now write the Markov property in an equivalent way.

Lemma 1. *$(X_n, n \in \mathbb{Z}_+)$ is Markov if and only if, for all $n \in \mathbb{Z}_+$ and all $i_0, \ldots, i_{n+1} \in S$,*

$$P(X_{n+1} = i_{n+1} \mid X_n = i_n, \ldots, X_0 = i_0) = P(X_{n+1} = i_{n+1} \mid X_n = i_n).$$

Proof. If (X_n) is Markov then the claim holds by the conditional independence of X_{n+1} and $(X_0, \ldots, X_{n-1})$ given X_n. On the other hand, if the claim holds, we can see that by applying it several times we obtain

$$P(\forall\ k \in [n+1, n+m]\ X_k = i_k \mid X_n = i_n, \ldots, X_0 = i_0) = \\ P(\forall\ k \in [n+1, n+m]\ X_k = i_k \mid X_n = i_n),$$

for any n, m and any choice of states $i_0, \ldots, i_{n+m}$, which precisely means that

$(X_{n+1}, \ldots, X_{n+m})$ and $(X_0, \ldots, X_{n-1})$ are independent conditionally on X_n.

This is true for any n and m, and so future is independent of past given the present, viz. the Markov property. □

3.2 Transition probability and initial distribution

We therefore see that joint probabilities can be expressed as

$$\begin{aligned} &P(X_0 = i_0, X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n) \\ &= P(X_0 = i_0) P(X_1 = i_1 \mid X_0 = i_0) P(X_2 = i_2 \mid X_1 = i_1) \cdots P(X_n = i_n \mid X_{n-1} = i_{n-1}), \end{aligned} \tag{1}$$

which means that the two functions

$$\mu(i) := P(X_0 = i) \quad , \quad i \in S$$
$$p_{i,j}(n, n+1) := P(X_{n+1} = j \mid X_n = i) \quad , \quad i, j \in S, \quad n \geq 0,$$

will specify all joint distributions[2]:

$$P(X_0 = i_0, X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n) \\ = \mu(i_0)\ p_{i_0,i_1}(0,1)\ p_{i_1,i_2}(1,2)\ \cdots\ p_{i_{n-1},i_n}(n-1,n).$$

The function $\mu(i)$ is called initial distribution. The function $p_{i,j}(n, n+1)$ is called transition probability from state i at time n to state j at time $n+1$. More generally,

$$p_{i,j}(m, n) := P(X_n = j \mid X_m = i)$$

is the transition probability from state i at time m to state j time $n \geq m$. Of course,

$$p_{i,j}(n, n) = \mathbf{1}(i = j)$$

and, by (1),

$$p_{i,j}(m, n) = \sum_{i_1 \in S} \cdots \sum_{i_{n-1} \in S} p_{i,i_1}(m, m+1) \cdots p_{i_{n-1},i}(n-1, n), \tag{2}$$

which is reminiscent of matrix multiplication. Therefore, remembering a bit of Algebra, we define, for each $m \leq n$ the matrix

$$\mathsf{P}(m, n) := [p_{i,j}(m, n)]_{i,j \in S},$$

viz., a square matrix whose rows and columns are indexed by S (and this, of course, requires *some* ordering on S) and whose entries are $p_{i,j}(m, n)$. If $|S| = d$ then we have $d \times d$ matrices, but if S is an infinite set, then the matrices are infinitely large.[3] In this new notation, we can write (2) as

$$\mathsf{P}(m, n) = \mathsf{P}(m, m+1)\mathsf{P}(m+1, m+2) \cdots \mathsf{P}(n-1, n),$$

or as

$$\mathsf{P}(m, n) = \mathsf{P}(m, \ell)\ \mathsf{P}(\ell, n) \quad \text{if } m \leq \ell \leq n,$$

something that can be called semigroup property or Chapman-Kolmogorov equation.

3.3 Time-homogeneous chains

We will specialise to the case of time-homogeneous chains, which, by definition, means that

the transition probabilities $p_{i,j}(n, n+1)$ do not depend on n ,

[2] And, by a result in Probability, all probabilities of events associated with the Markov chain

[3] This causes some analytical difficulties. But we will circumvent them by using Probability only.

and therefore we can simply write

$$p_{i,j}(n, n+1) =: p_{i,j} \;,$$

and call $p_{i,j}$ the (ONE-STEP) TRANSITION PROBABILITY from state i to state j, while

$$\mathsf{P} = [p_{i,j}]_{i,j \in S}$$

is called the (ONE-STEP) TRANSITION PROBABILITY MATRIX or, simply, TRANSITION MATRIX. Then

$$\mathsf{P}(m, n) = \underbrace{\mathsf{P} \times \mathsf{P} \times \cdots\cdots \times \mathsf{P}}_{n-m \text{ times}} = \mathsf{P}^{n-m},$$

for all $m \leq n$, and the semigroup property is now even more obvious because

$$\mathsf{P}^{a+b} = \mathsf{P}^a \mathsf{P}^b,$$

for all integers $a, b \geq 0$.

From now on, unless otherwise specified, when we say "Markov chain" we will mean "time-homogeneous Markov chain".

The matrix notation is useful. For example, let

$$\boldsymbol{\mu}_n := [\mu_n(i) = P(X_n = i)]_{i \in S}$$

be the distribution of X_n, arranged in a *row vector*. Then we have

$$\boldsymbol{\mu}_n = \boldsymbol{\mu}_0 \mathsf{P}^n \;, \tag{3}$$

as follows from (1) and the definition of P.

Notational conventions

Unit mass. The distribution that assigns value 1 to a state i and 0 to all other states is denoted by δ_i. In other words,

$$\delta_i(j) := \begin{cases} 1, & \text{if } j = i \\ 0, & \text{otherwise.} \end{cases}$$

We call δ_i the unit mass at i. It corresponds, of course, to picking state i with probability 1. Notice that any probability distribution on S can be written as a linear combination of unit masses. For example, if μ assigns probability p to state i_1 and $1 - p$ to state i_2, then

$$\mu = p\delta_{i_1} + (1-p)\delta_{i_2}.$$

Starting from a specific state i. A useful convention used for time-homogeneous chains is to write

$$P_i(A) := P(A \mid X_0 = i).$$

Thus P_i is the law of the chain when the initial distribution is δ_i. Similarly, if Y is a real random variable, we write

$$E_i Y := E(Y \mid X_0 = i) = \sum_y y P(Y = y \mid X_0 = i) = \sum_y y P_i(Y = y).$$

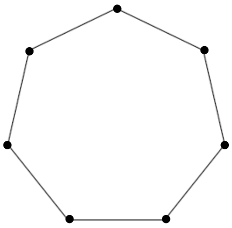
4 Stationarity

We say that the process $(X_n, n = 0, 1, \ldots)$ is STATIONARY if it has the same law as $(X_n, n = m, m+1, \ldots)$, for any $m \in \mathbb{Z}_+$.

In other words, the law of a stationary process does not depend on the origin of time.

Clearly, a sequence of i.i.d. random variables provides an example of a stationary process. We now investigate when a Markov chain is stationary. To provide motivation and intuition, let us look at the following example:

Example: Motion on a polygon. Consider a canonical heptagon and let a bug move on its vertices. If at time n the bug is on a vertex, then, at time $n+1$ it moves to one of the adjacent vertices with equal probability.



It should be clear that if initially place the bug at random on one of the vertices, i.e. selecting a vertex with probability $1/7$, then, at any point of time n, the distribution of the bug will still be the same. Hence the uniform probability distribution is stationary.

Example: the Ehrenfest chain. This is a model of gas distributed between two identical communicating chambers. Imagine there are N molecules of gas (where N is a large number, say $N = 10^{25}$) in a metallic chamber with a separator in the middle, dividing it into two identical rooms. The gas is placed partly in room 1 and partly in room 2. Then a hole is opened in the separator and we watch what happens as the molecules diffuse between the two rooms. It is clear, that if room 1 contains more gas than room 2 then there is a tendency to observe net motion from 1 to 2. We can model this by saying that the chance that a molecule move from room 1 to room 2 is proportional to the number of molecules in room 1, and vice versa. Let X_n be the number of molecules in room 1 at time n. Then

$$p_{i,i+1} = P(X_{n+1} = i+1 \mid X_n = i) = \frac{N-i}{N}$$
$$p_{i,i-1} = P(X_{n+1} = i-1 \mid X_n = i) = \frac{i}{N}.$$

If we start the chain with $X_0 = N$ (all molecules in room 1) then the process $(X_n, n \geq 0)$ will not be stationary: indeed, for a while we will be able to tell how long ago the process started, as it takes some time before the molecules diffuse from room to room. The question then is: can we distribute the initial number of molecules in a way that the origin of time plays no role? One guess might be to take $X_0 = N/2$. But this will not work, because it is impossible for the number of molecules to remain constant at all times. On the average, we indeed expect that we have $N/2$ molecules in each room. Another guess then is: Place each molecule at random either in room 1 or in room 2. If we do so, then the number of molecules in room 1 will be i with probability

$$\pi(i) = \binom{N}{i} 2^{-N}, \quad 0 \le i \le N,$$

(a binomial distribution). We can now verify that if $P(X_0 = i) = \pi(i)$ then $P(X_1 = i) = \pi(i)$ and, hence, $P(X_n = i) = \pi(i)$ for all n. Indeed,

$$\begin{aligned} P(X_1 = i) &= \sum_j P(X_0 = j) p_{j,i} \\ &= P(X_0 = i-1) p_{i-1,i} + P(X_0 = i+1) p_{i+1,i} \\ &= \pi(i-1) p_{i-1,i} + \pi(i+1) p_{i+1,i} \\ &= \binom{N}{i-1} 2^{-N} \frac{N-i+1}{N} + \binom{N}{i+1} 2^{-N} \frac{i+1}{N} = \cdots = \pi(i). \end{aligned}$$

(The dots signify some missing algebra, which you should go through by yourselves.)

The example above is typical: We found (by guessing!) some distribution π with the property that if X_0 has distribution π then every other X_n also has distribution π. If we can do so, then we have created a stationary Markov chain.

But we need to prove this.

Lemma 2. *A Markov chain $(X_n, n \in \mathbb{Z}_+)$ is stationary if and only if it is time-homogeneous and X_n has the same distribution as X_ℓ for all n and ℓ.*

Proof. The only if part is obvious. For the if part, use (1) to see that

$$P(X_0 = i_0, X_1 = i_1, X_2 = i_2, \dots, X_n = i_n) = \\ P(X_m = i_0, X_{m+1} = i_1, X_{m+2} = i_2, \dots, X_{m+n} = i_n),$$

for all $m \ge 0$ and $i_0, \dots, i_n \in S$. □

In other words, a Markov chain is stationary if and only if the distribution of X_n does not change with n. Since, by (3), the distribution μ_n at time n is related to the distribution at time 0 by $\mu_n = \mu_0 \mathsf{P}^n$, we see that the Markov chain is stationary if and only if the distribution μ_0 is chosen so that

$$\mu_0 = \mu_0 \mathsf{P}.$$

Changing notation, we are led to consider:

The stationarity problem: Given $[p_{ij}]$, find those distributions π such that

$$\pi \mathsf{P} = \pi. \tag{4}$$

Such a π is called . STATIONARY OR INVARIANT DISTRIBUTION. The equations (4) are called BALANCE EQUATIONS.

Eigenvector interpretation: The equation $\pi = \pi\mathsf{P}$ says that π is a left eigenvector of the matrix P with eigenvalue 1. In addition to that, π must be a probability:

$$\sum_{i \in S} \pi(i) = 1.$$

The matrix P has the eigenvalue 1 always because $\sum_{j \in S} p_{i,j} = 1$, which, in matrix notation, can be written as $\mathsf{P}\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ is a column whose entries are all 1, and this means that $\mathbf{1}$ is a (right) eigenvector of P corresponding to the eigenvalue 1.

Example: card shuffling. Consider a deck of $d = 52$ cards and shuffle them in the following (stupid) way: pick two cards at random and swap their positions. Keep doing it continuously. We can describe this by a Markov chain which takes values in the set S_d of $d!$ permutations of $\{1, \ldots, d\}$. Thus, each $x \in S_d$ is of the form $x = (x_1, \ldots, x_d)$, where all the x_i are distinct. It is easy to see that the stationary distribution is uniform:

$$\pi(x) = \frac{1}{d!}, \quad x \in S_d.$$

Example: waiting for a bus. After bus arrives at a bus stop, the next one will arrive in i minutes with probability $p(i)$, $i = 1, 2 \ldots$. The times between successive buses are i.i.d. random variables. Consider the following process: At time n (measured in minutes) let X_n be the time till the arrival of the next bus. Then $(X_n, n \geq 0)$ is a Markov chain with transition probabilities

$$\begin{aligned} p_{i+1,i} &= 1, \quad \text{if } i > 0 \\ p_{1,i} &= p(i), \quad i \in \mathbb{N}. \end{aligned}$$

The state space is $S = \mathbb{N} = \{1, 2, \ldots\}$. To find the stationary distribution, write the equations (4):

$$\pi(i) = \sum_{j \geq 1} \pi(j) p_{j,i} = \pi(0)p(i) + \pi(i+1), \quad i \geq 1.$$

These can easily be solved:

$$\pi(i) = \pi(1) \sum_{j \geq i} p(j), \quad i \geq 0.$$

To compute $\pi(1)$, we use the normalisation condition $\sum_{i \geq 1} \pi(i) = 1$, which gives

$$\pi(1) = \left(\sum_{i \geq 1} \sum_{j \geq i} p(j) \right)^{-1}.$$

But here, *we must be careful.* Who tells us that the sum inside the parentheses is finite? If the sum is infinite, then we run into trouble: indeed, $\pi(1)$ will be zero, and so each $\pi(i)$ will be zero. This means that the solution to the balance equations is identically zero, which in turn means that the normalisation condition cannot be satisfied. So we *must make an*:

ASSUMPTION: $\sum_{i \geq 1} \sum_{j \geq i} p(j) < \infty$.

Let us work out the sum to see if this assumption can be expressed in more "physical" terms:

$$\sum_{i\geq 1}\sum_{j\geq i} p(j) = \sum_{j\geq 1}\sum_{i\geq 1} \mathbf{1}(j \geq i)p(j) = \sum_{j\geq 1} jp(j),$$

and the latter sum is the expected time between two successive bus arrivals. So our assumption really means:

ASSUMPTION $\iff$ the expected time between successive bus arrivals is finite.

So, what we have really shown is that this assumption is a necessary and sufficient condition for the existence and uniqueness of a stationary distribution.

Example: bread kneading.* (This is slightly beyond the scope of these lectures and can be omitted.) Consider an infinite binary vector

$$\xi = (\xi_1, \xi_2, \ldots),$$

i.e. $\xi_i \in \{0, 1\}$ for all i, and transform it according to the following rule:

if $\xi_1 = 0$ then shift to the left to obtain $(\xi_2, \xi_3, \ldots)$;
if $\xi_1 = 1$ then shift to the left and flip all bits to obtain $(1 - \xi_2, 1 - \xi_3, \ldots)$.

We have thus defined a mapping φ from binary vectors into binary vectors. The Markov chain is obtained by applying the mapping φ again and again.

$$\xi(n + 1) = \varphi(\xi(n)).$$

Here $\xi(n) = (\xi_1(n), \xi_2(n), \ldots)$ is the state at time n. Note that there is nothing random in the transition mechanism of this chain! The only way to instill randomness is by choosing the initial state at random. So let us do so. You can check that , for any $n = 0, 1, 2, \ldots$, any $r = 1, 2, \ldots$, and any $\varepsilon_1, \ldots, \varepsilon_r \in \{0, 1\}$,

$$\begin{aligned} P(\xi_1(n+1) = \varepsilon_1, \ldots, \xi_r(n+1) = \varepsilon_r) \\ &= P(\xi_1(n) = 0, \xi_2(n) = \varepsilon_1, \ldots, \xi_{r+1}(n) = \varepsilon_r) \\ &\quad + P(\xi_1(n) = 1, \xi_2(n) = 1 - \varepsilon_1, \ldots, \xi_{r+1}(n) = 1 - \varepsilon_r). \end{aligned} \tag{5}$$

We can show that if we start with the $\xi_r(0)$, $r = 1, 2, \ldots$, i.i.d. with

$$P(\xi_r(0) = 1) = P(\xi_r(0) = 0) = 1/2$$

then the same will be true for all n.

$$P(\xi_r(n) = 1) = P(\xi_r(n) = 0) = 1/2$$

The proof is by induction: if the $\xi_1(n), \xi_2(n), \ldots$ are i.i.d. with $P(\xi_r(n) = 1) = P(\xi_r(n) = 0) = 1/2$ then, from (5), the same is true at $n + 1$. Hence choosing the initial state at random in this manner results in a stationary Markov chain.

Remarks: This is not a Markov chain with countably many states. So it goes beyond our theory. But to even think about it will give you some strength. As for the name of the chain (bread kneading) it can be justified as follows: to each ξ there corresponds a number x between 0 and 1 because ξ can be seen as the binary representation of the real number x. If $\xi_1 = 0$ then $x < 1/2$ and our shifting rule maps x into $2x$. If $\xi_1 = 1$ then $x \geq 1/2$, and the rule maps x into $2(1 - x)$. But the function $f(x) = \min(2x, 2(1 - x))$, applied to the interval $[0, 1]$ (think of $[0, 1]$ as a flexible bread dough arranged in a long rod) kneads the bread for it stretches it till it doubles in length and then folds it back in the middle.

Note on terminology: When a Markov chain is stationary we refer to it as being in STEADY-STATE.

4.1 Finding a stationary distribution

As explained above, a stationary distribution π satisfies

$$\begin{aligned} \pi &= \pi \mathsf{P} \quad &&\text{(balance equations)} \\ \pi \mathbf{1} &= 1 \quad &&\text{(normalisation condition).} \end{aligned}$$

In other words,

$$\begin{aligned} \pi(i) = \sum_{j \in S} \pi(j) p_{j,i}, \quad i \in S, \\ \sum_{j \in S} \pi(j) = 1. \end{aligned}$$

If the state space has d states, then the above are $d+1$ equations. But we only have d unknowns. This is OK, because the first d equations are linearly dependent: the sum of both sides equals 1, therefore one of them is always redundant.

Writing the equations in this form is not always the best thing to do. Instead of eliminating equations, we introduce more, expecting to be able to choose, amongst them, a set of d linearly independent equations which can be more easily solved. The convenience is often suggested by the topological structure of the chain, as we will see in examples.

Let us introduce the notion of "PROBABILITY FLOW" from a set A of states to its complement under a distribution π:

$$F(A, A^c) := \sum_{i \in A} \sum_{j \in A^c} \pi(i) p_{i,j}.$$

Think of $\pi(i)p_{i,j}$ as a "current" flowing from i to j. So $F(A, A^c)$ is the total current from A to A^c. We have:

Proposition 1. *π is a stationary distribution if and only if $\pi \mathbf{1} = 1$ and*

$$F(A, A^c) = F(A^c, A),$$

for all $A \subset S$.

Proof. If the latter condition holds, then choose $A = \{i\}$ and see that you obtain the balance equations. Conversely, suppose that the balance equations hold. Fix a set $A \subset S$ and write

$$\pi(i) = \sum_{j \in S} \pi(j) p_{j,i} = \sum_{j \in A} \pi(j) p_{j,i} + \sum_{j \in A^c} \pi(j) p_{j,i}$$

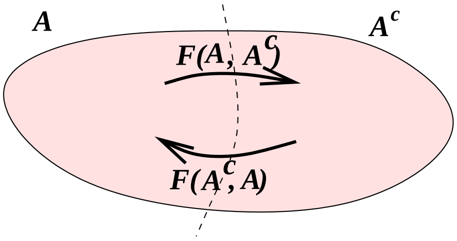
Multiply $\pi(i)$ by $\sum_{j \in S} p_{i,j}$ (which equals 1):

$$\pi(i) \sum_{j \in S} p_{i,j} = \sum_{j \in A} \pi(j) p_{j,i} + \sum_{j \in A^c} \pi(j) p_{j,i}$$

Now split the left sum also:

$$\sum_{j\in A}\pi(i)p_{i,j}+\sum_{j\in A^c}\pi(i)p_{i,j}=\sum_{j\in A}\pi(j)p_{j,i}+\sum_{j\in A^c}\pi(j)p_{j,i}$$

This is true for all $i \in S$. Summing over $i \in A$, we see that the first term on the left cancels with the first on the right, while the remaining two terms give the equality we seek. □



Schematic representation of the flow balance relations.

The extra degree of freedom provided by the last result, gives us flexibility. One can ask how to choose a minimal complete set of linearly independent equations, but we shall not do this here. Instead, here is an example:

Example: Consider the 2-state Markov chain with $p_{12} = \alpha$, $p_{21} = \beta$. (Consequently, $p_{11} = 1-\alpha$, $p_{22} = 1-\beta$.) Assume $0 < \alpha, \beta < 1$. We have

$$\pi(1)\alpha = \pi(2)\beta,$$

which immediately tells us that $\pi(1)$ is proportional to β and $\pi(2)$ proportional to α:

$$\pi(1) = c\beta, \quad \pi(2) = c\alpha.$$

The constant c is found from

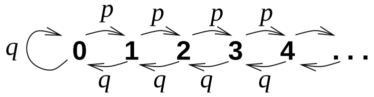
$$\pi(1) + \pi(2) = 1.$$

That is,

$$\pi(1) = \frac{\beta}{\alpha+\beta}, \quad \pi(2) = \frac{\alpha}{\alpha+\beta}.$$

If we take $\alpha = 0.05$, $\beta = 0.99$ (as in the mouse example), we find $\pi(1) = \frac{0.99}{1.04} \approx 0.952$, $\pi(2) = \frac{0.05}{1.04} \approx 0.048$. Thus, in *steady state*, the mouse spends only 95.2% of the time in room 1, and 4.8% of the time in room 2.

Example: a walk with a barrier. Consider the following chain, where $p + q = 1$:



The balance equations are:

$$\pi(i) = p\pi(i-1) + q\pi(i+1), \quad i = 1, 2, 3, \ldots$$

But we can make them simpler by choosing $A = \{0, 1, \ldots, i-1\}$. If $i \geq 1$ we have

$$F(A, A^c) = \pi(i-1)p = \pi(i)q = F(A^c, A).$$

These are much simpler than the previous ones. In fact, we can solve them immediately. For $i \geq 1$,

$$\pi(i) = (p/q)^i \pi(0).$$

Does this provide a stationary distribution? Yes, provided that $\sum_{i=0}^{\infty} \pi(i) = 1$. This is possible if and only if $p < q$, i.e. $p < 1/2$. (If $p \geq 1/2$ there is no stationary distribution.)

5 Topological structure

5.1 The graph of a chain

This refers to the structure of the process that does not depend on the exact values of the transition probabilities but only on which of them are positive. This is often cast in terms of a DIGRAPH (DIRECTED GRAPH), i.e. a pair (S, E) where S is the state space and $E \subset S \times S$ defined by

$$(i, j) \in E \iff p_{i,j} > 0.$$

In graph-theoretic terminology, S is a set of vertices, and E a set of directed edges. We can frequently (but not always) draw the graph (as we did in examples) and visualise the set E as the set of all edges with arrows.

5.2 The relation "leads to"

We say that a state i leads to state j if, starting from i, the chain will visit j at some finite time. In other words,

$$i \text{ LEADS TO } j \text{ (written as } i \rightsquigarrow j) \iff P_i(\exists n \geq 0 \; X_n = j) > 0.$$

Notice that this relation is REFLEXIVE):

$$\forall i \in S \quad i \rightsquigarrow i$$

Furthermore:

Theorem 1. *The following are equivalent:*

(i) $i \rightsquigarrow j$

(ii) $p_{i,i_1} p_{i_1,i_2} \cdots p_{i_{n-1},j} > 0$ *for some* $n \in \mathbb{N}$ *and some states* $i_1, \ldots, i_{n-1}$.

(iii) $P_i(X_n = j) > 0$ *for some* $n \geq 0$.

Proof. If $i = j$ there is nothing to prove. So assume $i \neq j$.
- Suppose that $i \rightsquigarrow j$. Since

$$0 < P_i(\exists n \geq 0 \; X_n = j) \leq \sum_{n \geq 0} P_i(X_n = j),$$

the sum on the right must have a nonzero term. In other words, (iii) holds.
• Suppose that (ii) holds. We have

$$P_i(X_n = j) = \sum_{i_1,\dots,i_{n-1}} p_{i,i_1} p_{i_1,i_2} \cdots p_{i_{n-1},j},$$

where the sum extends over all possible choices of states in S. By assumption, the sum contains a nonzero term. Hence $P_i(X_n = j) > 0$, and so (iii) holds.
• Suppose now that (iii) holds. Look at the last display again. Since $P_i(X_n = j) > 0$, we have that the sum is also positive and so one of its terms is positive, meaning that (ii) holds. In addition, (i) holds, because

$$P_i(X_n = j) \le P_i(\exists m \ge 0 \; X_m = j).$$

□

Corollary 1. *The relation* $\rightsquigarrow$ *is* TRANSITIVE*: If* $i \rightsquigarrow j$ *and* $j \rightsquigarrow k$ *then* $i \rightsquigarrow k$*.*

5.3 The relation "communicates with"

Define next the relation

i COMMUNICATES WITH j (written as $i \leftrightsquigarrow j$) $\iff$ $i \rightsquigarrow j$ and $j \rightsquigarrow i$.

Obviously, this is SYMMETRIC ($i \leftrightsquigarrow j \iff j \leftrightsquigarrow i$).

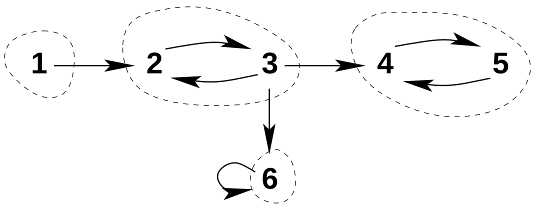
Corollary 2. *The relation* $\leftrightsquigarrow$ *is an* EQUIVALENCE RELATION*.*

Proof. Equivalence relation means that it is symmetric, reflexive and transitive. We just observed it is symmetric. It is reflexive and transitive because $\rightsquigarrow$ is so. □

Just as any equivalence relation, it partitions S into EQUIVALENCE CLASSES known as COMMUNICATING CLASSES. The communicating class corresponding to the state i is, by definition, the set of all states that communicate with i:

$$[i] := \{j \in S : \; j \leftrightsquigarrow i\}. \tag{6}$$

So, by definition, $[i] = [j]$ if and only if $i \leftrightsquigarrow j$. Two communicating classes are either identical or completely disjoint.



In the example of the figure, we see that there are four communicating classes:

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

The first two classes differ from the last two in character. The class $\{4,5\}$ *is closed: if the chain goes in there then it never moves out of it. However, the class* $\{2,3\}$ *is not closed.*

More generally, we say that a set of states $C \subset S$ is CLOSED if

$$\sum_{j \in C} p_{i,j} = 1, \quad \text{for all } i \in C.$$

Lemma 3. *A communicating class C is closed if and only if the following holds:*
If $i \in C$ and if $i \rightsquigarrow j$ then $j \in C$.

Proof. Suppose first that C is a closed communicating class. Suppose that $i \rightsquigarrow j$. We then can find states $i_1, \ldots, i_{n-1}$ such that $p_{i,i_1} p_{i_1,i_2} \cdots p_{i_{n-1},j} > 0$. Thus, $p_{i,i_1} > 0$. Since $i \in C$, and C is closed, we have $i_1 \in C$. Similarly, $p_{i_1,i_2} > 0$, and so $i_2 \in C$, and so on, we conclude that $j \in C$. The converse is left as an exercise. □

Closed communicating classes are particularly important because they decompose the chain into smaller, more manageable, parts.

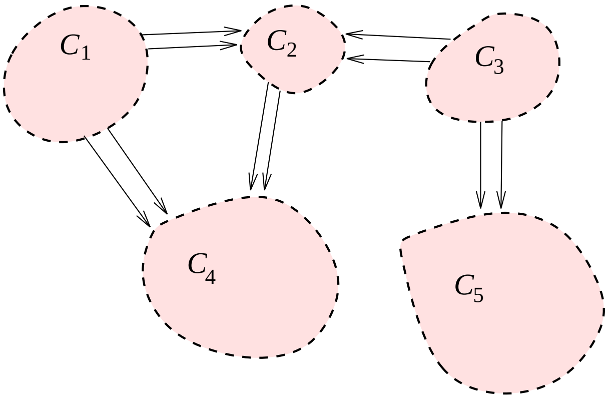
If all states communicate with all others, then the chain (or, more precisely, its transition matrix P) is called IRREDUCIBLE. In graph terms, starting from any state, we can reach any other state. In other words still, the state space S is a closed communicating class.

If a state i is such that $p_{i,i} = 1$ then we say that i is an ABSORBING state. To put it otherwise, the single-element set $\{i\}$ is a closed communicating class.

Note: *If we consider a Markov chain with transition matrix P and fix $n \in \mathbb{N}$ then the Markov chain with transition matrix P^n has exactly the same closed communicating classes. Why?*

A state i is ESSENTIAL if it belongs to a closed communicating class. Otherwise, the state is inessential.

Note: *An inessential state i is such that there exists a state j such that $i \rightsquigarrow j$ but $j \not\rightsquigarrow i$.*



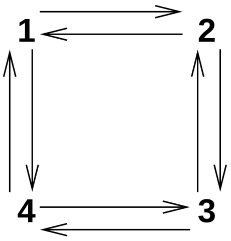
The general structure of a Markov chain is indicated in this figure. Note that there can be no arrows between closed communicating classes. The classes C_4, C_5 in the figure above are closed communicating classes. The communicating classes C_1, C_2, C_3 are not closed. There can be arrows between two non-closed communicating classes but they are always in the same direction.

A general Markov chain can have an infinite number of communicating classes.

If we remove the non-closed communicating classes then we obtain a collection of disjoint closed communicating classes with no arrows between them.

5.4 Period

Consider the following chain:



The chain is irreducible, i.e. all states form a single closed communicating class. Notice however, that we can further simplify the behaviour of the chain by noticing that if the chain is in the set $C_1 = \{1,3\}$ at some point of time then, it will move to the set $C_2 = \{2,4\}$ at the next step and vice versa. The chain alternates between the two sets. So if the chain starts with $X_0 \in C_1$ then we know that $X_1 \in C_2, X_2 \in C_1, X_3 \in C_2, X_4 \in C_1, \ldots$ We say that the chain has period 2. (Can you find an example of a chain with period 3?)

We will now give the definition of the period of a state of an arbitrary chain.

Let us denote by $p_{ij}^{(n)}$ the entries of P^n. Since

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

we have

$$p_{ij}^{(m+n)} \geq p_{ik}^{(m)} p_{kj}^{(n)}, \qquad \text{for all } m, n \geq 0,\ i, j \in S.$$

So $p_{ii}^{(2n)} \geq \left(p_{ii}^{(n)}\right)^2$, and, more generally,

$$p_{ii}^{(\ell n)} \geq \left(p_{ii}^{(n)}\right)^\ell, \quad \ell \in \mathbb{N}.$$

Therefore if $p_{ii}^{(n)} > 0$ then $p_{ii}^{(\ell n)} > 0$ for all $\ell \in \mathbb{N}$. Another way to say this is:

If the integer n DIVIDES m (denote this by: $n \mid m$) and if $p_{ii}^{(n)} > 0$ then $p_{ii}^{(m)} > 0$.

So, whether it is possible to return to i in m steps can be decided by one of the integer divisors of m.

The PERIOD of an essential state is defined as the greatest common divisor of all natural numbers n with such that it is possible to return to i in n steps:

$$d(i) := \gcd\{n \in \mathbb{N} : \ P_i(X_n = i) > 0\} \ .$$

A state i is called APERIODIC if $d(i) = 1$. The period of an inessential state is not defined.

Theorem 2 (the period is a class property)**.** *If $i \leftrightsquigarrow j$ then $d(i) = d(j)$.*

Proof. Consider two distinct essential states i, j. Let $D_i = \{n \in \mathbb{N} : \ p_{ii}^{(n)} > 0\}$, $D_j = \{n \in \mathbb{N} : \ p_{jj}^{(n)} > 0\}$, $d(i) = \gcd D_i$, $d(j) = \gcd D_j$. If $i \rightsquigarrow j$ there is $\alpha \in \mathbb{N}$ with $p_{ij}^{(\alpha)} > 0$ and if

$j \rightsquigarrow i$ there is $\beta \in \mathbb{N}$ with $p_{ji}^{(\beta)} > 0$. So $p_{jj}^{(\alpha+\beta)} \geq p_{ij}^{(\alpha)} p_{ji}^{(\beta)} > 0$, showing that $\alpha + \beta \in D_j$. Therefore

$$d(j) \mid \alpha + \beta.$$

Now let $n \in D_i$. Then $p_{ii}^{(n)} > 0$ and so $p_{jj}^{(\alpha+n+\beta)} \geq p_{ij}^{(\alpha)} p_{ii}^{(n)} p_{ji}^{(\beta)} > 0$, showing that $\alpha + n + \beta \in D_j$. Therefore

$$d(j) \mid \alpha + \beta + n \text{ for all } n \in D_i.$$

If a number divides two other numbers then it also divides their difference. From the last two displays we get

$$d(j) \mid n \text{ for all } n \in D_i.$$

So $d(j)$ is a divisor of all elements of D_i. Since $d(i)$ is the greatest common divisor we have $d(i) \geq d(j)$ (in fact, $d(j) \mid d(i)$). Arguing symmetrically, we obtain $d(i) \leq d(j)$ as well. So $d(i) = d(j)$. □

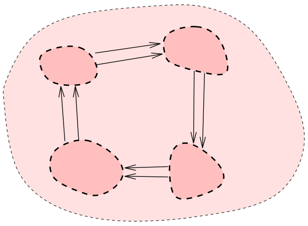
Theorem 3. *If i is an aperiodic state then there exists n_0 such that $p_{ii}^{(n)} > 0$ for all $n \geq n_0$.*

Proof. Pick $n_2, n_1 \in D_i$ such that $n_2 - n_1 = 1$. Let n be sufficiently large. Divide n by n_1 to obtain $n = qn_1 + r$, where the remainder $r \leq n_1 - 1$. Therefore $n = qn_1 + r(n_2 - n_1) = (q-r)n_1 + rn_2$. Because n is sufficiently large, $q - r > 0$. Since $n_1, n_2 \in D_i$, we have $n \in D_i$ as well. □

Of particular interest, are irreducible chains, i.e. chains where, as defined earlier, all states communicate with one another. An irreducible chain has period d if one (and hence all) of the states have period d. In particular, if $d = 1$, the chain is called aperiodic.

Corollary 3. *An irreducible chain with finitely many states is aperiodic if and only if there exists an n such that $p_{ij}^{(n)} > 0$ for all $i, j \in S$.*

More generally, if an irreducible chain has period d then we can decompose the state space into d sets $C_0, C_1, \ldots, C_{d-1}$ such that the chain moves cyclically between them.



This figure shows the internal structure of a closed communicating class with period $d = 4$.

Formally, this is the content of the following theorem.

Theorem 4. *Consider an irreducible chain with period d. Then we can uniquely partition the state space S into d disjoint subsets $C_0, C_1, \ldots, C_{d-1}$ such that*

$$\sum_{j \in C_{r+1}} p_{ij} = 1, \quad i \in C_r, \quad r = 0, 1, \ldots, d-1.$$

(Here $C_d := C_0$.)

Proof. Define the relation

$$i \overset{d}{\leftrightsquigarrow} j \iff p_{ij}^{(nd)} > 0 \text{ for some } n \geq 0.$$

Notice that this is an equivalence relation. Hence it partitions the state space S into d disjoint equivalence classes. We show that these classes satisfy what we need. Assume $d > 1$. Pick a state i_0 and let C_0 be its equivalence class (i.e the set of states j such that $i \overset{d}{\leftrightsquigarrow} j$). Then pick i_1 such that $p_{i_0 i_1} > 0$. Let C_1 be the equivalence class of i_1. Continue in this manner and define states

$$i_0, \quad i_1, \quad \ldots, \quad i_{d-1}$$

with corresponding classes

$$C_0, \quad C_1, \quad \ldots, \quad C_{d-1}.$$

It is easy to see that if we continue and pick a further state i_d with $p_{i_{d-1}, i_d} > 0$, then, necessarily, $i_d \in C_0$. We now show that if i belongs to one of these classes and if $p_{ij} > 0$ then, necessarily, j belongs to the next class. Take, for example, $i \in C_0$. Suppose $p_{ij} > 0$ but $j \notin C_1$ but, say, $j \in C_2$. Consider the path

$$i_0 \to i \to j \to i_2 \to i_3 \to \cdots \to i_d \to i_0.$$

Such a path is possible because of the choice of the $i_0, i_1, \ldots$, and by the assumptions that $i_0 \overset{d}{\leftrightsquigarrow} i$, $i_2 \overset{d}{\leftrightsquigarrow} j$. The existence of such a path implies that it is possible to go from i_0 to i_0 in a number of steps which is an integer multiple of $d - 1$ (why?), which contradicts the definition of d. □

6 Hitting times and first-step analysis

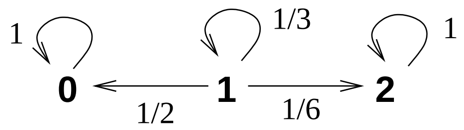
Consider a Markov chain with transition probability matrix P. We define the HITTING TIME[4] of a set of states A by

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

We are interested in deriving formulae for the probability that this time is finite as well as the expectation of this time.

We use a method that is based upon considering what the Markov chain does at time 1, i.e. after it takes one step from its current position; that is why we call it "FIRST-STEP ANALYSIS".

As an example, consider the chain



[4] We shall later consider the time $\inf\{n \geq 1 : \ X_n \in A\}$ which differs from T_A simply by considering $n \geq 1$ instead of $n \geq 0$. If $X_0 \notin A$ the two times coincide. We avoid excessive notation by using the same letter for both. The reader is warned to be alert as to which of the two variables we are considering at each time.

It is clear that $P_1(T_0 < \infty) < 1$, because the chain may end up in state 2. But what exactly is this probability equal to?

To answer this in its generality, fix a set of states A, and define

$$\varphi(i) := P_i(T_A < \infty).$$

We then have:

Theorem 5. *The function $\varphi(i)$ satisfies*

$$\begin{aligned} \varphi(i) &= 1, \quad i \in A \\ \varphi(i) &= \sum_{j \in S} p_{ij}\varphi(j), \quad i \notin A. \end{aligned}$$

Furthermore, if $\varphi'(i)$ is any other solution of these equations then $\varphi'(i) \geq \varphi(i)$ for all $i \in S$.

Proof. If $i \in A$ then $T_A = 0$, and so $\varphi(i) = 1$. If $i \notin A$, then $T_A \geq 1$. So $T_A = 1 + T'_A$, (where T'_A is the *remaining time* until A is hit). We first have

$$P_i(T_A < \infty) = \sum_{j \in S} P_i(1 + T'_A < \infty | X_1 = j) P_i(X_1 = j) = \sum_{j \in S} P_i(T'_A < \infty | X_1 = j) p_{ij}$$

But observe that the random variable T'_A is a function of the future after time 1 (i.e. a function of $X_1, X_2, \ldots$). Therefore, the event T'_A is independent of X_0, conditionally on X_1. Hence:

$$P(T'_A < \infty | X_1 = j, X_0 = i) = P(T'_A < \infty | X_1 = j).$$

But the Markov chain is homogeneous, which implies that

$$P(T'_A < \infty | X_1 = j) = P(T_A < \infty | X_0 = j) = P_j(T_A < \infty) = \varphi(j).$$

Combining the above, we have

$$P_i(T_A < \infty) = \sum_{j \in S} \varphi(j) p_{ij},$$

as needed. For the second part of the proof, let $\varphi'(i)$ be another solution. Then

$$\varphi'(i) = \sum_{j \in A} p_{ij} + \sum_{j \notin A} p_{ij}\varphi'(j).$$

By self-feeding this equation, we have

$$\begin{aligned} \varphi'(i) &= \sum_{j \in A} p_{ij} + \sum_{j \notin A} p_{ij} \left(\sum_{k \in A} p_{jk} + \sum_{k \notin A} p_{jk}\varphi'(k) \right) \\ &= \sum_{j \in A} p_{ij} + \sum_{j \notin A} p_{ij} \sum_{k \in A} p_{jk} + \sum_{j \notin A} p_{ij} \sum_{k \notin A} p_{jk}\varphi'(k) \end{aligned}$$

We recognise that the first term equals $P_i(T_A = 1)$, the second equals $P_i(T_A = 2)$, so the first two terms together equal $P_i(T_A \le 2)$. By omitting the last term we obtain $\varphi'(i) \ge P_i(T_A \le 2)$. By continuing self-feeding the above equation n times, we obtain

$$\varphi'(i) \ge P_i(T_A \le n).$$

Letting $n \to \infty$, we obtain

$$\varphi'(i) \ge P_i(T_A < \infty) = \varphi(i).$$

□

Example: In the example of the previous figure, we choose $A = \{0\}$, the set that contains only state 0. We let $\varphi(i) = P_i(T_0 < \infty)$, $i = 0, 1, 2$. We immediately have $\varphi(0) = 1$, and $\varphi(2) = 0$. (If the chain starts from 0 it takes no time to hit 0; if the chain starts from 2 it will never hit 0.) As for $\varphi(1)$, we have

$$\varphi(1) = \frac{1}{3}\varphi(1) + \frac{1}{2}\varphi(0),$$

so $\varphi(1) = 3/4$.

∼o∼o∼o∼o∼o∼o∼o∼

Next consider the mean time (mean number of steps) until the set A is hit for the first time.

$$\psi(i) := E_i T_A.$$

We have:

Theorem 6. *The function $\psi(i)$ satisfies*

$$\begin{aligned} &\psi(i) = 0, \quad i \in A \\ &\psi(i) = 1 + \sum_{j \notin A} p_{ij}\psi(j), \quad i \notin A. \end{aligned}$$

Furthermore, if $\psi'(i)$ is any other solution of these equations then $\psi'(i) \ge \psi(i)$ for all $i \in S$.

Proof. Start the chain from i. If $i \in A$ then, obviously, $T_A = 0$ and so $\psi(i) := E_i T_A = 0$. If $i \notin A$, then $T_A = 1 + T'_A$, as above. Therefore,

$$E_i T_A = 1 + E_i T'_A = 1 + \sum_{j \in S} p_{ij} E_j T_A = 1 + \sum_{j \notin A} p_{ij}\psi(j),$$

which is the second of the equations. The second part of the proof is omitted. □

Example: Continuing the previous example, let $A = \{0, 2\}$, and let $\psi(i) = E_i T_A$. Clearly, $\psi(0) = \psi(2) = 0$, because it takes no time to hit A if the chain starts from A. On the other hand,

$$\psi(1) = 1 + \frac{1}{3}\psi(1),$$

which gives $\psi(1) = 3/2$. (This should have been obvious from the start, because the underlying experiment is the flipping of a coin with "success" probability $2/3$, therefore the mean number of flips until the first success is $3/2$.)

∼o∼o∼o∼o∼o∼o∼o∼

Now let us consider two hitting times T_A, T_B for two disjoint sets of states A, B. We may ask to find the probabilities

$$\varphi_{AB}(i) = P_i(T_A < T_B).$$

It should be clear that the equations satisfied are:

$$\begin{aligned} \varphi_{AB}(i) &= 1, \quad i \in A \\ \varphi_{AB}(i) &= 0, \quad i \in B \\ \varphi_{AB}(i) &= \sum_{j \in A} p_{ij} \varphi_{AB}(j), \quad \text{otherwise.} \end{aligned}$$

Furthermore, φ_{AB} is the minimal solution to these equations.

∼o∼o∼o∼o∼o∼o∼o∼

As yet another application of the first-step analysis method, consider the following situation: Every time the state is x a reward $f(x)$ is earned. This happens up to the first hitting time T_A first hitting time of the set A. The total reward is $f(X_0) + f(X_1) + \cdots + f(X_{T_A})$. We are interested in the mean total reward

$$h(x) := E_x \sum_{n=0}^{T_A} f(X_n).$$

Clearly,

$$h(x) = f(x), \quad x \in A,$$

because when $X_0 \in A$ then $T_0 = 0$ and so the total reward is $f(X_0)$. Next, if $x \notin A$, as argued earlier,

$$T_A = 1 + T_A',$$

where T_A' is the remaining time, after one step, until set A is reached. Then

$$\begin{aligned} h(x) &= E_x \sum_{n=0}^{1+T_A'} f(X_n) \\ &= f(x) + E_x \sum_{n=1}^{1+T_A'} f(X_n) \\ &= f(x) + E_x \sum_{n=0}^{T_A'} f(X_{n+1}), \end{aligned}$$

where, in the last sum, we just changed index from n to $n+1$. Now the last sum is a function of the future after time 1, i.e. of the random variables $X_1, X_2, \ldots$. Hence, by the Markov property,

$$E_x \sum_{n=0}^{T_A'} f(X_{n+1}) = \sum_{y \in S} p_{xy} h(y).$$

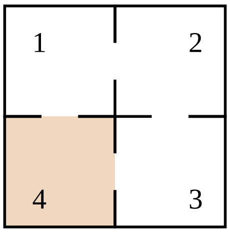
Thus, the set of equations satisfied by h, are

$$h(x) = 0, \quad x \in A$$

$$h(x) = f(x) + \sum_{y \in S} p_{xy} h(y), \quad x \notin A.$$

You can remember the latter ones by thinking that the expectecd total reward equals the immediate reward $f(x)$ plus the expected remaining reward.

Example: A thief enters sneaks in the following house and moves around the rooms, collecting "rewards": when in room i, he collects i pounds, for $i = 1, 2, 3$, unless he is in room 4, in which case he dies immediately.



The question is to compute the average total reward, if he starts from room 2, until he dies in room 4. (Sooner or later, he will die; why?) If we let $h(i)$ be the average total reward if he starts from room i, we have

$$\begin{aligned} h(4) &= 0 \\ h(1) &= 1 + \frac{1}{2}h(2) \\ h(3) &= 3 + \frac{1}{2}h(2) \\ h(2) &= 2 + \frac{1}{2}h(1) + \frac{1}{2}h(3). \end{aligned}$$

We solve and find $h(2) = 8$. (Also, $h(1) = 5$, $h(3) = 7$.)

7 Gambler's ruin

Consider the simplest game of chance of tossing an (possibly unfair) coin repeatedly and independently. Let p the probability of heads and $q = 1 - p$ that of tails. If heads come up then you win £1 per pound of stake. So, if just before the k-th toss you decide to bet Y_k pounds then after the realisation of the toss you will have earned Y_k pounds if heads came up or lost Y_k pounds if tails appeared. If ξ_k is the outcome of the k-th toss (where $\xi_k = +1$ means heads and $\xi_k = -1$ means tails), your fortune after the n-th toss is

$$X_n = X_0 + \sum_{k=1}^{n} Y_k \xi_k.$$

The successive stakes $(Y_k, k \in \mathbb{N})$ form the STRATEGY OF THE GAMBLER. Clearly, no gambler is assumed to have any information about the outcome of the upcoming toss, so if the gambler wants to have a strategy at all, she must base it on whatever outcomes have appeared thus far. In other words, Y_k cannot depend on ξ_k. It can only depend on $\xi_1, \dots, \xi_{k-1}$ and–as is often the case in practice–on other considerations of e.g. astrological nature.

We wish to consider the problem of finding the probability that the gambler will never lose her money.

This has profound actuarial applications: An insurance company must make sure that the strategies it employs are such that–at least–the company will not go bankrupt (and, in addition, it will make enormous profits). The difference between the gambler's ruin problem for a company and that for a mere mortal is that, in the first case, the company has control over the probability p.

Let us consider a concrete problem. The gambler starts with $X_0 = x$ pounds and has a target: To make b pounds ($b > x$). She will stop playing if her money fall below a ($a < x$).

To solve the problem, in simplest case $Y_k = 1$ for all k, is our next goal.

We use the notation P_x (respectively, E_x) for the probability (respectively, expectation) under the initial condition $X_0 = x$. Consider the hitting time of state y:

$$T_y := \inf\{n \geq 0 : X_n = y\}.$$

We are clearly dealing with a Markov chain in the state space

$$\{a, a+1, a+2, \dots, b-1, b\}$$

with transition probabilities

$$p_{i,i+1} = p, \; p_{i,i-1} = q = 1 - p, \quad a < i < b,$$

and where the states a, b are absorbing.

Theorem 7. *Consider a gambler playing a simple coin tossing game, as described above, betting £1 at each integer time against a coin which has $P(heads) = p$, $P(tails) = 1 - p$. Let £x be the initial fortune of the gambler. Then the probability that the gambler's fortune will reach level $b > x$ before reaching level $a < x$ equals:*

$$P_x(T_b < T_a) = \begin{cases} \dfrac{(q/p)^{x-a} - 1}{(q/p)^{b-a} - 1}, & \text{if } p \neq q \\[2ex] \dfrac{x-a}{b-a}, & \text{if } p = q = 1/2 \end{cases}, \quad a \leq x \leq b.$$

Proof. We first solve the problem when $a = 0$, $b = \ell > 0$ and let

$$\varphi(x, \ell) := P_x(T_\ell < T_0), \quad 0 \leq x \leq \ell.$$

The general solution will then obtained from

$$P_x(T_b < T_a) = \varphi(x - a, b - a).$$

(Think why!) To simplify things, write $\varphi(x)$ instead of $\varphi(x, \ell)$. We obviously have

$$\varphi(0) = 0, \quad \varphi(\ell) = 1,$$

and, by first-step analysis,

$$\varphi(x) = p\varphi(x+1) + q\varphi(x-1), \quad 0 < x < \ell.$$

Since $p + q = 1$, we can write this as

$$p[\varphi(x+1) - \varphi(x)] = q[\varphi(x) - \varphi(x-1)].$$

Let $\delta(x) := \varphi(x+1) - \varphi(x)$. We have

$$\delta(x) = \frac{q}{p}\delta(x-1) = \left(\frac{q}{p}\right)^2 \delta(x-2) = \cdots = \left(\frac{q}{p}\right)^x \delta(0).$$

Assuming that $\lambda := q/p \neq 1$, we find

$$\varphi(x) = \delta(x-1) + \delta(x-2) + \cdots + \delta(0) = [\lambda^{x-1} + \cdots + \lambda + 1]\delta(0) = \frac{\lambda^x - 1}{\lambda - 1}\delta(0).$$

Since $\varphi(\ell) = 1$, we have $\frac{\lambda^\ell - 1}{\lambda - 1}\delta(0) = 1$ which gives $\delta(0) = (\lambda - 1)/(\lambda^\ell - 1)$. We finally find

$$\varphi(x) = \frac{\lambda^x - 1}{\lambda^\ell - 1}, \quad 0 \leq x \leq \ell, \quad \lambda \neq 1.$$

If $\lambda = 1$ then $q = p$ and so

$$\varphi(x+1) - \varphi(x) = \varphi(x) - \varphi(x-1).$$

This gives $\delta(x) = \delta(0)$ for all x and $\varphi(x) = x\delta(0)$. Since $\varphi(\ell) = 1$, we find $\delta(0) = 1/\ell$ and so

$$\varphi(x) = \frac{x}{\ell}, \quad 0 \leq x \leq \ell, \quad \lambda = 1.$$

Summarising, we have obtained

$$\varphi(x, \ell) = \begin{cases} \frac{\lambda^x - 1}{\lambda^\ell - 1}, & \text{if } \lambda \neq 1 \\ \frac{x}{\ell}, & \text{if } \lambda = 1. \end{cases}$$

The general case is obtained by replacing ℓ by $b - a$ and x by $x - a$. □

Corollary 4. *The* RUIN PROBABILITY *of a gambler starting with £x is*

$$P_x(T_0 < \infty) = \begin{cases} (q/p)^x, & \textit{if } p > 1/2 \\ 1, & \textit{if } p \leq 1/2. \end{cases}$$

Proof. We have

$$P_x(T_0 < \infty) = \lim_{\ell \to \infty} P_x(T_0 < T_\ell) = 1 - \lim_{\ell \to \infty} P_x(T_\ell < T_0).$$

If $p > 1/2$, then $\lambda = q/p < 1$, and so

$$P_x(T_0 < \infty) = 1 - \lim_{\ell \to \infty} \frac{\lambda^x - 1}{\lambda^\ell - 1} = 1 - \frac{\lambda^x - 1}{0 - 1} = \lambda^x.$$

If $p < 1/2$, then $\lambda = q/p > 1$, and so

$$P_x(T_0 < \infty) = 1 - \lim_{\ell \to \infty} \frac{\lambda^x - 1}{\lambda^\ell - 1} = 1 - 0 = 1.$$

Finally, if $p = q$,

$$P_x(T_0 < \infty) = 1 - \lim_{\ell \to \infty} \frac{x}{\ell} = 1 - 0 = 1.$$

□

8 Stopping times and the strong Markov property

A random time is a random variable taking values in the time-set, including, possibly, the value $+\infty$. In particular, a STOPPING TIME τ is a random variable with values in $\mathbb{Z}_+ \cup \{+\infty\}$ such that $\mathbf{1}(\tau = n)$ is a deterministic function of $(X_0, \ldots, X_n)$ for all $n \in \mathbb{Z}_+$.

An example of a stopping time is the time T_A, the first hitting time of the set A, considered earlier.

An example of a random time which is not a stopping time is the last visit of a specific set.

Recall that the Markov property states that, for any n, the future process $(X_n, X_{n+1}, \ldots)$ is independent of the past process $(X_0, \ldots, X_n)$ if we know the present X_n.

We will now prove that this remains true if we replace the deterministic n by a random stopping time τ. This stronger property is referred to as the STRONG MARKOV PROPERTY.

Theorem 8 (strong Markov property)**.** *Suppose $(X_n, n \geq 0)$ is a time-homogeneous Markov chain. Let τ be a stopping time. Then, conditional on $\tau < \infty$ and $X_\tau = i$, the future process $(X_{\tau+n}, n \geq 0)$ is independent of the past process $(X_0, \ldots, X_\tau)$ and has the same law as the original process started from i.*

Proof. Let τ be a stopping time. Let A be any event determined by the past $(X_0, \ldots, X_\tau)$ if $\tau < \infty$. Since $(X_0, \ldots, X_\tau)\mathbf{1}(\tau < \infty) = \sum_{n \in \mathbb{Z}_+} (X_0, \ldots, X_n)\mathbf{1}(\tau = n)$, any such A must have the property that, for all $n \in \mathbb{Z}_+$, $A \cap \{\tau = n\}$ is determined by $(X_0, \ldots, X_n)$. We are going to show that for any such A,

$$\begin{aligned} &P(X_{\tau+1} = j_1, X_{\tau+2} = j_2, \ldots ; A \mid X_\tau, \tau < \infty) \\ &\qquad\qquad = P(X_{\tau+1} = j_1, X_{\tau+2} = j_2, \ldots \mid X_\tau, \tau < \infty) P(A \mid X_\tau, \tau < \infty) \end{aligned}$$

and that

$$P(X_{\tau+1} = j_1, X_{\tau+2} = j_2, \ldots \mid X_\tau = i, \tau < \infty) = p_{i,j_1} p_{j_1,j_2} \cdots$$

We have:

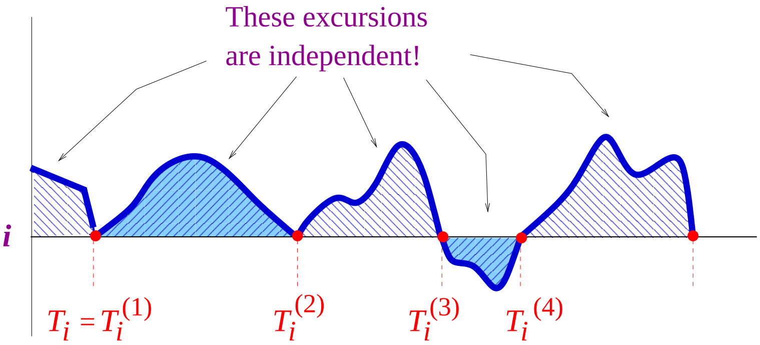
$$\begin{aligned} &P(X_{\tau+1} = j_1, X_{\tau+2} = j_2, \ldots ; A, X_\tau = i, \tau = n) \\ &\qquad \overset{(a)}{=} P(X_{n+1} = j_1, X_{n+2} = j_2, \ldots ; A, X_n = i, \tau = n) \\ &\qquad \overset{(b)}{=} P(X_{n+1} = j_1, X_{n+2} = j_2, \ldots \mid X_n = i, A, \tau = n) P(X_n = i, A, \tau = n) \\ &\qquad \overset{(c)}{=} P(X_{n+1} = j_1, X_{n+2} = j_2, \ldots \mid X_n = i) P(X_n = i, A, \tau = n) \\ &\qquad \overset{(d)}{=} p_{i,j_1} p_{j_1,j_2} \cdots P(X_n = i, A, \tau = n), \end{aligned}$$

where (a) is just logic, (b) is by the definition of conditional probability, (c) is due to the fact that $A \cap \{\tau = n\}$ is determined by $(X_0, \ldots, X_n)$ and the ordinary splitting property at n, and (d) is due to the fact that (X_n) is time-homogeneous Markov chain with transition probabilities p_{ij}. Our assertions are now easily obtained from this by first summing over all n (whence the event $\{\tau < \infty\}$ appears) and then by dividing both sides by $P(X_\tau = i, \tau < \infty)$. □

9 Regenerative structure of a Markov chain

A sequence of i.i.d. random variables is a (very) particular example of a Markov chain. The random variables X_n comprising a general Markov chain are not independent. However, by using the strong Markov property, we will show that we can split the sequence into i.i.d. "chunks".

The idea is that if we can find a state i that is visited again and again, then the behaviour of the chain between successive visits cannot be influenced by the past or the future (and that is due to the strong Markov property). Schematically,



Fix a state $i \in S$ and let T_i be the first time that the chain will visit this state:

$$T_i := \inf\{n \geq 1 : \ X_n = i\}.$$

Note the subtle difference between this T_i and the T_i of Section 6. Our T_i here is always positive. The T_i of Section 6 was allowed to take the value 0. If $X_0 \neq i$ then the two definitions coincide.

State i may or may not be visited again. Regardless, we let $T_i^{(2)}$ be the time of the second visit. (And we set $T_i^{(1)} := T_i$.) We let $T_i^{(3)}$ be the time of the third visit, and so on. Formally, we recursively define

$$T_i^{(r)} := \inf\{n > T_i^{(r-1)} : X_n = i\}, \quad r = 2, 3, \ldots$$

All these random times are stopping times. (Why?)

Consider the "trajectory" of the chain between two successive visits to the state i:

$$\mathscr{X}_i^{(r)} := \left(X_n, \ T_i^{(r)} \leq n < T_i^{(r+1)}\right), \quad r = 1, 2, \ldots$$

Let also

$$\mathscr{X}_i^{(0)} := \left(X_n, \ 0 \leq n < T_i^{(1)}\right).$$

We think of

$$\mathscr{X}_i^{(0)}, \quad \mathscr{X}_i^{(1)}, \quad \mathscr{X}_i^{(2)}, \quad \ldots$$

as "random variables", in a generalised sense. They are, in fact, random trajectories. We refer to them as excursions of the process. So $\mathscr{X}_i^{(r)}$ is the r-th excursion: the chain visits state i for the r-th time, and “goes for an excursion”; we “watch” it up until the next time it returns to state i.

An immediate consequence of the strong Markov property is:

Theorem 9. *The excursions*

$$\mathscr{X}_i^{(0)}, \quad \mathscr{X}_i^{(1)}, \quad \mathscr{X}_i^{(2)}, \quad \dots$$

are independent random variables. In particular, all, except, possibly, $\mathscr{X}_i^{(0)}$*, have the same distribution.*

Proof. Apply Strong Markov Property (SMP) at $T_i^{(r)}$: SMP says that, given the state at the *stopping time* $T_i^{(r)}$, the future after $T_i^{(r)}$ is independent of the past before $T_i^{(r)}$. But the state at $T_i^{(r)}$ is, by definition, equal to i. Therefore, the future is independent of the past, and this proves the first claim. The claim that $\mathscr{X}_i^{(1)}$, $\mathscr{X}_i^{(2)}$ have the same distribution follows from the fact that the Markov chain is time-homogeneous. Therefore, if it starts from state i at some point of time, it will behave in the same way as if it starts from the same state at another point of time. □

A trivial, but important corollary of the observation of this section is that:

Corollary 5. *(Numerical) functions* $g(\mathscr{X}_i^{(0)})$, $g(\mathscr{X}_i^{(1)})$, $g(\mathscr{X}_i^{(2)}), \dots$ *of the excursions are independent random variables.*

Example: Define

$$\Lambda_r := \sum_{n=T_i^{(r)}}^{T_i^{(r+1)}} \mathbf{1}(X_n = j)$$

The meaning of Λ_r is this: *it expresses the number of visits to a state j between two successive visits to the state i.* The last corollary ensures us that $\Lambda_1, \Lambda_2, \dots$ are independent random variables. Moreover, for time-homogeneous Markov chains, they are also identical in law (i.i.d.).

10 Recurrence and transience

When a Markov chain has finitely many states, at least one state must be visited infinitely many times. Perhaps some states (inessential ones) will be visited finitely many times, but there are always states which will be visited again and again, *ad infinitum.*

We abstract this trivial observation and give a couple of definitions:

First, we say that a state i is RECURRENT if, starting from i, the chain returns to i infinitely many times:

$$P_i(X_n = i \text{ for infinitely many } n) = 1.$$

Second, we say that a state i is TRANSIENT if, starting from i, the chain returns to i only finitely many times:

$$P_i(X_n = i \text{ for infinitely many } n) = 0.$$

Important note: *Notice that the two statements above are not negations of one another. Indeed, in principle, there could be yet another possibility:*

$$0 < P_i(X_n = i \text{ for infinitely many } n) < 1.$$

We will show that such a possibility does not exist, therefore concluding that every state is either recurrent or transient.

Consider the total number of visits to state i (excluding the possibility that $X_0 = i$):

$$J_i = \sum_{n=1}^{\infty} \mathbf{1}(X_n = i) = \sup\{r \geq 0 : T_i^{(r)} < \infty\} \ .$$

Suppose that the Markov chain starts from state i.

Notice that: state i is visited infinitely many times $\iff$ $J_i = \infty$.

Notice also that

$$f_{ii} := P_i(J_i \geq 1) = P_i(T_i < \infty).$$

We claim that:

Lemma 4. *Starting from $X_0 = i$, the random variable J_i is geometric:*

$$P_i(J_i \geq k) = f_{ii}^k \ , \quad k \geq 0.$$

Proof. By the strong Markov property at time $T_i^{(k-1)}$, we have

$$P_i(J_i \geq k) = P_i(J_i \geq k-1)P_i(J_i \geq 1).$$

Indeed, the event $\{J_i \geq k\}$ implies that $T_i^{(k-1)} < \infty$ and that $J_i \geq k-1$. But the past before $T_i^{(k-1)}$ is independent of the future, and that is why we get the product. Therefore,

$$P_i(J_i \geq k) = f_{ii}P_i(J_i \geq k-1) = f_{ii}^2 P_i(J_i \geq k-2) = \cdots = f_{ii}^k.$$

□

Notice that $f_{ii} = P_i(T_i < \infty)$ can either be equal to 1 or smaller than 1.

If $f_{ii} = 1$ we have that $P_i(J_i \geq k) = 1$ for all k, i.e. $P_i(J_i = \infty) = 1$. This means that the state i is recurrent.

If $f_{ii} < 1$, then $P_i(J_i < \infty) = 1$. This means that the state i is transient.

Because *either* $f_{ii} = 1$ *or* $f_{ii} < 1$, we conclude what we asserted earlier, namely, *every state is either recurrent or transient.*

We summarise this together with another useful characterisation of recurrence & transience in the following two lemmas.

Lemma 5. *The following are equivalent:*
1. State i is recurrent.
2. $f_{ii} = P_i(T_i < \infty) = 1$.
3. $P_i(J_i = \infty) = 1$.
4. $E_i J_i = \infty$.

Proof. State i is recurrent if, by definition, it is visited infinitely many times, which, by definition of J_i is equivalent to $P_i(J_i = \infty) = 1$. Hence $E_i J_i = \infty$ also. If $E_i J_i = \infty$, then, owing to the fact that J_i is a geometric random variable, we see that its expectation cannot be infinity without J_i itself being infinity with probability one. The equivalence of the first two statements was proved before the lemma. □

Lemma 6. *The following are equivalent:*
1. State i is transient.
2. $f_{ii} = P_i(T_i < \infty) < 1$.
3. $P_i(J_i < \infty) = 1$.
4. $E_i J_i < \infty$.

Proof. State i is transient if, by definition, it is visited finitely many times with probability 1. This, by the definition of J_i, is equivalent to $P_i(J_i < \infty) = 1$, and, since J_i is a geometric random variable, this implies that $E_i J_i < \infty$. On the other hand, if we assume $E_i J_i < \infty$, then, clearly, J_i cannot take value $+\infty$ with positive probability; therefore $P_i(J_i < \infty) = 1$. The equivalence of the first two statements was proved above. □

Corollary 6. *If i is a transient state then $p_{ii}^{(n)} \to 0$, as $n \to \infty$.*

Proof. If i is transient then $E_i J_i < \infty$. But

$$E_i J_i = E_i \sum_{n=1}^{\infty} \mathbf{1}(X_n = i) = \sum_{n=1}^{\infty} P_i(X_n = i) = \sum_{n=1}^{\infty} p_{ii}^{(n)}.$$

But if a series is finite then the summands go to 0, i.e. $p_{ii}^{(n)} \to 0$, as $n \to \infty$. □

10.1 First hitting time decomposition

Define

$$f_{ij}^{(n)} := P_i(T_j = n),$$

i.e. the probability to hit state j for the first time at time $n \geq 1$, assuming that the chain starts from state i. Notice the difference with $p_{ij}^{(n)}$. The latter is the probability to be in state j at time n (not necessarily for the first time) starting from i. Clearly,

$$f_{ij}^{(n)} \leq p_{ij}^{(n)},$$

but the two sequences are related in an exact way:

Lemma 7.

$$p_{ij}^{(n)} = \sum_{m=1}^{n} f_{ij}^{(m)} p_{jj}^{(n-m)}, \quad n \geq 1.$$

Proof. From Elementary Probability,

$$\begin{aligned} p_{ij}^{(n)} = P_i(X_n = j) &= \sum_{m=1}^{n} P_i(T_j = m, X_n = j) \\ &= \sum_{m=1}^{n} P_i(T_j = m) P_i(X_n = j \mid T_j = m) \end{aligned}$$

The first term equals $f_{ij}^{(m)}$, by definition. For the second term, we use the Strong Markov Property:

$$P_i(X_n = j \mid T_j = m) = P_i(X_n = j \mid X_m = j) = p_{jj}^{(n-m)}.$$

□

Corollary 7. *If j is a transient state then $p_{ij}^{(n)} \to 0$, as $n \to \infty$, for all states i.*

Proof. If j is transient, we have $\sum_{n=1}^{\infty} p_{jj} = E_j J_j < \infty$. Now, by summing over n the relation of Lemma 7, we obtain (after interchanging the order of the two sums in the right hand side)

$$\sum_{n=1}^{\infty} p_{ij}^{(n)} = \sum_{m=1}^{\infty} f_{ij}^{(m)} \sum_{n=m}^{\infty} p_{jj}^{(n-m)} = (1 + E_j J_j) \sum_{m=1}^{\infty} f_{ij}^{(m)} = (1 + E_j J_j) P_i(T_j < \infty),$$

which is finite, and hence $p_{ij}^{(n)}$ as $n \to \infty$. □

10.2 Communicating classes and recurrence/transience

Recall the definition of the relation "state i leads to state j", denoted by $i \rightsquigarrow j$: it means that, starting from i, the probability to go to j in some finite time is positive. In other words, if we let

$$f_{ij} := P_i(T_j < \infty) \,,$$

we have

$$i \rightsquigarrow j \iff f_{ij} > 0.$$

We now show that recurrence is a class property. (Remember: another property that was a class property was the property that the state have a certain period.)

Theorem 10. *If i is recurrent and if $i \rightsquigarrow j$ then j is recurrent and*

$$f_{ij} = P_i(T_j < \infty) = 1.$$

In particular, $j \rightsquigarrow i$, and

$$g_{ij} := P_i(T_j < T_i) > 0.$$

Proof. Start with $X_0 = i$. If i is recurrent and $i \rightsquigarrow j$ then there is a positive probability ($= f_{ij}$) that j will appear in one of the i.i.d. excursions $\mathscr{X}_i^{(0)}, \mathscr{X}_i^{(1)}, \ldots$, and so the probability that j will appear in a specific excursion is positive. So the random variables

$$\delta_{j,r} := \mathbf{1}\big(j \text{ appears in excursion } \mathscr{X}_i^{(r)}\big), \quad r = 0, 1, \ldots$$

are i.i.d. and since they take the value 1 with positive probability, infinitely many of them will be 1 (with probability 1), showing that j will appear in infinitely many of the excursions for sure. Hence, not only $f_{ij} > 0$, but also $f_{ij} = 1$. Hence, $j \rightsquigarrow i$. The last statement is simply an expression in symbols of what we said above. Indeed, the probability that j will appear in a specific excursion equals $g_{ij} = P_i(T_j < T_i)$. □

Corollary 8. *In a communicating class, either all states are recurrent or all states are transient.*

Proof. If i is recurrent then every other j in the same communicating class satisfies $i \leftrightsquigarrow j$, and hence, by the above theorem, every other j is also recurrent. If a communicating class contains a state which is transient then, necessarily, all other states are also transient. □

Theorem 11. *If i is inessential then it is transient.*

Proof. Suppose that i is inessential, i.e. there is a state j such that $i \rightsquigarrow j$ but $j \not\rightsquigarrow i$. Hence there is m such that

$$P_i(A) = P_i(X_m = j) > 0,$$

but

$$P_i(A \cap B) = P_i(X_m = j, \quad X_n = i \text{ for infinitely many } n) = 0.$$

But then,

$$0 < P_i(A) = P_i(A \cap B) + P_i(A \cap B^c) = P_i(A \cap B^c) \le P_i(B^c),$$

i.e.

$$P(B) = P_i(X_n = i \text{ for infinitely many } n) < 1.$$

Hence, $P_i(B) = 0$, and so i is a transient state. □

So if a communicating class is recurrent then it contains no inessential states and so it is closed.

Theorem 12. *Every finite closed communicating class is recurrent.*

Proof. Let C be the class. By closedness, X remains in C, if started in C. By finiteness, some state i must be visited infinitely many times. This state is recurrent. Hence all states are recurrent. □

11 Positive recurrence

Recall that if we fix a state i and let T_i be the first return time to i then either $P_i(T_i < \infty) = 1$ (in which case we say that i is recurrent) or $P_i(T_i < \infty) < 1$ (transient state).

We now take a closer look at recurrence. Recall that a finite random variable may not necessarily have finite expectation.

Example: Let $U_0, U_1, U_2, \dots$ be i.i.d. random variables, uniformly distributed in the unit interval $[0, 1]$. Let

$$T := \inf\{n \ge 1 : \; U_n > U_0\}.$$

Thus T represents the number of variables we need to draw to get a value larger than the initial one. It should be clear that this can, for example, appear in a certain game. What is the expectation of T? First, observe that T is finite. Indeed,

$$\begin{aligned} P(T > n) &= P(U_1 \le U_0, \dots, U_n \le U_0) \\ &= \int_0^1 P(U_1 \le x, \dots, U_n \le x \mid U_0 = x) dx \\ &= \int_0^1 x^n dx = \frac{1}{n+1}. \end{aligned}$$

Therefore,

$$P(T < \infty) = \lim_{n\to\infty} P(T \le n) = \lim_{n\to\infty} (1 - \frac{1}{n+1}) = 1.$$

But

$$ET = \sum_{n=1}^{\infty} P(T > n) = \sum_{n=1}^{\infty} \frac{1}{n+1} = \infty.$$

You are invited to think about the meaning of this: If the initial value U_0 is unknown, then it takes, on the average, an infinite number of steps to exceed it!

We now classify a recurrent state i as follows:

We say that i is POSITIVE RECURRENT if $E_i T_i < \infty$.

We say that i is NULL RECURRENT if $E_i T_i = \infty$.

Our goal now is to show that a chain possessing a positive recurrent state also possesses a stationary distribution. Before doing that, we shall discover what the form of this stationary distribution is by using the Law of Large Numbers

12 Law (=Theorem) of Large Numbers in Probability Theory

The following result is, arguably, the Fundamental Theorem of Probability Theory. It is traditional to refer to it as "Law". But this is misleading. It is not a Law, it is a Theorem that can be derived from the axioms of Probability Theory.

We state it here without proof.

Theorem 13. *Let $Z_1, Z_2, \dots$ be i.i.d. random variables.*
(i) Suppose $E|Z_1| < \infty$ and let $\mu = EZ_1$. Then

$$P\left(\lim_{n\to\infty} \frac{Z_1 + \cdots + Z_n}{n} = \mu\right) = 1.$$

(ii) Suppose $E\max(Z_1, 0) = \infty$, but $E\min(Z_1, 0) > -\infty$. Then

$$P\left(\lim_{n\to\infty} \frac{Z_1 + \cdots + Z_n}{n} = \infty\right) = 1.$$

13 Law of Large Numbers for Markov chains

The law of large numbers stated above requires *independence*. When we have a Markov chain, the sequence $X_0, X_1, X_2, \dots$ of successive states is not an independent sequence. To apply the law of large numbers we must create some kind of independence. To our rescue comes the *regenerative structure* of a Markov chain identified at an earlier section.

13.1 Functions of excursions

Namely, given a recurrent state $a \in S$ (which will be referred to as the GROUND STATE for no good reason other than that it has been fixed once and for all) the sequence of excursions

$$\mathscr{X}_a^{(1)}, \quad \mathscr{X}_a^{(2)}, \quad \mathscr{X}_a^{(3)}, \dots$$

forms an *independent* sequence of (generalised) random variables. Since functions of independent random variables are also independent, we have that

$$G(\mathscr{X}_a^{(1)}), \quad G(\mathscr{X}_a^{(2)}), \quad G(\mathscr{X}_a^{(3)}), \dots$$

are i.i.d. random variables, for reasonable choices of the functions G taking values in $\mathbb{R}$.

Example 1: To each state $x \in S$, assign a reward $f(x)$. Then, for an excursion $\mathscr{X}$, define $G(\mathscr{X})$ to be the maximum reward received during this excursion. Specifically, in this example,

$$G(\mathscr{X}_a^{(r)}) = \max\{f(X_t) : \; T_a^{(r)} \le t < T_a^{(r+1)}\}.$$

Example 2: Let $f(x)$ be as in example 1, but define $G(\mathscr{X})$ to be the total reward received over the excursion $\mathscr{X}$. Thus,

$$G(\mathscr{X}_a^{(r)}) = \sum_{T_a^{(r)} \le t < T_a^{(r+1)}} f(X_t). \tag{7}$$

Whichever the choice of G, the law of large numbers tells us that

$$\frac{G(\mathscr{X}_a^{(1)}) + \dots + G(\mathscr{X}_a^{(n)})}{n} \to EG(\mathscr{X}_a^{(1)}), \tag{8}$$

as $n \to \infty$, with probability 1. Note that

$$EG(\mathscr{X}_a^{(1)}) = EG(\mathscr{X}_a^{(2)}) = \dots = E\big[G(\mathscr{X}_a^{(0)}) \mid X_0 = a\big] = E_a G(\mathscr{X}_a^{(0)}).$$

because the excursions $\mathscr{X}_a^{(1)}, \quad \mathscr{X}_a^{(2)}, \dots$ are i.i.d., and because if we start with $X_0 = a$, the initial excursion $\mathscr{X}_a^{(0)}$ also has the same law as the rest of them. The final equality in the last display is merely our notation for the expectation of a random variable conditional on $X_0 = a$.

13.2 Average reward

Consider now a function $f : S \to \mathbb{R}$, which, as in Example 2, we can think of as *a reward* received when the chain is in state x. We are interested in the existence of the limit

$$\text{time-average reward} = \lim_{t \to \infty} \frac{1}{t} \sum_{k=0}^{t} f(X_k),$$

which represents the 'time-average reward' received. In particular, we shall assume that f is bounded and shall examine those conditions which yield a nontrivial limit.

Theorem 14. *Suppose that a Markov chain possesses a positive recurrent state $a \in S$ (i.e. $E_a T_a < \infty$) and assume that the distribution of X_0 is such that*

$$P(T_a < \infty) = 1.$$

Let $f : S \to \mathbb{R}_+$ be a bounded 'reward' function. Then, with probability one, the 'time-average reward' $\overline{f}$ exists and is a deterministic quantity:

$$\boxed{P\left(\lim_{t\to\infty} \frac{1}{t} \sum_{n=0}^{t} f(X_n) = \overline{f}\right) = 1},$$

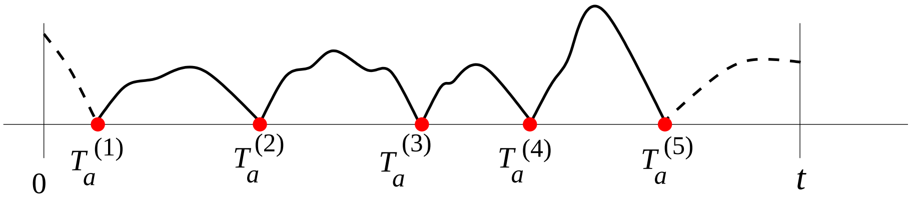
where

$$\boxed{\overline{f} = \frac{E_a \sum_{n=0}^{T_a-1} f(X_n)}{E_a T_a}} = \sum_{x\in S} f(x)\pi(x).$$

Proof. We shall present the proof in a way that the constant $\overline{f}$ will reveal itself–it will be discovered. The idea is this: Suppose that t is large enough. We are looking for the total reward received between 0 and t. We break the sum into the total reward received over the initial excursion, plus the total reward received over the next excursion, and so on. Thus we need to keep track of the number, N_t, say, of complete excursions from the ground state a that occur between 0 and t.

$$N_t := \max\{r \geq 1 : \; T_a^{(r)} \leq t\}.$$

For example, in the figure below, $N_t = 5$.



We write the total reward as a sum of rewards over complete excursions plus a first and a last term:

$$\sum_{n=0}^{t} f(X_n) = \sum_{r=1}^{N_t-1} G_r + G_t^{\text{first}} + G_t^{\text{last}}.$$

In other words, G_r is given by (7), G_t^{first} is the reward up to time $T_a = T_a^{(1)}$ or t (whichever is smaller), and G_t^{last} is the reward over the last incomplete cycle. We assumed that f is bounded, say $|f(x)| \leq B$. Hence $|G_t^{\text{first}}| \leq BT_a$. Since $P(T_a < \infty) = 1$, we have $BT_a/t \to 0$, as $t \to \infty$, and so

$$\frac{1}{t} G_t^{\text{first}} \to 0,$$

as $t \to \infty$, with probability one. A similar argument shows that

$$\frac{1}{t} G_t^{\text{last}} \to 0,$$

also. So we are left with the sum of the rewards over complete cycles. Write now

$$\frac{1}{t}\sum_{r=1}^{N_t-1} G_r = \frac{N_t}{t}\,\frac{1}{N_t}\sum_{r=1}^{N_t-1} G_r. \tag{9}$$

Look again what the Law of Large Numbers (8) tells us: that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{r=1}^{n} G_r = EG_1,$$

with probability one. Since $T_a^{(r)} \to \infty$, as $r \to \infty$, it follows that the number N_t of complete cycles before time t will also be tending to ∞, as $t \to \infty$. Hence,

$$\lim_{t\infty}\frac{1}{N_t}\sum_{r=1}^{N_t-1} G_r = EG_1. \tag{10}$$

Now look at the sequence

$$T_a^{(r)} = T_a^{(1)} + \sum_{k=1}^{r}(T_a^{(k)} - T_a^{(k-1)})$$

and see what the Law of Large Numbers again, Since $T_a^{(1)}/r \to 0$, and since $T_a^{(k)} - T_a^{(k-1)}$, $k = 1, 2, \ldots$, are i.i.d. random variables with common expectation E_aT_a, we have

$$\lim_{r\to\infty}\frac{T_a^{(r)}}{r} = E_aT_a. \tag{11}$$

By definition, we have that the visit with index $r = N_t$ to the ground state a is the last before t (see also figure above):

$$T_a^{(N_t)} \le t < T_a^{(N_t+1)}.$$

Hence

$$\frac{T_a^{(N_t)}}{N_t} \le \frac{t}{N_t} < \frac{T_a^{(N_t+1)}}{N_t}.$$

But (11) tells us that

$$\lim_{t\to\infty}\frac{T_a^{(N_t)}}{N_t} = \lim_{t\to\infty}\frac{T_a^{(N_t+1)}}{N_t} = E_aT_a.$$

Hence

$$\lim_{t\to\infty}\frac{N_t}{t} = \frac{1}{E_aT_a}. \tag{12}$$

Using (10) and (12) in (9) we have

$$\lim_{t\to\infty}\frac{1}{t}\sum_{r=1}^{N_t-1} G_r = \frac{EG_1}{E_aT_a}.$$

Thus

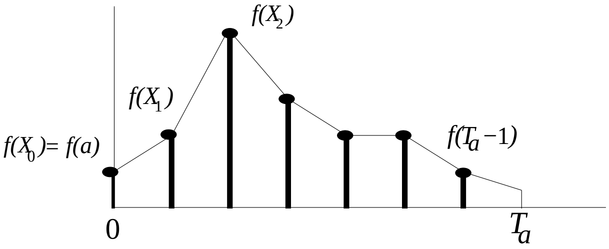
$$\lim_{t\to\infty}\frac{1}{t}\sum_{n=0}^{t} f(X_n) = \frac{EG_1}{E_aT_a} =: \overline{f}.$$

To see that $\overline{f}$ is as claimed, just observe that

$$EG_1 = E \sum_{T_a^{(1)} \le t < T_a^{(1+1)}} f(X_t) = E_a \sum_{0 \le t < T_a} f(X_t),$$

from the Strong Markov Property. □

Comment 1: It is important to understand what the formula says. The constant $\overline{f}$ is a sort of average over an excursion. The numerator of $\overline{f}$ is computed by summing up the rewards over the excursion (making sure that not both endpoints are included). The denominator is the duration of the excursion. Note that we include only one endpoint of the excursion, not both. We may write $EG_1 = E_a \sum_{0 \le t < T_a} f(X_t)$ or $EG_1 = E_a \sum_{0 < t \le T_a} f(X_t)$.



Comment 2: A very important particular case occurs when we choose

$$f(x) := \mathbf{1}(x = b),$$

i.e. we let f to be 1 at the state b and 0 everywhere else. The theorem above then says that, for all $b \in S$,

$$\boxed{\lim_{t\to\infty} \frac{1}{t} \sum_{n=1}^{t} \mathbf{1}(X_n = b) = \frac{E_a \sum_{k=0}^{T_a - 1} \mathbf{1}(X_k = b)}{E_a T_a},} \tag{13}$$

with probability 1.

Comment 3: If in the above we let $b = a$, the ground state, then we see that the numerator equals 1. Hence

$$\boxed{\lim_{t\to\infty} \frac{1}{t} \sum_{n=1}^{t} \mathbf{1}(X_n = a) = \frac{1}{E_a T_a},} \tag{14}$$

with probability 1.

The physical interpretation of the latter should be clear: If, on the average, it takes $E_a T_a$ units of time between successive visits to the state a, then the long-run proportion of visits to the state a should equal $1/E_a T_a$.

Comment 4: Suppose that two states, a, b, which communicate and are positive recurrent. We can use b as a ground state instead of a. Applying formula (14) with b in place of a we find that the limit in (13) also equals $1/E_b T_b$. The equality of these two limits gives:

$$\begin{pmatrix} \text{average no. of times state } b \text{ is visited} \\ \text{between two successive visits to state } a \end{pmatrix} = E_a \sum_{k=0}^{T_a - 1} \mathbf{1}(X_k = b) = \frac{E_a T_a}{E_b T_b}.$$

14 Construction of a stationary distribution

Our discussions on the law of large numbers for Markov chains revealed that the quantity

$$\nu^{[a]}(x) := E_a \sum_{n=0}^{T_a-1} \mathbf{1}(X_n = x) \quad , \quad x \in S. \tag{15}$$

should play an important role. Indeed, we saw (Comment 2 above) that if a is a positive recurrent state then

$$\frac{\nu^{[a]}(x)}{E_a T_a}$$

is the long-run fraction of times that state x is visited.

Note: *Although $\nu^{[a]}$ depends on the choice of the 'ground' state a, we will soon realise that this dependence vanishes when the chain is irreducible. Hence the superscript $^{[a]}$ will soon be dropped. In view of this, notice that the choice of notation $\nu^{[a]}$ is justifiable through the notation* (6) *for the communicating class of the state a.*

Clearly, this should have something to do with the concept of stationary distribution discussed a few sections ago.

Proposition 2. *Fix a ground state $a \in S$ and assume that it is recurrent*[5]*. Consider the function $\nu^{[a]}$ defined in* (15)*. Then $\nu^{[a]}$ satisfies the balance equations:*

$$\nu^{[a]} = \nu^{[a]} \mathsf{P},$$

i.e.

$$\nu^{[a]}(x) = \sum_{y \in S} \nu^{[a]}(y) p_{yx}, \quad x \in S.$$

Moreover, for any state b such that $a \to x$ (a leads to x), we have

$$0 < \nu^{[a]}(x) < \infty.$$

Proof. We start the Markov chain with $X_0 = a$, where a is the fixed recurrent ground state whose existence was assumed. Recurrence tells us that $T_a < \infty$ with probability one. Since

$$a = X_{T_a},$$

we can write

$$\nu^{[a]}(x) = E_a \sum_{k=1}^{T_a} \mathbf{1}(X_n = x). \tag{16}$$

This was a trivial but important step: We replaced the $n = 0$ term with the $n = T_a$ term. Both of them equal 1, so there is nothing lost in doing so. The real reason for doing so is because we wanted to replace a function of $X_0, X_1, X_2, \ldots$ by a function of $X_1, X_2, X_3, \ldots$

[5] We stress that we do not need the stronger assumption of positive recurrence here.

and thus be in a position to apply first-step analysis, which is precisely what we do next:

$$\begin{aligned}
\nu^{[a]}(x) &= E_a \sum_{n=1}^{\infty} \mathbf{1}(X_n = x, n \le T_a) \\
&= \sum_{n=1}^{\infty} P_a(X_n = x, n \le T_a) \\
&= \sum_{n=1}^{\infty} \sum_{y \in S} P_a(X_n = x, X_{n-1} = y, n \le T_a) \\
&= \sum_{n=1}^{\infty} \sum_{y \in S} p_{yx} P_a(X_{n-1} = y, n \le T_a) \\
&= \sum_{y \in S} p_{yx} \; E_a \sum_{n=1}^{T_a} \mathbf{1}(X_{n-1} = y) \\
&= \sum_{y \in S} p_{yx} \; \nu^{[a]}(y),
\end{aligned}$$

where, in this last step, we used (16). So we have shown $\nu^{[a]}(x) = \sum_{y \in S} p_{yx} \; \nu^{[a]}(y)$, for all $x \in S$, which are precisely the balance equations.
We just showed that $\nu^{[a]} = \nu^{[a]}\mathsf{P}$. Therefore, $\nu^{[a]} = \nu^{[a]}\mathsf{P}^n$, for all $n \in \mathbb{N}$.
Next, let x be such that $a \rightsquigarrow x$; so there exists $n \ge 1$, such that $p_{ax}^{(n)} > 0$. Hence

$$\nu^{[a]}(x) \ge \nu^{[a]}(a) p_{ax}^{(n)} = p_{ax}^{(n)} > 0.$$

From Theorem 10, $x \rightsquigarrow a$; so there exists $m \ge 1$, such that $p_{xa}^{(m)} > 0$. Hence

$$1 = \nu^{[a]}(a) \ge \nu^{[a]}(x) p_{xa}^{(m)} > \nu^{[a]}(x),$$

so, indeed, $\nu^{[a]}(x) < \infty$. □

Note: *The last result was shown under the assumption that a is a recurrent state. We are now going to assume more, namely, that a is positive recurrent. First note that, from* (16),

$$\sum_{x \in S} \nu^{[a]}(x) = E_a \sum_{n=1}^{T_a} \sum_{x \in S} \mathbf{1}(X_n = x) = E_a \sum_{n=1}^{T_a} 1 = E_a T_a,$$

which, by definition, is finite when a is positive recurrent.

We now have:

Corollary 9. *Suppose that the Markov chain possesses a positive recurrent state a. Define*

$$\pi^{[a]}(x) := \frac{\nu^{[a]}(x)}{E_a T_a} = \frac{E_a \sum_{n=0}^{T_a - 1} \mathbf{1}(X_n = x)}{E_a T_a}, \quad x \in S. \tag{17}$$

This π is a stationary distribution.

In other words: If there is a positive recurrent state, then the set of linear equations

$$\pi(x) = \sum_{y \in S} \pi(y) p_{yx}, \quad x \in S$$
$$\sum_{y \in S} \pi(y) = 1$$

have at least one solution.

Proof. Since a is positive recurrent, we have $E_a T_a < \infty$, and so $\pi^{[a]}$ is not trivially equal to zero. We have already shown, in Proposition 2, that $\nu^{[a]}\mathsf{P} = \nu^{[a]}$. But $\pi^{[a]}$ is simply a multiple of $\nu^{[a]}$. Therefore, $\pi^{[a]}\mathsf{P} = \pi^{[a]}$. It only remains to show that $\pi^{[a]}$ is a probability, i.e. that it sums up to one. But

$$\sum_{x \in S} \pi^{[a]}(x) = \frac{1}{E_a T_a} \sum_{x \in S} \nu^{[a]}(x) = 1.$$

□

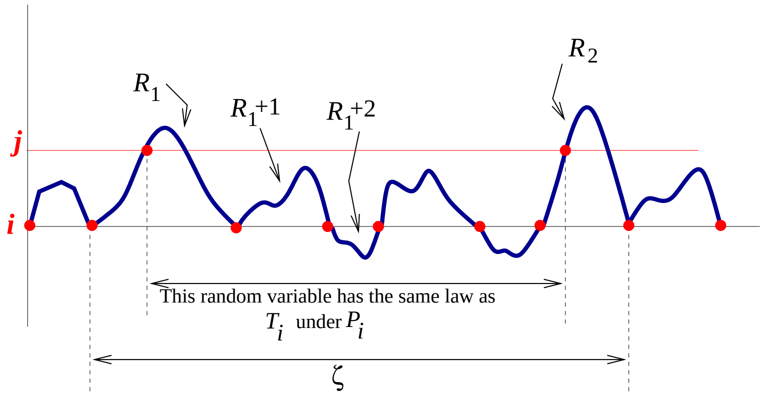
Pause for thought: *We have achieved something VERY IMPORTANT: if we have one, no matter which one, positive recurrent state, then we immediately have a stationary (probability) distribution! We DID NOT claim that this distribution is, in general, unique. This is what we will try to do in the sequel.*

15 Positive recurrence is a class property

We now show that (just like recurrence) positive recurrence is also a class property.

Theorem 15. *Let i be a positive recurrent state. Suppose that $i \rightsquigarrow j$. Then j is positive recurrent.*

Proof. Start with $X_0 = i$. Consider the excursions $\mathscr{X}_i^{(r)}$, $r = 0, 1, 2, \ldots$ We already know that j is recurrent. Hence j will occur in infinitely many excursions, and the probability $g_{ij} = P_i(T_j < \infty)$ that j occurs in a specific excursion is positive. Also, since the excursions are independent, the event that j occurs in a specific excursion is independent from the event that j occurs in the next excursion. In words, to decide whether j will occur in a specific excursion, we toss a coin with probability of heads $g_{ij} > 0$. The coins are i.i.d. If heads show up at the r-th excursion, then j occurs in the r-th excursion. Otherwise, if tails show up, then the r-th excursion does not contain state j. Let R_1 (respectively, R_2) be the index of the first (respectively, second) excursion that contains j.



We just need to show that the sum of the durations of excursions with indices $R_1, \dots, R_2$ has finite expectation. In other words, we just need to show that

$$\zeta := \sum_{r=1}^{R_2 - R_1 + 1} Z_r$$

has finite expectation, where the Z_r are i.i.d. with the same law as T_i. But, by Wald's lemma, the expectation of this random variable equals

$$E_i\zeta = (E_iT_i) \times E_i(R_2 - R_1 + 1) = (E_iT_i) \times (g_{ij}^{-1} + 1)$$

because $R_2 - R_1$ is a geometric random variable:

$$P_i(R_2 - R_1 = r) = (1 - g_{ij})^{r-1} g_{ij}, \quad r = 1, 2, \dots$$

Since $g_{ij} > 0$ and $E_iT_i < \infty$ we have that $E_i\zeta < \infty$. □

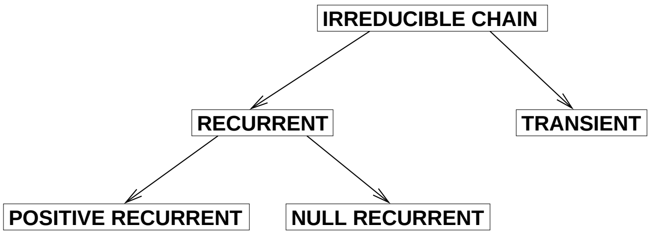
Corollary 10. *Let C be a communicating class. Then either all states in C are positive recurrent, or all states are null recurrent or all states are transient.*

An irreducible Markov chain is called recurrent if one (and hence all) states are recurrent.

An irreducible Markov chain is called positive recurrent if one (and hence all) states are positive recurrent.

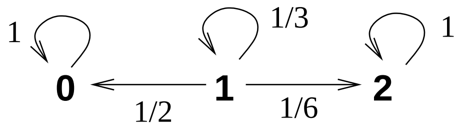
An irreducible Markov chain is called null recurrent if one (and hence all) states are null recurrent.

An irreducible Markov chain is called transient if one (and hence all) states are transient.



16 Uniqueness of the stationary distribution

At the risk of being too pedantic, we stress that a stationary distribution may not be unique.
Example: Consider the chain we've seen before:



Clearly, state 1 (and state 2) is positive recurrent, for totally trivial reasons (it is, after all, an absorbing state). Hence, there is a stationary distribution. But observe that there are many–infinitely many–stationary distributions: For each $0 \le \alpha \le 1$, the π defined by

$$\pi(0) = \alpha, \quad \pi(1) = 0, \quad \pi(2) = 1 - \alpha$$

is a stationary distribution. (You should think why this is OBVIOUS!) Notice that $1 \not\rightsquigarrow 2$. This lack of communication is, indeed, what prohibits uniqueness.

Theorem 16. *Consider positive recurrent irreducible Markov chain. Then there is a* <u>*unique*</u> *stationary distribution.*

Proof. Let π be a stationary distribution. Consider the Markov chain $(X_n, n \ge 0)$ and suppose that

$$P(X_0 = x) = \pi(x), \quad x \in S.$$

In other words, we start the Markov chain from the stationary distribution π. Then, for all n, we have

$$P(X_n = x) = \pi(x), \quad x \in S. \tag{18}$$

Since the chain is irreducible and positive recurrent we have $P_i(T_j < \infty) = 1$, for all states i and j. Fix a ground state a. Then

$$P(T_a < \infty) = \sum_{x \in S} \pi(x) P_x(T_a < \infty) = \sum_{x \in S} \pi(x) = 1.$$

By (13), following Theorem 14, we have, for all $x \in S$,

$$\frac{1}{t} \sum_{n=1}^{t} \mathbf{1}(X_n = x) \quad \to \quad \frac{E_a \sum_{k=0}^{T_a - 1} \mathbf{1}(X_k = x)}{E_a T_a} = \pi^{[a]}(x),$$

as $t \to \infty$, with probability one. By a theorem of Analysis (Bounded Convergence Theorem), recognising that the random variables on the left are nonnegative and bounded below 1, we can take expectations of both sides:

$$E\left(\frac{1}{t} \sum_{n=1}^{t} \mathbf{1}(X_n = x)\right) \quad \to \quad \pi^{[a]}(x),$$

as $t \to \infty$. But, by (18),

$$E\left(\frac{1}{t} \sum_{n=1}^{t} \mathbf{1}(X_n = x)\right) = \frac{1}{t} \sum_{n=1}^{t} P(X_n = x) = \pi(x).$$

Therefore $\pi(x) = \pi^{[a]}(x)$, for all $x \in S$. Thus, an *arbitrary* stationary distribution π must be equal to the *specific* one $\pi^{[a]}$. Hence there is only one stationary distribution. □

Corollary 11. *Consider a positive recurrent irreducible Markov chain and two states $a, b \in S$. Then*

$$\pi^{[a]} = \pi^{[b]}.$$

In particular, we obtain the CYCLE FORMULA*:*

$$\boxed{\frac{E_a \sum_{k=0}^{T_a-1} \mathbf{1}(X_k = x)}{E_a T_a} = \frac{E_b \sum_{k=0}^{T_b-1} \mathbf{1}(X_k = x)}{E_b T_b}, \qquad x \in S.}$$

Proof. There is a unique stationary distribution. Both $\pi^{[a]}$ and $\pi[b]$ are stationary distributions, so they are equal. The cycle formula merely re-expresses this equality. □

Corollary 12. *Consider a positive recurrent irreducible Markov chain with stationary distribution π. Then, for all $a \in S$,*

$$\boxed{\pi(a) = \frac{1}{E_a T_a}.}$$

Proof. There is only one stationary distribution. Hence $\pi = \pi^{[a]}$. In particular, $\pi(a) = \pi^{[a]}(a) = 1/E_a T_a$, from the definition of $\pi^{[a]}$. □

In Section 14 we assumed the existence of a positive recurrent state a and proved the existence of a stationary distribution π. The following is a sort of converse to that, which is useful because it is often used as a criterion for positive recurrence:

Corollary 13. *Consider an arbitrary Markov chain. Assume that a stationary distribution π exists. Then any state i such that $\pi(i) > 0$ is positive recurrent.*

Proof. Let i be such that $\pi(i) > 0$. Let C be the communicating class of i. Let $\pi(\cdot|C)$ be the distribution defined by restricting π to C:

$$\pi(x|C) := \frac{\pi(x)}{\pi(C)}, \qquad x \in C,$$

where $\pi(C) = \sum_{y \in C} \pi(y)$. Consider the restriction of the chain to C. Namely, delete all transition probabilities p_{xy} with $x \notin C$, $y \in C$. We can easily see that we thus obtain a Markov chain with state space C with stationary distribution $\pi(\cdot|C)$. This is in fact, the only stationary distribution for the restricted chain because C is irreducible. By the above corollary, $\pi(i|C) = \frac{1}{E_i T_i}$. But $\pi(i|C) > 0$. Therefore $E_i T_i < \infty$, i.e. i is positive recurrent. □

17 Structure of a stationary distribution

It is not hard to show now what an arbitrary stationary distribution looks like.

Consider a Markov chain. Decompose it into communicating classes. Then for each positive recurrent communicating class C there corresponds a unique stationary distribution π^C

which assigns positive probability to each state in C. This is defined by picking a state (any state!) $a \in C$ and letting $\pi^C := \pi^{[a]}$.

An arbitrary stationary distribution π must be a linear combination of these π^C.

Proposition 3. *To each positive recurrent communicating class C there corresponds a stationary distribution π^C. Any stationary distribution π is necessarily of the form*

$$\pi = \sum_{\substack{C:\ C \text{ positive recurrent} \\ \text{communicating class}}} \alpha_C \, \pi^C,$$

where $\alpha_C \geq 0$, such that $\sum_C \alpha_C = 1$.

Proof. Let π be an arbitrary stationary distribution. If $\pi(x) > 0$ then x belongs to some positive recurrent class C. For each such C define the conditional distribution

$$\pi(x|C) := \frac{\pi(x)}{\pi(C)}, \text{ where } \pi(C) := \sum_{x \in C} \pi(x).$$

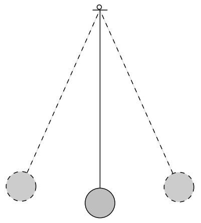
Then $(\pi(x|C), x \in C)$ is a stationary distribution. By our uniqueness result, $\pi(x|C) \equiv \pi^C(x)$. So we have that the above decomposition holds with $\alpha_C = \pi(C)$. □

18 Coupling and stability

18.1 Definition of stability

Stability refers to the convergence of the probabilities $P(X_n = x)$ as $n \to \infty$. So far we have seen, that, under irreducibility and positive recurrence conditions, the so-called CESARO AVERAGES $\frac{1}{n}\sum_{k=1}^{n} P(X_k = x)$ converge to $\pi(x)$, as $n \to \infty$. A sequence of real numbers a_n may not converge but its Cesaro averages $\frac{1}{n}(a_1 + \cdots + a_n)$ may converge; for example, take $a_n = (-1)^n$.

Why stability? There are physical reasons why we are interested in the stability of a Markov chain. A Markov chain is a simple model of a stochastic dynamical system. Stability of a deterministic dynamical system means that the system eventually converges to a fixed point (or, more generally, that it remains in a small region). For example, consider the motion of a pendulum under the action of gravity. We all know, from physical experience, that the pendulum will swing for a while and, eventually, it will settle to the vertical position.



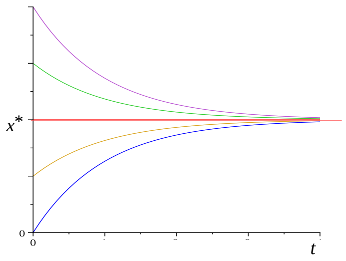
This is the stable position. There is dissipation of energy due to friction and that causes the settling down. In an ideal pendulum, without friction, the pendulum will perform an oscillatory motion ad infinitum. Still, we may it is stable because it does not (it cannot!) swing too far away.

The simplest differential equation As another example, consider the following differential equation

$$\dot{x} = -ax + b, \quad x \in \mathbb{R}.$$

There are myriads of applications giving physical interpretation to this, so I'll leave it upon the reader to decide whether he or she wants to consider this as an financial example, or an example in physics, or whatever. The thing to observe here is that there is an equilibrium point i.e. the one found by setting the right-hand side equal to zero:

$$x^* = b/a.$$



If we start with $x(t = 0) = x^*$, then $x(t) = x^*$ for all $t > 0$. If, moreover, $a > 0$, then regardless of the starting position, we have $x(t) \to x^*$, as $t \to \infty$. We say that x^* is a stable equilibrium.

A linear stochastic system We now pass on to stochastic systems. Since we are prohibited by the syllabus to speak about continuous time stochastic systems, let us consider a stochastic analogue of the above in discrete time:

$$X_{n+1} = \rho X_n + \xi_n.$$

Specifically, we here assume that $X_0, \xi_0, \xi_1, \xi_2, \ldots$ are given, mutually independent, random variables, and define a stochastic sequence $X_1, X_2, \ldots$ by means of the recursion above. We shall assume that the ξ_n have the same law. It follows that $|\rho| < 1$ is the condition for stability. But what does stability mean? In general, it cannot mean convergence to a fixed equilibrium point, simple because the ξ_n's depend on the time parameter n. However notice what happens when we solve the recursion. And we are in a happy situation here, for we can do it:

$$\begin{aligned}
X_1 &= \rho X_0 + \xi_0 \\
X_2 &= \rho X_1 + \xi_1 = \rho(\rho X_0 + \xi_0) + \xi_1 = \rho^2 X_0 + \rho\xi_0 + \xi_1 \\
X_3 &= \rho X_2 + \xi_2 = \rho^3 X_0 + \rho^2 \xi_0 + \rho\xi_1 + \xi_2 \\
&\cdots\cdots \\
X_n &= \rho^n X_0 + \rho^{n-1}\xi_0 + \rho^{n-2}\xi_1 + \cdots + \rho\xi_{n-2} + \xi_{n-1}.
\end{aligned}$$

Since $|\rho| < 1$, we have that the first term goes to 0 as $n \to \infty$. The second term,

$$\tilde{X}_n := \rho^{n-1}\xi_0 + \rho^{n-2}\xi_1 + \cdots + \rho\xi_{n-2} + \xi_{n-1},$$

which is a linear combination of $\xi_0, \ldots, \xi_{n-1}$ does not converge, but notice that if we let

$$\hat{X}_n := \rho^{n-1}\xi_{n-1} + \rho^{n-2}\xi_{n-2} + \cdots + \rho\xi_1 + \xi_0,$$

we have

$$P(\tilde{X}_n \le x) = P(\hat{X}_n \le x).$$

The nice thing is that, under nice conditions, (e.g. assume that the ξ_n are bounded) $\hat{X}_n$ does converge to

$$\hat{X}_\infty = \sum_{k=0}^{\infty} \rho^k \xi_k.$$

What this means is that, while the limit of X_n does not exists, the limit of its distribution does:

$$\lim_{n\to\infty} P(X_n \le x) = \lim_{n\to\infty} P(\tilde{X}_n \le x) = \lim_{n\to\infty} P(\hat{X}_n \le x) = P(\hat{X}_\infty \le x)$$

This is precisely the reason why we cannot, for time-homogeneous Markov chains, define stability in a way other than convergence of the distributions of the random variables.

18.2 The fundamental stability theorem for Markov chains

Theorem 17. *If a Markov chain is irreducible, positive recurrent and aperiodic with (unique) stationary distribution π then*

$$\lim_{n\to\infty} P(X_n = x) = \pi(x),$$

uniformly in x, for any initial distribution. In particular, the n-step transition matrix P^n, converges, as $n \to \infty$, to a matrix with rows all equal to π:

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi(j), \quad i, j \in S.$$

18.3 Coupling

To understand why the fundamental stability theorem works, we need to understand the notion of COUPLING.

Starting at an elementary level, suppose you are given the laws of two random variables X, Y but not their joint law. In other words, we know $f(x) = P(X = x)$, $g(y) = P(Y = y)$, but we are not given what $P(X = x, Y = y)$ is.

Coupling refers to the various methods for constructing this joint law.

A straightforward (often not so useful) construction is to assume that the random variables are independent and *define* $P(X = x, Y = y) = P(X = x)P(Y = y)$.

But suppose that we have some further requirement, for instance, to try to make the coincidence probability $P(X = Y)$ as large as possible. How should we then define what $P(X = x, Y = y)$ is?

Exercise: Let X, Y be random variables with values in $\mathbb{Z}$ and laws $f(x) = P(X = x)$, $g(y) = P(Y = y)$, respectively. Find the joint law $h(x, y) = P(X = x, Y = y)$ which respects the marginals, i.e. $f(x) = \sum_y h(x, y)$, $g(y) = \sum_x h(x, y)$, and which maximises $P(X = Y)$.

The coupling we are interested in is a coupling of processes. We are given the law of a stochastic process $X = (X_n, n \geq 0)$ and the law of a stochastic process $Y = (Y_n, n \geq 0)$, but we are not told what the joint probabilities are. A coupling of them refers to a joint construction on the same probability space.

Suppose that we have two stochastic processes as above and suppose that they have been coupled. We say that ransom time T is MEETING TIME of the two processes if

$$X_n = Y_n \text{ for all } n \geq T.$$

Note that it makes sense to speak of a meeting time, precisely because of the assumption that the two processes have been constructed together. This T is a random variables that takes values in the set of times or it may take values $+\infty$, on the event that the two processes never meet. The following result is the so-called COUPLING INEQUALITY:

Proposition 4. *Let T be a meeting of two coupled stochastic processes X, Y. Then, for all $n \geq 0$, and all $x \in S$,*

$$|P(X_n = x) - P(Y_n = x)| \leq P(T > n).$$

If, in particular, T is finite with probability one, i.e.

$$P(T < \infty) = 1,$$

then

$$|P(X_n = x) - P(Y_n = x)| \to 0, \quad \text{as } n \to \infty,$$

uniformly in $x \in S$.

Proof. We have

$$\begin{aligned} P(X_n = x) &= P(X_n = x, n < T) + P(X_n = x, n \geq T) \\ &= P(X_n = x, n < T) + P(Y_n = x, n \geq T) \\ &\leq P(n < T) + P(Y_n = x), \end{aligned} \tag{19}$$

and hence

$$P(X_n = x) - P(Y_n = x) \leq P(T > n).$$

Repeating (19), but with the roles of X and Y interchanged, we obtain

$$P(Y_n = x) - P(X_n = x) \leq P(T > n).$$

Combining the two inequalities we obtain the first assertion:

$$|P(X_n = x) - P(Y_n = x)| \leq P(T > n), \quad n \geq 0, \quad x \in S.$$

Since this holds for all $x \in S$ and since the right hand side does not depend on x, we can write it as

$$\sup_{x \in S} |P(X_n = x) - P(Y_n = x)| \leq P(T > n), \quad n \geq 0.$$

Assume now that $P(T < \infty) = 1$. Then

$$\lim_{n\to\infty} P(T > n) = P(T = \infty) = 0.$$

Therefore,

$$\sup_{x\in S} |P(X_n = x) - P(Y_n = x)| \to 0,$$

as $n \to \infty$. $\square$

18.4 Proof of the fundamental stability theorem

First, review the assumptions of the theorem: the chain is irreducible, positive recurrent and aperiodic. Let p_{ij} denote, as usual, the transition probabilities. We know that there is a unique stationary distribution, denoted by π. We shall consider the laws of two processes. The first process, denoted by $X = (X_n, n \geq 0)$ has the law of a Markov chain with transition probabilities p_{ij} and X_0 distributed according to some arbitrary distribution μ. The second process, denoted by $Y = (Y_n, n \geq 0)$ has the law of a Markov chain with transition probabilities p_{ij} and Y_0 distributed according to the stationary distribution μ. So both processes are positive recurrent aperiodic Markov chains with the same transition probabilities; they differ only in how the initial states are chosen. It is very important to note that we have only defined the law of X and the law of Y but we have not defined a joint law; in other words, we have not defined joint probabilities such as $P(X_n = x, X_m = y)$. We now couple them by doing the straightforward thing: we assume that $X = (X_n, n \geq 0)$ is independent of $Y = (Y_n, n \geq 0)$. Having coupled them, we can define joint probabilities, and it also makes sense to consider their first meeting time:

$$T := \inf\{n \geq 0 : \; X_n = Y_n\}.$$

Consider the process

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

Then $(W_n, n \geq 0)$ is a Markov chain with state space $S \times S$. Its initial state $W_0 = (X_0, Y_0)$ has distribution $P(W_0 = (x, y) = \mu(x)\pi(y)$. Its (1-step) transition probabilities are

$$\begin{aligned} q_{(x,x'),(y,y')} &:= P\big(W_{n+1} = (x', y') \mid W_n = (x, y)\big) \\ &= P(X_{n+1} = x' \mid X_n = x) \; P(Y_{n+1} = y' \mid Y_n = y) = p_{x,x'} \; p_{y,y'}. \end{aligned}$$

Its n-step transition probabilities are

$$q^{(n)}_{(x,x'),(y,y')} = p^{(n)}_{x,x'} \; p^{(n)}_{y,y'}.$$

From Theorem 3 and the aperiodicity assumption, we have that $p^{(n)}_{x,x'} > 0$ and $p^{(n)}_{y,y'}$ for all large n, implying that $q^{(n)}_{(x,x'),(y,y')}$ for all large n, and so $(W_n, n \geq 0)$ is an irreducible chain. Notice that

$$\sigma(x, y) := \pi(x)\pi(y), \quad (x, y) \in S \times S,$$

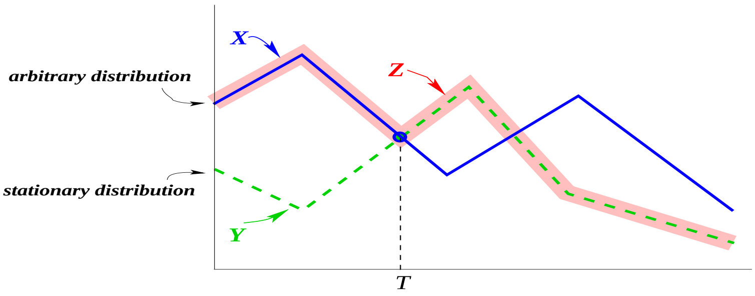
is a stationary distribution for $(W_n, n \geq 0)$. (Indeed, had we started with both X_0 and Y_0 independent and distributed according to π, then, for all $n \geq 1$, X_n and Y_n would be independent and distributed according to π.) By positive recurrence, $\pi(x) > 0$ for all $x \in S$.

Therefore $\sigma(x,y) > 0$ for all $(x,y) \in S \times S$. Hence $(W_n, n \geq 0)$ is positive recurrent. In particular,

$$P(T < \infty) = 1.$$

Define now a third process by:

$$Z_n := X_n \mathbf{1}(n < T) + Y_n \mathbf{1}(n \geq T).$$



Thus, Z_n equals X_n before the meeting time of X and Y; after the meeting time, $Z_n = Y_n$. Obviously, $Z_0 = X_0$ which has an arbitrary distribution μ, by assumption. Moreover, $(Z_n, \geq 0)$ is a Markov chain with transition probabilities p_{ij}.

$$\text{Hence } (Z_n, n \geq 0) \text{ is identical in law to } (X_n, n \geq 0).$$

In particular, for all n and x,

$$P(X_n = x) = P(Z_n = x).$$

Clearly, T is a meeting time between Z and Y. By the coupling inequality,

$$|P(Z_n = x) - P(Y_n = x)| \leq P(T > n).$$

Since $P(Z_n = x) = P(X_n = x)$, and since $P(Y_n = x) = \pi(x)$, we have

$$|P(X_n = x) - \pi(x)| \leq P(T > n),$$

which converges to zero, as $n \to \infty$, uniformly in x, precisely because $P(T < \infty) = 1$. □

Example showing that the aperiodicity assumption is essential for the method used in the proof: If p_{ij} are the transition probabilities of an irreducible chain X and if Y is an independent copy of X, then the process $W = (X, Y)$ may fail to be irreducible. For example, suppose that $S = \{0, 1\}$ and the p_{ij} are

$$p_{01} = p_{10} = 1.$$

Then the product chain has state space $S \times S = \{(0,0), (0,1), (1,0), (1,1)\}$ and the transition probabilities are

$$q_{(0,0),(1,1)} = q_{(1,1),(0,0)} = q_{(0,1),(1,0)} = q_{(1,0),(0,1)} = 1.$$

This is not irreducible. However, if we modify the original chain and make it aperiodic by, say,

$$p_{00} = 0.01, \ p_{01} = 0.99, \ p_{10} = 1,$$

then the product chain becomes irreducible.

Example showing that the aperiodicity assumption is essential for the limiting result: Consider the same chain as above:

$$p_{01} = p_{10} = 1.$$

Suppose $X_0 = 0$. Then $X_n = 0$ for even n and $= 1$ for odd n. Therefore $p_{00}^{(n)} = 1$ for even n and $= 0$ for odd n. Thus, $\lim_{n\to\infty} p_{00}^{(n)}$ does not exist. However, if we modify the chain by

$$p_{00} = 0.01, \; p_{01} = 0.99, \; p_{10} = 1,$$

then

$$\lim_{n\to\infty} p_{00}^{(n)} = \lim_{n\to\infty} p_{10}^{(n)} = \pi(0), \quad \lim_{n\to\infty} p_{11}^{(n)} = \lim_{n\to\infty} p_{01}^{(n)} = \pi(1),$$

where $\pi(0), \pi(1)$ satisfy

$$0.99\pi(0) = \pi(1), \quad \pi(0) + \pi(1) = 1,$$

i.e. $\pi(0) \approx 0.503$, $\pi(1) \approx 0.497$. In matrix notation,

$$\lim_{n\to\infty} \begin{pmatrix} p_{00}^{(n)} & p_{01}^{(n)} \\ p_{10}^{(n)} & p_{11}^{(n)} \end{pmatrix} = \begin{pmatrix} 0.503 & 0.497 \\ 0.503 & 0.497 \end{pmatrix}.$$

18.5 Terminology

A Markov chain which is irreducible and positive recurrent will be called ERGODIC.

A Markov chain which is irreducible and positive recurrent and aperiodic will be called MIXING.

The reader is asked to pay attention to this terminology which is different from the one used in the Markov chain literature. Indeed, most people use the term "ergodic" for an irreducible, positive recurrent and aperiodic Markov chain. But this is "wrong". [6] *The reason is that the term "ergodic" means something specific in Mathematics (c.f. Parry, 1981), and this is expressed by our terminology.*

18.6 Periodic chains

Suppose now that we have a positive recurrent irreducible chain. By the results of Section 14, we know we have a unique stationary distribution π and that $\pi(x) > 0$ for all states x.

Suppose that the period of the chain is d. By the results of §5.4 and, in particular, Theorem 4, we can decompose the state space into d classes $C_0, C_1, \ldots, C_{d-1}$, such that the chain moves cyclically between them: from C_0 to C_1, from C_1 to C_2, and so on, from C_d back to C_0.

It is not difficult to see that

[6] We put "wrong" in quotes. Terminology cannot be, *per se* wrong. It is, however, customary to have a consistent terminology.

Lemma 8. *The chain $(X_{nd}, n \geq 0)$ consists of d closed communicating classes, namely $C_0, \ldots, C_{d-1}$. All states have period 1 for this chain.*

Hence if we restrict the chain $(X_{nd}, n \geq 0)$ to any of the sets C_r we obtain a positive recurrent irreducible aperiodic chain, and the previous results apply.

We shall show that

Lemma 9.

$$\sum_{i \in C_r} \pi(i) = 1/d,$$

for all $r = 0, 1, \ldots, d-1$.

Proof. Let $\alpha_r := \sum_{i \in C_r} \pi(i)$. We have that π satisfies the balance equations:

$$\pi(i) = \sum_{j \in S} \pi(j) p_{ji}, \quad i \in S.$$

Suppose $i \in C_r$. Then $p_{ji} = 0$ unless $j \in C_{r-1}$. So

$$\pi(i) = \sum_{j \in C_{r-1}} \pi(j) p_{ji}, \quad i \in C_r.$$

Summing up over $i \in C_r$ we obtain

$$\alpha_r = \sum_{i \in C_r} \pi(i) = \sum_{j \in C_{r-1}} \pi(j) \sum_{i \in C_r} p_{ji} = \sum_{j \in C_{r-1}} \pi(j) = \alpha_{r-1}.$$

Since the α_r add up to 1, each one must be equal to $1/d$. □

Suppose that $X_0 \in C_r$. Then, the (unique) stationary distribution of $(X_{nd}, n \geq 0)$ is $d\pi(i), i \in C_r$. Since $(X_{nd}, n \geq 0)$ is aperiodic, we have

$$p_{ij}^{(nd)} = P_i(X_{nd} = j) \to d\pi(j), \text{ as } n \to \infty,$$

for all $i, j \in C_r$.

Can we also find the limit when i and j do not belong to the same class? Suppose that $i \in C_k$, $j \in C_\ell$. Then, starting from i, the original chain enters C_ℓ in $r = \ell - k$ steps (where r is taken to be a number between 0 and $d-1$, after reduction by d) and, thereafter, if sampled every d steps, it remains in C_ℓ. This means that:

Theorem 18. *Suppose that p_{ij} are the transition probabilities of a positive recurrent irreducible chain with period d. Let $i \in C_k$, $j \in C_\ell$, and let $r = \ell - k$ (modulo d). Then*

$$p_{ij}^{(r+nd)} \to d\pi(j), \text{ as } n \to \infty,$$

where π is the stationary distribution.

19 Limiting theory for Markov chains*

We wish to complete our discussion about the possible limits of the n-step transition probabilities $p_{ij}^{(n)}$, as $n \to \infty$.

Recall that we know what happens when j is positive recurrent and i is in the same communicating class as j If the period is 1, then $p_{ij}^{(n)} \to \pi(j)$ (Theorems 17). If the period is d then the limit exists only along a selected subsequence (Theorem 18). We also saw in Corollary 7 that, if j is transient, then $p_{ij}^{(n)} \to 0$ as $n \to \infty$, for all $i \in S$.

19.1 Limiting probabilities for null recurrent states

Theorem 19. *If j is a <u>null recurrent state</u> then*

$$p_{ij}^{(n)} \to 0, \qquad \textit{as } n \to \infty,$$

for all $i \in S$.

The proof will be based on two results from Analysis, the first of which is known as Helly-Bray Lemma:

Lemma 10. *Consider, for each $n = 1, 2, \ldots$, a probability distribution on $\mathbb{N}$, i.e. $p^{(n)} = \left(p_1^{(n)}, p_2^{(n)}, p_3^{(n)}, \ldots\right)$, where $\sum_{i=1}^{\infty} p_i^{(n)} = 1$ and $p_i^{(n)} \geq 0$ for all i. Then we can find a sequence $n_1 < n_2 < n_3 < \cdots$ of integers such that $\lim_{k\to\infty} p_i^{(n_k)}$ exists for all $i \in \mathbb{N}$.*

Proof. Since the numbers $p_1^{(n)}$ are contained between 0 and 1, there must be a subsequence $n_1^1 < n_2^1 < n_3^1 < \cdots$, such that $\lim_{k\to\infty} p_1^{(n_k^1)}$ exists and equals, say, p_1. Now consider the numbers $p_2^{(n_k^1)}$, $k = 1, 2, \ldots$. Since they are contained between 0 and 1 there must be a subsequence n_k^2 of n_k^1 such that $\lim_{k\to\infty} p_1^{(n_k^2)}$ exists and equals, say, p_2. It is clear how we continue. For each i we have found subsequence n_k^i, $k = 1, 2, \ldots$ such that $\lim_{k\to\infty} p_i^{(n_k^i)}$ exists and equals, say, p_i. We have, schematically,

$$\begin{array}{cccccc}
\boxed{p_1^{(n_1^1)}} & p_1^{(n_2^1)} & p_1^{(n_3^1)} & \cdots & \to & p_1 \\
p_1^{(n_1^2)} & \boxed{p_1^{(n_2^2)}} & p_1^{(n_3^2)} & \cdots & \to & p_2 \\
p_1^{(n_1^3)} & p_1^{(n_2^3)} & \boxed{p_1^{(n_3^3)}} & \cdots & \to & p_3 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots
\end{array}$$

Now pick the numbers in the diagonal. By construction, the sequence n_k^2 of the second row is a subsequence of n_k^1 of the first row; the sequence n_k^3 of the third row is a subsequence of n_k^2 of the second row; and so on. So the diagonal subsequence $(n_k^k, k = 1, 2, \ldots)$ is a subsequence of each $(n_k^i, k = 1, 2, \ldots)$, for each i. Hence $p_i^{(n_k^k)} \to p_i$, as $k \to \infty$, for each i. $\square$

The second result is Fatou's lemma, whose proof we omit[7]:

[7] See Brémaud (1999).

Lemma 11. *If* $(y_i, i \in \mathbb{N})$, $(x_i^{(n)}, i \in \mathbb{N})$, $n = 1, 2, \ldots$ *are sequences of nonnegative numbers then*

$$\sum_{i=1}^{\infty} y_i \liminf_{n\to\infty} x_i^{(n)} \le \liminf_{n\to\infty} \sum_{i=1}^{\infty} y_i x_i^{(n)}.$$

Proof of Theorem 19. Let j be a null recurrent state. Suppose that for some i, the sequence $p_{ij}^{(n)}$ does not converge to 0. Hence we can find a subsequence n_k such that

$$\lim_{k\to\infty} p_{ij}^{(n_k)} = \alpha_j > 0.$$

Consider the probability vector

$$\big(p_{ix}^{(n_k)},\ x \in S\big).$$

By Lemma 10, we can pick subsequence (n_k') of (n_k) such that

$$\lim_{k\to\infty} p_{ix}^{(n_k')} = \alpha_x, \quad \text{for all } x \in S.$$

Since $a_j > 0$, we have

$$0 < \sum_{x\in S} \alpha_x \le \liminf_{k\to\infty} \sum_{x\in S} p_{ix}^{(n_k')} = 1.$$

The inequality follows from Lemma 11, and the last equality is obvious since we have a probability vector. But $\mathsf{P}^{n+1} = \mathsf{P}^n\mathsf{P}$, i.e.

$$p_{ix}^{(n_k'+1)} = \sum_{y\in S} p_{iy}^{(n_k')} p_{yx}.$$

Therefore, by Lemma 11,

$$\lim_{k\to\infty} p_{ix}^{(n_k'+1)} \ge \sum_{y\in S} \lim_{k\to\infty} p_{iy}^{(n_k')} p_{yx}.$$

Hence

$$\alpha_x \ge \sum_{y\in S} \alpha_y p_{yx}, \quad x \in S.$$

We wish to show that these inequalities are, actually, equalities. Suppose that one of them is not, i.e. it is a strict inequality:

$$\alpha_{x_0} > \sum_{y\in S} \alpha_y p_{yx_0}, \text{ for some } x_0 \in S.$$

This would imply that

$$\sum_{x\in S} \alpha_x > \sum_{x\in S}\sum_{y\in S} \alpha_y p_{yx} = \sum_{y\in S} \alpha_y \sum_{x\in S} p_{yx} = \sum_{y\in S} \alpha_y,$$

and this is a contradiction (the number $\sum_{x\in S} \alpha_x$ cannot be strictly larger than itself). Thus,

$$\alpha_x = \sum_{y\in S} \alpha_y p_{yx}, \quad x \in S.$$

Since $\sum_{x\in S} \alpha_x \le 1$, we have that

$$\pi(x) := \frac{\alpha_x}{\sum_{y\in S} \alpha_y}$$

satisfies the balance equations and has $\sum_{x\in S} \pi(x) = 1$, i.e. π is a stationary distribution. Now $\pi(j) > 0$ because $\alpha_j > 0$, by assumption. By Corollary 13, state j is positive recurrent: Contradiction. □

19.2 The general case

It remains to see what the limit of $p_{ij}^{(n)}$ is, as $n \to \infty$, when j is a positive recurrent state but i and j do not necessarily belong to the same communicating class. The result when the period j is larger than 1 is a bit awkward to formulate, so we will only look that the aperiodic case.

Theorem 20. *Let j be a positive recurrent state with period 1. Let C be the communicating class of j and let π^C be the stationary distribution associated with C. Let i be an arbitrary state. Let*

$$H_C := \inf\{n \geq 0 : \ X_n \in C\}.$$

Let

$$\varphi_C(i) := P_i(H_C < \infty).$$

Then

$$\lim_{n\to\infty} p_{ij}^{(n)} = \varphi_C(i)\pi^C(j).$$

Proof. If i also belongs to C then $\varphi^C(i) = 1$ and so the result follows from Theorem 17. In general, we have

$$\begin{aligned} p_{ij}^{(n)} = P_i(X_n = j) &= P_i(X_n = j, H_C \leq n) \\ &= \sum_{m=1}^{n} P_i(H_C = m) P_i(X_n = j \mid H_C = m). \end{aligned}$$

Let μ be the distribution of X_{H_C} given that $X_0 = i$ and that $\{H_C < \infty\}$. From the Strong Markov Property, we have $P_i(X_n = j \mid H_C = m) = P_\mu(X_{n-m} = j)$. But Theorem 17 tells us that the limit exists regardless of the initial distribution, that is, $P_\mu(X_{n-m} = j) \to \pi^C(j)$, as $n \to \infty$. Hence

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi^C(j) \sum_{m=1}^{\infty} P_i(H_C = m) = \pi^C(j) P_i(H_C < \infty),$$

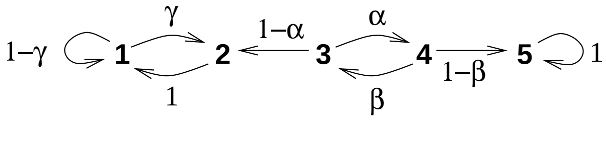
which is what we need. □

Remark: We can write the result of Theorem 20 as

$$\lim_{n\to\infty} p_{ij}^{(n)} = \frac{f_{ij}}{E_j T_j},$$

where $T_j = \inf\{n \geq 1 : \ X_n = j\}$, and $f_{ij} = P_i(T_j < \infty)$, as in §10.2. Indeed, $\pi^C(j) = 1/E_j T_j$, and $f_{ij} = \varphi_C(j)$. In this form, the result can be obtained by the so-called Abel's Lemma of Analysis, starting with the first hitting time decomposition relation of Lemma 7. This is a more old-fashioned way of getting the same result.

Example: Consider the following chain ($0 < \alpha, \beta, \gamma < 1$):



The communicating classes are $I = \{3, 4\}$ (inessential states), $C_1 = \{1, 2\}$ (closed class), $C_2 = \{5\}$ (absorbing state). The stationary distributions associated with the two closed classes are:

$$\pi^{C_1}(1) = \frac{1}{1+\gamma}, \quad \pi^{C_1}(2) = \frac{\gamma}{1+\gamma}, \qquad \pi^{C_2}(5) = 1.$$

Let $\varphi(i)$ be the probability that class C_1 will be reached starting from state i. We have

$$\varphi(1) = \varphi(2) = 1, \quad , \varphi(5) = 0,$$

while, from first-step analysis,

$$\begin{aligned}\varphi(3) &= (1-\alpha)\varphi(2) + \alpha\varphi(4) \\ \varphi(4) &= (1-\beta)\varphi(5) + \beta\varphi(3),\end{aligned}$$

whence

$$\varphi(3) = \frac{1-\alpha}{1-\alpha\beta}, \quad \varphi(4) = \frac{\beta(1-\alpha)}{1-\alpha\beta}.$$

Both C_1, C_2 are aperiodic classes. Hence, as $n \to \infty$,

$$p_{31}^{(n)} \to \varphi(3)\pi^{C_1}(1), \quad p_{32}^{(n)} \to \varphi(3)\pi^{C_1}(2), \quad p_{33}^{(n)} \to 0, \quad p_{34}^{(n)} \to 0, \quad p_{35}^{(n)} \to (1-\varphi(3))\pi^{C_2}(5).$$

Similarly,

$$p_{41}^{(n)} \to \varphi(4)\pi^{C_1}(1), \quad p_{42}^{(n)} \to \varphi(4)\pi^{C_1}(2), \quad p_{43}^{(n)} \to 0, \quad p_{44}^{(n)} \to 0, \quad p_{45}^{(n)} \to (1-\varphi(4))\pi^{C_2}(5).$$

20 Ergodic theorems

An ergodic theorem is a result about time-averages of random processes. The subject of Ergodic Theory has its origins in Physics. The ergodic hypothesis in physics says that the proportion of time that a particle spends in a particular region of its "phase space" is proportional to the volume of this region. The phase (or state) space of a particle is not just a physical space but it includes, for example, information about its velocity. Pioneers in the are were, among others, the 19-th century French mathematicians Henri Poincaré and Joseph Liouville.

Ergodic Theory is nowadays an area that combines many fields of Mathematics and Science, ranging from Mathematical Physics and Dynamical Systems to Probability and Statistics.

We will avoid to give a precise definition of the term 'ergodic' but only mention what constitutes a result in the area. For example, Theorem 14 is a type of ergodic theorem. And so is the Law of Large Numbers itself (Theorem 13).

Recall that, in Theorem 14, we assumed the existence of a positive recurrent state. We now generalise this.

Theorem 21. *Suppose that a Markov chain possesses a recurrent state $a \in S$ and starts from an initial state X_0 such that $P(T_a < \infty) = 1$. Consider the quantity $\nu^{[a]}(x)$ defined as the average number of visits to state x between two successive visits to state a–see 16. Let f be a reward function such that*

$$\sum_{x \in S} |f(x)| \nu^{[a]}(x) < \infty. \tag{20}$$

Then

$$P\left(\lim_{t\to\infty}\frac{\sum_{n=0}^{t} f(X_n)}{\sum_{n=0}^{t}\mathbf{1}(X_n=a)}=\widehat{f}\right)=1 \, , \tag{21}$$

where

$$\widehat{f}:=\sum_{x\in S} f(x)\nu^{[a]}(x).$$

Proof. The reader is asked to revisit the proof of Theorem 14. Note that the denominator $\sum_{n=0}^{t}\mathbf{1}(X_n=a)$ equals the number N_t of visits to state a up to time t. As in Theorem 14, we break the total reward as follows:

$$\sum_{n=0}^{t} f(X_n)=\sum_{r=1}^{N_t-1} G_r+G_t^{\text{first}}+G_t^{\text{last}},$$

where $G_r=\sum_{T_a^{(r)}\le t<T_a^{(r+1)}} f(X_t)$, while G_t^{first}, G_t^{last} are the total rewards over the first and last incomplete cycles, respectively. As in the proof of Theorem 14, we have

$$\lim_{t\to\infty}\frac{1}{t}G_t^{\text{last}}=\lim_{t\to\infty}\frac{1}{t}G_t^{\text{first}}=0,$$

with probability one. The Law of Large Numbers tells us that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{r=0}^{n-1} G_r=EG_1,$$

with probability 1 as long as $E|G_1|<\infty$. But this is precisely the condition (20). Replacing n by the subsequence N_t, we obtain the result. We only need to check that $EG_1=\widehat{f}$, and this is left as an exercise. □

Remark 1: The difference between Theorem 14 and 21 is that, in the former, we assumed positive recurrence of the ground state a. If so, then, as we saw in the proof of Theorem 14, we have $N_t/t\to 1/E_aT_a$. So if we divide by t the numerator and denominator of the fraction in (21), we have that the denominator equals N_t/t and converges to $1/E_aT_a$, so the numerator converges to $\widehat{f}/E_aT_a$, which is the quantity denoted by $\overline{f}$ in Theorem 14.

Remark 2: If a chain is irreducible and positive recurrent then the Ergodic Theorem Theorem 14 holds, and this justifies the terminology of §18.5.

21 Finite chains

We take a brief look at some things that are specific to chains with finitely many states. Then at least one of the states must be visited infinitely often. So there are always recurrent states. From proposition 2 we know that the balance equations

$$\nu=\nu\mathsf{P}$$

have at least one solution. Now, since the number of states is finite, we have

$$C:=\sum_{i\in S}\nu(i)<\infty.$$

Hence

$$\pi(i) := \frac{1}{C}\nu(i), \quad i \in S,$$

satisfies the balance equations and the normalisation condition $\sum_{i\in S}\pi(i) = 1$. Therefore π is a stationary distribution. At least some state i must have $\pi(i) > 0$. By Corollary 13, this state must be positive recurrent.

Hence, a finite Markov chain always has positive recurrent states. If, in addition, the chain is irreducible, then all states are positive recurrent, because positive recurrence is a class property (Theorem 15). In addition, by Theorem 16, the stationary distribution is unique.

We summarise the above discussion in:

Theorem 22. *Consider an irreducible Markov chain with finitely many states. Then the chain is positive recurrent with a unique stationary distribution π. If, in addition, the chain is aperiodic, we have*

$$\mathsf{P}^n \to \Pi, \qquad \textit{as } n \to \infty,$$

where P is the transition matrix and Π is a matrix all the rows of which are equal to π.

22 Time reversibility

Consider a Markov chain $(X_n, n \geq 0)$. We say that it is TIME-REVERSIBLE, or, simply, reversible if it has the same distribution when the arrow of time is reversed. More precisely,

$$(X_0, X_1, \ldots, X_n) \text{ has the same distribution as } (X_n, X_{n-1}, \ldots, X_0), \text{ for all } n.$$

Another way to put this is by saying that observing any part of the process will not reveal any information about whether the process runs forwards or backwards in time. It is, in a sense, like playing a film backwards, without being able to tell that it is, actually, running backwards. For example, if we film a standing man who raises his arms up and down successively, and run the film backwards, it will look the same. But if we film a man who walks and run the film backwards, then we instantly know that the film is played in reverse.

Lemma 12. *A reversible Markov chain is stationary.*

Proof. Reversibility means that $(X_0, X_1, \ldots, X_n)$ has the same distribution as $(X_n, X_{n-1}, \ldots, X_0)$ for all n. In particular, the first components of the two vectors must have the same distribution; that is, X_n has the same distribution as X_0 for all n. Because the process is Markov, this means that it is stationary. □

Theorem 23. *Consider an irreducible Markov chain with transition probabilities p_{ij} and stationary distribution π. Then the chain is reversible if and only if the* DETAILED BALANCE EQUATIONS *hold:*

$$\pi(i)p_{ij} = \pi(j)p_{ji}, \quad i, j \in S.$$

Proof. Suppose first that the chain is reversible. Taking $n = 1$ in the definition of reversibility, we see that (X_0, X_1) has the same distribution as (X_1, X_0), i.e.

$$P(X_0 = i, X_1 = j) = P(X_1 = i, X_0 = j),$$

for all i and j. But

$$P(X_0 = i, X_1 = j) = \pi(i)p_{ij}, \quad P(X_0 = j, X_1 = i) = \pi(j)p_{ji}.$$

So the detailed balance equations (DBE) hold.
Conversely, suppose the DBE hold. The DBE imply immediately that (X_0, X_1) has the same distribution as (X_1, X_0). Now pick a state k and write, using the DBE,

$$\pi(i)p_{ij}p_{jk} = [\pi(j)p_{ji}]p_{jk} = [\pi(j)p_{jk}]p_{ji} = \pi(k)p_{kj}p_{ji},$$

and this means that

$$P(X_0 = i, X_1 = j, X_2 = k) = P(X_0 = k, X_1 = j, X_2 = i),$$

for all i, j, k, and hence (X_0, X_1, X_2) has the same distribution as (X_2, X_1, X_0). By induction, it is easy to show that, for all n, $(X_0, X_1, \dots, X_n)$ has the same distribution as $(X_n, X_{n-1}, \dots, X_0)$, and the chain is reversible. □

Theorem 24. *Consider an aperiodic irreducible Markov chain with transition probabilities p_{ij} and stationary distribution π. Then the chain is reversible if and only if the* Kolmogorov's loop criterion *holds: for all $i_1, i_2, \dots, i_m \in S$ and all $m \geq 3$,*

$$p_{i_1 i_2} p_{i_2 i_3} \cdots p_{i_{m-1} i_m} = p_{i_1 i_m} p_{i_m i_{m-1}} \cdots p_{i_2 i_1}. \tag{22}$$

Proof. Suppose first that the chain is reversible. Hence the DBE hold. We will show the Kolmogorov's loop criterion (KLC). Pick 3 states i, j, k, and write, using DBE,

$$\begin{aligned}\pi(i)p_{ij}p_{jk}p_{ki} &= [\pi(j)p_{ji}]p_{jk}p_{ki} = [\pi(j)p_{jk}]p_{ji}p_{ki} = [\pi(k)p_{kj}]p_{ji}p_{ki} \\ &= [\pi(k)p_{ki}]p_{ji}p_{kj} = [\pi(i)p_{ik}]p_{ji}p_{kj} = \pi(i)p_{ik}p_{kj}p_{ji}\end{aligned}$$

Since the chain is irreducible, we have $\pi(i) > 0$ for all i. Hence cancelling the $\pi(i)$ from the above display we obtain the KLC for $m = 3$. The same logic works for any m and can be worked out by induction.
Conversely, suppose that the KLC (22) holds. Summing up over all choices of states $i_3, i_4, \dots, i_{m-1}$ we obtain

$$p_{i_1 i_2} p^{(m-3)}_{i_2 i_1} = p^{(m-3)}_{i_1 i_2} p_{i_2 i_1}$$

If the chain is aperiodic, then, letting $m \to \infty$, we have $p^{(m-3)}_{i_2 i_1} \to \pi(i_1)$, and $p^{(m-3)}_{i_1 i_2} \to \pi(i_2)$, implying that

$$\pi(i_1) p_{i_1 i_2} = \pi(i_2) p_{i_2 i_1}.$$

These are the DBE. So the chain is reversible. □

Notes:
1) A necessary condition for reversibility is that $p_{ij} > 0$ if and only if $p_{ji} > 0$. In other words, if there is an arrow from i to j in the graph of the chain but no arrow from j to i, then the chain cannot be reversible.
2) The DBE are equivalent to saying that the ratio p_{ij}/p_{ji} can be written as a product of a function of i and a function of j; we have separation of variables:

$$\frac{p_{ij}}{p_{ji}} = f(i)g(j).$$

(Convention: we form this ratio only when $p_{ij} > 0$.) Necessarily, $f(i)g(i)$ must be 1 (why?), so we have

$$\frac{p_{ij}}{p_{ji}} = \frac{g(j)}{g(i)}.$$

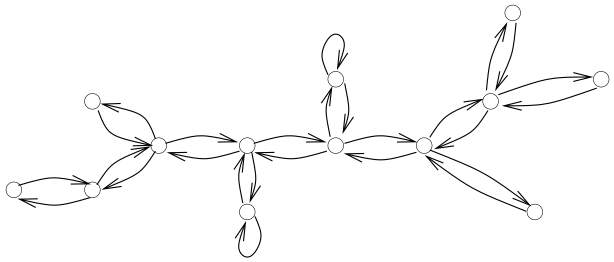
Multiplying by $g(i)$ and summing over j we find

$$g(i) = \sum_{j \in S} g(j) p_{ji}.$$

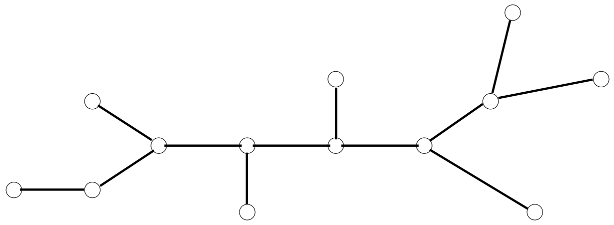
So g satisfies the balance equations. This gives the following method for showing that a finite irreducible chain is reversible:

- Form the ratio p_{ij}/p_{ji} and show that the variables separate.
- If the chain has infinitely many states, then we need, in addition, to guarantee, using another method, that the chain is positive recurrent.

Example 1: Chain with a tree-structure. Consider a positive recurrent irreducible Markov chain such that if $p_{ij} > 0$ then $p_{ji} > 0$. Form the graph by deleting all self-loops, and by replacing, for each pair of distinct states i, j for which $p_{ij} > 0$, the two oriented edges from i to j and from j to i by a single edge without arrow. If the graph thus obtained is a TREE, meaning that it has no loops, then the chain is reversible. For example, consider the chain:



Replace it by:



This is a tree. Hence the original chain is reversible. The reason is as follows. Pick two distinct states i, j for which $p_{ij} > 0$. If we remove the link between them, the graph becomes disconnected. (If it didn't, there would be a loop, and, by assumption, there isn't any.) Let A, A^c be the sets in the two connected components. There is obviously only one arrow from $AtoA^c$, namely the arrow from i to j; and the arrow from j to j is the only one from A^c to A. The balance equations are written in the form $F(A, A^c) = F(A^c, A)$ (see §4.1) which gives

$$\pi(i)p_{ij} = \pi(j)p_{ji},$$

i.e. the detailed balance equations. Hence the chain is reversible. And hence π can be found very easily.

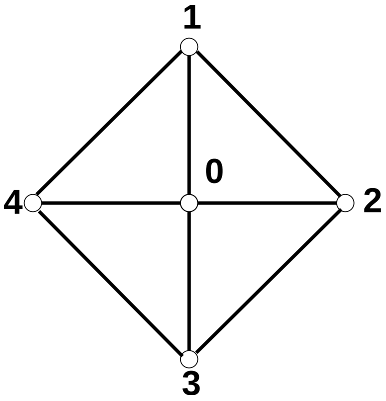
Example 2: Random walk on a graph. Consider an undirected graph G with vertices S, a finite set, and a set E of undirected edges. Each edge connects a pair of distinct vertices (states) without orientation. Suppose the graph is connected. States j are i are called NEIGHBOURS if they are connected by an edge. For each state i define its DEGREE

$$\delta(i) = \text{ number of neighbours of } i.$$

Now define the the following transition probabilities:

$$p_{ij} = \begin{cases} 1/\delta(i), & \text{if } j \text{ is a neighbour of } i \\ 0, & \text{otherwise.} \end{cases}$$

The Markov chain with these transition probabilities is called a RANDOM WALK ON A GRAPH (RWG). For example,



Here we have $p_{12} = p_{10} = p_{14} = 1/3$, $p_{02} = 1/4$, etc.

The stationary distribution of a RWG is very easy to find. We claim that

$$\pi(i) = C\delta(i),$$

where C is a constant. To see this let us verify that the detailed balance equations hold.

$$\pi(i)p_{ij} = \pi(j)p_{ji}.$$

There are two cases: If i, j are not neighbours then both sides are 0. If i, j are neighbours then $\pi(i)p_{ij} = C\delta(i)\frac{1}{\delta(i)} = C$ and $\pi(j)p_{ji} = C\delta(j)\frac{1}{\delta(j)} = C$; so both members are equal to C. In all cases, the DBE hold. Hence we conclude that:

1. A RWG is a reversible chain.

2. The stationary distribution assigns to a state probability which is proportional to its degree.

The constant C is found by normalisation, i.e.

$$\sum_{i \in S} \pi(i) = 1.$$

Since

$$\sum_{i \in S} \delta(i) = 2|E|,$$

where $|E|$ is the total number of edges, we have that $C = 1/2|E|$.

23 New Markov chains from old ones*

We now deal with the question: if we have a Markov chain (X_n), how can we create new ones?

23.1 Subsequences

The most obvious way to create a new Markov chain is by picking a subsequence, i.e. a sequence of integers $(n_k)_{k\geq 0}$ such that

$$0 \leq n_0 < n_1 < n_2 < \cdots$$

and by letting

$$Y_k := X_{n_k}, \quad k = 0, 1, 2, \ldots$$

The sequence $(Y_k, k = 0, 1, 2, \ldots)$ has the Markov property but it is not time-homogeneous, in general. If the subsequence forms an arithmetic progression, i.e. if

$$n_k = n_0 + km, \quad k = 0, 1, 2, \ldots,$$

then $(Y_k, k = 0, 1, 2, \ldots)$ has time-homogeneous transition probabilities:

$$P(Y_{k+1} = j | Y_k = i) = p_{ij}^{(m)},$$

where $p_{ij}^{(m)}$ are the m-step transition probabilities of the original chain. In this case, if, in addition, π is a stationary distribution for the original chain, then it is a stationary distribution for the new chain.

23.2 Watching a Markov chain when it visits a set

Suppose that the original chain is ergodic (i.e. irreducible and positive recurrent). Let A be a set of states and let $T_A^{(r)}$ be the r-th visit to A. Define

$$Y_r := X_{T_A^{(r)}}, \quad r = 1, 2, \ldots$$

Due to the Strong Markov Property, this process is also a Markov chain and it does have time-homogeneous transition probabilities. The state space of (Y_r) is A. If π is the stationary distribution of the original chain then $\pi/\pi(A)$ is the stationary distribution of (Y_r).

23.3 Subordination

Let $(X_n, n \geq 0)$ be Markov with transition probabilities p_{ij}. Let $(S_t, t \geq 0)$ be independent from (X_n) with

$$S_0 = 0, \quad S_{t+1} = S_t + \xi_t, \quad t \geq 0,$$

where ξ_t are i.i.d. random variables with values in $\mathbb{N}$ and distribution

$$\lambda(n) = P(\xi_0 = n), \quad n \in \mathbb{N}.$$

Then

$$Y_t := X_{S_t}, \quad t \geq 0,$$

is a Markov chain with

$$P(Y_{t+1} = j | Y_t = i) = \sum_{n=1}^{\infty} \lambda(n) p_{ij}^{(n)}.$$

23.4 Deterministic function of a Markov chain

If (X_n) is a Markov chain with state space S and if $f : S \to S'$ is some function from S to some countable set S', then the process

$$Y_n = f(X_n), \quad n \geq 0$$

is NOT, in general, a Markov chain.

If, however, f is a one-to-one function then (Y_n) is a Markov chain. Indeed, knowledge of $f(X_n)$ implies knowledge of X_n and so, conditional on Y_n the past is independent of the future.

(Notation: We will denote the elements of S by $i, j, \ldots$, and the elements of S' by $\alpha, \beta, \ldots$.)

More generally, we have:

Lemma 13. *Suppose that $P(Y_{n+1} = \beta \mid X_n = i)$ depends on i only through $f(i)$, i.e. that there is a function $Q : S' \times S' \to \mathbb{R}$ such that*

$$P(Y_{n+1} = \beta \mid X_n = i) = Q(f(i), \beta).$$

Then (Y_n) has the Markov property.

Proof. Fix $\alpha_0, \alpha_1, \ldots, \alpha_{n-1} \in S'$ and let $\mathscr{Y}_{n-1}$ be the event

$$\mathscr{Y}_{n-1} := \{Y_0 = \alpha_0, \ldots, Y_{n-1} = \alpha_{n-1}\}.$$

We need to show that

$$P(Y_{n+1} = \beta \mid Y_n = \alpha, \mathscr{Y}_{n-1}) = P(Y_{n+1} = \beta \mid Y_n = \alpha),$$

for all choices of the variables involved. But

$$\begin{aligned}
&P(Y_{n+1} = \beta \mid Y_n = \alpha, \mathscr{Y}_{n-1}) \\
&\qquad = \sum_{i \in S} P(Y_{n+1} = \beta \mid X_n = i, Y_n = \alpha, \mathscr{Y}_{n-1}) P(X_n = i \mid Y_n = \alpha, \mathscr{Y}_{n-1}) \\
&\qquad = \sum_{i \in S} Q(f(i), \beta) P(X_n = i \mid Y_n = \alpha, \mathscr{Y}_{n-1}) \\
&\qquad = \sum_{i \in S} Q(f(i), \beta) \frac{P(Y_n = \alpha | X_n = i, \mathscr{Y}_{n-1}) P(X_n = i | \mathscr{Y}_{n-1})}{P(Y_n = \alpha | \mathscr{Y}_{n-1})} \\
&\qquad = \sum_{i \in S} Q(f(i), \beta) \frac{\mathbf{1}(\alpha = f(i)) P(X_n = i | \mathscr{Y}_{n-1})}{P(Y_n = \alpha | \mathscr{Y}_{n-1})} \\
&\qquad = Q(\alpha, \beta) \sum_{i \in S} \frac{P(Y_n = \alpha | X_n = i, \mathscr{Y}_{n-1}) P(X_n = i | \mathscr{Y}_{n-1})}{P(Y_n = \alpha | \mathscr{Y}_{n-1})} \\
&\qquad = Q(\alpha, \beta) \frac{\sum_{i \in S} P(Y_n = \alpha, X_n = i | \mathscr{Y}_{n-1}) P(X_n = i | \mathscr{Y}_{n-1})}{P(Y_n = \alpha | \mathscr{Y}_{n-1})} \\
&\qquad = Q(\alpha, \beta) \frac{P(Y_n = \alpha | \mathscr{Y}_{n-1})}{P(Y_n = \alpha | \mathscr{Y}_{n-1})} \\
&\qquad = Q(\alpha, \beta).
\end{aligned}$$

Thus (Y_n) is Markov with transition probabilities $Q(\alpha, \beta)$. □

Example: Consider the symmetric drunkard's random walk, i.e.

$$P(X_{n+1} = i \pm 1 | X_n = i) = 1/2, \quad i \in \mathbb{Z}.$$

Define

$$Y_n := |X_n|.$$

Then (Y_n) is a Markov chain. To see this, observe that the function $|\cdot|$ is not one-to-one. But we will show that

$$P(Y_{n+1} = \beta | X_n = i), \quad i \in \mathbb{Z}, \beta \in \mathbb{Z}_+,$$

depends on i only through $|i|$. Indeed, conditional on $\{X_n = i\}$, and if $i \neq 0$, the absolute value of next state will either increase by 1 or decrease by 1 with probability 1/2, regardless of whether $i > 0$ or $i < 0$, because the probability of going up is the same as the probability of going down. On the other hand, if $i = 0$, then $|X_{n+1}| = 1$ for sure. In other words, if we let

$$Q(\alpha, \beta) := \begin{cases} 1/2, & \text{if } \beta = \alpha \pm 1, \ \alpha > 0 \\ 1, & \text{if } \beta = 1, \ \alpha = 0, \end{cases}$$

We have that

$$P(Y_{n+1} = \beta | X_n = i) = Q(|i|, \beta),$$

for all $i \in \mathbb{Z}$ and all $\beta \in \mathbb{Z}_+$, and so (Y_n) is, indeed, a Markov chain with transition probabilities $Q(\alpha, \beta)$.

24 Applications

24.1 Branching processes: a population growth application

We describe a basic model for population growth known as the Galton-Watson process. The idea is this: At each point of time n we have a number X_n of individuals each of which gives birth to a random number of offspring, independently from one another, and following the same distribution. The parent dies immediately upon giving birth. We let p_k be the probability that an individual will bear k offspring, $k = 0, 1, 2, \ldots$. We find the population at time $n+1$ by adding the number of offspring of each individual. Let $\xi_k^{(n)}$ be the number of offspring of the k-th individual in the n-th generation. Then the size of the $(n+1)$-st generation is:

$$X_{n+1} = \xi_1^{(n)} + \xi_2^{(n)} + \cdots + \xi_{X_n}^{(n)}.$$

If we let

$$\xi^{(n)} = \left(\xi_1^{(n)}, \xi_2^{(n)}, \ldots\right)$$

then we see that the above equation is of the form

$$X_{n+1} = f(X_n, \xi^{(n)}),$$

and, since $\xi^{(0)}$, $\xi^{(1)}, \ldots$ are i.i.d. (and independent of the initial population size X_0–a further assumption), we have that $(X_n, n \geq 0)$ has the Markov property. It is a Markov chain with time-homogeneous transitions. Computing the transition probabilities $p_{ij} = P(X_{n+1} = j | X_n = i)$ is, in general, a hard problem, so we will find some different method to proceed. But here is an interesting special case.

Example: Suppose that $p_1 = p$, $p_0 = 1 - p$. Thus, each individual gives birth to at one child with probability p or has children with probability $1-p$. If we know that $X_n = i$ then X_{n+1} is the number of successful births amongst i individuals, which has a binomial distribution:

$$p_{ij} = \binom{i}{j} p^j (1-p)^{i-j}, \quad 0 \leq j \leq i.$$

In this case, the population cannot grow: it will eventually reach 0, and 0 is always an absorbing state.

Back to the general case, one question of interest is whether the process will become extinct or not. We will not attempt to compute the p_{ij} neither draw the transition diagram because they are both futile.

If $p_1 = P(\xi = 1) = 1$ then $X_n = X_0 + n$ and this is a trivial case. So assume that $p_1 \neq 1$. Then

$$P(\xi = 0) + P(\xi \geq 2) > 0.$$

We distinguish two cases:
Case I: If $p_0 = P(\xi = 0) = 0$ then the population can only grow. Hence all states $i \geq 1$ are transient. The state 0 is irrelevant because it cannot be reached from anywhere.
Case II: If $p_0 = P(\xi = 0) > 0$ then state 0 can be reached from any other state: indeed, if the current population is i there is probability p_0^i that everybody gives birth to zero offspring. But 0 is an absorbing state. Hence all states $i \geq 1$ are inessential; therefore transient.
Therefore, with probability 1, if 0 is never reached then $X_n \to \infty$. In other words, if the

population does not become extinct then it must grow to infinity. Let D be the event of 'death' or extinction:

$$D := \{X_n = 0 \text{ for some } n \geq 0\}.$$

We wish to compute the extinction probability

$$\varepsilon(i) := P_i(D),$$

when we start with $X_0 = i$. Obviously, $\varepsilon(0) = 1$. For $i \geq 1$, we can argue that

$$\varepsilon(i) = \varepsilon^i, \quad \text{where } \varepsilon = P_1(D).$$

Indeed, if we start with $X_0 = i$ in ital ancestors, each one behaves independently of one another. For each $k = 1, \ldots, i$, let D_k be the event that the branching process corresponding to the k-th initial ancestor dies. We then have that

$$D = D_1 \cap \cdots \cap D_i,$$

and $P(D_k) = \varepsilon$, so $P(D|X_0 = i) = \varepsilon^i$. Therefore the problem reduces to the study of the branching process starting with $X_0 = 1$. The result is this:

Proposition 5. *Consider a Galton-Watson process starting with $X_0 = 1$ initial ancestor. Let*

$$\mu = E\xi = \sum_{k=1}^{\infty} k p_k$$

be the mean number of offspring of an individual. Let

$$\varphi(z) := Ez^{\xi} = \sum_{k=0}^{\infty} z^k p_k$$

be the probability generating function of ξ. Let ε be the extinction probability of the process. If $\mu \leq 1$ then the $\varepsilon = 1$. If $\mu > 1$ then $\varepsilon < 1$ is the unique solution of

$$\varphi(z) = z.$$

Proof. We let

$$\varphi_n(z) := Ez^{X_n} = \sum_{k=0}^{\infty} z^k P(X_n = k)$$

be the probability generating function of X_n. This function is of interest because it gives information about the extinction probability ε. Indeed,

$$\varphi_n(0) = P(X_n = 0),$$

and, since

$$X_n = 0 \text{ implies } X_m = 0 \text{ for all } m \geq n,$$

we have

$$\begin{aligned}
\lim_{n\to\infty} \varphi_n(0) &= \lim_{n\to\infty} P(X_m = 0 \text{ for all } m \geq n) \\
&= P(\text{there exists } n \text{ such that } X_m = 0 \text{ for all } m \geq n) \\
&= P(D) = \varepsilon.
\end{aligned}$$

Note also that

$$\varphi_0(z) = 1, \quad \varphi_1(z) = \varphi(z).$$

We then have

$$\begin{aligned}
\varphi_{n+1}(z) = Ez^{X_{n+1}} &= E\left[z^{\xi_1^{(n)}} z^{\xi_2^{(n)}} \cdots z^{\xi_{X_n}^{(n)}}\right] \\
&= \sum_{i=0}^{\infty} E\left[z^{\xi_1^{(n)}} z^{\xi_2^{(n)}} \cdots z^{\xi_{X_n}^{(n)}} \;\; \mathbf{1}(X_n = i)\right] \\
&= \sum_{i=0}^{\infty} E\left[z^{\xi_1^{(n)}} z^{\xi_2^{(n)}} \cdots z^{\xi_i^{(n)}} \;\; \mathbf{1}(X_n = i)\right] \\
&= \sum_{i=0}^{\infty} E[z^{\xi_1^{(n)}}]\; E[z^{\xi_2^{(n)}}] \; \cdots E[z^{\xi_i^{(n)}}]\; E[\mathbf{1}(X_n = i)] \\
&= \sum_{i=0}^{\infty} \varphi(z)^i P(X_n = i) \\
&= E\left[\varphi(z)^{X_n}\right] \\
&= \varphi_n(\varphi(z)).
\end{aligned}$$

Hence $\varphi_2(z) = \varphi_1(\varphi(z)) = \varphi(\varphi(z))$, $\varphi_3(z) = \varphi_2(\varphi(z)) = \varphi(\varphi(\varphi(z))) = \varphi(\varphi_2(z))$, and so on,

$$\varphi_{n+1}(z) = \varphi(\varphi_n(z)).$$

In particular,

$$\varphi_{n+1}(0) = \varphi(\varphi_n(0)).$$

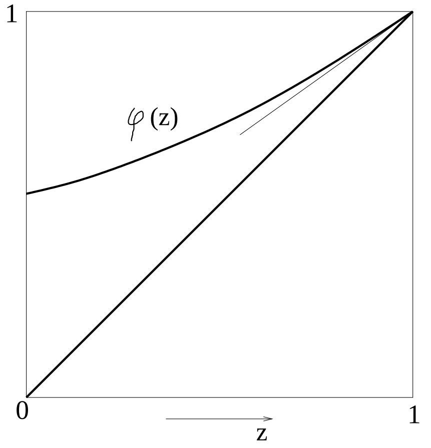
But $\varphi_{n+1}(0) \to \varepsilon$ as $n \to \infty$. Also $\varphi_n(0) \to \varepsilon$, and, since φ is a continuous function, $\varphi(\varphi_n(0)) \to \varphi(\varepsilon)$. Therefore

$$\varepsilon = \varphi(\varepsilon).$$

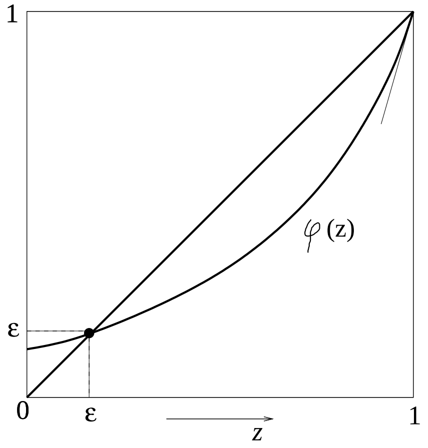
Notice now that

$$\mu = \frac{d}{dz}\varphi(z)\bigg|_{z=1}.$$

Also, recall that φ is a convex function with $\varphi(1) = 1$. Therefore, if the slope μ at $z = 1$ is ≤ 1, the graph of $\varphi(z)$ lies above the diagonal (the function z) and so the equation $z = \varphi(z)$ has no solutions other than $z = 1$. This means that $\varepsilon = 1$ (see left figure below). On the other hand, if $\mu > 1$, then the graph of $\varphi(z)$ does intersect the diagonal at some point $\varepsilon^* < 1$. Let us show that, in fact, $\varepsilon = \varepsilon^*$. Since both ε and ε^* satisfy $z = \varphi(z)$, and since the latter has only two solutions ($z = 0$ and $z = \varepsilon^*$), all we need to show is that $\varepsilon < 1$. But $0 \leq \varepsilon^*$. Hence $\varphi(0) \leq \varphi(\varepsilon^*) = \varepsilon^*$. By induction, $\varphi_n(0) \leq \varepsilon^*$. But $\varphi_n(0) \to \varepsilon$. So $\varepsilon \leq \varepsilon^* < 1$. Therefore $\varepsilon = \varepsilon^*$. (See right figure below.)



Case: $\mu \le 1$. *Then* $\varepsilon = 1$.



Case: $\mu > 1$. *Then* $\varepsilon < 1$.

24.2 The ALOHA protocol: a computer communications application

This is a protocol used in the early days of computer networking for having computer hosts communicate over a shared medium (radio channel). The problem of communication over a common channelis that if more than one hosts attempt to transmit a packet, there is collision and the transmission is unsuccessful. Thus, for a packet to go through, only one of the hosts should be transmitting at a given time. One obvious solution is to allocate time slots to hosts in a specific order, periodically. So if there are, say, 3 hosts, then time slots $0, 1, 2, 3, 4, 5, 6, \ldots$ are allocated to hosts $1, 2, 3, 1, 2, 3, 1, \ldots$. But this is not a good method: it is a waste of time to allocate a slot to a host if that host has nothing to transmit. Since things happen fast in a computer network, there is no time to make a prior query whether a host has something to transmit or not. Abandoning this idea, we let hosts transmit whenever they like. If collision happens, then the transmission is not successful and the transmissions must be attempted anew.

In a simplified model, assume that on a time slot there are x packets amongst all hosts that are waiting for transmission (they are 'backlogged'). During this time slot assume there is a number a of new packets that need to be transmitted ($a = 0, 1, 2, \ldots$). Each of the backlogged or newly arrived packets is independently selected for transmission with (a small) probability p. Let s be the number of selected packets. If $s = 1$ there is a successful transmission and, on the next times slot, there are $x + a - 1$ packets. If $s \ge 2$, there is a collision and, on the next times slot, there are $x + a$ packets.

If we let X_n be the number of backlogged packets at the beginning of the n-th time slot, A_n the number of new packets on the same time slot, and S_n the number of selected packets, then

$$X_{n+1} = \begin{cases} \max(X_n + A_n - 1, 0), & \text{if } S_n = 1, \\ X_n + A_n, & \text{otherwise.} \end{cases}$$

We assume that the A_n are i.i.d. random variables with some common distribution:

$$P(A_n = k) = \lambda_k, \quad k \ge 0,$$

where $\lambda_k \ge 0$ and $\sum_{k\ge0} k\lambda_k = 1$. We also let

$$\mu := \sum_{k\ge1} k\lambda_k$$

be the expectation of A_n. The random variable S_n has, conditional on X_n and A_n (and all past states and arrivals) a distribution which is Binomial with parameters $X_n + A_n$, and p. Indeed, selections are made amongst the $X_n + A_n$ packets. So

$$P(S_n = k | X_n = x, A_n = a, X_{n-1}, A_{n-1}, \ldots) = \binom{x+a}{k} p^k q^{x+a-k}, \quad 0 \le k \le x+a,$$

where $q = 1 - p$. The reason we take the maximum with 0 in the above equation is because, if $X_n + A_n = 0$, we should not subtract 1. The process $(X_n, n \ge 0)$ is a Markov chain. Let us find its transition probabilities. To compute $p_{x,x-1}$ when $x > 0$, we think as follows: to have a reduction of x by 1, we must have no new packets arriving, and exactly one selection:

$$p_{x,x-1} = P(A_n = 0, S_n = 1 | X_n = x) = \lambda_0 x p q^{x-1}.$$

To compute $p_{x,x+k}$ for $k \ge 0$, we think like this: to have an increase by k we either need exactly k new packets and no successful transmissions (i.e. S_n should be 0 or ≥ 2) or we need $k+1$ new packets and one selection ($S_n = 1$). So:

$$p_{x,x+k} = \lambda_k (1 - (x+k) p q^{x+k-1}) + \lambda_{k+1} (x+k+1) p q^{x+k}, \quad k \ge 1.$$

(One can easily check that $\sum_{k=-1}^{\infty} p_{x,x+k} = 1$.) The main result is:

Proposition 6. *The Aloha protocol is unstable, i.e. the Markov chain (X_n) is transient.*

Idea of proof. Proving this is beyond the scope of the lectures, but we mention that it is based on the so-called drift criterion which says that if the expected change $X_{n+1} - X_n$ conditional on $X_n = x$ is bounded below from a positive constant for all large x, then the chain is transient. We here only restrict ourselves to the computation of this expected change, defined by‘

$$\Delta(x) := E[X_{n+1} - X_n | X_n = x].$$

For $x > 0$, we have

$$\begin{aligned}
\Delta(x) &= (-1) p_{x,x-1} + \sum_{k \ge 1} k p_{x,x+k} \\
&= -\lambda_0 x p q^{x-1} + \sum_{k \ge 1} k \lambda_k - \sum_{k \ge 1} k \lambda_k (x+k) p q^{x+k-1} + \sum_{k \ge 1} k \lambda_{k+1} (x+k+1) p q^{x+k}. \\
&= -\lambda_0 x p q^{x-1} + \mu - \lambda_1 (x+1) p q^x - \sum_{k \ge 2} \lambda_k (x+k) p q^{x+k-1} \\
&= \mu - q^x G(x),
\end{aligned}$$

where $G(x)$ is a function of the form $\alpha + \beta_x$. Since $p > 0$, we have $q < 1$, and so q^x tends to 0 much faster than $G(x)$ tends to ∞, so $q^x G(x)$ tends to 0 as $x \to \infty$. So we can make $q^x G(x)$ as small as we like by choosing x sufficiently large, for example we can make it $\le \mu/2$. Therefore $\Delta(x) \ge \mu/2$ for all large x. Unless $\mu = 0$ (which is a trivial case: no arrivals!), we have that the drift is positive for all large x and this can be used to prove transience. □

24.3 PageRank (trademark of Google): A World Wide Web application

PageRank is an algorithm used by the popular search engine Google in order to assign importance to a web page. The way it does that is by computing the stationary distribution of a Markov chain that we now describe.

The web graph is a a graph set of vertices V consisting of web pages and edges representing hyperlinks: thus, if page i has a link to page j then we put an arrow from i to j and consider it as a directed edge of the graph. For each page i let $L(i)$ be the set of pages it links to. If $L(i) = \varnothing$, then i is a 'dangling page'. Define transition probabilities as follows:

$$\widehat{p}_{ij} = \begin{cases} 1/|L(i)|, & \text{if } j \in L(i) \\ 1/|V|, & \text{if } L(i) = \varnothing \\ 0, & \text{otherwise.} \end{cases}$$

So if $X_n = i$ is the current location of the chain, the next location X_{n+1} is picked uniformly at random amongst all pages that i links to, unless i is a dangling page; in the latter case, the chain moves to a random page.

It is not clear that the chain is irreducible, neither that it is aperiodic.

So we make the following modification. We pick α between 0 and 1 (typically, $\alpha \approx 0.2$) and let

$$p_{ij} := (1-\alpha)\widehat{p}_{ij} + \frac{\alpha}{|V|}.$$

These are new transition probabilities. The interpretation is that of a 'bored surfer': a surfer moves according the the original Markov chain, unless he gets bored–and this happens with probability α–in which case he clicks on a random page.

So this is how Google assigns ranks to pages: It uses an algorithm, PageRank, to compute the stationary distribution π corresponding to the Markov chain with transition matrix $\mathsf{P} = [p_{ij}]$:

$$\pi = \pi \mathsf{P}.$$

Then it says:

$$\text{page } i \text{ has higher rank than } j \text{ if } \pi(i) > \pi(j).$$

The idea is simple. Its implementation though is one of the largest matrix computations in the world, because of the size of the problem. The algorithm used (PageRank) starts with an initial guess $\pi(0)$ for π and updates it by

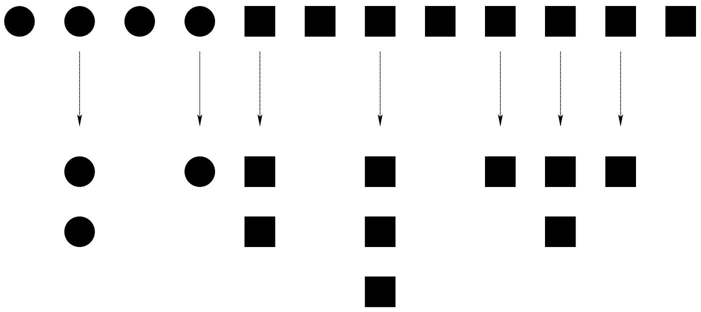
$$\pi^{(k)} = \pi^{(k-1)} \mathsf{P}.$$

It uses advanced techniques from linear algebra to speed up computations.

24.4 The Wright-Fisher model: an example from biology

In an organism, a gene controlling a specific characteristic (e.g. colour of eyes) appears in two forms, a dominant (say A) and a recessive (say a). The alleles in an individual come in pairs and express the individual's genotype for that gene. This means that the external appearance of a characteristic is a function of the pair of alleles. An individual can be of type AA (meaning that all its alleles are A) or aa or Aa. When two individuals mate, their offspring inherits an allele from each parent. Thus, two AA individuals will produce and AA child. But an AA mating with a Aa may produce a child of type AA or one of type Aa.

Imagine that we have a population of N individuals who mate at random producing a number of offspring. We will assume that a child chooses an allele from each parent at random. We will also assume that the parents are replaced by the children. The process repeats, generation after generation. We would like to study the genetic diversity, meaning how many characteristics we have. To simplify, we will assume that, in each generation, the genetic diversity depends only on the total number of alleles of each type. So let x be the number of alleles of type A (so and $2N - x$ is the number of those of type a). If so, then, at the next generation the number of alleles of type A can be found as follows: Pick an allele at random; this allele is of type A with probability $x/2N$; maintain this allele for the next generation. Do the same thing $2N$ times.



In this figure there are 4 alleles of type A (circles) and 6 of type a (squares). One of the square alleles is selected 4 times. One of the circle alleles is selected twice. Three of the square alleles are never selected. And so on. Alleles that are never picked up may belong to individuals who never mate or may belong to individuals who mate but their offspring did not inherit this allele. Since we don't care to keep track of the individuals' identities, all we need is the information outlined in this figure.

We are now ready to define the Markov chain X_n representing the number of alleles of type A in generation n. We have a transition from x to y if y alleles of type A were produced. Each allele of type A is selected with probability $x/2N$. Since selections are independent, we see that

$$p_{xy} = \binom{2N}{y} \left(\frac{x}{2N}\right)^y \left(1 - \frac{x}{2N}\right)^{2N-y}, \quad 0 \le y \le 2N.$$

Notice that there is a chance ($= (1 - \frac{x}{2N})^{2N}$) that no allele of type A is selected at all, and the population becomes one consisting of alleles of type a only. Similarly, there is a chance ($= (\frac{x}{2N})^{2N}$) that only alleles of type A are selected, and the population becomes one consisting of alleles of type A only. Clearly, $p_{00} = p_{2N,2N} = 1$, so both 0 and $2N$ are absorbing states. Since $p_{xy} > 0$ for all $1 \le x, y \le 2N - 1$, the states $\{1, \dots, 2N - 1\}$ form a communicating class of inessential states. Therefore

$$P(X_n = 0 \text{ or } X_n = 2N \text{ for all large } n) = 1,$$

i.e. the chain gets absorbed by either 0 or $2N$. Let $M = 2N$. The fact that M is even is irrelevant for the mathematical model. Another way to represent the chain is as follows:

$$X_{n+1} = \xi_1^n + \cdots + \xi_M^n,$$

where, conditionally on $(X_0, \dots, X_n)$, the random variables $\xi_1^n, \dots, \xi_M^n$ are i.i.d. with

$$P(\xi_k^n = 1 | X_n, \dots, X_0) = \frac{X_n}{M}, \quad P(\xi_k^n = 0 | X_n, \dots, X_0) = 1 - \frac{X_n}{M},$$

Therefore,

$$E(X_{n+1} | X_n, \dots, X_0) = E(X_{n+1} | X_n) = M\frac{X_n}{M} = X_n.$$

This implies that, for all $n \geq 0$,

$$EX_{n+1} = EX_n$$

and thus, on the average, the number of alleles of type A won't change. Clearly, we would like to know the probability

$$\varphi(x) = P_x(T_0 > T_M)P(T_0 > T_M | X_0 = x)$$

that the chain will eventually be absorbed by state M. (Here, as usual, $T_x := \inf\{n \geq 1 : \ X_n = x\}$.)

One can argue that, since, on the average, the expectation doesn't change and, since, at "the end of time" the chain (started from x) will be either M with probability $\varphi(x)$ or 0 with probability $1 - \varphi(x)$, we should have

$$M \times \varphi(x) + 0 \times (1 - \varphi(x)) = x,$$

i.e.

$$\varphi(x) = x/M.$$

The result is correct, but the argument needs further justification because "the end of time" is not a deterministic time. We show that our guess is correct by proceeding via the mundane way of verifying that it satisfies the first-step analysis equations. These are:

$$\varphi(0) = 0, \quad \varphi(M) = 1, \quad \varphi(x) = \sum_{y=0}^{M} p_{xy}\varphi(y).$$

The first two are obvious. The right-hand side of the last equals

$$\begin{aligned}
\sum_{y=0}^{M} \binom{M}{y} \left(\frac{x}{M}\right)^y \left(1 - \frac{x}{M}\right)^{M-y} \frac{y}{M} &= \sum_{y=1}^{M} \frac{M!}{y!(M-y)!} \left(\frac{x}{M}\right)^y \left(1 - \frac{x}{M}\right)^{M-y} \frac{y}{M} \\
&= \sum_{y=1}^{M} \frac{(M-1)!}{(y-1)!(M-y)!} \left(\frac{x}{M}\right)^{y-1+1} \left(1 - \frac{x}{M}\right)^{M-y-1+1} \\
&= \frac{x}{M} \sum_{y=1}^{M} \binom{M-1}{y-1} \left(\frac{x}{M}\right)^{y-1} \left(1 - \frac{x}{M}\right)^{(M-1)-(y-1)} \\
&= \frac{x}{M} \left(\frac{x}{M} + 1 - \frac{x}{M}\right)^{M-1} = \frac{x}{M},
\end{aligned}$$

as needed.

24.5 A storage or queueing application

A storage facility stores, say, X_n electronic components at time n. A number A_n of new components are ordered and added to the stock, while a number of them are sold to the market. If the demand is for D_n components, the demand is met instantaneously, provided that enough components are available, and so, at time $n+1$, there are $X_n + A_n - D_n$ components in the stock. Otherwise, the demand is only partially satisfied, and the stock reached level zero at time $n+1$. We can express these requirements by the recursion

$$X_{n+1} = (X_n + A_n - D_n) \vee 0$$

where $a \vee b := \max(a, b)$. Here are our assumptions: The pairs of random variables $((A_n, D_n), n \geq 0)$ are i.i.d. and

$$E(A_n - D_n) = -\mu < 0.$$

Thus, the demand is, on the average, larger than the supply per unit of time. We have that $(X_n, n \geq 0)$ is a Markov chain. We will show that this Markov chain is positive recurrent. We will apply the so-called Loynes' method.

Let

$$\xi_n := A_n - D_n,$$

so that $X_{n+1} = (X_n + \xi_n) \vee 0$, for all $n \geq 0$. Since the Markov chain is represented by this very simple recursion, we will try to solve the recursion. Working for $n = 1, 2, \ldots$ we find

$$\begin{aligned}
X_1 &= (X_0 + \xi_0) \vee 0 \\
X_2 &= (X_1 + \xi_1) \vee 0 = (X_0 + \xi_0 + \xi_1) \vee \xi_1 \vee 0 \\
X_3 &= (X_2 + \xi_2) \vee 0 = (X_0 + \xi_0 + \xi_1 + \xi_2) \vee (\xi_1 + \xi_2) \vee \xi_2 \vee 0 \\
&\cdots \\
X_n &= (X_0 + \xi_0 + \cdots + \xi_{n-1}) \vee (\xi_1 + \cdots + \xi_{n-1}) \vee \cdots \vee \xi_{n-1} \vee 0.
\end{aligned}$$

The correctness of the latter can formally be proved by showing that it does satisfy the recursion. Let us consider

$$g_{xy}^{(n)} := P_x(X_n > y) = P\big((x + \xi_0 + \cdots + \xi_{n-1}) \vee (\xi_1 + \cdots + \xi_{n-1}) \vee \cdots \vee \xi_{n-1} \vee 0 > y\big).$$

Thus, the event whose probability we want to compute is a function of $(\xi_0, \ldots, \xi_{n-1})$. Notice, however, that

$$(\xi_0, \ldots, \xi_{n-1}) \overset{\mathrm{d}}{=} (\xi_{n-1}, \ldots, \xi_0).$$

Indeed, the random variables (ξ_n) are i.i.d. so we can shuffle them in any way we like without altering their joint distribution. In particular, instead of considering them in their original order, we reverse it. Therefore,

$$g_{xy}^{(n)} = P\big((x + \xi_{n-1} + \cdots + \xi_0) \vee (\xi_{n-2} + \cdots + \xi_0) \vee \cdots \vee \xi_0 \vee 0 > y\big).$$

Let

$$\begin{aligned}
Z_n &:= \xi_0 + \cdots + \xi_{n-1} \\
M_n &:= Z_n \vee Z_{n-1} \vee \cdots \vee Z_1 \vee 0.
\end{aligned}$$

Then

$$g_{0y}^{(n)} = P(M_n > y).$$

Since

$$M_{n+1} \geq M_n$$

we have

$$g_{0y}^{(n)} \leq g_{0y}^{(n+1)},$$

for all n. The limit of a bounded and increasing sequence certainly exists:

$$g(y) := \lim_{n\to\infty} g_{0y}^{(n)} = \lim_{n\to\infty} P_0(X_n > y).$$

We claim that

$$\pi(y) := g(y) - g(y-1)$$

satisfies the balance equations. But

$$\begin{aligned} M_{n+1} &= 0 \vee \max_{1\leq j\leq n+1} Z_j \\ &= 0 \vee \big(\max_{1\leq j\leq n+1} (Z_j - Z_1) + \xi_0 \big) \\ &\overset{\mathrm{d}}{=} 0 \vee (M_n + \xi_0), \end{aligned}$$

where the latter equality follows from the fact that, in computing the distribution of M_n which depends on $\xi_0, \ldots, \xi_{n-1}$, we may replace the latter random variables by $\xi_1, \ldots, \xi_n$. Hence, for $y \geq 0$,

$$\begin{aligned} P(M_{n+1} = y) &= P(0 \vee (M_n + \xi_0) = y) \\ &= \sum_{x\geq 0} P(M_n = x) P(0 \vee (x + \xi_0) = y) \\ &= \sum_{x\geq 0} P(M_n = x)\ p_{xy}. \end{aligned}$$

Since the n-dependent terms are bounded, we may take the limit of both sides as $n \to \infty$, and, using $\pi(y) = \lim_{n\to\infty} P(M_n = y)$, we find

$$\pi(y) = \sum_{x\geq 0} \pi(x) p_{xy}.$$

So π satisfies the balance equations. We must make sure that it also satisfies $\sum_y \pi(y) = 1$. This will be the case if $g(y)$ is not identically equal to 1. Here, we will use the assumption that $E\xi_n = -\mu < 0$. From the Law of Large Numbers we have

$$P(\lim_{n\to\infty} Z_n/n = -\mu) = 1.$$

This implies that

$$P(\lim_{n\to\infty} Z_n = -\infty) = 1.$$

But then

$$P(\lim_{n\to\infty} M_n = 0 \vee \max\{Z_1, Z_2, \ldots\} < \infty) = 1.$$

Hence

$$g(y) = P(0 \vee \max\{Z_1, Z_2, \ldots\} > y)$$

is not identically equal to 1, as required.

It can also be shown (left as an exercise) that if $E\xi_n > 0$ then the Markov chain is transient.

The case $E\xi_n = 0$ is delicate. Depending on the actual distribution of ξ_n, the Markov chain may be transient or null recurrent.

PART II: RANDOM WALKS

25 Random walk

The purpose of this part of the notes is to take a closer look at a special kind of Markov chain known as random walk. Our Markov chains, so far, have been processes with time-homogeneous transition probabilities. A random walk possesses an additional property, namely, that of spatial homogeneity. This means that the transition probability p_{xy} should depend on x and y only through their relative positions in space. To be able to say this, we need to give more structure to the state space S, a structure that enables us to say that

$$p_{x,y} = p_{x+z,y+z}$$

for any translation z. For example, we can let $S = \mathbb{Z}^d$, the set of d-dimensional vectors with integer coordinates. More general spaces (e.g. groups) are allowed, but, here, we won't go beyond $\mathbb{Z}^d$.

So, given a function

$$p(x), \quad x \in \mathbb{Z}^d,$$

a random walk is a Markov chain with time- and space-homogeneous transition probabilities given by

$$p_{x,y} = p(y-x).$$

Example 1: Let

$$p(x) = C 2^{-|x_1|-\cdots-|x_d|},$$

where C is such that $\sum_x p(x) = 1$.

Example 2: Let where e_j be the vector that has 1 in the j-th position and 0 everywhere else. Define

$$p(x) = \begin{cases} p_j, & \text{if } x = e_j, \quad j = 1, \ldots, d \\ q_j, & \text{if } x = -e_j, \quad j = 1, \ldots, d \\ 0, & \text{otherwise} \end{cases}$$

where $p_1 + \cdots + p_d + q_1 + \cdots + q_d = 1$. A random walk of this type is called SIMPLE RANDOM WALK or NEAREST NEIGHBOUR RANDOM WALK. If $d = 1$, then

$$p(1) = p, \quad p(-1) = q, \quad p(x) = 0, \text{ otherwise},$$

where $p + q = 1$. This is the drunkard's walk we've seen before. In dimension $d = 1$, this walk enjoys, besides time- and space-homogeneity, the SKIP-FREE PROPERTY, namely that to go from state x to state y it must pass through all intermediate states because its value can change by at most 1 at each step.

Example 3: Define

$$p(x) = \begin{cases} \frac{1}{2d}, & \text{if } x \in \{\pm e_1, \ldots, \pm e_d\} \\ 0, & \text{otherwise} \end{cases}$$

This is a simple random walk, which, in addition, has equal probabilities to move from one state x to one of its "neighbouring states" $x \pm e_1, \ldots, x \pm e_d$. We call it SIMPLE SYMMETRIC RANDOM WALK. If $d = 1$, then

$$p(1) = p(-1) = 1/2, \quad p(x) = 0, \text{ otherwise},$$

and this is the symmetric drunkard's walk.

The special structure of a random walk enables us to give more detailed formulae for its behaviour. Let us denote by S_n the state of the walk at time n. It is clear that S_n can be represented as

$$S_n = S_0 + \xi_1 + \cdots + \xi_n,$$

where $\xi_1, \xi_2, \ldots$ are i.i.d. random variables in $\mathbb{Z}^d$ with common distribution

$$P(\xi_n = x) = p(x), \quad x \in \mathbb{Z}^d.$$

We refer to ξ_n as the INCREMENTS of the random walk. Furthermore, the initial state S_0 is independent of the increments. Obviously,

$$p_{xy}^{(n)} = P_x(S_n = y) = P(x + \xi_1 + \cdots + \xi_n = y) = P(\xi_1 + \cdots + \xi_n = y - x).$$

The random walk $\xi_1 + \cdots + \xi_n$ is a random walk starting from 0. So, we can reduce several questions to questions about this random walk.

Furthermore, notice that

$$P(\xi_1 + \xi_2 = x) = \sum_y P(\xi_1 = x, x + \xi_2 = y) = \sum_y p(x)p(y - x).$$

(When we write a sum without indicating explicitly the range of the summation variable, we will mean a sum over all possible values.) Now the operation in the last term is known as CONVOLUTION. The convolution of two probabilities $p(x), p'(x)$, $x \in \mathbb{Z}^d$ is defined as a new probability $p * p'$ given by

$$(p * p')(x) = \sum_y p(x)p'(y - x).$$

The reader can easily check that

$$p * p' = p' * p, \quad p * (p' * p'') = (p * p') * p'', \quad p * \delta_z = p.$$

(Recall that δ_z is a probability distribution that assigns probability 1 to the point z.) In this notation, we have

$$P(\xi_1 + \xi_2 = x) = (p * p)(x).$$

We will let p^{*n} be the convolution of p with itself n times. We easily see that

$$P(\xi_1 + \cdots + \xi_n = x) = p^{*n}(x).$$

One could stop here and say that we have a formula for the n-step transition probability. But this would be nonsense, because all we have done is that we dressed the original problem in more fancy clothes. Indeed, computing the n-th order convolution for all n is a notoriously hard problem, in general. We shall resist the temptation to compute and look at other methods. After all, who cares about the exact value of $p^{*n}(x)$ for all n and x? Perhaps it is the case (a) that different questions we need to ask or (b) that approximate methods are more informative than exact ones.

Here are a couple of meaningful questions:
A. Will the random walk ever return to where it started from?
B. If yes, how long will that take?
C. If not, where will it go?

26 The simple symmetric random walk in dimension 1: path counting

This is a random walk with values in $\mathbb{Z}$ and transition probabilities $p_{i,i+1} = p_{i,i-1} = 1/2$ for all $i \in \mathbb{Z}$. When the walk starts from 0 we can write

$$S_n = \xi_1 + \cdots + \xi_n,$$

where the ξ_n are i.i.d. random variables (random signs) with $P(\xi_n = \pm 1) = 1/2$.

Look at the distribution of $(\xi_1, \ldots, \xi_n)$, a random vector with values in $\{-1, +1\}^n$. Clearly, all possible values of the vector are equally likely. Since there are 2^n elements in $\{0, 1\}^n$, we have

$$P(\xi_1 = \varepsilon_1, \ldots, \xi_n = \varepsilon_n) = 2^{-n}.$$

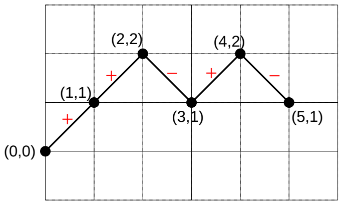
So, for fixed n, we are dealing with a uniform distribution on the sample space $\{-1, +1\}^n$. Events A that depend only on these the first n random signs are subsets of this sample space, and so

$$P(A) = \frac{\#A}{2^n}, \quad A \subset \{-1, +1\}^n,$$

where $\#A$ is the number of elements of A.

Therefore, if we can count the number of elements of A we can compute its probability. The principle is trivial, but its practice may be hard. In the sequel, we shall learn some counting methods.

First, we should assign, to each initial position and to each sequence of n signs, a path of length n. So if the initial position is $S_0 = 0$, and the sequence of signs is $\{+1, +1, -1, +1, -1\}$ then we have a path in the plane that joins the points $(0, 0) \to (1, 1) \to (2, 2) \to (3, 1) \to (4, 2) \to (5, 1)$:



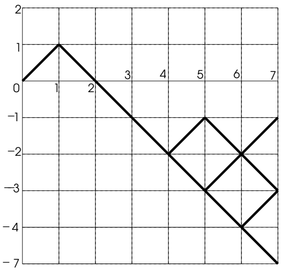
We can thus think of the elements of the sample space as paths.

As a warm-up, consider the following:

Example: Assume that $S_0 = 0$, and compute the probability of the event

$$A = \{S_2 \geq 0,\ S_4 = -1,\ S_6 < 0,\ S_7 < 0\}.$$

The paths comprising this event are shown below. We can easily count that there are 6 paths in A. So $P(A) = 6/2^6 = 6/128 = 3/64 \approx 0.047$.



26.1 Paths that start and end at specific points

Now consider a path that starts from a specific point and ends up at a specific point. We use the notation

$$\{(m,x) \rightsquigarrow (n,y)\} =: \{\text{all paths that start at } (m,x) \text{ and end up at } (n,y)\}.$$

Let us first show that

Lemma 14. *The number of paths that start from* $(0,0)$ *and end up at* (n,y) *is given by:*

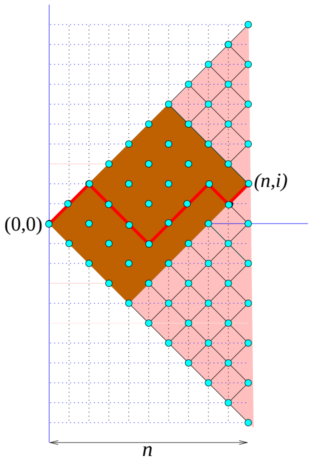
$$\boxed{\#\{(0,0) \rightsquigarrow (n,i)\} = \binom{n}{\frac{n+i}{2}}}$$

More generally, the number of paths that start from (m,x) *and end up at* (n,y) *is given by:*

$$\boxed{\#\{(m,x) \rightsquigarrow (n,y)\} = \binom{n-m}{(n-m+y-x)/2}}. \tag{23}$$

Remark. In if $n+i$ is not an even number then there is no path that starts from $(0,0)$ and ends at (n,i). Hence $\binom{n}{(n+i)/2} = 0$, in this case.

Proof. Consider a path that starts from $(0,0)$ and ends at (n,i).



The lightly shaded region contains all paths of length n *that start at* $(0,0)$*). (There are* 2^n *of them.)*
The darker region shows all paths that start at $(0,0)$ *and end up at the specific point* (n,i).

Let u be the number of $+$'s in this path (the number of upward segments) and d the number of $-$'s (the number of downward segments). Then

$$u + d = n, \quad u - d = i.$$

Hence

$$u = (n+i)/2, \quad d = (n-i)/2.$$

So if we know where the path starts from and where it ends at we know how many times it has gone up and how many down. The question then becomes: given a sequence of n signs out of which u are $+$ and the rest are $-$, in how many ways can we arrange them? Well, in exactly $\binom{n}{u}$ ways, and this proves the formula.

Notice that $u = (n+i)/2$ must be an integer. This is true if $n+i$ is even or, equivalently if n and i are simultaneously even or simultaneously odd. If this is not the case, then there is no path that can go from $(0,0)$ to (n,i). For example (look at the figure), there is no path that goes from $(0,0)$ to $(3,2)$. We shall thus define $\binom{n}{u}$ to be zero if u is not an integer.

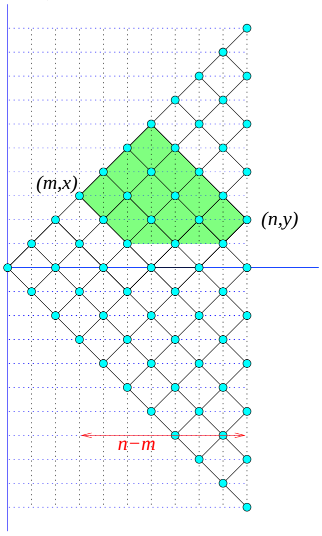
More generally, the number of paths that start from (m,x) and end up at (n,y) equals the number of paths that start from $(0,0)$ and end up at $(n-m, y-x)$. So, applying the previous formula, we obtain (23). □

26.2 Paths that start and end at specific points and do not touch zero at all

Lemma 15. *Suppose that $x, y > 0$. The number of paths $(m,x) \rightsquigarrow (n,y)$ that never become zero (i.e. they do not touch the horizontal axis) is given by:*

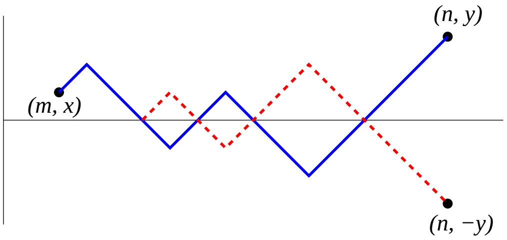
$$\#\{(m,x) \rightsquigarrow (n,y); \ \ \textit{remain} \ > 0\} = \binom{n-m}{\frac{1}{2}(n-m+y-x)} - \binom{n-m}{\frac{1}{2}(n-m-y-x)}. \tag{24}$$

Proof.



The shaded region contains all possible paths that start at (m,x) and end up at the specific point (n,y) which, in addition, never touch the horizontal axis.

Consider, instead, a path that starts at (m,x) and ends up at point (n,y) which DOES touch the horizontal axis, like the path indicated by the solid line below. There is a trick, called REFLECTION, which is useful here. Note that to each such path there corresponds another path obtained by following it up to the first time it hits zero and then reflecting it around the horizontal axis. In the figure below, follow the solid path starting at (m, x) up to the first time it hits zero and then follow the dotted path: this is the reflected path.



A path and its reflection around a horizontal line. Notice that the reflection starts only at the point where the path touches the horizontal line.

Note that if we know the reflected path we can recover the original path. Hence we may as well count the number of reflected paths. But each reflected path starts from (m, x) and ends at $(n, -y)$ and so it necessarily crosses the horizontal axis. We thus apply (23) with $-y$ in place of y:

$$\begin{aligned}\#\{(m,x) \rightsquigarrow (n,y); \text{ touch zero}\} &= \#\{(m,x) \rightsquigarrow (n,-y)\} \\ &= \binom{n-m}{(n-m-y-x)/2}.\end{aligned}$$

So, for $x > 0$, $y > 0$,

$$\begin{aligned}&\#\{(m,x) \rightsquigarrow (n,y); \text{ remain } > 0\} \\ &\quad = \#\{(m,x) \rightsquigarrow (n,y)\} - \#\{(m,x) \rightsquigarrow (n,y); \text{ touch zero}\} \\ &\quad = \binom{n-m}{\frac{1}{2}(n-m+y-x)} - \binom{n-m}{\frac{1}{2}(n-m-y-x)}.\end{aligned}$$

□

Corollary 14. *For any $y > 0$,*

$$\#\{(1,1) \rightsquigarrow (n,y); \text{ remain } > 0\} = \left(\frac{y}{n}\right) \#\{(0,0) \rightsquigarrow (n,y)\}. \tag{25}$$

Let $m = 1, x = 1$ in (24):

$$\begin{aligned}\#\{(1,1) \rightsquigarrow (n,y); \text{ remain } > 0\} &= \binom{n-1}{\frac{1}{2}(n+y)-1} - \binom{n-1}{\frac{1}{2}(n-y)-1} \\ &= \frac{(n-1)!}{\left(\frac{n+y}{2}-1\right)!\left(\frac{n-y}{2}\right)!} - \frac{(n-1)!}{\left(\frac{n-y}{2}-1\right)!\left(\frac{n-y}{2}\right)!} \\ &= \left(\frac{1}{n}\frac{n+y}{2} - \frac{1}{n}\frac{n-y}{2}\right)\frac{n!}{\left(\frac{n+y}{2}\right)!\left(\frac{n-y}{2}\right)!} \\ &= \frac{y}{n}\binom{n}{\frac{n+y}{2}},\end{aligned}$$

and the latter term equals the total number of paths from $(0, 0)$ to (n, y). □

Corollary 15. *The number of paths from* (m,x) *to* (n,y) *that do not touch level* $z < \min(x,y)$ *is given by:*

$$\#\{(m,x) \rightsquigarrow (n,y);\ \textit{remain}\ > z\} = \binom{n-m}{\frac{1}{2}(n-m+y-x)} - \binom{n-m}{\frac{1}{2}(n-m-y-x)-z}. \tag{26}$$

Proof. Because it is only the relative position of x and y with respect to z that plays any role, we have

$$\#\{(m,x) \rightsquigarrow (n,y);\ \text{remain}\ > z\} = \#\{(m,x+z) \rightsquigarrow (n,y+z);\ \text{remain}\ > 0\}$$

and so the formula is obtained by an application of (24). □

Corollary 16. *The number of paths from* $(0,0)$ *to* (n,y)*, where* $y \geq 0$*, that remain* ≥ 0 *is given by:*

$$\#\{(0,0) \rightsquigarrow (n,y);\ \textit{remain}\ \geq 0\} = \binom{n}{\frac{n+y}{2}} - \binom{n}{\frac{n+y}{2}+1} \tag{27}$$

Proof.

$$\begin{aligned}\#\{(0,0) \rightsquigarrow (n,y);\ \text{remain}\ \geq 0\} &= \#\{(0,0) \rightsquigarrow (n,y);\ \text{remain}\ > -1\} \\ &= \#\{(0,1) \rightsquigarrow (n,y+1);\ \text{remain}\ > 0\} \\ &= \binom{n}{\frac{n+y}{2}} - \binom{n}{\frac{n-y}{2}-1} \\ &= \binom{n}{\frac{n+y}{2}} - \binom{n}{\frac{n+y}{2}+1},\end{aligned}$$

where we used (26) with $z = -1$. □

Corollary 17 (Catalan numbers)**.** *The number of paths from* $(0,0)$ *to* $(n,0)$ *that remain* ≥ 0 *is given by:*

$$\#\{(0,0) \rightsquigarrow (n,0);\ \textit{remain}\ \geq 0\} = \frac{1}{n+1}\binom{n+1}{n/2}.$$

Proof. Apply (27) with $y = 0$:

$$\#\{(0,0) \rightsquigarrow (n,0);\ \text{remain}\ \geq 0\} = \binom{n}{\frac{n}{2}} - \binom{n}{\frac{n}{2}-1} = \frac{1}{n+1}\binom{n+1}{n/2},$$

where the latter calculation is as in the proof of (25). □

Corollary 18. *The number of paths that start from* $(0,0)$*, have length* n*, and remain nonnegative is given by:*

$$\#\{(0,0) \rightsquigarrow (n,\bullet);\ \textit{remain}\ \geq 0\} = \binom{n}{\lceil n/2 \rceil}. \tag{28}$$

where $\lceil x \rceil$ *denotes the smallest integer that is* $\geq x$*.*

Proof. To see, this, just use (27) and sum over all $y \geq 0$. If $n = 2m$ then we need to set $y = 2z$ in (27) (otherwise (27) is just zero) and so we have

$$\#\{(0,0) \rightsquigarrow (2m, \bullet);\ \text{remain } \geq 0\} = \sum_{z \geq 0} \left[\binom{2m}{m+z} - \binom{2m}{m+z+1} \right]$$

$$= \binom{2m}{m} - \binom{2m}{m+1} + \binom{2m}{m+1} - \binom{2m}{m+2} + \cdots + \binom{2m}{2m-1} - \binom{2m}{2m} + \binom{2m}{2m} \pm \cdots$$

$$= \binom{2m}{m} = \binom{n}{n/2},$$

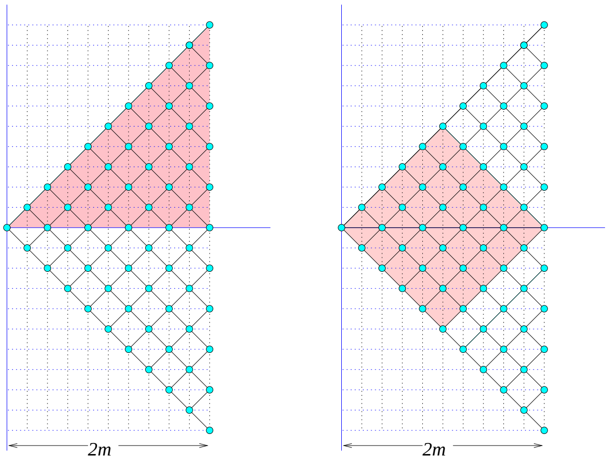
because all the terms that are denoted by $\cdots$ are equal to 0, while the rest cancel with one another, and only the first one remains. If $n = 2m+1$ then we need to set $y = 2z+1$ in (27) (otherwise (27) is just zero) and so we have

$$\#\{(0,0) \rightsquigarrow (2m+1, \bullet);\ \text{remain } \geq 0\} = \sum_{z \geq 0} \left[\binom{2m+1}{m+z+1} - \binom{2m+1}{m+z+2} \right]$$

$$= \binom{2m+1}{m+1} - \binom{2m+1}{m+2} + \binom{2m+1}{m+2} - \binom{2m+1}{m+3} + \cdots + \binom{2m+1}{2m} - \binom{2m+1}{2m+1} + \binom{2m+1}{2m+1} \pm \cdots$$

$$= \binom{2m+1}{m+1} = \binom{n}{\lceil n/2 \rceil}.$$

□



We have shown that the shaded regions contain exactly the same number of paths. Was this, a posteriori, *obvious?*

27 The simple symmetric random walk in dimension 1: simple probability calculations

Since all paths of certain length have the same probability, we can translate the path counting formulae into corresponding probability formulae.

27.1 Distribution after n steps

Lemma 16.

$$P_0(S_n = y) = \binom{n}{(n+y)/2} 2^{-n}.$$

$$P(S_n = y | S_m = x) = \binom{n-m}{(n-m+y-x)/2} 2^{-n-m}.$$

In particular, if n is even,

$$u(n) := P_0(S_n = 0) = \binom{n}{n/2} 2^{-n}. \tag{29}$$

Proof. We have

$$P_0(S_n = y) = \frac{\#\{(0,0) \rightsquigarrow (n,y)\}}{2^n},$$

and

$$P(S_n = y | S_m = x) = \frac{\#\{\text{paths } (m,x) \rightsquigarrow (n,y)\}}{2^{n-m}},$$

and we apply (23). □

27.2 The ballot theorem

Theorem 25. *The probability that, given that $S_0 = 0$ and $S_n = y > 0$, the random walk never becomes zero up to time n is given by*

$$P_0(S_1 > 0, S_2 > 0, \ldots, S_n > 0 \mid S_n = y) = \frac{y}{n}. \tag{30}$$

Proof. Such a simple answer begs for a physical/intuitive explanation. But, for now, we shall just compute it. The left side of (30) equals

$$\frac{\#\{\text{paths } (1,1) \rightsquigarrow (n,y) \text{ that do not touch zero}\} \times 2^{-n}}{\#\{\text{paths } (0,0) \rightsquigarrow (n,y)\} \times 2^{-n}}$$
$$= \frac{\binom{n-1}{\frac{1}{2}(n-1+y-1)} - \binom{n-1}{\frac{1}{2}(n-1-y-1)}}{\binom{n}{\frac{1}{2}(n+y)}} = \frac{y}{n}.$$

□

This is called the BALLOT THEOREM and the reason will be given in Section 29.

27.3 Some conditional probabilities

Lemma 17. *If $n, y \geq 0$ are both even or both odd then*

$$P_0(S_1 \geq 0 \ldots, S_n \geq 0 \mid S_n = y) = 1 - \frac{P_0(S_n = y+2)}{P_0(S_n = y)} = \frac{y+1}{\frac{1}{2}(n+y)+1}.$$

In particular, if n is even,

$$P_0(S_1 \geq 0 \ldots, S_n \geq 0 \mid S_n = 0) = \frac{1}{(n/2)+1}.$$

Proof. We use (27):

$$P_0(S_n \geq 0 \ldots, S_n \geq 0 \mid S_n = y) = \frac{P_0(S_n \geq 0 \ldots, S_n \geq 0, S_n = y)}{P_0(S_n = y)}$$
$$= \frac{\binom{n}{\frac{n+y}{2}} - \binom{n}{\frac{n+y}{2}+1}}{\binom{n}{\frac{n+y}{2}}} = 1 - \frac{P_0(S_n = y+2)}{P_0(S_n = y)}.$$

This is further equal to

$$1 - \frac{n!}{(\frac{n+y}{2}+1)!\,(\frac{n-y}{2}-1)!} \frac{(\frac{n+y}{2})!\,(\frac{n-y}{2})!}{n!} = 1 - \frac{\frac{n-y}{2}}{\frac{n+y}{2}+1} = \frac{y+1}{\frac{n+y}{2}+1}.$$

□

27.4 Some remarkable identities

Theorem 26. *If n is even,*

$$P_0(S_1 \geq 0, \ldots, S_n \geq 0) = P_0(S_1 \neq 0, \ldots, S_n \neq 0) = P_0(S_n = 0).$$

Proof. For even n we have

$$P_0(S_1 \geq 0, \ldots, S_n \geq 0) = \frac{\#\{(0,0) \rightsquigarrow (n, \bullet);\ \text{remain } \geq 0\}}{2^n} = \binom{n}{n/2} 2^{-n} = P_0(S_n = 0),$$

where we used (28) and (29). We next have

$$\begin{aligned}
P_0(S_1 \neq 0, \ldots, S_n \neq 0) &\overset{(a)}{=} P_0(S_1 > 0, \ldots, S_n > 0) + P_0(S_1 < 0, \ldots, S_n < 0) \\
&\overset{(b)}{=} 2P_0(S_1 > 0, \ldots, S_n > 0) \\
&\overset{(c)}{=} 2P_0(\xi_1 = 1)P_0(S_2 > 0, \ldots, S_n > 0 \mid \xi_1 = 1) \\
&\overset{(d)}{=} P_0(1 + \xi_2 > 0, 1 + \xi_2 + \xi_3 > 0, \ldots, 1 + \xi_2 + \cdots + \xi_n > 0 \mid \xi_1 = 1) \\
&\overset{(e)}{=} P_0(\xi_2 \geq 0, \xi_2 + \xi_3 \geq 0, \ldots, \xi_2 + \cdots + \xi_n \geq 0) \\
&\overset{(f)}{=} P_0(S_1 \geq 0, \ldots, S_{n-1} \geq 0) \\
&\overset{(g)}{=} P_0(S_1 \geq 0, \ldots, S_n \geq 0),
\end{aligned}$$

where (a) is obvious, (b) follows from symmetry, (c) is just conditioning on $\{\xi_1 = 1\}$ and using the fact that $S_1 = \xi_1$, (d) follows from $P_0(\xi_1 = 1) = 1/2$ and from the substitution of the value of ξ_1 on the left of the conditioning, (e) follows from the fact that ξ_1 is independent of $\xi_2, \xi_3, \ldots$ and from the fact that if m is integer then $1 + m > 0$ is equivalent to $m \geq 0$, (f) follows from the replacement of $(\xi_2, \xi_3, \ldots)$ by $(\xi_1, \xi_2, \ldots)$, and finally (g) is trivial because $S_{n-1} \geq 0$ means $S_{n-1} > 0$ because S_{n-1} cannot take the value 0, and $S_{n-1} > 0$ implies that $S_n \geq 0$. □

27.5 First return to zero

In (29), we computed, for even n, the probability $u(n)$ that the walk (started at 0) attains value 0 in n steps. We now want to find the probability $f(n)$ that the walk will return to 0 for the first time in n steps.

Theorem 27. *Let n be even, and $u(n) = P_0(S_n = 0)$. Then*

$$f(n) := P_0(S_1 \neq 0, \dots, S_{n-1} \neq 0, S_n = 0) = \frac{u(n)}{n-1}.$$

Proof. Since the random walk cannot change sign before becoming zero,

$$f(n) = P_0(S_1 > 0, \dots, S_{n-1} > 0, S_n = 0) + P_0(S_1 < 0, \dots, S_{n-1} < 0, S_n = 0).$$

By symmetry, the two terms must be equal. So

$$\begin{aligned} f(n) &= 2P_0(S_1 > 0, \dots, S_{n-1} > 0, S_n = 0) \\ &= 2P_0(S_{n-1} > 0, S_n = 0)\ P_0(S_1 > 0, \dots, S_{n-2} > 0 | S_{n-1} > 0, S_n = 0) \end{aligned}$$

Now,

$$P_0(S_{n-1} > 0, S_n = 0) = P_0(S_n = 0)\ P(S_{n-1} > 0 | S_n = 0).$$

We know that $P_0(S_n = 0) = u(n) = \binom{n}{n/2} 2^{-n}$. Also, given that $S_n = 0$, we either have $S_{n-1} = 1$ or -1, with equal probability. So the last term is 1/2. To take care of the last term in the last expression for $f(n)$, we see that $S_{n-1} > 0, S_n = 0$ means $S_{n-1} = 1, S_n = 0$. By the Markov property at time $n - 1$, we can omit the last event. So

$$f(n) = 2u(n)\frac{1}{2}P_0(S_1 > 0, \dots, S_{n-2} > 0 | S_{n-1} = 1),$$

and the last probability can be computed from the ballot theorem–see (30): it is equal to $1/(n-1)$. So

$$f(n) = \frac{u(n)}{n-1}.$$

□

28 The reflection principle for a simple random walk in dimension 1

The REFLECTION PRINCIPLE says the following. Fix a state a. Put a two-sided mirror at a. If a particle is to the right of a then you see its image on the left, and if the particle is to the left of a then its mirror image is on the right. So if the particle is in position s then its mirror image is in position $\widetilde{s} = 2a - s$ because $s - a = a - \widetilde{s}$.

It is clear that if a particle starts at 0 and performs a simple symmetric random walk then its mirror image starts at a and performs a simple symmetric random walk. This is not so interesting.

What is more interesting is this: Run the particle till it hits a (it will, for sure–Theorem 35) and after that consider its mirror image $\widetilde{S}_n = 2a - S_n$. In other words, define

$$\widetilde{S}_n = \begin{cases} S_n, & n < T_a \\ 2a - S_n, & n \geq T_a, \end{cases}.$$

where

$$T_a = \inf\{n \geq 0 : \ S_n = a\}$$

is the first hitting time of a. Then

Theorem 28. *If (S_n) is a simple symmetric random walk started at 0 then so is $(\widetilde{S}_n)$.*

Proof. Trivially,

$$S_n = \begin{cases} S_n, & n < T_a \\ a + (S_n - a), & n \geq T_a \end{cases}.$$

By the strong Markov property, the process $(S_{T_a+m} - a, m \geq 0)$ is a simple symmetric random walk, independent of the past before T_a. But a *symmetric* random walk has the same law as its negative. So $S_n - a$ can be replaced by $a - S_n$ in the last display and the resulting process, called $(\widetilde{S}_n, n \in \mathbb{Z}_+)$ has the same law as $(S_n, n \in \mathbb{Z}_+)$. □

28.1 Distribution of hitting time and maximum

We can take advantage of this principle to establish the distribution function of T_a:

Theorem 29. *Let $(S_n, n \geq 0)$ be a simple symmetric random walk and $a > 0$. Then*

$$P_0(T_a \leq n) = 2P_0(S_n \geq a) - P_0(S_n = a) = P_0(|S_n| \geq a) - \frac{1}{2}P_0(|S_n| = a).$$

Proof. If $S_n \geq a$ then $T_a \leq n$. So:

$$P_0(S_n \geq a) = P_0(T_a \leq n, S_n \geq a).$$

In the last event, we have $n \geq T_a$, so we may replace S_n by $2a - S_n$ (Theorem 28). Thus,

$$P_0(S_n \geq a) = P_0(T_a \leq n, 2a - S_n \geq a) = P_0(T_a \leq n, S_n \leq a).$$

But

$$P_0(T_a \leq n) = P_0(T_a \leq n, S_n \leq a) + P_0(T_a \leq n, S_n > a) = P_0(T_a \leq n, S_n \leq a) + P_0(S_n > a).$$

The last equality follows from the fact that $S_n > a$ implies $T_a \leq n$. Combining the last two displays gives

$$P_0(T_a \leq n) = P_0(S_n \geq a) + P_0(S_n > a) = P_0(S_n \geq a) + P_0(S_n \geq a) - P_0(S_n = a).$$

Finally, observing that S_n has the same distribution as $-S_n$ we get that this is also equal to $P_0(|S_n| \geq a) - \frac{1}{2}P_0(|S_n| = a)$. □

Define now the RUNNING MAXIMUM

$$M_n = \max(S_0, S_1, \dots, S_n),$$

Then, as a corollary to the above result, we obtain

Corollary 19.

$$P_0(M_n \geq x) = P_0(T_x \leq n) = 2P_0(S_n \geq x) - P_0(S_n = x) = P_0(|S_n| \geq x) - \frac{1}{2}P_0(|S_n| = x).$$

Proof. Observe that $M_n \geq x$ if and only if $S_k \geq x$ for some $k \leq n$ which is equivalent to $T_x \leq n$ and use Theorem 29. □

We can do more: we can derive the joint distribution of M_n and S_n:

Theorem 30.

$$P_0(M_n < x, S_n = y) = P_0(S_n = y) - P_0(S_n = 2x - y) \quad , \qquad x > y.$$

Proof. Since $M_n < x$ is equivalent to $T_x > n$, we have

$$P_0(M_n < x, S_n = y) = P_0(T_x > n, S_n = y) = P_0(S_n = y) - P_0(T_x \leq n, S_n = y). \tag{31}$$

If $T_x \leq n$, then, by applying the reflection principle (Theorem 28), we can replace S_n by $2x - S_n$ in the last part of the last display and get

$$P_0(T_x \leq n, S_n = y) = P_0(T_x \leq n, S_n = 2x - y).$$

But if $x > y$ then $2x - y > x$ and so $\{S_n = 2x - y\} \subset \{T_x \leq n\}$, which results into

$$P_0(T_x \leq n, S_n = y) = P_0(S_n = 2x - y).$$

This, combined with (31), gives the result. □

29 Urns and the ballot theorem

Suppose that an urn[8] contains n items, of which a are coloured azure and b black; $n = a+b$. A sample from the urn is a sequence $\eta = (\eta_1, \dots, \eta_n)$, where η_i indicates the colour of the i-th item picked. The set of values of η contains $\binom{n}{a}$ elements and η is uniformly distributed in this set.

The original ballot problem, posed in 1887[9] asks the following: if we start picking the items from the urn one by one without replacement, what is the probability that the number of azure items is constantly ahead of the number of black ones? The problem was solved[10] in the same year. The answer is very simple:

Theorem 31. *If an urn contains a azure items and $b = n - a$ black items, then the probability that, during sampling without replacement, the number of selected azure items is always ahead of the black ones equals $(a - b)/n$, as long as $a \geq b$.*

[8] An URN is a vessel, usually a vase furnished with a foot or pedestal, employed for different purposes, as for holding liquids, ornamental objects, the ashes of the dead after cremation, and anciently for holding ballots to be drawn. It is the latter use we are interested in in Probability and Statistics.

[9] Bertrand, J. (1887). Solution d'un problème. *Compt. Rend. Acad. Sci. Paris*, **105**, 369.

[10] André, D. (1887). Solution directe du problème résolu par M. Bertrand. *Compt. Rend. Acad. Sci. Paris*, **105**, 436-437.

We shall call $(\eta_1, \dots, \eta_n)$ an urn process with parameters n, a, always assuming (without any loss of generality) that $a \ge b = n - a$.

An urn can be realised quite easily be a random walk:

Theorem 32. *Consider a simple symmetric random walk $(S_n, n \ge 0)$ with values in $\mathbb{Z}$ and let y be a positive integer. Then, conditional on the event $\{S_n = y\}$, the sequence $(\xi_1, \dots, \xi_n)$ of the increments is an urn process with parameters n and $a = (n+y)/2$.*

Proof. The values of each ξ_i are ± 1, so 'azure' items are $+1$ and 'black' items are -1. We need to show that, conditional on $\{S_n = y\}$, the sequence $(\xi_1, \dots, \xi_n)$ is uniformly distributed in a set with $\binom{n}{a}$ elements. But if $S_n = y$ then, letting a, b be defined by $a + b = n$, $a - b = y$, we have that a out of the n increments will be $+1$ and b will be -1. So the number of possible values of $(\xi_1, \dots, \xi_n)$ is, indeed, $\binom{n}{a}$. It remains to show that all values are equally likely. Let $\varepsilon_1, \dots, \varepsilon_n$ be elements of $\{-1, +1\}$ such that $\varepsilon_1 + \cdots + \varepsilon_n = y$. Then, using the definition of conditional probability and Lemma 16, we have:

$$\begin{aligned} P_0(\xi_1 = \varepsilon_1, \dots, \xi_n = \varepsilon_n \mid S_n = y) &= \frac{P_0(\xi_1 = \varepsilon_1, \dots, \xi_n = \varepsilon_n)}{P_0(S_n = y)} \\ &= \frac{2^{-n}}{\binom{n}{(n+y)/2} 2^{-n}} = \frac{1}{\binom{n}{a}}. \end{aligned}$$

Thus, conditional on $\{S_n = y\}$, $(\xi_1, \dots, \xi_n)$ has a uniform distribution. □

Proof of Theorem 31. Since an urn with parameters n, a can be realised by a simple symmetric random walk starting from 0 and ending at $S_n = y$, where $y = a - (n - a) = 2a - n$, we can translate the original question into a question about the random walk. Since the 'azure' items correspond to $+$ signs (respectively, the 'black' items correspond to $-$ signs), the event that the azure are always ahead of the black ones during sampling is precisely the event $\{S_1 > 0, \dots, S_n > 0\}$ that the random walk remains positive. But in (30) we computed that

$$P_0(S_1 > 0, \dots, S_n > 0 \mid S_n = y) = \frac{y}{n}.$$

Since $y = a - (n - a) = a - b$, the result follows. □

30 The asymmetric simple random walk in dimension 1

Consider now a simple random walk with values in $\mathbb{Z}$, but which is not necessarily symmetric. (The word "asymmetric" will always mean "not necessarily symmetric".) So

$$p_{i,i+1} = p, \quad p_{i,i-1} = q = 1 - p, \quad i \in \mathbb{Z}.$$

We have

$$P_0(S_n = k) = \binom{n}{\frac{n+k}{2}} p^{(n+k)/2} q^{(n-k)/2}, \quad -n \le k \le n.$$

30.1 First hitting time analysis

We will use probabilistic + analytical methods to compute the distribution of the hitting times

$$T_x = \inf\{n \geq 0 : \; S_n = x\}, \quad x \in \mathbb{Z},$$

This random variable takes values in $\{0, 1, 2, \ldots\} \cup \{\infty\}$.

Lemma 18. *For all $x, y \in \mathbb{Z}$, and all $t \geq 0$,*

$$P_x(T_y = t) = P_0(T_{y-x} = t).$$

Proof. This follows from the spatial homogeneity of the transition probabilities: to reach y starting from x, only the relative position of y with respect to x is needed. □

In view of this, it is no loss of generality to consider the random walk starting from 0.

Lemma 19. *Consider a simple random walk starting from 0. If $x > 0$, then the summands in*

$$T_x = T_1 + (T_2 - T_1) + \cdots + (T_x - T_{x-1})$$

are i.i.d. random variables.

Proof. To prove this, we make essential use of the skip-free property, namely the obvious fact that a simple random walk starting from 0 must pass through levels $1, 2, \ldots, x-1$ in order to reach x. The rest is a consequence of the strong Markov property. □

Corollary 20. *Consider a simple random walk starting from 0. The process $(T_x, x \geq 0)$ is also a random walk.*

In view of Lemma 19 we need to know the distribution of T_1. We will approach this using probability generating functions. Let

$$\varphi(s) := E_0 s^{T_1}.$$

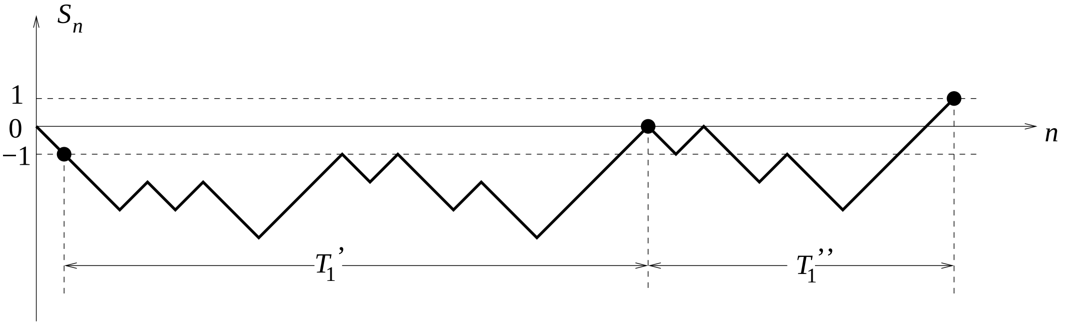
(We tacitly assume that $S_0 = 0$.) Then, for $x > 0$,

$$E_0 s^{T_x} = \varphi(s)^x.$$

Theorem 33. *Consider a simple random walk starting from 0. Then the probability generating function of the first time that level 1 is reached is given by*

$$\varphi(s) = E_0 s^{T_1} = \frac{1 - \sqrt{1 - 4pqs^2}}{2qs}.$$

Proof. Start with $S_0 = 0$ and watch the process until the first $n = T_1$ such that $S_n = 1$. If $S_1 = 1$ then $T_1 = 1$. If $S_1 = -1$ then T_1 is the sum of three times: one unit of time spent to go from 0 to -1, plus a time T_1' required to go from -1 to 0 for the first time, plus a further time T_1'' required to go from 0 to $+1$ for the first time. See figure below.



If $S_1 = -1$ then, in order to hit 1, the RW must first hit 0 (it takes time T_1' to do this) and then 1 (it takes a further time of T_1'').

Hence

$$T_1 = \begin{cases} 1 & \text{if } S_1 = 1 \\ 1 + T_1' + T_1'' & \text{if } S_1 = -1 \end{cases}$$

This can be used as follows:

$$\begin{aligned} \varphi(s) &= E_0[s\mathbf{1}(S_1 = 1) + s^{1+T_1'+T_1''}\mathbf{1}(S_1 = -1)] \\ &= sP(S_1 = 1) + sE_0[s^{T_1'}s^{T_1''}|S_1 = -1]P(S_1 = -1) \\ &= sp + sE_0[s^{T_1'}s^{T_1''}|S_1 = -1]q \end{aligned}$$

But, from the strong Markov property, T_1' is independent of T_1'' conditional on $S_1 = -1$ and each with distribution that of T_1. We thus have

$$\varphi(s) = ps + qs\varphi(s)^2 \,.$$

The formula is obtained by solving the quadratic. □

Corollary 21. *Consider a simple random walk starting from 0. Then the probability generating function of the first time that level −1 is reached is given by*

$$E_0 s^{T_{-1}} = \frac{1 - \sqrt{1 - 4pqs^2}}{2ps}.$$

Proof. The first time n such that $S_n = -1$ is the first time n such that $-S_n = 1$. But note that $-S_n$ is a simple random walk with upward transition probability q and downward transition probability p. Hence the previous formula applies with the roles of p and q interchanged. □

Corollary 22. *Consider a simple random walk starting from 0. Then the probability generating function of the first time that level $x \in \mathbb{Z}$ is reached is given by*

$$E_0 s^{T_x} = \begin{cases} \left(E_0 s^{T_1}\right)^x = \left(\frac{1-\sqrt{1-4pqs^2}}{2qs}\right)^x, & \textit{if } x > 0 \\ \left(E_0 s^{T_{-1}}\right)^{|x|} = \left(\frac{1-\sqrt{1-4pqs^2}}{2ps}\right)^{|x|}, & \textit{if } x < 0 \end{cases} \tag{32}$$

Proof. Use Lemma 19 and the previous formulae of Theorem 33 and Corollary 21. □

30.2 First return to the origin

Now let us ask: If the particle starts at 0, how long will it take till it first returns to 0? We are talking about the stopping time

$$T_0' := \inf\{n \geq 1 : S_n = 0\}.$$

(Pay attention: it is important to write $n \geq 1$ inside the set!)

Theorem 34. *Consider a simple random walk starting from* 0*. Then the probability generating function of the first return time to* 0 *is given by*

$$E_0 s^{T_0'} = 1 - \sqrt{1 - 4pqs^2}.$$

Proof. We again do first-step analysis.

$$\begin{aligned} E_0 s^{T_0'} &= E_0\big(s^{T_0'}\,\mathbf{1}(S_1 = -1)\big) + E_0\big(s^{T_0'}\,\mathbf{1}(S_1 = 1)\big) \\ &= qE_0\big(s^{T_0'} \mid S_1 = -1\big) + pE_0\big(s^{T_0'} \mid S_1 = 1\big). \end{aligned}$$

But if $S_1 = -1$ then T_0' equals 1 plus the time required for the walk to hit 0 starting from -1. The latter is, in distribution, the same as the time required for the walk to hit 1 starting from 0. Similarly, if $S_1 = 1$, then T_0' equals 1 plus a time whose distribution that of the time required for the walk to hit -1 starting from 0.

$$\begin{aligned} E_0 s^{T_0'} &= qE_0(s^{1+T_1}) + pE_0(s^{1+T_{-1}}) \\ &= qs\frac{1-\sqrt{1-4pqs^2}}{2qs} + ps\frac{1-\sqrt{1-4pqs^2}}{2ps}. \\ &= 1 - \sqrt{1-4pqs^2}. \end{aligned}$$

□

30.3 The distribution of the first return to the origin

We wish to find the exact distribution of the first return time to the origin. For a symmetric random walk, this was found in §30.2: $P_0(T_0' = n) = f(n) = u(n)/(n-1)$ is n is even. For the general case, we have computed the probability generating function $E_0 s^{T_0}$–see Theorem 34

The tool we need here is Taylor's theorem for the function $f(x) = (1+x)^\alpha$, where α is a real number. If $\alpha = n$ is a positive integer then

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

by the binomial theorem. If we define $\binom{n}{k}$ to be equal to 0 for $k > n$, then we can omit the upper limit in the sum, and simply write

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

Recall that

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

It turns out that the formula is formally true for general α, and it looks exactly the same:

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

as long as we define

$$\binom{\alpha}{k} = \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!}.$$

The adverb 'formally' is supposed to mean that this formula holds for some x around 0, but that we won't worry about which ones.

For example, we will show that

$$(1-x)^{-1} = 1 + x + x^2 + x^3 + \cdots$$

which is of course the all too-familiar geometric series. Indeed,

$$(1-x)^{-1} = \sum_{k\geq 0} \binom{-1}{k} (-x)^k$$

But, by definition,

$$\binom{-1}{k} = \frac{(-1)(-2)\cdots(-1-k+1)}{k!} = \frac{(-1)^k k!}{k!} = (-1)^k,$$

so

$$(1-x)^{-1} = \sum_{k\geq 0} (-1)^k (-x)^k = \sum_{k\geq 0} (-1)^{2k} x^k = \sum_{k\geq 0} x^k,$$

as needed.

As another example, let us compute $\sqrt{1-x}$. We have

$$\sqrt{1-x} = (1-x)^{1/2} = \sum_{k\geq 0} \binom{1/2}{k} (-1)^k x^k.$$

We can write this in more familiar symbols if we observe that

$$\binom{1/2}{k} (-1)^k = \frac{1}{2k-1} \binom{2k}{k} 4^{-k},$$

and this is shown by direct computation. So

$$\sqrt{1-x} = \sum_{k\geq 0} \frac{1}{2k-1} \binom{2k}{k} x^k = 1 + \sum_{k\geq 1} \frac{1}{2k-1} \binom{2k}{k} x^k$$

We apply this to $E_0 s^{T_0'} = 1 - \sqrt{1-4pqs^2}$:

$$E_0 s^{T_0'} = \sum_{k\geq 1} \frac{1}{2k-1} \binom{2k}{k} (pq)^k s^{2k}.$$

But, by the definition of $E_0 s^{T_0'}$,

$$E_0 s^{T_0'} = \sum_{n \geq 0} P_0(T_0' = n) s^n.$$

Comparing the two 'infinite polynomials' we see that we must have $n = 2k$ (we knew that– the random walk can return to 0 only in an even number of steps) and then

$$P_0(T_0' = 2k) = \frac{1}{2k-1} \binom{2k}{k} p^k q^k = \frac{1}{2k-1} P_0(S_{2k} = 0).$$

31 Recurrence properties of simple random walks

Consider again a simple random walk S_n with values in $\mathbb{Z}$ with upward probability p and downward probability $q = 1 - p$. By applying the law of large numbers we have that

$$P_0\left(\lim_{n\to\infty} \frac{S_n}{n} = p - q \right) = 1.$$

- If $p > 1/2$ then $p - q > 0$ and so S_n tends to $+\infty$ with probability one. Hence it is transient.
- If $p < 1/2$ then $p - q < 0$ and so S_n tends to $-\infty$ with probability one. Hence it is again transient.
- If $p = 1/2$ then $p - q = 0$ and the law of large numbers only tells us that the walk is not positive recurrent. That it is actually *null-recurrent* follows from Markov chain theory.

We will see how the results about the generating functions of the hitting and return times help us answer these questions directly, without appeal to the Markov chain theory. The key fact needed here is:

$$P_0(T_x < \infty) = \lim_{s \uparrow 1} E_0 s^{T_x},$$

which is true for *any* probability generating function. We apply this to (32). The quantity inside the square root becomes $1 - 4pq$. But remember that $q = 1 - p$, so

$$\sqrt{1-4pq} = \sqrt{1 - 4p + 4p^2} = \sqrt{(1-2p)^2} = |1 - 2p| = |p - q|.$$

So

$$P_0(T_x < \infty) = \begin{cases} \left(\frac{1-|p-q|}{2q} \right)^x, & \text{if } x > 0 \\ \left(\frac{1-|p-q|}{2p} \right)^{|x|}, & \text{if } x < 0 \end{cases} \tag{33}$$

This formula can be written more neatly as follows:

Theorem 35.

$$P_0(T_x < \infty) = \begin{cases} \left(\frac{p}{q} \wedge 1\right)^x, & \textit{if } x > 0 \\ \left(\frac{q}{p} \wedge 1\right)^{|x|}, & \textit{if } x < 0 \end{cases}$$

In particular, if $p = q$ *then* $P_0(T_x < \infty) = 1$ *for all* $x \in \mathbb{Z}$.

Proof. If $p > q$ then $|p - q| = p - q$ and so (33) gives

$$P_0(T_x < \infty) = \begin{cases} 1, & \text{if } x > 0 \\ \frac{q}{p}, & \text{if } x < 0 \end{cases},$$

which is what we want. If $p < q$ then $|p - q| = q - p$ and, again, we obtain what we are looking for. □

Corollary 23. *The simple symmetric random walk with values in $\mathbb{Z}$ is null-recurrent.*

Proof. The last part of Theorem 35 tells us that it is recurrent. We know that it cannot be positive recurrent, and so it is null-recurrent. □

31.1 Asymptotic distribution of the maximum

Suppose now that $p < 1/2$. Then $\lim_{n\to\infty} S_n = -\infty$ with probability 1. This means that there is a largest value that will be ever attained by the random walk. This value is random and is denoted by

$$\boxed{M_\infty = \sup(S_0, S_1, S_2, \ldots)}$$

If we let

$$M_n = \max(S_0, \ldots, S_n),$$

then we have $M_\infty = \lim_{n\to\infty} M_n$. Indeed, the sequence M_n is nondecreasing; every nondecreasing sequence has a limit, except that the limit may be ∞; but here it is $< \infty$, with probability 1 because $p < 1/2$.

Theorem 36. *Consider a simple random walk with values in $\mathbb{Z}$. Assume $p < 1/2$. Then the overall maximum M_∞ has a geometric distribution:*

$$P_0(M_\infty \geq x) = (p/q)^x, \quad x \geq 0.$$

Proof.

$$P_0(M_\infty \geq x) = \lim_{n\to\infty} P_0(M_n \geq x) = \lim_{n\to\infty} P_0(T_x \leq n) = P_0(T_x < \infty) = (p/q)^x, \quad x \geq 0.$$

□

31.2 Return probabilities

Only in the symmetric case $p = q = 1/2$ will the random walk return to the point where it started from. Otherwise, with some positive probability, it will never ever return to the point where it started from. What is this probability?

Theorem 37. *Consider a random walk with values in $\mathbb{Z}$. Let*

$$\boxed{T'_0 = \inf\{n \geq 1 : \; S_n = 0\}}$$

be the first return time to 0. Then

$$P_0(T'_0 < \infty) = 2(p \wedge q).$$

Unless $p = 1/2$, this probability is always strictly less than 1.

Proof. The probability generating function of T_0' was obtained in Theorem 34. But

$$P_0(T_0' < \infty_) = \lim_{s \uparrow 1} E s^{T_0'} = 1 - \sqrt{1-4pq} = 1 - |p-q|.$$

If $p = q$ this gives 1. If $p > q$ this gives $1-(p-q) = 1-(1-q-q) = 2q$. And if $p < q$ this gives $1-(q-p) = 1-(1-p-p) = 2p$. □

31.3 Total number of visits to a state

If the random walk is recurrent then it visits any state infinitely many times. But if it is transient, the total number of visits to a state will be finite. The total number of visits to state i is given by

$$\boxed{J_i = \sum_{n=1}^{\infty} \mathbf{1}(S_n = i),} \tag{34}$$

a random variable first considered in Section 10 where it was also shown to be geometric. We now compute its exact distribution.

Theorem 38. *Consider a simple random walk with values in* $\mathbb{Z}$. *Then, starting from zero, the total number* J_0 *of visits to state* 0 *is geometrically distributed with*

$$P_0(J_0 \geq k) = (2(p \wedge q))^k, \quad k = 0, 1, 2, \ldots,$$
$$E_0 J_0 = \frac{2(p \wedge q)}{1 - 2(p \wedge q)}.$$

Proof. In Lemma 4 we showed that, if a Markov chain starts from a state i then J_i is geometric. And so

$$P_0(J_0 \geq k) = P_0(J_0 \geq 1)^k, \quad k = 0, 1, 2, \ldots$$

But $J_0 \geq 1$ if and only if the first return time to zero is finite. So

$$P_0(J_0 \geq 1) = P_0(T_0' < \infty) = 2(p \wedge q),$$

where the last probability was computed in Theorem 37. This proves the first formula. For the second formula we have

$$E_0 J_0 = \sum_{k=1}^{\infty} P_0(J_0 \geq k) = \sum_{k=1}^{\infty} (2(p \wedge q))^k = \frac{2(p \wedge q)}{1 - 2(p \wedge q)},$$

this being a geometric series. □

Remark 1: By spatial homogeneity, we obviously have that the same formulae apply for any i:

$$P_i(J_i \geq k) = (2(p \wedge q))^k, \quad k = 0, 1, 2, \ldots$$
$$E_i J_i = \frac{2(p \wedge q)}{1 - 2(p \wedge q)}.$$

Remark 2: The formulae work for all values of $p \in [0, 1]$, even for $p = 1/2$. In the latter case, $P_i(J_i \geq k) = 1$ for all k, meaning that $P_i(J_i = \infty) = 1$, as already known.

We now look at the the number of visits to some state but if we start from a different state.

Theorem 39. *Consider a random walk with values in $\mathbb{Z}$. If $p \le 1/2$ then*

$$P_0(J_x \ge k) = (p/q)^{x \vee 0}\ (2p)^{k-1}, \quad k \ge 1$$
$$E_0 J_x = \frac{(p/q)^{x \vee 0}}{1-2p}.$$

If $p \ge 1/2$ then

$$P_0(J_x \ge k) = (q/p)^{(-x) \vee 0}\ (2q)^{k-1}, \quad k \ge 1$$
$$E_0 J_x = \frac{(q/p)^{(-x) \vee 0}}{1-2q}.$$

Proof. Suppose $x > 0$, $k \ge 1$. Then

$$P_0(J_x \ge k) = P_0(T_x < \infty,\ J_x \ge k),$$

simply because if $J_x \ge k \ge 1$ then $T_x < \infty$. Apply now the strong Markov property at T_x. Then

$$P_0(T_x < \infty,\ J_x \ge k) = P_0(T_x < \infty) P_x(J_x \ge k-1).$$

The reason that we replaced k by $k-1$ is because, given that $T_x < \infty$, there has already been one visit to state x and so we need at least $k-1$ *remaining visits* in order to have at least k visits in total. The first term was computed in Theorem 35, and the second in Theorem 38. So we obtain

$$P_0(J_x \ge k) = \left(\frac{p}{q} \wedge 1\right)^x\ (2(p \wedge q))^{k-1}.$$

If $p \le 1/2$ then this gives $P_0(J_x \ge k) = (p/q)^x (2p)^{k-1}$. If $p \ge 1/2$ then this gives $P_0(J_x \ge k) = (2q)^{k-1}$. We repeat the argument for $x < 0$ and find

$$P_0(J_x \ge k) = \left(\frac{q}{p} \wedge 1\right)^{|x|}\ (2(p \wedge q))^{k-1}.$$

If $p \le 1/2$ then this gives $P_0(J_x \ge k) = (2p)^{k-1}$. If $p \ge 1/2$ then this gives $P_0(J_x \ge k) = (q/p)^{|x|}(2q)^{k-1}$.
As for the expectation, we apply $E_0 J_x = \sum_{k=1}^{\infty} P_0(J_x \ge k)$ and perform the sum of a geometric series. □

Remark 1: The expectation $E_0 J_x$ is finite if $p \ne 1/2$. Consider the case $p < 1/2$. For $x > 0$, then $E_0 J_x$ drops down to zero geometrically fast as x tends to infinity. But for $x < 0$, we have $E_0 J_x = 1/(1-2p)$ regardless of x. In other words, if the random walk "drifts down" (i.e. $p < 1/2$) then, starting from 0, it will visit any point below 0 exactly the same number of times on the average.

Remark 2: If the initial state is not 0 then use spatial homogeneity to obtain the analogous formulae, i.e. $P_y(J_x \ge k) = P_0(J_{x-y} \ge k)$ and $E_y J_x = E_0 J_{x-y}$.

32 Duality

Consider random walk $S_k = \sum_{i=1}^{k} \xi_i$, i.e. a sum of i.i.d. random variables. Fix an integer n. Then the DUAL $(\widetilde{S}_k,\ 0 \le k \le n)$ of $(S_k,\ 0 \le k \le n)$ is defined by

$$\widetilde{S}_k := S_n - S_{n-k}.$$

Theorem 40. *$(\widetilde{S}_k,\ 0 \le k \le n)$ has the same distribution as $(S_k,\ 0 \le k \le n)$.*

Proof. Use the obvious fact that $(\xi_1, \xi_2, \dots, \xi_n)$ has the same distribution as $(\xi_n, \xi_{n-1}, \dots, \xi_1)$. □

Thus, every time we have a probability for S we can translate it into a probability for $\widetilde{S}$, which is basically another probability for S, because the two have the same distribution.

Here is an application of this:

Theorem 41. *Let (S_k) be a simple symmetric random walk and x a positive integer. Let $T_x = \inf\{n \ge 1 : S_n = x\}$ be the first time that x will be visited. Then*

$$\boxed{P_0(T_x = n) = \frac{x}{n} P_0(S_n = x)}\ .$$

Proof. Rewrite the ballot theorem terms of the dual:

$$P_0(\widetilde{S}_1 > 0, \dots, \widetilde{S}_n > 0 \mid \widetilde{S}_n = x) = \frac{x}{n}, \quad x > 0.$$

But the left-hand side is

$$P_0(S_n - S_{n-k} > 0,\ 1 \le k \le n \mid S_n = x) = \frac{P_0(S_k < x,\ 0 \le k \le n-1;\ S_n = x)}{P_0(S_n = x)} = \frac{P_0(T_x = n)}{P_0(S_n = x)}.$$

□

Pause for reflection: We have been dealing with the random times T_x for simple random walks (symmetric or asymmetric) from the beginning of the lectures.

In Lemma 19 we observed that, for a simple random walk and for $x > 0$, T_x is the sum of x i.i.d. random variables, each distributed as T_1, and thus derived the generating function of T_x:

$$E_0 s^{T_x} = \left(\frac{1 - \sqrt{1 - 4pqs^2}}{2qs} \right)^x.$$

Specialising to the symmetric case we have

$$E_0 s^{T_x} = \left(\frac{1 - \sqrt{1 - s^2}}{s} \right)^x. \tag{35}$$

In Theorem 29 we used the reflection principle and derived the distribution function of T_x for a simple symmetric random walk:

$$P_0(T_x \le n) = P_0(|S_n| \ge x) - \frac{1}{2} P_0(|S_n| = x). \tag{36}$$

Lastly, in Theorem 41, we derived using duality, for a simple symmetric random walk, the probabilities

$$P_0(T_x = n) = \frac{x}{n} P_0(S_n = x). \tag{37}$$

We remark here something methodological: It was completely different techniques that led to results about, essentially, the same thing: the distribution of T_x. This is an important lesson to learn.

Of course, the three formulae should be compatible: the number (37) should be the coefficient of s^n in (35); summing up (37) should give (36), or taking differences in (36) should give (37). But on trying to verify these things algebraically, the reader will realise that it is not easy.

33 Amount of time that a SSRW is positive*

Consider a SSRW (S_n) starting from 0. Define its trajectory as the polygonal line obtained by joining the points (n, S_n) in the plane, as described earlier. Watch this trajectory up to time $2n$. Let $T_+(2n)$ be the number of sides of the polygonal line that are above the horizontal axis and $T_-(2n) = 2n - T_+(2n)$ the number of sides that are below. Clearly, $T_\pm(2n)$ are both even numbers because, if the RW is at 0, it takes an even number of steps to return to 0. We can think of $T_+(2n)$ as the number of leads in $2n$ coin tosses, where we have a lead at some time if the number of heads till that time exceeds the number of tails. We are dealing with a fair coin which is tossed $2n$ times, so the event that the number of heads equals k has maximum probability when $k = n$ (heads=tails).

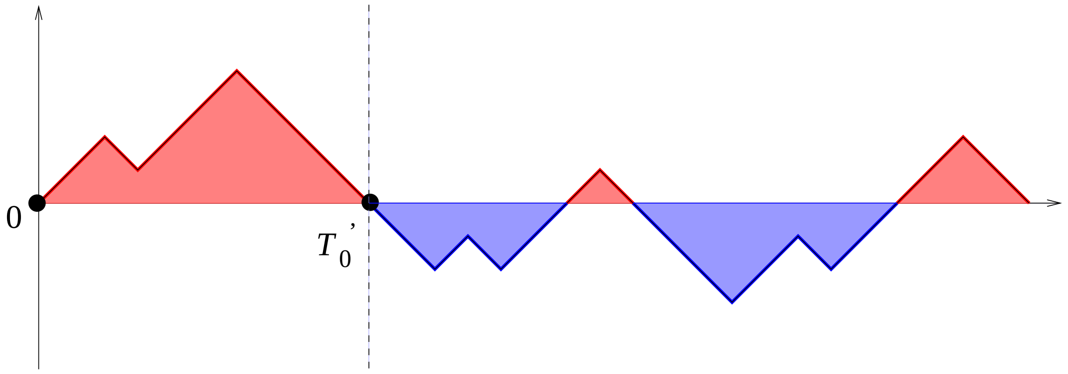


Figure 1: *Part of the time the SSRW is positive and part negative. After it returns to* 0 *at time* T_0', *it behaves again like a SSRW and is independent of the past.*

Most people take this statement and translate it into a statement about $T_+(2n)$ arguing that $P(T_+(2n) = m)$ is maximised when $m = n$, i.e. when $T_+(2n) = T_-(2n)$. This is wrong. In fact, as the theorem below shows, the most likely value of $P(T_+(2n) = m)$ is when $m = 0$ or $m = 2n$.

Theorem 42 (distribution of the total time the random walk is positive)**.**

$$P(T_+(2n) = 2k) = P(S_{2k} = 0)P(S_{2n-2k} = 0) \quad , \quad 0 \le k \le n.$$

Proof. The idea is to condition at the first time T_0' that the RW returns to 0:

$$P(T_+(2n) = 2k) = \sum_{r=1}^{n} P(T_+(2n) = 2k, T_0' = 2r).$$

Between 0 and T_0' the random walk is either above the horizontal axis or below; call the first event A_+ and the second A_- and write

$$\begin{aligned} P(T_+(2n) = 2k) = &\sum_{r=1}^{n} P(A_+, T_0' = 2r) P(T_+(2n) = 2k \mid A_+, T_0' = 2r) \\ &+ \sum_{r=1}^{n} P(A_-, T_0' = 2r) P(T_+(2n) = 2k \mid A_-, T_0' = 2r) \end{aligned}$$

In the first case,

$$P(T_+(2n) = 2k \mid A_+, T_0' = 2r) = P(T_+(2n) = 2k-2r \mid A_+, T_0' = 2r) = P(T_+(2n-2r) = 2k-2r),$$

because, if $T_0' = 2r$ and A_+ occurs then the RW has already spent $2r$ units of time above the axis and so $T_+(2n)$ is reduced by $2r$. Furthermore, we may remove the conditioning because the future evolution of the RW after T_0' is independent of the past. In the second case, a similar argument shows

$$P(T_-(2n) = 2k \mid A_-, T_0' = 2r) = P(T_-(2n) = 2k \mid A_-, T_0' = 2r) = P(T_-(2n-2r) = 2k).$$

We also observe that, by symmetry,

$$P(A_+, T_0' = 2r) = P(A_-, T_0' = 2r) = \frac{1}{2} P(T_0' = 2r) = f_{2r}.$$

Letting $p_{2k,2n} = P(T_+(2n) = 2k)$, we have shown that

$$p_{2k,2n} = \frac{1}{2} \sum_{r=1}^{k} f_{2r} p_{2k-2r,2n-2r} + \frac{1}{2} \sum_{r=1}^{n-k} f_{2r} p_{2k,2n-2r}$$

We want to show that $p_{2k,2n} = u_{2k} u_{2n-2k}$, where $u_{2k} = P(S_{2k} = 0)$. With this hypothesis, we can start playing with the recursion of the display above and using induction on n we will realise that

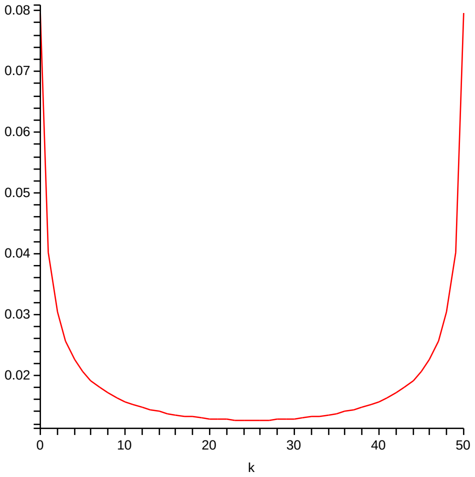
$$p_{2k,2n} = \frac{1}{2} u_{2n-2k} \sum_{r=1}^{k} f_{2r} u_{2k-2r} + \frac{1}{2} u_{2k} \sum_{r=1}^{n-k} f_{2r} u_{2n-2k-2r}.$$

□

Explicitly, we have

$$P(T_+(2n) = 2k) = \binom{2k}{k} \binom{2n-2k}{n-k} 2^{-2n}, \quad 0 \le k \le n. \tag{38}$$

With $2n = 100$, we plot the function $P(T_{100} = 2k)$ for $0 \le k \le 50$ below. Notice that it is maximised at the ends of the interval.



The arcsine law*

Use of Stirling's approximation in formula (38) gives

$$\boxed{P(T_+(2n) = 2k) \sim \frac{1/\pi}{\sqrt{k(n-k)}}}, \qquad \text{as } k \to \infty \text{ and } n-k \to \infty.$$

Now consider the calculation of the following probability:

$$\begin{aligned} P\left(\frac{T_+(2n)}{2n} \in [a,b]\right) &= \sum_{na \le k \le nb} P\left(\frac{T_+(2n)}{2n} = \frac{k}{n}\right) \\ &\sim \sum_{a \le k/n \le b} \frac{1/\pi}{\sqrt{k(n-k)}} = \sum_{a \le k/n \le b} \frac{1/\pi}{\sqrt{\frac{k}{n}(n-\frac{k}{n})}} \frac{1}{n}. \end{aligned}$$

This is easily seen to have a limit as $n \to \infty$, the limit being

$$\int_a^b \frac{1/\pi}{\sqrt{t(1-t)}} dt,$$

because the last sum is simply a Riemann approximation to the last integral. This shows the celebrated ARCSINE LAW:

$$\lim_{n\to\infty} P\left(\frac{T_+(2n)}{2n} \le x\right) = \int_0^x \frac{1/\pi}{\sqrt{t(1-t)}} dt = \frac{2}{\pi} \arcsin \sqrt{x}.$$

We thus have that the limiting distribution of $T_+(2n)$ (the fraction of time that the RW is positive), as $n \to \infty$, is the distribution function $\frac{2}{\pi} \arcsin \sqrt{x}$, $0 < x < 1$. This is known as the arcsine law. This distribution function has density $\frac{1/\pi}{\sqrt{x(1-x)}}$, $0 < x < 1$, the plot of which is amazingly close to the plot of the figure above.

Now consider again a SSRW and consider the hitting times $(T_x, x \ge 0)$. This is another random walk because the summands in $T_x = \sum_{y=1}^{x}(T_y - T_{y-1})$ are i.i.d. all distributed as T_1:

$$P(T_1 \ge 2n) = P(S_{2n} = 0) = \binom{2n}{n} 2^{-2n} \sim \frac{1}{\sqrt{\pi n}}.$$

From this, we immediately get that the expectation of T_1 is infinity:

$$ET_1 = \infty.$$

Many other quantities can be approximated in a similar manner.

34 Simple symmetric random walk in two dimensions

Consider a drunkard in a city whose roads are arranged in a rectangular grid. The corners are described by coordinates (x, y) where $x, y \in \mathbb{Z}$. He starts at $(0, 0)$ and, being totally drunk, he moves east, west, north or south, with equal probability (1/4). When he reaches a corner he moves again in the same manner, going at random east, west, north or south.

We describe this motion as a random walk in $\mathbb{Z}^2 = \mathbb{Z}\times\mathbb{Z}$: Let $\xi_1, \xi_2, \ldots$ i.i.d. random vectors such that

$$P(\xi_n = (0,1)) = P(\xi_n = (-1,0)) = P(\xi_n = (1,0)) = P(\xi_n = (-1,0)) = 1/4.$$

We then let

$$S_0 = (0,0), \quad S_n = \sum_{i=1}^{n} \xi_i, \quad n \geq 1.$$

If $x \in \mathbb{Z}^2$, we let x^1 be its horizontal coordinate and x^2 its vertical. So $S_n = (S_n^1, S_n^2) = \left(\sum_{i=1}^n \xi_i^1, \sum_{i=1}^n \xi_i^2\right)$. We have

Lemma 20. *$(S_n^1, n \in \mathbb{Z}_+)$ is a random walk. $(S_n^2, n \in \mathbb{Z}_+)$ is also a random walk. The two random walks are not independent.*

Proof. Since $\xi_1, \xi_2, \ldots$ are independent, so are $\xi_1^1, \xi_2^1, \ldots$, the latter being functions of the former. They are also identically distributed with

$$\begin{aligned}
&P(\xi_n^1 = 0) = P(\xi_n = (0,1)) + P(\xi_n = (0,-1)) = 1/2,\\
&P(\xi_n^1 = 1) = P(\xi_n = (1,0)) = 1/4\\
&P(\xi_n^1 = -1) = P(\xi_n = (-1,0)) = 1/4.
\end{aligned}$$

Hence $(S_n^1, n \in \mathbb{Z}_+)$ is a sum of i.i.d. random variables and hence a random walk. It is not a simple random walk because there is a possibility that the increment takes the value 0.

Same argument applies to $(S_n^2, n \in \mathbb{Z}_+)$ showing that it, too, is a random walk.

However, the two stochastic processes are not independent. Indeed, for each n, the random variables ξ_n^1, ξ_n^2 are not independent simply because if one takes value 0 the other is nonzero. □

Consider a change of coordinates: rotate the axes by 45^o (and scale them), i.e. map (x^1, x^2) into $(x^1 + x^2, x^1 - x^2)$. So let

$$R_n = (R_n^1, R_n^2) := (S_n^1 + S_n^2, S_n^1 - S_n^2).$$

We have:

Lemma 21. *$(R_n, n \in \mathbb{Z}_+)$ is a random walk in two dimensions. $(R^1_n, n \in \mathbb{Z}_+)$ is a random walk in one dimension. $(R^2_n, n \in \mathbb{Z}_+)$ is a random walk in one dimension. The last two are independent.*

Proof. We have $R^1_n = \sum_{i=1}^n (\xi^1_i + \xi^2_i)$, $R^2_n = \sum_{i=1}^n (\xi^1_i - \xi^2_i)$. So $R_n = \sum_{i=1}^n \eta_n$ where η_n is the vector $\eta_n = (\xi^1_n + \xi^2_n, \xi^1_n - \xi^2_n)$. The sequence of vectors $\eta_1, \eta_2, \ldots$ is i.i.d., being obtained by taking the same function on the elements of the sequence $\xi_1, \xi_2, \ldots$ Hence $(R_n, n \in \mathbb{Z}_+)$ is a random walk in two dimensions. Similar argument shows that each of $(R^1_n, n \in \mathbb{Z}_+)$, $(R^2_n, n \in \mathbb{Z}_+)$ is a random walk in one dimension. To show that the last two are independent, we check that $\eta^1_n = \xi^1_n + \xi^2_n$, $\eta^2_n = \xi^1_n - \xi^2_n$ are independent for each n. The possible values of each of the η^1_n, η^2_n are ±1. We have

$$\begin{aligned}
P(\eta^1_n = 1, \eta^2_n = 1) &= P(\xi_n = (1,0)) = 1/4 \\
P(\eta^1_n = -1, \eta^2_n = -1) &= P(\xi_n = (-1,0)) = 1/4 \\
P(\eta^1_n = 1, \eta^2_n = -1) &= P(\xi_n = (0,1)) = 1/4 \\
P(\eta^1_n = -1, \eta^2_n = 1) &= P(\xi_n = (0,=1)) = 1/4.
\end{aligned}$$

Hence $P(\eta^1_n = 1) = P(\eta^1_n = -1) = 1/2$, Hence $P(\eta^2_n = 1) = P(\eta^2_n = -1) = 1/2$. And so $P(\eta^1_n = \varepsilon, \eta^2_n = \varepsilon') = P(\eta^1_n = \varepsilon)P(\eta^2_n = \varepsilon')$ for all choices of $\varepsilon, \varepsilon' \in \{-1, +1\}$. □

Lemma 22.

$$P(S_{2n} = 0) = \binom{2n}{n}^2 2^{-4n}.$$

Proof.

$$P(S_{2n} = 0) = P(R_{2n} = 0) = P(R^1_{2n} = 0, R^2_{2n} = 0) = P(R^1_{2n} = 0)(R^2_{2n} = 0),$$

where the last equality follows from the independence of the two random walks. But (R^1_n) is a simple symmetric random walk. So $P(R^1_{2n} = 0) = \binom{2n}{n} 2^{-2n}$. The same is true for (R^2_n).

Corollary 24. *The two-dimensional simple symmetric random walk is recurrent.*

Proof. By Lemma 5, we need to show that $\sum_{n=0}^\infty P(S_n = 0) = \infty$. But by the previous lemma, and by Stirling's approximation, $P(S_{2n} = 0) \sim \frac{c}{n}$, in the sense that the ratio of the two sides goes to 1, as $n \to \infty$, where c is a positive constant. □

35 Skorokhod embedding*

We take a closer look at the gambler's ruin problem for a simple symmetric random walk in one dimension. Let $a < 0 < b$. Recall, from Theorem 7, that for a simple symmetric random walk (started from $S_0 = 0$),

$$P(T_b < T_a) = \frac{-a}{b-a}, \quad P(T_a < T_b) = \frac{b}{b-a}.$$

We can read this as follows:

$$P(S_{T_a \wedge T_b} = b) = \frac{-a}{b-a}, \quad P(S_{T_a \wedge T_b} = a) = \frac{b}{b-a}.$$

This last display tells us the distribution of $S_{T_a \wedge T_b}$. Note, in particular, that $ES_{T_a \wedge T_b} = 0$.

Conversely, given a 2-point probability distribution $(p, 1-p)$, such that p is rational, we can always find integers a, b, $a < 0 < b$ such that $pa + (1-p)b = 0$. Indeed, if $p = M/N$, $M < N$, $M, N \in \mathbb{N}$, choose, e.g. $a = M - N$, $b = N$. This means that we can *simulate* the distribution by generating a simple symmetric random walk starting from 0 and watching it till the first time it exits the interval $[a, b]$. The probability it exits it from the left point is p, and the probability it exits from the right point is $1 - p$.

How about more complicated distributions? Let suppose we are given a probability distribution $(p_i, i \in \mathbb{Z})$. Can we find a stopping time T such that S_T has the given distribution? The SKOROKHOD EMBEDDING theorem tells us how to do this.

Theorem 43. *Let (S_n) be a simple symmetric random walk and $(p_i, i \in \mathbb{Z})$ a given probability distribution on $\mathbb{Z}$ with zero mean: $\sum_{i\in\mathbb{Z}} ip_i = 0$. Then there exists a stopping time T such that*

$$P(S_T = i) = p_i, \quad i \in \mathbb{Z}.$$

Proof. Define a pair of random variables (A, B) taking values in $\{(0,0)\} \cup (\mathbb{N} \times \mathbb{N})$ with the following distribution:

$$\begin{aligned} P(A = i, B = j) &= c^{-1}(j - i)p_i p_j, \quad i < 0 < j, \\ P(A = 0, B = 0) &= p_0, \end{aligned}$$

where c^{-1} is chosen by normalisation:

$$P(A = 0, B = 0) + \sum_{i<0} \sum_{j>0} P(A = i, B = j) = 1.$$

This leads to:

$$c^{-1} \sum_{j>0} jp_j \sum_{i<0} p_i + c^{-1} \sum_{i<0} (-i)p_i \sum_{j>0} p_j = 1 - p_0$$

Since $\sum_{i\in\mathbb{Z}} ip_i = 0$, we have $\sum_{i<0}(-i)p_i = \sum_{i>0} ip_i$ and, using this, we get that c is precisely this common value:

$$c = \sum_{i<0}(-i)p_i = \sum_{i>0} ip_i.$$

Letting $T_i = \inf\{n \geq 0 : S_n = i\}$, $i \in \mathbb{Z}$, and assuming that (A, B) are independent of (S_n), we consider the random variable

$$Z = S_{T_A \wedge T_B}.$$

The claim is that

$$P(Z = k) = p_k, \quad k \in \mathbb{Z}.$$

To see this, suppose first $k = 0$. Then $Z = 0$ if and only if $A = B = 0$ and this has probability p_0, by definition. Suppose next $k > 0$. We have

$$\begin{aligned} P(Z = k) &= P(S_{T_A \wedge T_B} = k) \\ &= \sum_{i<0} \sum_{j>0} P(S_{T_i \wedge T_j} = k) P(A = i, B = j) \\ &= \sum_{i<0} P(S_{T_i \wedge T_k} = k) P(A = i, B = k) \\ &= \sum_{i<0} \frac{-i}{k-i}\ c^{-1}(k-i) p_i p_k \\ &= c^{-1} \sum_{i<0} (-i) p_i\ p_k = p_k. \end{aligned}$$

Similarly, $P(Z = k) = p_k$ for $k < 0$. □

PART III: APPENDICES

This section contains some *sine qua non* elements from standard Mathematics, as taught in high schools or universities. They are not supposed to replace textbooks or basic courses; they merely serve as reminders.

Arithmetic

Arithmetic deals with integers. The set $\mathbb{N}$ of natural numbers is the set of positive integers:

$$\mathbb{N} = \{1, 2, 3, \ldots\}.$$

The set $\mathbb{Z}$ of integers consists of $\mathbb{N}$, their negatives and zero. Zero is denoted by 0.

$$\mathbb{Z} = \{\cdots, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6, \cdots\}.$$

Given integers a, b, with $b \neq 0$, we say that b divides a if there is an integer k such that $a = kb$. We write this by $b \mid a$. If $c \mid b$ and $b \mid a$ then $c \mid a$. If $a \mid b$ and $b \mid a$ then $a = b$.

Now, if a, b are arbitrary integers, with at least one of them nonzero, we can define their GREATEST COMMON DIVISOR $d = \gcd(a, b)$ as the unique positive integer such that

(i) $d \mid a$, $d \mid b$;
(ii) if $c \mid a$ and $c \mid b$, then $d \mid c$.

(That it is unique is obvious from the fact that if d_1, d_2 are two such numbers then d_1 would divide d_2 and vice-versa, leading to $d_1 = d_2$.) Notice that:

$$\gcd(a, b) = \gcd(a, b - a).$$

Indeed, let $d = \gcd(a, b)$. We will show that $d = \gcd(a, b - a)$. But $d \mid a$ and $d \mid b$. Therefore $d \mid b - a$. So (i) holds. Now suppose that $c \mid a$ and $c \mid b - a$. Then $c \mid a + (b - a)$, i.e. $c \mid b$. Because $c \mid a$ and $c \mid b$, and $d = \gcd(a, b)$, we have $c \mid d$. So (ii) holds, and we're done. This gives us an algorithm for finding the gcd between two numbers: Suppose $0 < a < b$ are integers. Replace b by $b - a$. If $b - a > a$, replace $b - a$ by $b - 2a$. Keep doing this until you obtain $b - qa < a$. Then interchange the roles of the numbers and continue. For instance,

$$\begin{gathered}\gcd(120, 38) = \gcd(82, 38) = \gcd(44, 38) = \gcd(6, 38) \\ = \gcd(6, 32) = \gcd(6, 26) = \gcd(6, 20) = \gcd(6, 14) = \gcd(6, 8) = \gcd(6, 2) \\ = \gcd(4, 2) = \gcd(2, 2) = 2.\end{gathered}$$

But in the first line we subtracted 38 exactly 3 times. But 3 is the maximum number of times that 38 "fits" into 120. Indeed $4 \times 38 > 120$. In other words, we may summarise the first line by saying that we replace the largest of the two numbers by the remainder of the division of the largest with the smallest. Similarly, with the second line. So we work faster, if we also invoke Euclid's algorithm (i.e. the process of long division that one learns in elementary school).

The theorem of division says the following: Given two positive integers b, a, with $a > b$, we can find an integer q and an integer $r \in \{0, 1, \ldots, a-1\}$, such that

$$a = qb + r.$$

Its proof is obvious. The long division algorithm is a process for finding the quotient q and the remainder r. For example, to divide 3450 with 26 we do

$$\begin{array}{r|l}
 & \underline{132} \\
26 & 3450 \\
 & \underline{26} \\
 & \;\;85 \\
 & \;\;\underline{78} \\
 & \;\;\;\;70 \\
 & \;\;\;\;\underline{52} \\
 & \;\;\;\;18
\end{array}$$

and obtain $q = 26$, $r = 18$.

Here are some other facts about the greatest common divisor:

<u>First</u>, we have the fact that if $d = \gcd(a, b)$ then there are integers x, y such that

$$d = xa + yb.$$

To see why this is true, let us follow the previous example again:

$$\begin{aligned}
\gcd(120, 38) &= \gcd(6, 38) \\
&= \gcd(6, 2) \\
&= 2.
\end{aligned}$$

In the process, we used the divisions

$$\begin{aligned}
120 &= 3 \times 38 + 6 \\
38 &= 6 \times 6 + 2.
\end{aligned}$$

Working backwards, we have

$$\begin{aligned}
2 &= 38 - 6 \times 6 \\
&= 38 - 6 \times (120 - 3 \times 38) \\
&= 19 \times 38 - 6 \times 120.
\end{aligned}$$

Thus, $\gcd(38, 120) = 38x + 120y$, with $x = 19$, $y = 6$. The same logic works in general.

<u>Second</u>, we can show that, for integers $a, b \in \mathbb{Z}$, not both zero, we have

$$\gcd(a, b) = \min\{xa + yb : \; x, y \in \mathbb{Z}, \; xa + yb > 0\}.$$

To see this, let d be the right hand side. Then $d = xa + yb$, for some $x, y \in \mathbb{Z}$. Suppose that $c \mid a$ and $c \mid b$. Then c divides any linear combination of a, b; in particular, it divides d. This shows (ii) of the definition of a greatest common divisor. We need to show (i), i.e.

that $d \mid a$ and $d \mid b$. Suppose this is not true, i.e. that d does not divide both numbers, say, e.g., that d does not divide b. Since $d > 0$, we can divide b by d to find

$$b = qd + r,$$

for some $1 \le r \le d - 1$. But then

$$b = q(xa + yb) + r,$$

yielding

$$r = (-qx)a + (1 - qy)b.$$

So r is expressible as a linear combination of a and b and is strictly smaller than d (being a remainder). So d is not minimum as claimed; and this is a contradiction; so $r = 0$; so d divides both a and b.

The greatest common divisor of 3 integers can now be defined in a similar manner. Finally, the greatest common divisor of any subset of the integers also makes sense.

Relations

A RELATION on a set A is a mathematical notion that establishes a way of pairing elements of A. For example, the relation $<$ between real numbers can be thought of as consisting of pairs of numbers (x, y) such that x is smaller than y. In general, a relation can be thought of as a set of pairs of elements of the set A. A relation is called REFLEXIVE if it contains pairs (x, x). The relation $<$ is not reflexive. The equality relation $=$ is reflexive. A relation is called SYMMETRIC if it cannot contain a pair (x, y) without contain its symmetric (y, x). The equality relation is symmetric. The relation $<$ is not symmetric. A relation is TRANSITIVE if when it contains (x, y) and (y, z) it also contains (x, z). Both $<$ and $=$ are transitive. If we take A to be the set of students in a classroom, then the relation "x is a friend of y" is symmetric, it can be thought as reflexive (unless we do not want to accept the fact that one may not be afriend of oneself), but is not necessarily transitive. A relation which possesses all the three properties is called EQUIVALENCE RELATION. Any equivalence relation splits a set into equivalence classes: the equivalence class of x contains all y that are related (are equivalent) to x.

We encountered the equivalence relation $i \leftrightsquigarrow j$ (i communicates with j) in Markov chain theory.

Functions

We a function f from a set X into a set Y is denoted by $f : X \to Y$. The range of f is the set of its values, i.e. the set $\{f(x) : x \in X\}$. A function is called one-to-one (or injective) if two different $x_1, x_2 \in X$ have different values $f(x_1), f(x_2)$. A function is called onto (or surjective) if its range is Y. If $B \subset Y$ then $f^{-1}(B)$ consists of all $x \in X$ with $f(x) \in B$. The set of functions from X to Y is denoted by Y^X.

Counting

Let A, B be finite sets with n, m elements, respectively.

The number of functions from A to B (i.e. the cardinality of the set B^A) is m^n. Indeed, we may think of the elements of A as animals and those of B as names. A function is then an assignment of a name to each animal. (The same name may be used for different animals.) So the first animal can be assigned any of the m available names; so does the second; and the third, and so on. Hence the total number of assignments is

$$\underbrace{m \times m \times \cdots \times m}_{n \text{ times}}$$

Suppose now we are not allowed to give the same name to different animals. To be able to do this, we need more names than animals: $n \leq m$. Each assignment of this type is a one-to-one function from A to B. Since the name of the first animal can be chosen in m ways, that of the second in $m-1$ ways, and so on, the total number of one-to-one functions from A to B is

$$(m)_n := m(m-1)(m-2)\cdots(m-n+1).$$

In the special case $m = n$, we denote $(n)_n$ also by $n!$; in other words, $n!$ stands for the product of the first n natural numbers:

$$n! = 1 \times 2 \times 3 \times \cdots \times n.$$

Incidentally, a one-to-one function from the set A into A is called a PERMUTATION of A. Thus, $n!$ is the total number of permutations of a set of n elements.

Any set A contains several subsets. If A has n elements, the number of subsets with k elements each can be found by designating which of the n elements of A will be chosen. The first element can be chosen in n ways, the second in $n-1$, and so on, the k-th in $n-k+1$ ways. So we have $(n)_k = n(n-1)(n-2)\cdots(n-k+1)$ ways to pick k elements if we care about their order. If we don't then we divide by the number $k!$ of permutations of the k elements concluding that the number of subsets with k elements of a set with n elements equals

$$\frac{(n)_k}{k!} =: \binom{n}{k},$$

where the symbol $\binom{n}{k}$ is merely a name for the number computed on the left side of this equation. Clearly,

$$\binom{n}{k} = \binom{n}{n-k} = \frac{n!}{k!(n-k)!}.$$

Since there is only one subset with no elements (it is the empty set $\varnothing$), we have

$$\binom{n}{0} = 1.$$

The total number of subsets of A (regardless of their number of elements) is thus equal to

$$\sum_{k=0}^{n} \binom{n}{k} = (1+1)^n = 2^n,$$

and this follows from the binomial theorem. We thus observe that there are as many subsets in A as number of elements in the set $\{0,1\}^n$ of sequences of length n with elements 0 or 1.

We shall let $\binom{n}{k} = 0$ if $n < k$.
Also, by convention, if x is a real number, we define

$$\binom{x}{m} = \frac{x(x-1)\cdots(x-m+1)}{m!}.$$

If y is not an integer then, by convention, we let $\binom{n}{y} = 0$.

The Pascal triangle relation states that

$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}.$$

Indeed, in choosing k animals from a set of $n+1$ ones, consider a specific animal, the pig, say. If the pig is not included in the sample, we choose k animals from the remaining n ones and this can be done in $\binom{n}{k}$ ways. If the pig is included, then there are $k-1$ remaining animals to be chosen and this can be done in $\binom{n}{k-1}$ ways. Since the sentences "the pig is included" and "the pig is not included" cannot be simultaneously true, we add the two numbers together in order to find the total number of ways to pick k animals from $n+1$ animals.

Maximum and minimum

The MINIMUM of two numbers a, b is denoted by

$$a \wedge b = \min(a, b).$$

Thus $a \wedge b = a$ if $a \le b$ or $= b$ if $b \le a$.

The MAXIMUM is denoted by

$$a \vee b = \max(a, b).$$

Note that $-(a \vee b) = (-a) \wedge (-b)$ and $(a \wedge b) + c = (a+c) \wedge (b+c)$.

In particular, we use the following notations:

$$a^+ = \max(a, 0), \quad a^- = -\min(a, 0).$$

Both quantities are nonnegative; a^+ is called the positive part of a; and a^- is called the negative part of a. Note that it is always true that

$$a = a^+ - a^-.$$

The absolute value of a is defined by

$$|a| = a^+ + a^-.$$

The notation extends to more than one numbers. For example,

$$a \vee b \vee c$$

refers to the maximum of the three numbers a, b, c, and we don't care to put parantheses because the order of finding a maximum is irrelevant.

Indicator function

$\mathbf{1}_A$ stands for the INDICATOR function of a set A, meaning a function that takes value 1 on A and 0 on the complement A^c of A. We also use the notation $\mathbf{1}(\text{clause})$ for a number that is 1 whenever the clause is true or 0 otherwise. This is very useful notation because conjunction of clauses is transformed into multiplication. It is easy to see that

$$\mathbf{1}_{A\cap B} = \mathbf{1}_A \mathbf{1}_B, \quad 1 - \mathbf{1}_A = \mathbf{1}_{A^c}.$$

Several quantities of interest are expressed in terms of indicators. For instance, if $A_1, A_2, \ldots$ are sets or clauses then the number $\sum_{n\geq 1} \mathbf{1}_{A_n}$ is the number true clauses.

Thus, if I go to a bank N times and A_i is the clause "I am robbed the i-th time" then $\sum_{n=1}^{N} \mathbf{1}_{A_i}$ is the total number of times I am robbed.

If A is an event then $\mathbf{1}_A$ is a random variable. It takes value 1 with probability $P(A)$ and 0 with probability $P(A^c)$. Therefore its expectation is

$$E\mathbf{1}_A = 1 \times P(A) + 0 \times P(A^c) = P(A).$$

It is worth remembering that.

An example of its application: Since

$$\begin{aligned} \mathbf{1}_{A\cup B} &= 1 - \mathbf{1}_{(A\cup B)^c} = 1 - \mathbf{1}_{A^c \cap B^c} = 1 - \mathbf{1}_{A^c}\mathbf{1}_{B^c} \\ &= 1 - (1-\mathbf{1}_A)(1-\mathbf{1}_B) = \mathbf{1}_A + \mathbf{1}_B - \mathbf{1}_A\mathbf{1}_B, = \mathbf{1}_A + \mathbf{1}_B - \mathbf{1}_{A\cap B}, \end{aligned}$$

we have

$$E[\mathbf{1}_{A\cup B}] = E[\mathbf{1}_A] + E[\mathbf{1}_B] - E[\mathbf{1}_{A\cap B}],$$

which means

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Another example: Let X be a nonnegative random variable with values in $\mathbb{N} = \{1, 2, \ldots\}$. Then

$$X = \sum_{n=1}^{X} 1 \text{ (the sum of } X \text{ ones is } X).$$

So

$$X = \sum_{n\geq 1} \mathbf{1}(X \geq n).$$

So

$$E[X] = \sum_{n\geq 1} E\mathbf{1}(X \geq n) = \sum_{n\geq 1} P(X \geq n).$$

Matrices

Let $\mathbb{R}^d$ be the set of vectors with d components. A linear function φ from $\mathbb{R}^n$ to $\mathbb{R}^m$ is such that

$$\varphi(\alpha x + \beta y) = \alpha\varphi(x) + \beta\varphi(y),$$

for all $x, y \in \mathbb{R}^n$, and all $\alpha, \beta \in \mathbb{R}$. Let $u_1, \dots, u_n$ be a basis for $\mathbb{R}^n$. Then any $x \in \mathbb{R}^n$ can be *uniquely* written as

$$x = \sum_{j=1}^{n} x^j u_j,$$

where $x^1, \dots, x^n \in \mathbb{R}$. The column $\begin{pmatrix} x^1 \\ \vdots \\ x^n \end{pmatrix}$ is a column representation of x with respect to the given basis. Then, because φ is linear,

$$y := \varphi(x) = \sum_{j=1}^{n} x^j \varphi(u_j).$$

But the $\varphi(u_j)$ are vectors in $\mathbb{R}^m$. So if we choose a basis $v_1, \dots, v_m$ in $\mathbb{R}^n$ we can express each $\varphi(u_j)$ as a linear combination of the basis vectors:

$$\varphi(u_j) = \sum_{i=1}^{m} a_{ij} v_i.$$

Combining, we have

$$\varphi(x) = \sum_{i=1}^{m} v_i \sum_{j=1}^{n} a_{ij} x^j.$$

Let

$$y^i := \sum_{j=1}^{n} a_{ij} x^j, \quad i = 1, \dots, m.$$

The column $\begin{pmatrix} y^1 \\ \vdots \\ y^m \end{pmatrix}$ is a column representation of $y := \varphi(x)$ with respect to the given basis $v_1, \dots, v_m$. The matrix A of φ with respect to the choice of the two basis is defined by

$$A := \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

It is a $m \times n$ matrix. The relation between the column representation of y and the column representation of x is written as

$$\begin{pmatrix} y^1 \\ \vdots \\ y^m \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x^1 \\ \vdots \\ x^n \end{pmatrix}$$

Suppose that ψ, φ are linear functions:

$$\mathbb{R}^n \xrightarrow{\psi} \mathbb{R}^m \xrightarrow{\varphi} \mathbb{R}^\ell$$

Then $\varphi \circ \psi$ is a linear function:

$$\mathbb{R}^n \xrightarrow{\varphi \circ \psi} \mathbb{R}^\ell$$

Fix three sets of basis on $\mathbb{R}^n$, $\mathbb{R}^m$ and $\mathbb{R}^\ell$ and let A, B be the matrix representations of φ, ψ, respectively, with respect to these bases. Then the matrix representation of $\varphi \circ \psi$ is AB, where

$$(AB)_{ij} = \sum_{k=1}^{m} A_{ik} B_{kj}, \quad 1 \le i \le \ell, \quad 1 \le j \le n.$$

So A is a $\ell \times m$ matrix, B is a $m \times n$ matrix, and AB is a $\ell \times n$ matrix.

A matrix A is called square if it is a $n \times n$ matrix. If we fix a basis in $\mathbb{R}^n$, a square matrix represents a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^n$, i.e. it is its matrix representation with respect to the fixed basis.

Usually, the choice of the basis is the so-called standard basis. The standard basis in $\mathbb{R}^d$ consists of the vectors

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad e_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \dots, \quad e_d = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}.$$

In other words, $e^i_j = \mathbf{1}(i = j)$.

Supremum and infimum

When we have a set of numbers I (for instance, a sequence $a_1, a_2, \dots$ of numbers) then the minimum of I is denoted by $\min I$ and refers to a number $c \in I$ such that $c \le x$ for all the numbers $x \in I$. Similarly, we define $\max I$. This minimum (or maximum) may not exist. For instance $I = \{1, 1/2, 1/3, \dots\}$ has no minimum. In these cases we use the notation $\inf I$, standing for the "INFIMUM" of I, defined as the maximum of all the numbers y with the property $y \le x$ for all $x \in I$. A property of real numbers says that if the set I is bounded from below then this maximum of all lower bounds exists, and it is this that we called infimum. Likewise, we define the SUPREMUM $\sup I$ to be the minimum of all upper bounds. If I has no lower bound then $\inf I = -\infty$. If I has no upper bound then $\sup I = +\infty$. If I is empty then $\inf \emptyset = +\infty$, $\sup \emptyset = -\infty$.

$\inf I$ is characterised by: $\inf I \le x$ for all $x \in I$ and, for any $\varepsilon > 0$, there is $x \in I$ with $x \le \varepsilon + \inf I$. $\sup I$ is characterised by: $\sup I \ge x$ for all $x \in I$ and, for any $\varepsilon > 0$, there is $x \in I$ with $x \ge \varepsilon + \inf I$.

Limsup and liminf

If $x_1, x_2, \dots$ is a sequence of numbers then $\inf\{x_1, x_2, \dots\} \le \inf\{x_2, x_3, \dots\} \le \inf\{x_3, x_4, \dots\} \le \cdots$ and, since any increasing sequence has a limit (possibly $+\infty$) we take this limit and call it $\limsup x_n$ (limit superior) of the sequence. So $\limsup x_n = \lim_{n\to\infty} \inf\{x_n, x_{n+1}, x_{n+2}, \cdots\}$. Similarly, we define $\limsup x_n$. We note that $\liminf(-x_n) = -\limsup x_n$. The sequence x_n has a limit ℓ if and only if $\limsup x_n = \liminf x_n = \ell$.

Probability Theory

I do not require[11] knowledge of axiomatic Probability Theory (with Measure Theory). But I would like the reader to keep in mind one thing: Every mathematical system has a set of axioms from which we can prove (useful) theorems, derive formulae, and apply them. Probability Theory is exceptionally simple in terms of its axioms, for (besides $P(\Omega) = 1$) there is essentially ONE AXIOM , namely:

$$\text{If } A_1, A_2, \ldots \text{ are mutually disjoint events, then } P\left(\bigcup_{n\geq 1} A_n\right) = \sum_{n\geq 1} P(A_n).$$

Everything else, in these notes, and everywhere else, follows from this axiom. That is why Probability Theory is very rich. In contrast, Linear Algebra that has many axioms is more 'restrictive'. (That is why Probability Theory is very rich. In contrast, Linear Algebra that has many axioms is more 'restrictive'.)

If A, B are disjoint events then $P(A \cup B) = P(A) + P(B)$. This can be generalised to any finite number of events. If the events A, B are not disjoint then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. This is the inclusion-exclusion formula for two events.

If A is an event then $\mathbf{1}_A$ or $\mathbf{1}(A)$ is the indicator random variable, i.e. the function that takes value 1 on A and 0 on A^c. Therefore, $E\mathbf{1}_A = P(A)$. Note that $\mathbf{1}_A \mathbf{1}_B = \mathbf{1}_{A\cap B}$, and $\mathbf{1}_{A^c} = 1 - \mathbf{1}_A$.

If A_n are events with $P(A_n) = 0$ then $P(\cup_n A_n) = 0$. Indeed, $P(\cup_n A_n) \leq \sum_n P(A_n)$. Similarly, if A_n are events with $P(A_n) = 1$ then $P(\cap_n A_n) = 1$.

Usually, showing that $P(A) \leq P(B)$ is more of a logical problem than a numerical one: we need to show that A implies B (i.e. A is a subset of B). Similarly, to show that $EX \leq EY$, we often need to argue that $X \leq Y$. For example, Markov's inequality for a positive random variable X states that $P(X > x) \leq EX/x$. This follows from the logical statement $X\mathbf{1}(X > x) \leq X$, which holds since X is positive.

In Markov chain theory, we work a lot with conditional probabilities. Recall that $P(B|A)$ is defined by $P(A \cap B)/P(A)$ and is a very motivated definition. Often, to compute $P(A \cap B)$ we read the definition as $P(A \cap B) = P(A)P(B|A)$. This is generalisable for a number of events:

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)$$

If we interchange the roles of A, B in the definition, we get $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$ and this is called Bayes' theorem.

Quite often, to compute $P(A)$, it is useful to find disjoint events $B_1, B_2, \ldots$ such that $\cup_n B_n = \Omega$, in which case we write

$$P(A) = \sum_n P(A \cap B_n) = \sum_n P(A|B_n)P(B_n).$$

If the events $A_1, A_2, \ldots$ are increasing with $A = \cup_n A_n$ then $P(A) = \lim_{n\to\infty} P(A_n)$. Similarly, if he events $A_1, A_2, \ldots$ are decreasing with $A = \cap_n A_n$ then $P(A) = \lim_{n\to\infty} P(A_n)$.

[11] Of course, by not requiring this, I do sometimes "cheat" from a strict mathematical point of view, but not too much...

An application of this is:

$$P(X < \infty) = \lim_{x \to \infty} P(X \leq x).$$

In Markov chain theory, we encounter situations where $P(X = \infty)$ may be positive, so it is useful to keep this in mind.

The expectation of a nonnegative random variable X may be defined by

$$EX = \int_0^\infty P(X > x) dx.$$

The formula holds for *any* positive random variable. For a random variable which takes positive and negative values, we define $X^+ = \max(X, 0)$, $X^- = -\min(X, 0)$, which are both nonnegative and then define

$$EX = EX^+ - EX^-,$$

which is motivated from the algebraic identity $X = X^+ - X^-$. This definition makes sense only if $EX^+ < \infty$ or if $EX^- < \infty$. In case the random variable is discrete, the formula reduces to the known one, viz., $EX = \sum_x xP(X = x)$. In case the random variable is absoultely continuous with density f, we have $EX = \int_{-\infty}^\infty xf(x)dx$. For a random variable that takes values in $\mathbb{N}$ we can write $EX = \sum_{n=1}^\infty P(X \geq n)$. This is due to the fact that $P(X \geq n) = E\mathbf{1}(X \geq n)$.

Generating functions

A GENERATING FUNCTION of a sequence $a_0, a_1, \ldots$ is an infinite polynomial having the a_i as coefficients:

$$G(s) = a_0 + a_1 s + a_2 s^2 + \cdots .$$

We do not a priori know that this exists for any s other than, obviously, for $s = 0$ for which $G(0) = 0$. If it exists for some s_1 then it exists uniformly for all s with $|s| < |s_1|$ (and this needs a proof, taught in calculus courses).

As an example, the generating function of the geometric sequence $a_n = \rho^n, n \geq 0$ is

$$\sum_{n=0}^\infty \rho^n s^n = \frac{1}{1 - \rho s}, \tag{39}$$

as long as $|\rho s| < 1$, i.e. $|s| < 1/|\rho|$.

Note that $|G(s)| \leq \sum_n |a_n||s|^n$ and this is $\leq \sum_n |a_n|$ if $s \in [-1, 1]$. So if $\sum_n a_n < \infty$ then $G(s)$ exists for all $s \in [-1, 1]$, making it a useful object when the $a_0, a_1, \ldots$ is a probability distribution on $Z_+ = \{0, 1, 2, \ldots\}$ because then $\sum_n a_n = 1$.

Suppose now that X is a random variable with values in $\mathbb{Z}_+ \cup \{+\infty\}$. The generating function of the sequence

$$p_n = P(X = n), \quad n = 0, 1, \ldots$$

is called PROBABILITY GENERATING FUNCTION of X:

$$\varphi(s) = Es^X = \sum_{n=0}^\infty s^n P(X = n)$$

This is defined for all $-1 < s < 1$. Note that

$$\sum_{n=0}^{\infty} p_n \leq 1,$$

and

$$p_\infty := P(X = \infty) = 1 - \sum_{n=0}^{\infty} p_n.$$

We do allow X to, possibly, take the value $+\infty$. Since the term $s^\infty p_\infty$ could formally appear in the sum defining $\varphi(s)$, but, since $|s| < 1$, we have $s^\infty = 0$.

Note that $\varphi(0) = P(X = 0)$, while

$$\varphi(1) = \sum_{n=0}^{\infty} P(X = n) = P(X < +\infty),$$

We can recover the probabilities p_n from the formula

$$p_n = \frac{\varphi^{(n)}(0)}{n!},$$

as follows from Taylor's theorem, and this is why φ is called *probability* generating function. Also, we can recover the expectation of X from

$$EX = \lim_{s \to 1} \varphi'(s),$$

something that can be formally seen from

$$\frac{d}{ds} Es^X = E \frac{d}{ds} s^X = EXs^{X-1},$$

and by letting $s = 1$.

Generating functions are useful for solving linear recursions. As an example, consider the FIBONACCI SEQUENCE defined by the recursion

$$x_{n+2} = x_{n+1} + x_n, \quad n \geq 0, \text{ with } x_0 = x_1 = 1.$$

If we let

$$X(s) = \sum_{n \geq 0} x_n s^n$$

be the generating of $(x_n, n \geq 0)$ then the generating function of $(x_{n+1}, n \geq 0)$ is

$$\sum_{n \geq 0} x_{n+1} s^n = s^{-1}(X(s) - x_0)$$

and the generating function of $(x_{n+2}, n \geq 0)$ is

$$\sum_{n \geq 0} x_{n+2} s^n = s^{-2}(X(s) - x_0 - sx_1).$$

From the recursion $x_{n+2} = x_{n+1} + x_n$ we have (and this is were linearity is used) that the generating function of $(x_{n+2}, n \geq 0)$ equals the sum of the generating functions of $(x_{n+1}, n \geq 0)$ and $(x_n, n \geq 0)$, namely,

$$s^{-2}(X(s) - x_0 - sx_1) = s^{-1}(X(s) - x_0) + X(s).$$

Since $x_0 = x_1 = 1$, we can solve for $X(s)$ and find

$$X(s) = \frac{-1}{s^2 + s - 1}.$$

But the polynomial $s^2 + s - 1$ has two roots:

$$a = (\sqrt{5} - 1)/2, \quad b = -(\sqrt{5} + 1)/2.$$

Hence

$$s^2 + s - 1 = (s - a)(s - b),$$

and so

$$X(s) = \frac{1}{a-b}\Big(\frac{1}{s-b} - \frac{1}{s-a}\Big) = \frac{1}{b-a}\Big(\frac{1}{b-s} - \frac{1}{a-s}\Big).$$

But (39) tells us that ρ^n has generating function $\frac{1}{1-\rho s} = \frac{1/\rho}{(1/\rho)-s}$. Thus $\frac{1}{(1/\rho)-s}$ is the generating function of ρ^{n+1}. This tells us that $\frac{1}{s-b}$ is the generating function of $(1/b)^{n+1}$ and that $\frac{1}{s-a}$ is the generating function of $(1/a)^{n+1}$. Hence $X(s)$ is the generating function of

$$x_n = \frac{1}{b-a}((1/b)^{n+1} - (1/a)^{n+1}).$$

Noting that $ab = -1$, we can also write this as

$$x_n = \frac{(-a)^{n+1} - (-b)^{n+1}}{b-a}.$$

is the solution to the recursion, i.e. the n-th term of the Fibonacci sequence. Despite the square roots in the formula, this number is always an integer (it has to be!)

Distribution or law of a random object

When we speak of the "DISTRIBUTION" or "LAW" of a random variable X with values in $\mathbb{Z}$ or $\mathbb{R}$ or $\mathbb{R}^n$ or even in a much larger space, we speak of some function which enables us to specify the probabilities of the form

$$P(X \in B)$$

where B is a set of values of X.

So, for example, if X is a random variable with values in $\mathbb{R}$, then the so-called distribution function, i.e. the function

$$P(X \leq x), \quad x \in \mathbb{R},$$

is such an example. But I could very well use the function

$$P(X > x),$$

or the 2-parameter function

$$P(a < X < b).$$

If X is absolutely continuous, I could use its density

$$f(x) = \frac{d}{dx} P(X \le x).$$

All these objects can be referred to as "distribution" or "law" of X. Even the generating function

$$Es^X$$

of a $\mathbb{Z}_+$-valued random variable X is an example, for it does allow us to specify the probabilities $P(X = n)$, $n \in \mathbb{Z}_+$. Warning: the verb "specify" should not be interpreted as "compute" in the sense of existence of an analytical formula or algorithm.

We frequently encounter random objects which belong to larger spaces, for example they could take values which are themselves functions. For instance, we encountered the concept of an excursion of a Markov chain from a state i. Such an excursion is a random object which should be thought of as a random function. Specifying the distribution of such an object may be hard, but using or merely talking about such an object presents no difficulty other than overcoming a psychological barrier.

Large products: Stirling's formula

If we want to compute $n! = 1 \times 2 \times \cdots \times n$ for large n then, since the number is so big, it's better to use logarithms:

$$n! = e^{\log n} = \exp \sum_{k=1}^{n} \log k.$$

Now, to compute a large sum we can, instead, compute an integral and obtain an approximation. Since $\log x$ does not vary too much between two successive positive integers k and $k+1$ it follows that

$$\int_k^{k+1} \log x \; dx \approx \log k.$$

Hence

$$\sum_{k=1}^{n} \log k \approx \int_1^n \log x \; dx.$$

Since $\frac{d}{dx}(x \log x - x) = \log x$, it follows that

$$\int_1^n \log x \; dx = n \log n - n + 1 \approx n \log n - n.$$

Hence we have

$$n! \approx \exp(n \log n - n) = n^n e^{-n}.$$

We call this the rough Stirling approximation. It turns out to be a good one. But there are two problems: (a) We do not know in what sense the symbol $\approx$ should be interpreted. (b) The approximation is not good when we consider ratios of products. For example (try it!) the rough Stirling approximation give that the probability that, in $2n$ fair coin tosses we have exactly n heads equals $\binom{2n}{n}2^{-2n} \approx 1$, and this is terrible, because we know that the probability goes to zero as $n \to \infty$.

To remedy this, we shall prove STIRLING'S APPROXIMATION:

$$\boxed{n! \sim n^n e^{-n}\sqrt{2\pi n}}$$

where the symbol $a_n \sim b_n$ means that $a_n/b_n \to 0$, as $n \to \infty$.

The geometric distribution

We say that the random variable X with values in $\mathbb{Z}_+$ is geometric if it has the MEMORYLESS PROPERTY:

$$P(X - n \ge m \mid X \ge n) = P(X \ge m), \quad 0 \le n \le m.$$

If we think of X as the time of occurrence of a random phenomenon, then X is such that knowledge that X has exceeded n does not make the distribution of the remaining time $X - n$ any different from X.

We have already seen a geometric random variable. That was the random variable $J_0 = \sum_{n=1}^{\infty} \mathbf{1}(S_n = 0)$, defined in (34). It was shown in Theorem 38 that J_0 satisfies the memoryless property. In fact it was shown there that $P(J_0 \ge k) = P(J_0 \ge 1)^k$, a geometric function of k.

The same process shows that a geometric random variable has a distribution of the form

$$P(X \ge n) = \lambda^n, \quad n \in \mathbb{Z}_+,$$

where $\lambda = P(X \ge 1)$. We baptise this GEOMETRIC(λ) DISTRIBUTION.

Markov's inequality

Arguably this is the most useful inequality in Probability, and is based on the concept of *monotonicity*. Surely, you have seen it before, so this little section is a reminder.

Consider an increasing and nonnegative function $g : \mathbb{R} \to \mathbb{R}_+$. We can case both properties of g neatly by writing

$$\text{For all } x, y \in \mathbb{R} \quad g(y) \ge g(x)\mathbf{1}(y \ge x).$$

Indeed, if $y \ge x$ then $\mathbf{1}(y \ge x) = 1$ and the display tells us $g(y) \ge g(x)$, expressing monotonicity. And if $y < x$ then $\mathbf{1}(y < x) = 0$ and the display reads $g(y) \ge 0$, expressing non-negativity. Let X be a random variable. Then, for all x,

$$g(X) \ge g(x)\mathbf{1}(X \ge x),$$

and so, by taking expectations of both sides (expectation is an increasing operator),

$$\boxed{Eg(X) \ge g(x)P(X \ge x)}\,.$$

This completely trivial thought gives the very useful MARKOV'S INEQUALITY.

Bibliography

1. Brémaud P (1997) *An Introduction to Probabilistic Modeling.* Springer.

2. Brémaud P (1999) *Markov Chains.* Springer.

3. Chung, K L and Aitsahlia F (2003) *Elementary Probability Theory (With Stochastic Processes and an Introduction to Mathematical Finance).* Springer.

4. Feller, W (1968) *An Introduction to Probability Theory and its Applications.* Wiley.

5. Grinstead, C M and Snell, J L (1997) *Introduction to Probability.* Amer. Math. Soc.
   Also available at: `http://www.dartmouth.edu/~chance/teaching_aids/books_articles/` `/probability_book/amsbook.mac.pdf`

6. Norris, J R (1998) *Markov Chains.* Cambridge University Press.

7. Parry, W (1981) *Topics in Ergodic Theory.* Cambridge University Press.

Index