Markov chains -MCMC (Lecture material)

Chapter · July 2020

CITATIONS
0

READS
4

1 author:

Marc Fischer

22 PUBLICATIONS **79** CITATIONS

SEE PROFILE

Saint-Étienne School of Mines

Markov chains - MCMC algorithms

Dr. Marc Fischer (Assistant professor)

Markov chains and processes

Marc Fischer

July 31, 2020

Contents

Chapter 1

Introduction

Named after the great Russian mathematician Andrey Andreyevich Markov, a Markov chain is a specific type of stochastic process. It is defined in such a way that knowing only a small part of the past history of that process is enough to make predictions that are as good as if we knew the entire history. We make a distinction between Markov chains of different orders. In case of a Markov chain of first order, the future state of the process only depends on the current state and it is not influenced by any past states. We also say that a Markov chain of first order is *memoryless* in that it doesn't remember the past and the trajectory that got it where it is now.

In the case of a discrete state space, the mathematical formalism only requires the notion of discrete distribution as well as that of conditional probability, whereas it presupposes the concept of filtration and conditional expectation in the case of a continuous state space. The primal goal of the application of Markov chains consists of determining the probability of the occurrence of future events. A very simple example of a Markov chain of first order is represented in Figure 1.1. The state space consists of two discrete values 1 and 2 with the following transition probabilities:

$$p(X_{i+1} = 1|X_i = 1) = 0.3$$

, and

$$p(X_{i+1} = 2|X_i = 1) = 0.7$$

.

$$p(X_{i+1} = 2|X_i = 2) = 0.6$$

Figure 1.1: Simple example of a Markov chain

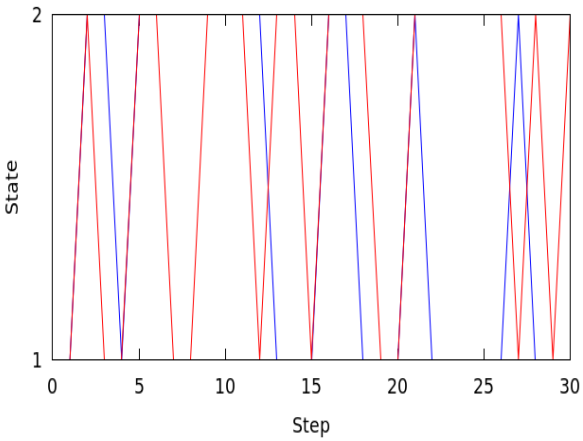


Figure 1.2: Two random Markov chains

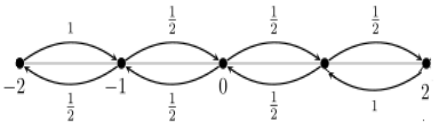


Figure 1.3: Finite random walk with reflecting states

,

$$p(X_{i+1} = 1 | X_i = 2) = 0.4$$

, An example of two realisations of that Markov chain are shown in Figure 1.2.

Another example is shown in Figure 1.3. Three realisations of this Markov chain can be seen in Figure 1.4. It is a random walk in a finite space with discrete values. It can be seen it is periodic with a period equal to 2. If a point was visited at t = 3, it can henceforth only be visited at $t = 3 + 2k$, where k is an integer.

An additional example can be visualised in Figure 1.5. It represents a random walk on a finite state space with two absorbing states -2 and 2. Four typical Markov chains can be seen in Figure 1.5. Once such an absorbing state is reached, the chain is stuck in it and can no longer take on any other values. Otherwise, the transition probability between two neighbouring states is always equal to 50 %. We feel intuitively that the chain is bound to converge to one of the extreme values in a finite time and that it will remain immobile forever thereafter.

Biologists often use a Markov chain called the Galton-Watson process to model the population of organisms which reproduce asexually. We suppose that the number of offspring of any organism i is a random variable Z_i that follows a Poisson distribution so that $p(Z_i = k) = P_\lambda(k) = \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda}$. The

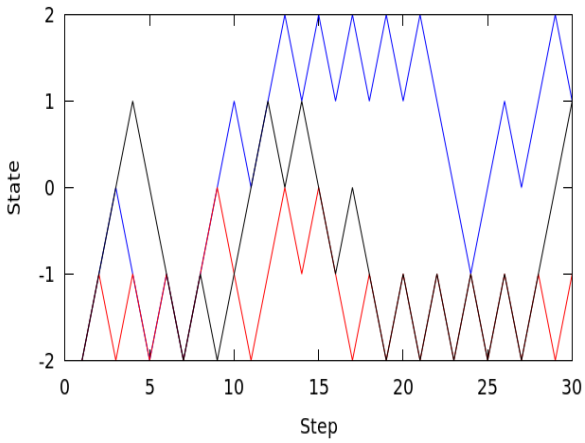


Figure 1.4: Two random Markov chains

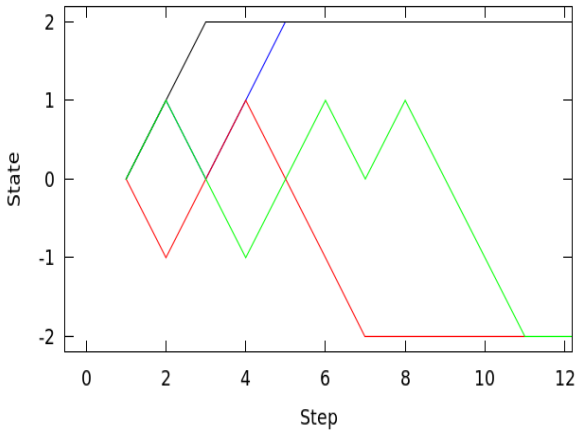


Figure 1.5: Finite random walk with two absorbing states

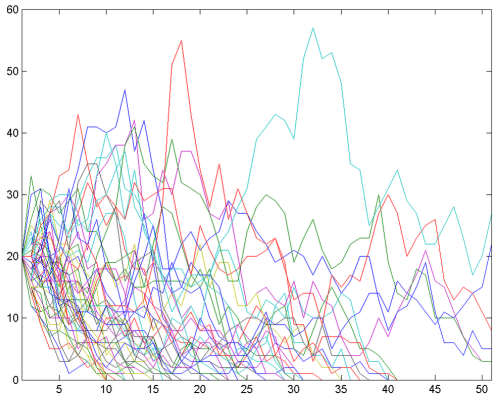


Figure 1.6: Galton process

number of organisms at time i is a homogeneous discrete Markov chain defined by $X_0 = a$ and $X_{i+1} = X_i + \sum_{j=1}^{X_i} Z_j$. The evolution of 50 populations starting with 20 individuals is shown in Figure 1.6. It can be seen that all but six populations died out before the end of the experiment.

While many authors employ "Markov chain" and "Markov process" as synonyms, a Markov chain is often characterised by a discrete state space and by discrete time steps whereas a Markov process tends to be characterised by a continuous state space and an evolution in continuous time..

The purpose of this lecture is to present Markov chains along with several important applications. This lecture material is organised as follows. In Chapter 2, basic notions necessary to understand Markov chains are presented. In Chapter 3, the simplest case, namely Markov chains in discrete time taking values in a discrete state space is examined. In Chapter 4, Markov chains in discrete time taking values in a *continuous* state space are presented. In Chapter 5, the Markov-Chain Monte-Carlo method is presented. It is very useful for computing complicated integrals which are ubiquitous in Bayesian statistics and reasoning.

Chapter 2

Basic notions

2.1 Probability space

Definition 2.1 A probability space is a triplet $(\Omega, F, \mathbb{P})$ made of a sample space, a σ-algebra and a probability measure $\mathbb{P}$.

To illustrate this concept, we can consider the example of the wheel of fortune which can be seen in Figure 2.1.

Definition 2.2 Ω is the sample space, i.e. the space of all possible *elementary* outcomes.

In our case, we have $\Omega = \{1, 2, 3\}$.

Definition 2.3 F is a σ-algebra which must contain all possible combinations of elementary events defined in Ω. It is mathematically defined through three properties.

- F always contains Ω and $\emptyset$.
- F is closed under the formation of complements: if $A \in F$, $A^c \in F$.
- F is closed under countable union: if $A_1, A_2, ..., A_n \in F$, $\bigcup_i A_i \in F$.

For the wheel of fortune, we have F $= \{\emptyset, 1, 2, 3, (1, 2), (1, 3), (2, 3), (1, 2, 3)\}$. In general, for a finite space with n possible values, the σ-algebra is given by the power set of the sample space: $F = \mathcal{P}(\Omega)$. In the case of a continuous sample space (such as $\mathbb{R}$), it is not possible to write the content of the σ-algebra. In that case, we resort to the notion of Borel set. We consider a **generator** such as $E_0 = \{A \subset \mathbb{R} \mid A \text{ is open}\}$ or $E_1 = \{[a, b] \subset \mathbb{R} \mid a \leq b\}$ or $E_2 = \{[a, b] \subset \mathbb{R} \mid a \leq b,\ a, b \in Q\}$, or $E_3 = \{(a, b) \subset \mathbb{R} \mid a < b\}$. F_i is then defined as the smallest σ-algebra containing E_i.

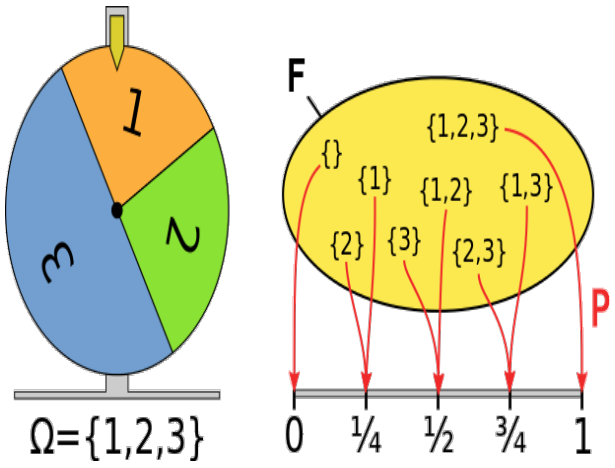


Figure 2.1: Wheel of fortune

Definition 2.4 $\mathbb{P}$ is a probability distribution which must fulfil the Kolmogorov axioms.

$$\begin{aligned} \forall A \in F, \quad & \mathbb{P}(A) \geq 0 \\ & \mathbb{P}(\Omega) = 1 \\ \forall A, B \in F, \quad & \mathbb{P}(A \bigcup B) = \mathbb{P}(A) + \mathbb{P}(B) \text{ if } A \bigcap B = \emptyset \end{aligned} \tag{2.1}$$

2.2 Random variable and stochastic process

2.2.1 Random variable

Definition 2.5 A random variable is a function $X : \Omega \to E$ that binds the sample space Ω to a measurable space E and which must be itself *measurable*. Examples of measures are the distance (in $\mathbb{R}$), the surface (in $\mathbb{R}^2$) and the volume (in $\mathbb{R}^3$).

Let us call $\mathcal{E}$ the σ-algebra associated to E. X is measurable if and only if for all $A \in \mathcal{E}$, we have $X^{-1}(A) = \{\omega \in \Omega, X(\omega) \in A\} \in F$. For all $B \in \mathcal{E}$, the probability of B is defined as $\mathbb{P}(B) = \mathbb{P}(\omega, X(\omega) \in B)$. We see that it is the measurability of a random variable that allows us to associate a probability distribution to it.

2.2.2 Stochastic process

Definition 2.6 A stochastic process $(X_t)_{t \in T}$ taking values in the state space (E,$\mathcal{E}$) is a family of random variables ($X_t \to E, \quad t \in T$) defined on the same probability space (Ω, F, P).

For each $\omega \in \Omega$, we call the function $X(\omega) : T \to E,\ t \to (X(\omega))(t) := X_t(\omega)$ pathway, trajectory or realisation of the random process. If ω is know, it is a deterministic function of t. We interpret the index set T as time so that it will most often be equal to $\mathcal{N}_0$ or $[0, +\infty]$. It is worth noting things could also be more complicated. For example, we could have $T = \mathcal{Z}^d$. In that case, we speak of *stochastic fields* which could, for example, be used to describe geographical or geological phenomena such as the propagation of an earthquake.

As we just noticed, the pathways of an E-valued stochastic process are functions from T into E and are consequently elements of E^{T}. We can especially conceive of $X : \Omega \to E^T$ as an E^T-valued **random function**, if we introduce a σ-algebra in this space. Most of the time, we use the product σ-algebra $\mathcal{E}^T$. This is the smallest σ-algebra that contains all finite rectangles of the form $\{f \in E^T\ :\ f(t_1)\ \in\ E_1, ..., f(t_k)\ \in\ E_k\}$, $k \in \mathcal{N}$, $t_1, ..., t_k\ \in\ T$, $E_1, ..., E_k\ \in\ \mathcal{E}\}$. It is possible to prove that X is measurable with respect to $\mathcal{E}^T$ so that we can finally speak of an E^{T}-valued random variable. Like we did in 2.2.1, for $A \in \mathcal{E}^T$, we pose that $p_X(A) := \mathcal{P}(X^{-1}(A))$. It is in general much harder to describe such random functions than the random variables we're used to. One way to do this is to use *finite-dimensional* distributions. In what follows, we write that $J \subset T$ if J is a finite non-empty part of the index set T.

Definition 2.7 (Finite-dimensional distributions). Let us call $\mathrm{X} = (X_t)_{t \in T}$ a stochastic process. For $J \subset T$, we define $\hat{P}_j$ as the joint distribution of the vector $(X_j)_{j \in J}$. The family $(\hat{P}_J)_{J \subset T}$ is called the *family of finite-dimensional distributions of X*.

It is easy to see that the family of the finite-dimensional distributions are unambiguously determined by the distribution X itself, since they are merely projections from T onto J. If these projections satisfy certain conditions, the Kolmogorov extension theorem guarantees the existence of a unique all-encompassing probability measure on $(E^T, \mathcal{E}^T)$.

Definition 2.8 (Consistency).

Let us consider T and $(E, \mathcal{E})$ like above. For each $J \subset T$, the probability distribution is given by $\overline{P}_J$ on $(E^J, \mathcal{E}^J)$. The family $(\overline{P}_J)_{J \subset T}$ is said to be consistent if and only if for all $J_1 \subset T$ and $J_2 \subset T$ such that $J_1 \subset J_2$, $\overline{P}_{J_1}$ coincides completely with the restriction of $\overline{P}_{J_2}$ to J_1.

Definition 2.9 (Kolmogorov extension theorem).

Let E be a Polish space (a separable completely metrisable topological space such as Z^n and $\mathcal{R}^n$), T a non-empty index set and $(\overline{P}_J)_{J \subset T}$ a consistent family of probability distributions. There exists then exactly one probability distribution $\overline{P}$ on $(E^T, \mathcal{E}^T)$ so that $(\overline{P}_J)_{J \subset T}$ is the family of

finite-dimensional distributions belonging to $\overline{P}$. We further know that there exists a probability space (Ω, $\mathcal{F}$, $\mathcal{P}$) and a stochastic process X = $(X_t)_{t\in T}$ taking values in the state space (E, $\mathcal{E}$) and having the distribution $\overline{P}$.

If the index set is discrete, in order to define a stochastic process $(X_n)_{n\subset\mathcal{N}_0}$ or to prove its existence, it is sufficient to write the distribution of $(X_0, ..., X_n)$ for every $n \in \mathcal{N}_0$ and to show that the different distributions are consistent.

Among other things, this theorem allows us to prove the existence of an endless sequence of independent and identically distributed (i.i.d.) random variables, since its projection on $[|0; m|], m < n$ trivially coincides with the (m+1) first members of the chain with n elements. Later on, we shall see how to prove the existence of full-blown Markov chains with the Kolmogorov extension theorem.

2.3 Useful theorems and lemmas

2.3.1 Matrix operations

We presuppose that the reader has a good knowledge of matrices and linear algebra. In what follows, we'll merely summarise the main definitions and theorems we'll need to deal with Markov chains.

Definition 2.10 (Rank of a matrix)
The rank of a quadratic matrix $A \in \mathbb{R}^{n\times n}$ is the number of linearly independent row or column vectors we can form out of it.

Theorem 2.1 (Invertible matrix)
A is invertible if and only if $rank(A) = n \leftrightarrow det(A) \neq 0$.

Definition 2.11 (Eigenvectors and eigenvalues)
A (row) eigenvector is a vector $x \in \mathbb{C}^n \setminus 0$ such that

$$xA = \lambda x.$$

This is equivalent to

$$x(A - \lambda I_n) = 0$$

where $I_n \in \mathbb{R}^{n\times n}$ is the identity matrix.
λ is called the eigenvalue of A [1].

Corollary 2.1 All eigenvalues are roots of the characteristic polynomial $P(\lambda) = |A - \lambda I_n|$.

[1] "Eigen" is a German word meaning intrinsic.

Definition 2.12 (Spectrum of a matrix)
The spectrum of a matrix A is the ensemble of all its eigenvalues.

$$\sigma(A) = \{\lambda \in \mathbb{C} : \lambda \text{ is an eigenvalue of } A \ \leftrightarrow |A - \lambda I_n| = 0\}$$

Definition 2.13 (Radius of a matrix)
$r(A) = max(|\lambda|, \lambda \in \sigma(A))$ is the radius of A.

Theorem 2.2 (Radius of a stochastic matrix)
If $\forall i \in [|1; n|] \sum_{j=1}^{n} A_{ij} = 1$, $r(A) \leq 1$.

Definition 2.14 (Eigenspace of a matrix)
For a given eigenvalue λ of the matrix A, the eigenspace $E(\lambda)$ is the ensemble of all eigenvectors corresponding to λ:

$$E(\lambda) = \{x \in \mathbb{C}^n | x \neq 0, Ax = \lambda x\} \cup \{0\}.$$

Definition 2.15 (Algebraic multiplicity)
The algebraic multiplicity of λ_i is its multiplicity as a root of the characteristic polynomial, that is, the largest integer k such that $(\lambda - \lambda_i)^k$ divides evenly that polynomial.

Definition 2.16 (Geometrical multiplicity)
The geometrical multiplicity of λ_i is the dimension of the eigenspace $E(\lambda_i)$, i.e. the number of linearly independent vectors different from 0 belonging to that eigenspace.

Theorem 2.3 The geometrical multiplicity can never be greater than the algebraic multiplicity.

Theorem 2.4 Eigenvectors corresponding to different eigenvalues, i.e. belonging to different eigenspaces are linearly independent.

Definition 2.17 A vector x_m is a **generalised eigenvector** of rank m of the n x n matrix A corresponding to the eigenvalue λ if

$$x_m(A - \lambda I_n)^m = 0 \text{ where } I_n \text{ is the identity matrix.}$$

Definition 2.18 The **generalised eigenspace** of the n x n matrix A corresponding to the eigenvalue λ is given by

$$F(\lambda) = \{x_m \in \mathbb{C}^n | m > 0, x_m(A - \lambda I_n)^m = 0\}.$$

Theorem 2.5 The *algebraic* multiplicity of an eigenvalue λ is equal to the dimension of $F(\lambda)$.

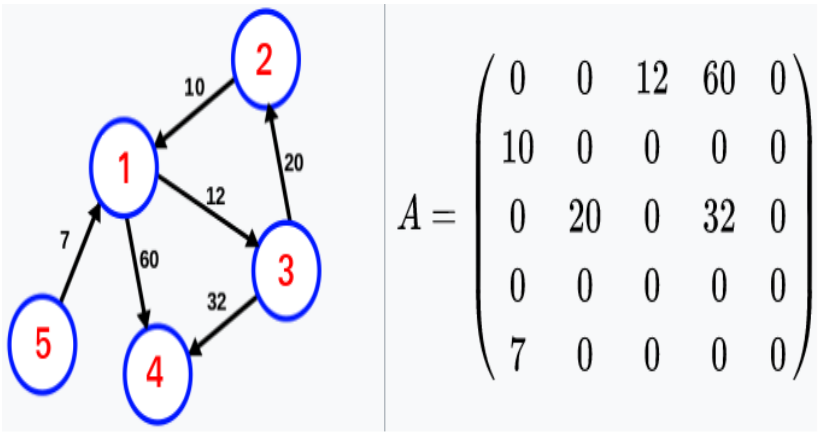


Figure 2.2: A matrix and its graph

Theorem 2.6 If the algebraic multiplicity of A corresponding to the eigenvalue λ is equal to l_1 and if the geometric multiplicity of A corresponding to the eigenvalue λ is $l_2 < l_1$, then there are $l_1 - l_2$ linearly independent *generalised* eigenvectors having ranks greater than 1.

Theorem 2.7 If A is a n x n symmetrical matrix, the following statements are true:

1 All eigenvalues of A are real.

2 The algebraic and geometrical multiplicity of every eigenvalue are the same.

3 If the algebraic multiplicity of λ is equal to n_λ, there are n_λ linearly independent eigenvectors belonging to the eigenspace $E(\lambda_i)$.

4 A can be diagonalised.

5 There is an orthogonal basis of $\mathbb{R}^n$ consisting of eigenvectors of A.

Definition 2.19 (Graph of a matrix)
For a quadratic matrix A of dimension n, we consider the set of all vertices $I = \{1, ..., n\}$. A vertex $i \in I$ is connected to a vertex $j \neq i$ if and only if $A_{ij} \neq 0$. In this way, we can generate a graph as illustrated in Figure 2.2.

Definition 2.20 (Irreducibility of a matrix)
A matrix A is irreducible if and only if for any pair $i, j \in I$, there exists a path going from i to j. This means that

$$\exists i_1, i_2, ..., i_m \in [|1; n|], A_{ii_1} A_{i_1 i_2} ... A_{i_{m-1} i_m} A_{i_m j} \neq 0, 1 \leq m \leq n$$

The matrix shown in Figure 2.2 is **reducible**. If we start off in state 4, we cannot reach any other state.

Theorem 2.8 A matrix $A \in \mathbb{R}^{n\times n}$ is reducible if and only if it can be brought into block upper-triangular form by simultaneous row/column permutations. This means that there exists a permutation matrix $Q \in \mathbb{R}^{n\times n}$ such that

$$QAQ^T = \begin{pmatrix} B_{MM} & 0 \\ C_{NM} & D_{NN} \end{pmatrix}.$$

with $B_{MM} \in \mathbb{R}^{M\times M}$, $C_{NM} \in \mathbb{R}^{N\times M}$, $D_{NN} \in \mathbb{R}^{N\times N}$,and $N = n - M$.

Definition 2.21 A matrix A (which could be a vector) is positive or nonnegative if and only if all its elements are positive or nonnegative, respectively. We write that $A > 0$ and $A \geq 0$.

Theorem 2.9 (Theorem of Perron-Frobenius)

Let $A \in \mathbb{R}^{n\times n}, A \geq 0$ be an irreducible matrix. Let $r := r(A)$ be the radius of that matrix. The following statements can be shown to be true:

(i) $r > 0$

(ii) $r \in \sigma(A)$, i.e. the radius is a real eigenvalue of A which is called the Perron-root of A.

(iii) There is one eigenvector $x \in \mathbb{R}^n > 0$ whose eigenvalue is r.

(iv) There is an unambiguous eigenvector $p \in \mathbb{R}^n$ of A with

$$pA = rp, p > 0, \|p\|_1 = \sum_{j=1}^{n} p_j = 1.$$

p is called the Perron eigenvector of A.

(vi) (Formula of Collatz-Wieland)

$$r = max_{x\in\mathbb{E}} f(x)$$

with

$$f(x) = min^{1\leq i\leq n}_{x_i\neq 0} \frac{(xA)_i}{x_i}$$

and

$$\mathbb{E} = \{x \in \mathbb{R}^n,\ x \geq 0 \text{ and } x \neq 0\}.$$

Definition 2.22 A matrix A is positive semidefinite when for every row vector $x \neq 0$,

$$xAx^T \geq 0.$$

2.3.2 Miscellanous

Theorem 2.10 (Limit of an infinite sequence of values with an upper bound)

Let $(A_n)_{n\geq 0}$ be a sequence of positive numbers.

$$\sum_{n=0}^{+\infty} A_n < +\infty \Rightarrow \lim_{n\to+\infty} A_n = 0.$$

Proof. Let's suppose instead that

$$A_n \overset{n+\infty}{\nrightarrow} 0.$$

This could mean three different things.

(i) $\lim_{n\to+\infty} A_n = a > 0$.

$$\text{For } \epsilon = \frac{a}{10} > 0, \exists n_0 \in \mathbb{N}, n > n_0 \Rightarrow \frac{9a}{10} < A_n < \frac{11a}{10}$$

Hence

$$\sum_{n=0}^{+\infty} A_n \geq \sum_{n=n_0+1}^{+\infty} A_n > \sum_{n=n_0+1}^{+\infty} \frac{9a}{10} = \infty.$$

This contradicts our main postulate.

(ii) $\lim_{n\to+\infty} A_n = \infty$. In that case, it's easy to show in the same manner that $\sum_{n=0}^{+\infty} A_n = \infty$.

(iii) A_n keeps fluctuating without ever approaching 0. This means that

$$\exists \epsilon > 0, \nexists n_0 \in \mathbb{N}, n > n_0 \Rightarrow A_n < \epsilon.$$

As a consequence, there exists an infinite subsequence $n' \subset \mathbb{N}$ such that $A_{n'} > \epsilon$ everywhere because otherwise the sequence would converge to 0 eventually.

$$\sum_{n=0}^{\infty} A_n \geq \sum_{n'} A_{n'} \geq \sum_{n'} \epsilon = +\infty.$$

We also get a contradiction. It is worth noting that the demonstration offered for (iii) is also valid for (i) and (ii).

As a consequence, if $\sum_{n=0}^{+\infty} A_n$ is finite, $A_n \geq 0$ must go to zero.

Theorem 2.11 (Law of total probability)

(i) We consider a probability space $(\Omega, F, \mathbb{P})$, if $\{B_n : n \in I\}$ is a finite or countably infinite partition of Ω such that the B_n are **disjoint**, for any event A, we have

$$p(A) = \sum_{n\in I} p(A \cap B_n) = \sum_{n\in I} p(A|B_n)p(B_n) \tag{2.2}$$

(ii) For every non-empty event C,

$$p(A|C) = \sum_{n\in I} p(A \cap B_n|C) \tag{2.3}$$

Proof of (i).

$p(A) = p(A \cap \Omega) = p(A \cap \cup_{n\in I} B_n) = p(\cup_{n\in I}\{A \cap B_n\})$

Since the B_n are disjoint, the $A \cap B_n$ are disjoint as well $\Rightarrow$

$$p(A) = \sum_{n\in I} p(A \cap B_n) = \sum_{n} p(A|B_n)p(B_n)$$

Proof of (ii).
$p(A|C) = \frac{p(A\cap C)}{p(C)}$. By applying (i) to $A \cap C$;

$$p(A \cap C) = \sum_{n\in I} p(A \cap C \cap B_n)$$
$$p(A|C) = \sum_{n\in I} \frac{p(A \cap C \cap B_n)}{p(C)}$$
$$p(A|C) = \sum_{n\in I} p(A \cap B_n|C)$$

Lemma 2.1 (Markov inequality)
If X is a nonnegative random variable and $a > 0$, then the probability that X is at least equal to a is at most the expectation of X divided by a:

$$\mathrm{P}(X \geq a) \leq \frac{\mathrm{E}(X)}{a}$$

Proof

For any event E, let I_E be the indicator random variable of E, that is, $I_E = 1$ if E occurs and $I_E = 0$ otherwise. Using this notation, we have $I_{(X\geq a)} = 1$ if the event $X \geq a$ occurs, and $I_{(X\geq a)} = 0$ if $X < a$. Then, given $a > 0$, $aI_{(X\geq a)} \leq X$ which is clear if we consider the two possible values of $X \geq a$. If $X < a$, then $I_{(X\geq a)} = 0$, and so $aI_{(X\geq a)} = 0 \leq X$. Otherwise, we have $X \geq a$, for which $I_{X\geq a} = 1$ and so $aI_{X\geq a} = a \leq X$. Since E is a monotonically increasing function, taking expectation of both sides of an inequality cannot reverse it. Therefore, $\mathrm{E}(aI_{(X\geq a)}) \leq \mathrm{E}(X)$. Now, using linearity of expectations, the left side of this inequality is the same as $a\,\mathrm{E}(I_{(X\geq a)}) = a(1 \cdot \mathrm{P}(X \geq a) + 0 \cdot \mathrm{P}(X < a)) = a\,\mathrm{P}(X \geq a)$. Thus we have :$a\,\mathrm{P}(X \geq a) \leq \mathrm{E}(X)$ and since $a > 0$, we can divide both sides by a.

Lemma 2.2 Let X be a nonnegative random variable.

$$E(X) < \infty \Rightarrow p(X < \infty) = 1$$

Proof.

$$p(X < \infty) = 1 - lim_{a\to\infty} p(X \geq a)$$

Because of the Markov inequality,

$$0 \leq lim_{a\to\infty} p(X \geq a) \leq lim_{a\to\infty} \frac{\mathrm{E}(X)}{a}$$

Since $E(X)$ is finite, this implies that $lim_{a\to\infty} p(X \geq a) = 0$.

We'll now state the lemma of Borel-Antelli which turns out to be very useful to study probabilistic convergence.

Lemma 2.3 (Lemma of Borel-Cantelli)

Let us consider a sequence A_n of events.

$$A = \limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m = \{A_n \text{ infinitely often}\}$$
$$= \{\omega \in \Omega : \omega \in A_n \text{ for infinitely many n} \in \mathbb{N}\}$$

we have

$$\sum_{n\geq 1} P(A_n) < \infty \Rightarrow P(A) = 0$$

$$\sum_{n\geq 1} P(A_n) = \infty \text{ and the } A_n \text{ are pairwise stochastic ally independent events } \Rightarrow P(A) = 1$$

Proof of the first part.

Let $I_n = I(A_n)$ be the Bernoulli variable equal to 1 if A_n occurs and to 0 otherwise. Let

$$N = \sum_{n=1}^{\infty} I_n,$$

be the total number of events which were realised. Then $p(A) = 0 \Rightarrow A = \{\varnothing\}$ is equivalent to saying that it is impossible that an endless number of events happen and thus also to $p(N < \infty) = \sum_{n=0}^{\infty} p(N = n) = 1$.

We also know that the expected value of N is given by $E(N) = \sum_{n=1}^{\infty} p(A_n) < \infty$. Because of lemma 2.2, this entails that $p(X < \infty) = 1$.

Proof of the second part.

$$\sum_{n=k}^{\infty} p(A_n) = \infty, k \geq 1.$$

Let $\overline{A_n}$ denote the complement of the set A_n.

$$p(A) = lim_{k\to\infty} p(\cup_{n=k}^{\infty}) A_n = 1 - lim_{k\to\infty} p(\cap_{n=k}^{\infty} \overline{A_n}).$$

To complete the proof, we will show that

$$p(\cap_{n=k}^{\infty} \overline{A_n}) = 0, k \geq 1.$$

Because of the mutual independence of the events and of the basic fact that $1 - x \leq e^{-x}, x \geq 0$,

$$\begin{aligned} p(\cap_{n=k}^{\infty} \overline{A_n}) = \prod_{n=k}^{\infty} p(\overline{A_n}) & \\ &= \prod_{n=k}^{\infty} (1 - p(A_n)) \\ &\leq \prod_{n=k}^{\infty} e^{-p(A_n)} \\ &= e^{-\sum_{n=k}^{\infty} p(A_n)} \\ &= e^{-\infty} = 0, \end{aligned}$$

where the last equality comes from the condition.

Chapter 3

Discrete Markov chains in discrete time

In this chapter, we'll be dealing with Markov chains in discrete time taking values in a countable (finite or countably infinite) state space I. The elements of I are called *states*. A stochastic process $(X_n)_{n\geq 0}$ is a family of I-valued random variables defined on the same probability space (Ω, F, P) which we'll consider to be given.

3.1 Definition and first examples

A Markov chain taking values in a discrete state space I is a stochastic process which could, for example, model the motion of a particle making discrete jumps in I at discrete times belonging to T. The particle "decides" randomly at every new time step how it moves from its current position. The Markov property stipulates then that this decision only depends on the particle's position in the **present** and not in the past. The trajectory (also called pathway) of the particle is thus completely irrelevant for predicting its behaviour in the future *if we know its position at the last time point.*

Example 3.1 (Random walk in a group).

Let I be a commutative countable group (such as $\mathcal{Z}^d$) and $\mathcal{E}_1$, $\mathcal{E}_2$,... a sequence of independent and identically distributed (i.i.d.) I-valued random variables. For $x \in I$, we define $(X_n)_{n\geq 0}$ in such a way that

$$X_n = x + \sum_{i=1}^{n} \mathcal{E}_i.$$

Since $X_{n+1} = X_n + \mathcal{E}_{n+1}$ and because of the independence of $\mathcal{E}_{n+1}$ from $\mathcal{E}_1$, $\mathcal{E}_2$,..., $\mathcal{E}_n$ and thus also from X_0,X_1, X_2,...,X_n, X_{n+1} does not depend on the values of $X_0, X_1, ..., X_{n-1}$ if we know X_n. As a consequence, $(X_n)_{n\geq 0}$ is a Markov chain with $X_0 = x$. We call it a *random walk* on the group I.

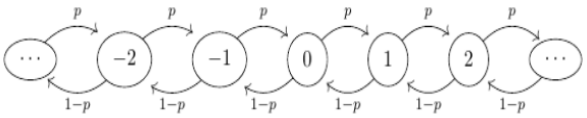


Figure 3.1: Diagram of a random walk on $\mathcal{Z}$

In the case of $I = \mathcal{Z}$, we have $p(\mathcal{E}_i = 1) = p$ and $p(\mathcal{E}_i = -1) = 1 - p$ Its diagram can be seen in Figure 3.1.

Thanks to the Markov property, in order to describe a Markov chain, it is sufficient to know its initial state and the rules governing the transition from one state to the next. They are given by the *initial probability distribution* and the *transition matrix*.

Definition 3.1 (Stochastic matrix). A matrix $P = (p_{ij})_{ij \in I}$ is called a stochastic matrix if and only if the lines of P are probability distributions, that is

$$p_{ij} \geq 0 \text{ for all } i, j \ \in \ I \text{ and } \sum_{j \in I} p_{ij} = 1 \text{ for all } i \in I. \tag{3.1}$$

Definition 3.2 (Markov chain in discrete time)

Let $\upsilon = (\upsilon(i))_{i \in I}$ be a probability distribution on I. An I-valued stochastic process $(X_n)_{n \geq 0}$ is a Markov chain in discrete time with the initial distribution υ, if and only if for all $n \in \mathcal{N}$ and all $i_0, i_1, ..., i_n, i_{n+1} \in I$ we have

$$p(X_0 = i_0) = \upsilon(i_0) \tag{3.2}$$

and

$$p(X_{n+1} = i_{n+1} | X_n = i_n, ..., X_0 = i_0) = p(X_{n+1} = i_{n+1} | X_n = i_n) \tag{3.3}$$

insofar as $p(X_n = i_n, ..., X_0 = i_0) > 0$. We call $(X_n)_{n \geq 0}$ a Markov chain. 3.3 is called the **Markov property**.

Definition 3.3 (Homogeneous Markov chain in discrete time)

We say that $(X_n)_{n \geq 0}$ is an **homogeneous** (ν, P)- Markov chain if and only if it satisfies the Markov property and

$$\forall n \geq 0, p(X_{n+1} = j | X_n = j) = p(X_1 = j | X_0 = i) = P_{ij} = p_{ij}.$$

The stochastic matrix P is called the transition matrix.

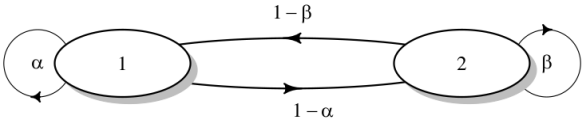


Figure 3.2: Standard Markov chain with two states

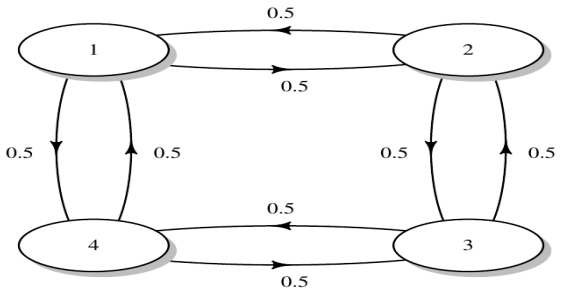


Figure 3.3: Random walk on a graph

Thanks to this definition, we see among other things that the i-th line of the transition matrix is the conditional probability distribution of X_{n+1} given that $X_n = i$.

As we saw in Example 3.1, Markov chain can be easily visualised through transition diagrams. They can be readily built out of the transition matrices. Conversely, it is also quite simple to write up the transition matrix based on the transition diagram.

Example 3.2 (Standard Markov chain with two states). Figure 3.2 shows the diagram of a standard Markov chain with the state space $I = \{1, 2\}$ with $\alpha, \beta \in [0, 1]$. If the system is in state 1 at time n, it will remain in that state with a probability equal to α or shift into state 2 with a probability equal to $1 - \alpha$ at time $n + 1$. Likewise, if the system is in state 2 at time n, it will remain in that state with a probability equal to β or shift into state 1 with a probability equal to $1 - \beta$ at time $n + 1$. The transition matrix is

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

Example 3.3 (Random walk on a graph).

Let the state space be $I = \{1, 2, 3, 4\}$. The transition diagram is represented in Figure 3.3. The corresponding transition matrix is given by

$$P = \begin{pmatrix} 0 & 1/2 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 \end{pmatrix}.$$

A quick look at the diagram is enough to realise that the Markov chain is periodic (a definition of periodicity will be given later on). If the random walker is at time 0 in $\{1, 3\}$, then he will certainly be in $\{1, 3\}$ at even times and in $\{2, 4\}$ at odd times.

The next theorem provides us with an equivalent characterisation of Markov chains. We don't immediately recognise the Markov property but we can directly apply the Kolmogorov extension theorem (2.9) to this situation.

Theorem 3.1 A stochastic process $X = (X_n)_{n\geq 0}$ taking values in I is a Markov process if and only if for all $n \in \mathbb{N}$ and all $i_0, ..., i_n \in I$

$$p(X_n = i_n, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = \nu(i_0)p_{i_0 i_1}p_{i_1 i_2}...p_{i_{n-1} i_n}.$$

Proof. Let X be a (ν, P)-MC. For $i_0, ..., i_n \in I$ with $p(X_{n-1} = i_{n-1}, ..., X_0 = i_0) \geq 0$, we have

$$\begin{aligned} &p(X_n = i_n, ..., X_0 = i_0) \\ &\quad = p(X_n | X_{n-1} = i_{n-1}, ..., X_0 = i_0)p(X_{n-1} = i_{n-1}, ..., X_0 = i_0) \\ &\quad\quad = p(X_n = i_n | X_{n-1} = i_{n-1})p(X_{n-1} = i_{n-1}, ..., X_0 = i_0) \end{aligned}$$

We can go from the second line to the third line thanks to the Markov property: given X_{n-1}, the earlier values of the chain have no influence on X_n. Reasoning inductively, we find that

$$\begin{aligned} p(X_n = i_n, ..., X_0 = i_0) = p(X_n = i_n | X_{n-1})...p(X_1 = i_1 | X_0 = i_0)p(X_0 = i_0) \\ = p_{i_{n-1} i_n}...p_{i_1 i_2}p_{i_0 i_1}\nu(i_0) \end{aligned}$$

We must now also consider the case when $p(X_{n-1} = i_{n-1}, ..., X_0 = i_0) = 0$. This logically entails that $p(X_n = i_n, ..., X_0 = i_0) = 0$ since the second event can never happen if the first event is impossible. We further know that $\{X_{n-1} = i_{n-1}, ..., X_0 = i_0\} = \{X_0 = i_0, i_0 \to i_1, i_1 \to i_2, ..., i_{n-2} \to i_{n-1}\}$ So we know that the probability of at least one member of the right hand side is equal to 0. This implies that $p_{i_n i_{n+1}}p_{i_{n-1} i_n}...p_{i_1 i_2}p_{i_0 i_1}\nu(i_0) = 0$. In both cases, the left side of the theorem naturally leads to the right one.

We now want to prove that its reciprocal is true as well. We know that for all $n \in \mathbb{N}$ and all $i_0, ..., i_n \in I$,

$$p(X_n = i_n, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = \nu(i_0)p_{i_0 i_1}p_{i_1 i_2}...p_{i_{n-1} i_n}.$$

Applying this equality to $n = 0$ directly leads to

$$p(X_0 = i_0) = \nu(i_0) \text{ for all } i_0 \in I.$$

Furthermore, for all $n \in \mathbb{N}$ and all $i_0, i_1, ..., i_{n+1} \in I$;

$$\begin{aligned} p(X_{n+1} = i_{n+1}, ..., X_0 = i_0) &= \frac{p(X_{n+1} = i_{n+1}, X_n = i_n, ..., X_0 = i_0)}{p(X_n = i_n, ..., X_0 = i_0)} \\ &= \frac{\nu(i_0)p_{i_0 i_1}p_{i_1 i_2}...p_{i_{n-1}i_n}p_{i_n i_{n+1}}}{\nu(i_0)p_{i_0 i_1}p_{i_1 i_2}...p_{i_{n-1}i_n}} \\ &= p_{i_n i_{n+1}} \end{aligned}$$

if $p(X_n = i_n, X_{n-1} = i_{n-1}, ..., X_0 = i_0) \geq 0$. If it is equal to 0, both terms are necessarily equal to 0. We can thus conclude that X is a (ν, P) Markov chain.

Theorem 3.2 (Existence of Markov chains) Let's consider the discrete probability distribution $\nu = (\nu(i))_{i \in I}$ and a transition matrix $P = (p_{ij})_{ij \in I}$. There exists then a probability space $(\Omega, F, \mathbb{P})$ and a discrete stochastic process characterised by the function

$$X_n : \Omega \to I, \forall n \in \mathbb{N}$$

which is a Markov chain whose initial distribution is ν and whose transition matrix is P.

Proof. For each $n \in \mathbb{N}_0$, we define a probability measure $\overline{P_n}$ on I^{n+1} through

$$\overline{P_n}(i_0, ..., i_n) = \nu(i_0)p_{i_0 i_1}p_{i_1 i_2}...p_{i_{n-1}i_n}, i_0, ..., i_n \in I.$$

We also have

$$\overline{P_{n-1}}(i_0, ..., i_{n-1}) = \nu(i_0)p_{i_0 i_1}p_{i_1 i_2}...p_{i_{n-2}i_{n-1}}, i_0, ..., i_{n-1} \in I.$$

The restriction of $\overline{P_n}$ to $\{i_0, ..., i_{n-1}\}$ is given by

$$\begin{aligned} \overline{P_n}(i_0, ..., i_{n-1}) &= \sum_{i_n \in I} \overline{P_n}(i_0, ..., i_{n-1}, i_n) \\ &= \sum_{i_n \in I} \nu(i_0)p_{i_0 i_1}p_{i_1 i_2}...p_{i_{n-2}i_{n-1}}p_{i_{n-1}i_n} \\ &= \nu(i_0)p_{i_0 i_1}p_{i_1 i_2}...p_{i_{n-2}i_{n-1}} \sum_{i_n \in I} p_{i_{n-1}i_n} \\ &= \overline{P_{n-1}}(i_0, ..., i_{n-1}) \end{aligned}$$

Therefore, we can apply Kolmogorov extension theorem which proves the existence of an all-encompassing stochastic process $(X_n)n \geq 0$. According to Theorem 3.1, it is a Markov chain.

For a given event $A \in F$, we write that

$$p_i(A) = p(A|X_0 = i), i \in I \tag{3.4}$$

and

$$p_\nu(A) = \sum_{i \in I} \nu(i) p_i(A). \tag{3.5}$$

$p_\nu(A)$ is the probability of A given that ν is the initial probability distribution. The Dirac measure is defined by

$$\delta_i(j) = \begin{cases} 1 : j = i \\ 0, i \neq j \end{cases} . \tag{3.6}$$

We then have $\nu_j = p_{\delta_i(j)}$, which means that the initial probability distribution is equal to 0 save for $X_0 = i$.

Theorem 3.3 (Variant of the Markov property) Let $(X_n)_{n\geq 0}$ be a (ν, P) Markov Chain (MC). For $m \in \mathbb{N}_0$, given that $X_m = i$, $(X_{m+n})_{n\geq 0}$ is a (δ_i, P)-MC independent of $X_0, ..., X_{m-1}$.

Proof. By setting $m = 0$, it is obvious that any stochastic process satisfying this condition will be a Markov chain. We shall now demonstrate that any Markov chain fulfils the property 3.3.

Relying on theorem 3.1, we must show that $\forall n \geq 0, \forall m \geq 0$ and $\forall i_m, i_{m+1}, ..., i_{m+n-1}, i_{m+n} \in I$, we have

$$p(X_{m+n} = i_{m+n}, X_{m+n-1} = i_{m+n-1}, ..., X_m = i_m) = p(X_m = i_m) p_{i_m i_{m+1}} p_{i_{m+1} i_{m+2}} ... p_{i_{m+n-1} i_{m+n}}.$$

Again, this is trivial for $m = 0$. For $m > 0$, we have

$$\begin{aligned} &p(X_{m+n} = i_{m+n}, X_{m+n-1} = i_{m+n-1}, ..., X_m = i_m) = \\ &\sum_{i_0,...,i_{m-1} \in I} p(X_{m+n} = i_{m+n}, X_{m+n-1} = i_{m+n-1}, ..., X_m = i_m, X_{m-1} = i_{m-1}, ..., X_0 = i_0) \\ &= \sum_{i_0,...,i_{m-1} \in I} \nu(i_0) p_{i_0 i_1} p_{i_1 i_2} ... p_{i_{m-1} i_m} p_{i_m i_{m+1}} ... p_{i_{n+m-1} i_{n+m}} \\ &= (\sum_{i_0,...,i_{m-1} \in I} \nu(i_0) p_{i_0 i_1} p_{i_1 i_2} ... p_{i_{m-1} i_m}) p_{i_m i_{m+1}} ... p_{i_{n+m-1} i_{n+m}} \\ &= (\sum_{i_0,...,i_{m-1} \in I} p(X_0 = i_0, ..., X_{m-1} = i_{m-1}, X_m = i_m) p_{i_m i_{m+1}} ... p_{i_{n+m-1} i_{n+m}} \\ &= p(X_m = i_m) p_{i_m i_{m+1}} ... p_{i_{n+m-1} i_{n+m}} \end{aligned}$$

Theorem 3.4 (Conditional independence of past and future)

Let's consider the events $A \in \sigma(X_0, ..., X_m)$ and $B \in \sigma(X_m, ..., X_{m+n})$. This implies that at time m, A lies in the past whereas B lies in the future.

$(X_k)_{k\geq 0}$ is a Markov chain if and only for all $n, m \in \mathbb{N}_0$,

$$p(A \bigcap B | X_m = i) = p(A | X_m = i) p(B | X_m = i)$$

.

This means that given the present, the past and the future are independent.

Proof.

Let us first consider that $A \Rightarrow X_m \neq i$. In that case, $p(A \bigcap B|X_m = i) = p(A|X_m = i) = 0$ so that $p(A|X_m = i)p(B|X_m = i) = 0$. The same holds true if $B \Rightarrow X_m \neq i$. In both cases, the equality is trivially satisfied.

Let us now consider the more frequent case where $A \not\Rightarrow X_m \neq i$ and $B \not\Rightarrow X_m \neq i$.

$$p(A \bigcap B|X_m = i) = \frac{p(A, B, X_m = i)}{p(X_m = i)}$$

$$p(A \bigcap B|X_m = i) = \frac{p(B|A, X_m = i)}{p(X_m = i)} p(A, X_m = i)$$

Since $A \in \sigma(X_0, ..., X_m)$, the Markov property allows us to **intuitively** write that $p(B|A, X_m = i) = p(B|X_m = i)$

More rigorously, we can always decompose A into the union of a finite number of **disjoint** sets A_k such that

$$A_k = \{X_0 = a_{k0}, ..., X_m = a_{km}\}.$$

$$p(A \bigcap B) = p(\{\bigcup_k A_k\} \bigcap B)$$

.

$$p(A \bigcap B) = p(\bigcup_k \{A_k \bigcap B\})$$

.

$$p(A \bigcap B|X_m = i) = p(\bigcup_k \{A_k, B, X_m = i\}|X_m = i)$$

.

$$p(A \bigcap B|X_m = i) = \sum_k \frac{p(A_k, B, X_m = i)}{p(X_m = i)}$$

.

$$p(A \bigcap B|X_m = i) = \sum_k \frac{p(B|A_k, X_m = i)}{p(X_m = i)} p(A_k, X_m = i)$$

. Since $A_k = \{X_0 = a_{k0}, ..., X_m = a_{km}\}.$, $p(B|A_k, X_m = i) = p(B|X_m = i)$ because of the Markov property.

$$p(A \bigcap B|X_m = i) = \sum_k \frac{p(B|X_m = i)}{p(X_m = i)} p(A_k, X_m = i)$$

.

$$p(A \bigcap B|X_m = i) = \frac{p(B|X_m = i)}{p(X_m = i)} p(A, X_m = i)$$

.

$$p(A \bigcap B|X_m = i) = p(B|X_m = i)p(A|X_m = i)$$

.

Let us now prove the reciprocal. We have a stochastic chain $(X_{k\geq 0})$ such that for all $A \in \sigma(X_0, ..., X_m)$ and $B \in \sigma(X_m, ..., X_{m+n})$,

$$p(A \bigcap B | X_m = i) = p(A|X_m = i)p(B|X_m = i)$$

. Let's choose $B = \{X_{n+1} = i_{n+1}, X_n = i_n\}$ and $A = \{X_0 = i_1, X_1 = i_1, ..., X_n = i_n\}$. We have

$$\begin{aligned} p(X_0 = i_0, ..., X_n = i_n, X_{n+1} = i_{n+1} | X_n = i_n) &= p(X_0 = i_1, X_1 = i_1, ..., X_n = i_n | X_n = i_n) \\ &\quad p(X_{n+1} = i_{n+1}, X_n = i_n | X_n = i_n) \\ &= p(X_0 = i_1, X_1 = i_1, ..., X_n = i_n | X_n = i_n) p(X_{n+1} = i_{n+1} | X_n = i_n) \end{aligned}$$

$$\begin{aligned} \frac{p(X_0 = i_0, ..., X_n = i_n, X_{n+1} = i_{n+1})}{p(X_n = i_n)} &= \frac{p(X_0 = i_1, X_1 = i_1, ..., X_n = i_n)}{p(X_n = i_n)} \\ &\quad \times p(X_{n+1} = i_{n+1} | X_n = i_n) \\ p(X_0 = i_0, ..., X_n = i_n, X_{n+1} = i_{n+1}) &= p(X_0 = i_1, X_1 = i_1, ..., X_n = i_n) \\ &\quad \times p(X_{n+1} = i_{n+1} | X_n = i_n) \end{aligned}$$

$$\begin{aligned} &p(X_{n+1} = i_{n+1} | X_0 = i_0, ..., X_n = i_n) p(X_0 = i_0, ..., X_n = i_n) \\ &\quad = p(X_0 = i_1, X_1 = i_1, ..., X_n = i_n) p(X_{n+1} = i_{n+1} | X_n = i_n) \\ &\quad p(X_{n+1} = i_{n+1} | X_0 = i_0, ..., X_n = i_n) = p(X_{n+1} = i_{n+1} | X_n = i_n) \end{aligned}$$

Therefore, $(X_{k\geq 0})$ is a Markov chain.

We shall now see how to determine the probability distribution at time n of X_n if we know the initial distribution and the transition matrix. We write discrete and finite probability distributions over I in the form of a line vector. The usual rules governing matrix algebra can be extended to the case of infinite vectors and matrices.

Lemma 3.1 If ν is a probability distribution over I and P is a stochastic matrix,

$$(\nu P)_j = \sum_{i \in I} \nu(i) p_{ij}$$

is a probability distribution.

Proof : $(\nu P)_j \geq 0$ and we have

$$\sum_{j \in I} (\nu P)_j = \sum_{j \in I} \sum_{i \in I} \nu(i) p_{ij} = \sum_{i \in I} \nu(i) \sum_{j \in I} p_{ij} = \sum_{i \in I} \nu(i) = 1.$$

Lemma 3.2 Since the lines of a stochastic matrix are themselves probability distributions over I,

$$(P^2)_{ij} = \sum_{k\in I} p_{ik}p_{kj}$$

is itself a stochastic matrix.

Proof : obviously, $(P^2)_{ij} \geq 0$.

$$\sum_{j\in I}(P^2)_{ij} = \sum_{j\in I}\sum_{k\in I} p_{ik}p_{kj} = \sum_{k\in I} p_{ik}\sum_{j\in I} p_{kj}$$

$$\sum_{j\in I}(P^2)_{ij} = \sum_{k\in I} p_{ik} = 1.$$

Lemma 3.3 In general, P^n, $n \in \mathbb{N}$ is a stochastic matrix.

Proof.

We'll first have to derive the general expression of P^n_{ij}, which only demands a good knowledge of matrix algebra. Let d be the dimension of the state space.

$$(P^2)_{ij} = \sum_{i_1=1}^{d} p_{ii_1}p_{i_1j}$$

$$(P^3)_{ij} = \sum_{i_2=1}^{d}(P^2)_{ii_2}p_{i_2j}$$

$$(P^3)_{ij} = \sum_{i_1=1}^{d}\sum_{i_2=1}^{d} p_{ii_1}p_{i_1i_2}p_{i_2j}$$

By recurrence, we can easily prove that for $n > 3$,

$$(P^n)_{ij} = \sum_{i_1=1}^{d}\sum_{i_2=1}^{d}...\sum_{i_{n-1}=1}^{d} p_{ii_1}p_{i_1i_2}...p_{i_{n-2}i_{n-1}}p_{i_{n-1}j}.$$

Obviously, $(P^n)_{ij} \geq 0$.

$$\sum_{j=1}^{d}(P^n)_{ij} = \sum_{j=1}^{d}\sum_{i_1=1}^{d}\sum_{i_2=1}^{d}...\sum_{i_{n-1}=1}^{d} p_{ii_1}p_{i_1i_2}...p_{i_{n-2}i_{n-1}}p_{i_{n-1}j}$$

$$\sum_{j=1}^{d}(P^n)_{ij} = \sum_{i_1=1}^{d}\sum_{i_2=1}^{d}...\sum_{i_{n-1}=1}^{d} p_{ii_1}p_{i_1i_2}...p_{i_{n-2}i_{n-1}}\sum_{j=1}^{d} p_{i_{n-1}j}$$

$$\sum_{j=1}^{d}(P^n)_{ij} = \sum_{i_1=1}^{d}\sum_{i_2=1}^{d}...\sum_{i_{n-1}=1}^{d} p_{ii_1}p_{i_1i_2}...p_{i_{n-2}i_{n-1}}$$

$$\sum_{j=1}^{d}(P^n)_{ij} = \sum_{i_1=1}^{d} p_{ii_1}\sum_{i_2=1}^{d} p_{i_1i_2}...\sum_{i_{n-1}=1}^{d} p_{i_{n-2}i_{n-1}}$$

$$\sum_{j=1}^{d}(P^n)_{ij} = 1$$

.

What does P^n represent? To answer this question, let ν be the initial distribution.

$$(\nu P)_i = \sum_{i_0=1}^{d} \nu_{i_0} p_{i_0 i} = \sum_{i_0=1}^{d} p(X_0 = i_0)p(X_1 = i|X_0 = i_0)$$

$$(\nu P)_i = \sum_{i_0=1}^{d} p(X_1 = i, X_0 = i_0) = p(X_1 = i)$$

Consequently, νP is the probability distribution of X_1.

$(\nu P)P =$

Since this results holds for all discrete probability distributions and all stochastic matrices and thanks to the Markov property, we can conclude that νP^2 is the probability distribution of X_2, νP^3 is the probability distribution of X_3, ..., and that νP^n is the probability distribution of X_n. For $k \in I$, if $\nu_k = \delta_{ik}$ (which means that the chain always start in state i,

$$p(X_n = j|X_0 = i) = p_i(X_n = j) = 1(P^n)_{ij}.$$

Hence

$$(P^n)_{ij} = p(X_n = j|X_0 = i).$$

The i-th line of P^n is the probability distribution of X_n if $X_0 = i$. What is more, $P^{m+n} = P^m P^n$ because of the basic properties of matrix multiplication.

Let's write $p_{ij}^{(n)} = (P^n)_{ij}$ for $n \geq 0$ whereby P^0 is the identity matrix. This directly leads to the following theorem:

Theorem 3.5 (Chapman-Komologorov equation) For all $i, j \in I$, $n, m \in \mathbb{N}_0$, we have

$$p_{ij}^{(n+m)} = \sum_{k \in I} p_{ik}^{(n)} p_{kj}^{(m)}$$

and for all $k \in I$,

$$p_{ij}^{(n+m)} \geq p_{ik}^{(n)} p_{kj}^{(m)}.$$

Theorem 3.6 (Transition matrices for n steps)

If $(X_n)_{n\geq 0}$ is a (ν, P) Markov chain, then we have $p_\nu(X_n = j) = (\nu P^n)_j$. We especially have $p_i(X_n = j) = p_{ij}^{(n)}$.

Proof. See above.

Example 3.4 (Urn. Example of a temporally inhomogeneous Markov chain). Let us consider a urn which contains r_0 red and g_0 green balls at time 0. At time n, we randomly draw a ball from the urn and put it back in. If it is green, we put an additional green ball in the urn. If it is red, we put an additional red ball in the urn. It is easy to see that at time n, the urn will contain $n + r_0 + g_0$ balls. Let R_n be the number of red balls in the urn at time n. If we assume that the drawing is unbiased, $R := (R_n)_{n\geq 0}$ is then a Markov chain whose state space is $\{r_0, r_0 + 1, r_0 + 2, ..., \}$ and whose transition probabilities are given by

$$p(R_{n+1} = j | R_n = i) = \begin{cases} \frac{i}{n+r_0+g_0} & if \quad j = i+1 \\ 1 - \frac{i}{n+r_0+g_0} & if \quad j = i \\ 0 \quad else \end{cases} . \tag{3.7}$$

Since the transition probabilities depend on n, $R := (R_n)_{n\geq 0}$ is obviously not homogeneous in time.

We can, however, define the sequence $(R_n, G_n)_{n\geq 0}$, whereby G_n is the number of green balls at time n. The state space of $(R_n, G_n)_{n\geq 0}$ is

$$\{r_0, r_0 + 1, r_0 + 2, ..., \}\{g_0, g_0 + 1, g_0 + 2, ..., \}$$

. We further have $p((R_0, G_0) = (r_0, g_0)) = 1$.

The transition matrix of $(R_n, G_n)_{n\geq 0}$ is given by

$$p((R_{n+1}, G_{n+1}) = (r', g') | (R_n, G_n) = (r, g)) := \begin{cases} \frac{r}{r+g} & if \quad (r', g') = (r+1, g) \\ \frac{g}{r+g} & if \quad (r', g') = (r, g+1) \\ 0 \quad else \end{cases} . \tag{3.8}$$

In general, we can always convert an inhomogeneous Markov chain into a homogeneous MC by explicitly or implicitly including the time step n into the new variable. This fact can be very useful because it allows us to indirectly apply the theorems for homogeneous Markov chains to inhomogeneous ones so that we'll from now on only consider homogeneous MC.

3.2 Communicating classes and periods

In this section, we'll be talking about the irreducibility and the aperiodicity of a Markov Chain and of its corresponding transition matrix. These notions play especially an important role for Markov chains taking values in a finite state space, because in that case the combination of these two properties guarantee the *existence* and the *uniqueness* of the invariant distribution of the Markov chain.

Example 3.5 Figure 3.4 shows the diagram of a Markov chain taking values in $I = \{1, 2, ..., 11\}$. An arrow from i to j means that $p_{ij} \geq 0$ without

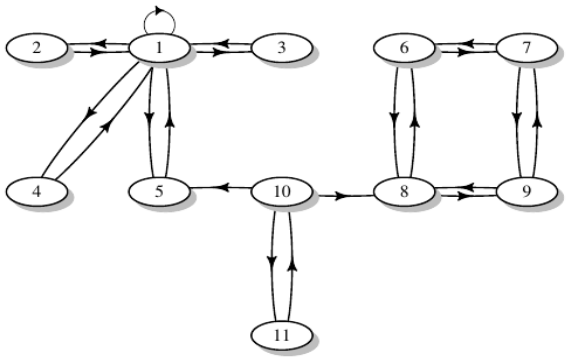


Figure 3.4: Random walk on a graph

revealing any information about its precise value. We just have to read the diagram to discover several important features of the chain:

- If the chain begins in $\{1, ..., 5\}$, it never leaves that set.
- If the chain begins in $\{6, ..., 9\}$, it never leaves that set. What's more, every state can only be revisited after an **even** number of steps.
- If the chain begins in $\{10, 11\}$, it can remain in those states for some amount of time but it is bound to leave them sooner or later and join either $\{1, ..., 5\}$ or $\{6, ..., 9\}$. Consequently, $\{10, 11\}$ play no role in the asymptotic behaviour of the chain.

Definition 3.4 (Communicating states, essential states)
We say that state j can be reached from state i ($i \to j$) if and only if

$$\exists n \geq 0, p_{ij}^{(n)} > 0. \tag{3.9}$$

We say that i and j communicate ($i \leftrightarrow j$) if and only if ($i \to j$) and ($j \to i$).

A state i is called **inessential** if there exists one state j such that $i \to j$ and $j \nrightarrow i$. Otherwise, the state i is **essential**.

Naturally, the asymptotic behaviour of a Markov chain is entirely governed by the essential states.

Lemma 3.4 ($\leftrightarrow$) is an equivalence relation.

Proof. Reflexivity: obviously true, $p_{ii}^{(0)} = 1 > 0$. Symmetry: if $i \leftrightarrow j$, $\exists n_1, n_2 > 0, p_{ij}^{(n_1)} > 0, p_{ji}^{(n_2)} > 0$. It immediately follows that $j \leftrightarrow i$. Transitivity: Let's suppose that $i \leftrightarrow j$ and $j \leftrightarrow k$. $\exists n_1, n_2 > 0, p_{ij}^{(n_1)} >$

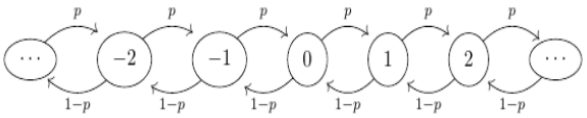


Figure 3.5: Diagram of a random walk on $\mathcal{Z}$

$0, p_{ji}^{(n_2)} > 0$. $\exists n_3, n_4 > 0, p_{jk}^{(n_3)} > 0, p_{kj}^{(n_4)} > 0$. Thanks to the Chapman-Kolmogorov equation, we know that $p_{ik}^{(n_1+n_3)} \geq p_{ij}^{(n_1)} p_{jk}^{(n_3)} > 0$. Likewise $p_{ki}^{(n_4+n_2)} \geq p_{kj}^{(n_4)} p_{ji}^{(n_2)} > 0$.

Definition 3.5 (Closeness, irreducibility) A subset $C \subset I$ is *closed*, if and only if when $i \in C$ and $i \to j$, j belongs to C. A state i is *absorbing* if i is a closed class on its own. A Markov chain (and its associated transition matrix) is *irreducible*, if the state space I itself is a closed communicating class.

Following this definition and thanks to lemma 3.4, we can say that a state $k \in I$ is either inessential or essential, in which case it can be reached from any other state $j \in I$. If $k_1 \in C_1$ and $k_2 \in C_2$ communicate, they belong to the same class $C = C_1 \bigcup C_2$, otherwise C_1 and C_2 are two **disjoint** communicating classes.

Therefore, we can always decompose the state space into

$$I = U \bigcup C_0 \bigcup C_1 \bigcup C_2 \bigcup ...$$

U is the set of inessential states whereas $C_0, C_1, C_2, ...$ are disjoint closed communicating classes.

Example 3.6 In Figure 3.4, $U = \{10, 11\}$ is the set of inessential states, $C_0 = \{1, 2, 3, 4, 5\}$ and $C_1 = \{6, 7, 8, 9\}$ are the communicating classes. C_0 and C_1 are closed but not U because $10 \to 5$ but $5 \nrightarrow 10$.

C_1 and C_2 are essential, U is inessential.

We'll now consider the concept of **periodicity** and the situation in Figure 3.5.

We decompose the state space I into the set of even number K_1 and the set of odd numbers K_2. It is clear that when $X_n \in K_1$, X_{n+1} will belong to K_2, and that when $X_n \in K_2$, X_{n+1} will be part of K_1, because the random walk takes one step either left or right but cannot remain at the same position. Importantly, it means that the Markov chain can only return to a state after an even number of steps.

Corollary 3.1 A Markov chain is called irreducible if and only if every state $i \in I$ can be reached from any other state $j \in I$, i.e. $\exists m > 0, p_{ij}^m > 0$. Following Definition 2.20, this is equivalent to the irreducibility of the transition matrix P.

Proof.

Let's suppose

$$\exists i_1, i_2, ..., i_m \in [|1; n|], p_{ii_1} p_{i_1 i_2} ... p_{i_{m-1} i_m} p_{i_m j} \neq 0, m \leq n.$$

Because of the Chapman-Komolgorov equation,

$$p_{ij}^{(m+2)} \geq p_{ii_1} p_{i_1 i_2} ... p_{i_{m-1} i_m} p_{i_m j} > 0.$$

(The result is even more trivial if i can switch to j directly or after one step).

If, on the other hand, no such series of connections can be generated, the Chapman-Komolgorov equality tells us that the transition probability p_{ij}^m is always equal to 0 for all m.

In what follows, $gcd\{A\}$ designates the greatest common divisor of set A.

Definition 3.6 (Periodicity and aperiodicity) The period of a state i is defined as

$$d_i = gcd\{n \geq 1 : p_{ii}^{(n)} > 0\}$$

with $d_i = +\infty$, if $p_{ii}^{(n)} = 0$, for all $n \geq 1$. If $d_i = 1$, state i is called aperiodic.

According to this definition, each state of example 3.5 has a period equal to 2.

Example 3.7 (Periodic and aperiodic Markov chains)

Figure 3.6 shows the transition diagrams of two Markov chains taking values in the same state space $I = \{1, ..., 5\}$. Both Markov chains are irreducible and we'll see later on that the period is a *class property*. Let us first consider graph (A). It is clear that for all $i \in I$ and all $n \geq 1$, $p_{ii}^{(2n)} > 0$ and $p_{ii}^{(2n+1)} = 0$. Therefore, all states have a period equal to 2.

We're now considering graph (B). The crucial difference with (A) lies in the fact in (A), $p_{11} = 0$ whereas in B $p_{11} > 1$. As a consequence, $d = 1$, the Markov chain is aperiodic because it can remain in state 1 for an arbitrarily long period of time.

Remark: it is often easy to determine whether a Markov chain is aperiodic by looking at the graph. We just need to find two pathways from i to i whose lengths don't have a common divisor greater than 1.

Lemma 3.5 If $M \subset \mathbb{N}$ is a set which is closed with respect to addition with $gcd\{M\} = 1$, then there exists a n_0 with $n \in M$ for all $n \geq n_0$.

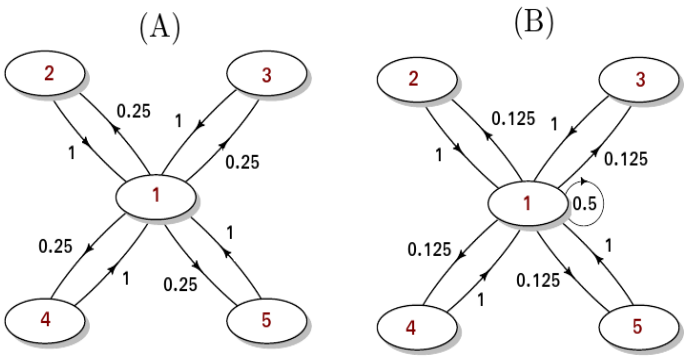


Figure 3.6: Markov chain with a period of 2 and its aperiodic counterpart

Theorem 3.7 Let j be a state with period d. Then the following statements are valid.

(i) If $d = 1$, there exists a $n_0 = n_0(j)$ with $p_{jj}^{(n)} > 0$ for all $n \geq n_0$.

(ii) If d > 1, there is a $n_0 = n_0(j, d)$ with $p_{jj}^{(nd)} > 0$ for all $n \geq n_0$.

(iii) If $d \geq 1$ and $p_{ij}^{(m)} > 0$ for $i \in I$ and $m \geq 1$, there is a $n_0 = n_0(j, d, m)$ with $p_{ij}^{(nd+m)} > 0$ for all $n \geq n_0$.

Proof.

(i) Let $M = \{m\}$ be the set of all integers such that $p_{jj}^{(m)} > 0$. Let us consider $m_1, m_2 \in M$. We have $p_{jj}^{(m1)} > 0$ and $p_{jj}^{(m2)} > 0$. Because of the Chapman-Komolgorov equation,

$$p_{jj}^{(m1+m2)} \geq p_{jj}^{(m1)} p_{jj}^{(m2)} > 0$$

$m_1 + m_2 \in M$, M is closed with respect to addition. We can thus apply lemma 3.5 to it.

(ii) Let $M_d = \{m\}$ be the set of all integers such that $p_{jj}^{(md)} > 0$. Given that d is the greatest common divisor of the lengths *allowing* the chain to return to j, all such numbers can be written as md with $m \in M_d$. Using the Chapman-Komolgorov equation, it is easy to show that M is closed with respect to addition. We can thus apply lemma 3.5 to M_d.

(iii) We know thanks to (ii) that there is a $n_0 = n_0(j, d)$ with $p_{jj}^{(nd)} > 0$ for all $n \geq n_0$. According to the Chapman-Komolgorov equation, $p_{ij}^{(nd+m)} \geq p_{ij}^{(m)} p_{jj}^{(nd)} > 0$.

We can now prove that elements belonging to the same communicating class have the same period.

Theorem 3.8 (The period is constant on a class).

If C is a communicating class, then we always have $d_i = d_j$ for all $i, j \in C$.

Proof. If $i \neq j$, there exists $k, l \geq 1$ such that $p_{ij}^{(k)} > 0, p_{ji}^{(l)} > 0$. Thanks to the Chapman-Komolgorov equation, we have

$$p_{ii}^{(k+l)} \geq p_{ij}^{(k)} p_{ji}^{(l)},$$

so that d_i must be a divisor of $k+l$. Let us now consider n such that $p_{jj}^{(n)} > 0$. d_j is a divisor of n. Because of the Chapman-Komolgorov equation, we know that

$$p_{ii}^{(k+l+n)} \geq p_{ij}^{(k)} p_{jj}^{(n)} p_{ji}^{(l)} > 0,$$

d_i must be a divisor of $k + l + n$. Since d_i is a divisor of $k + l$, it must be a divisor of any n such that $p_{jj}^{(n)} > 0$.. Since $d_j = gcd\{n \geq 1 : p_{jj}^{(n)} > 0\}$, $d_i \leq d_j$. If we switch the role of i and j, we can similarly prove that $d_j \leq d_i$, which shows that the two states have the same period.

Definition 3.7 If C is a communicating class, then we call $d(C)$ the (common) period of the elements of that class. We say that the class is *aperiodic* if $d(C) = 1$.

Theorem 3.9 If C is a communicating class with the period $d = d(C) > 1$, then there are d *disjoint* cyclical subclasses $K_0, ..., K_{d-1}$ with $\bigcup_{l<d} K_l = C$. For $i \in K_l$ and $l \in \{0, 1, ..., d-1\}$,

$$\sum_{j \in K_{l+1}} p_{ij} = 1,$$

whereby $K_d = K_0$. If the Markov chain starts in K_0, at time $l + kd, l \in \{0, 1, ..., d-1\}, k \in \mathbb{N}_0$ the chain will be in K_l, at the next time step in K_{l+1} etc.

Proof.

Let us consider a given $i_0 \in C$ and let us define the following sets:

$$K_0 = \{j \in C : p_{i_0 j}^{(n)} > 0 \Rightarrow n = kd, \text{ for } k \in \mathbb{N}\}$$

$$K_1 = \{j \in C : p_{i_0 j}^{(n)} > 0 \Rightarrow n = kd + 1, \text{ for } k \in \mathbb{N}_0\}$$

. . .

$$K_{d-1} = \{j \in C : p_{i_0 j}^{(n)} > 0 \Rightarrow n = kd + (d-1), \text{ for } k \in \mathbb{N}_0\}$$

The union of all these sets is C. For each $j \in C$, there is a $n \in \mathbb{N}$ with $p_{i_0 j}^{(n)} > 0$ (because C is communicating) and it is possible to find an appropriate $l \in \{0, 1, ..., d-1\}$ and an appropriate $k \in \mathbb{N}_0$ such that $n = kd + l$.

Let us now prove that the classes $K_0, K_1, ..., K_{d-1}$ are disjoint. Let's consider a state i. We know that there exists a $n \in \mathbb{N}$ such that $p_{i_0 i}^{(n)} > 0$.

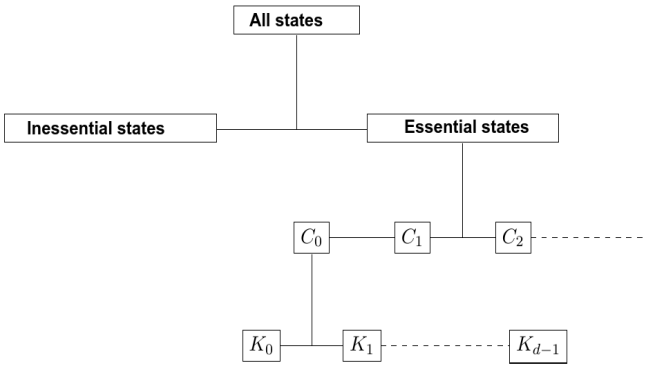


Figure 3.7: Classification of all states. $C_0, C_1, ...$ are communicating classes. $K_0, ..., K_{d-1}$ are cyclical subclasses of C_0 with $d = d(C_0)$

By dividing n by the period d, there is a unique $k \in \mathbb{N}_0$ and a **unique** $l \in \{0, 1, ..., d-1\}$ such that $n = kd + l$. i belongs thus to K_l and it cannot belong to any other class.

Let us now consider another state j such that $p_{ij} > 0$. Since $p_{i_0 i}^{(kd+l)} > 0$, according to Chapman-Komolgorov, $p_{i_0 j}^{(kd+l+1)} > 0$ which means that $j \in K_{l+1}$.

The differences between periodicity and the concept of essentiality can be seen in Figure 3.7

Example 3.8 (Transition matrices with periodic classes) We're now considering the first Markov chain of Figure 3.6. The transition matrix is given by

$$P = \begin{pmatrix} 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

so that

$$P^2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}.$$

We can recognise thanks to P^2 that after two time steps, the chain is back in the underclass where it started.

More generally, if $d = 5$, the transition matrices of a MC can be written

as

$$P = \begin{pmatrix} 0 & P_{12} & 0 & 0 & 0 \\ 0 & 0 & P_{23} & 0 & 0 \\ 0 & 0 & 0 & P_{34} & 0 \\ 0 & 0 & 0 & 0 & P_{45} \\ P_{51} & 0 & 0 & 0 & 0 \end{pmatrix}.$$

and

$$P^5 = \begin{pmatrix} E_{11} & 0 & 0 & 0 & 0 \\ 0 & E_{22} & 0 & 0 & 0 \\ 0 & 0 & E_{33} & 0 & 0 \\ 0 & 0 & 0 & E_{44} & 0 \\ 0 & 0 & 0 & 0 & E_{55} \end{pmatrix}.$$

3.3 Markov times and the strong Markov property

The variant of the Markov property which we mentioned in 3.1 stipulates that at any (deterministic) time m, if we know the state $X_m = i$, the Markov chain after this time behaves like a newly started Markov chain. In this section, we shall see that this property is also valid for a particular kind of **random time** called the Markov time.

Definition 3.8 (Filtration)
For each $n \in \mathbb{N}$, let's define $F_n := \sigma(X_m, m \leq n)$ which is the σ-algebra produced by the finite sequence $X_0, ..., X_n$. Obviously, $F_0 \subset F_1 \subset ... \subset F_n \subset ... \subset F$. We call the sequence $(F_n)_{n\geq 0}$ the filtration belonging to the Markov chain. For instance, F_3 is the ensemble of all possible events involving the random variables X_0, X_1, X_2, X_3.

Definition 3.9 (Markov time)
An application $T : \Omega \to \mathbb{N}_0 \bigcup \{+\infty\}$ is a **Markov time** of $(X_n)_{n\geq 0}$ if and only if for each $n \in \mathbb{N}_0$, we have $\{T = n\} \in F_n$. This means that the event $\{T = n\}$ can only depend on $X_0, X_1, ..., X_n$ and not on X_{n+1} or any other subsequent value of the Markov chain. Therefore, the information we've gathered up until time n, that is the values of $X_0, X_1, ..., X_n$, are sufficient to determine whether the Markov time takes on the value n.

Example 3.9 (Examples of Markov times.)

1. Deterministic times, i.e. $T \equiv N$, are Markov times.

2. For $A \subset I$ let us call

$$T_A := inf\{n \geq 1 : X_n \in A\} \tag{3.10}$$

the first return time in A and

$$S_A := inf\{n \geq 0 : X_n \in A\} \tag{3.11}$$

the first arrival time in A. We employ the usual convention $inf\{\emptyset\} = +\infty$. It is worth noting that T_A and S_A are identical if the Markov chain doesn't start in A. In both cases, we have

$$\{T_A = n\} = \{X_1 \notin A, ..., X_{n-1} \notin A, X_n \in A\}$$

and

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

The two events $\{T_A = n\}$ and $\{S_A = n\}$ are made of the values of the Markov chain before time $n + 1$. As a consequence, the variables T_A and S_A are Markov times. For $A = \{i\}$, we write T_i and S_i, respectively.

3. Through

$$T_A^{(0)} = 0 \text{ and } T_A^{(k)} := inf\{n > T_A^{(k-1)} : X_n \in A, k \geq 1\},$$

we can recursively define a sequence of Markov times. We call $T_A^{(k)}$ the k-th return time into A.

4. The exit time from A, which is defined by

$$L^A := sup\{n \geq 0 : X_n \in A\},$$

is in general not a Markov time, because the event $\{L^A = n\}$ depends on all subsequent states, namely on whether or not A is visited again at a time $m > n$.

5. The sums, the maxima and the minima of Markov times are themselves Markov times. However, differences between Markov times are not generally Markov times. For example, $\{T_i - 1\}$ is not a Markov time because

$$\{T_i - 1 = n\} = \{T_i = n + 1\}.$$

The information is contained in F_{n+1} but not only in $F_0, F_1, ..., F_n$.

Theorem 3.10 (Strong Markov property) Let $(X_n)_{n\geq 0}$ be a (ν, P)-MC and T be a Markov time of $(X_n)_{n\geq 0}$ (i.e. a random variable). Given $T < +\infty$ and $X_T = i$, $(X_{T+n})_{n\geq 0}$ is a (δ_i, P)-MC independent of $X_0, ..., X_T$.

As we saw, $T_i - 1$ isn't a Markov time. We can also easily understand why the strong Markov property cannot be valid at that time. Indeed, after $T_i - 1$, the chain jumps **deterministically** into state i and doesn't behave like a newly started Markov chain.

Proof of the strong Markov property

Let's define $A_{i_0,i_1,...,i} = \{X_0 = i_0, X_1 = i_1, ..., X_{T-1} = i_{T-1}, X_T = i\}$ and $B = \{X_T = i, X_{T+1} = j_1, ..., X_{T+n} = j_n\}$. For an homogeneous Markov chain, we must prove that

$$p(B|T < \infty, X_T = i) = p_{ij_1}p_{j_1j_2}...p_{j_{n-1}j_n}.$$

$$p(B|T < \infty, X_T = i) = \sum_{i_0,i_1,...,i_{T-1}} p(A_{i_0,i_1,...,i_{T-1},i}, B|T < \infty, X_T = i)$$

$$p(A_{i_0,i_1,...,i_{T-1},i}, B|T < \infty, X_T = i) = \frac{p(A_{i_0,i_1,...,i_{T-1},i}, B, T < \infty, X_T = i)}{p(T < \infty, X_T = i)}$$

$$p(A_{i_0,i_1,...,i_{T-1},i}, B, T < \infty, X_T = i) = \sum_{m=0}^{\infty} p(A_{i_0,i_1,...,i_{m-1},i}, B, T = m, X_m = i)$$

$$p(A_{i_0,i_1,...,i_{T-1},i}, B, T = m, X_m = i) = p(A_{i_0,i_1,...,i_{m-1},i}, B_m, T = m|X_m = i)p(X_m = i)$$

with $B_m = \{X_m = i, X_{m+1} = j_1, ..., X_{m+n} = j_n\}$. Since $\{A_{i_0,i_1,...,i_{T-1},i}, T = m\} \in \sigma(X_0, ..., X_m)$ and $B_m \in \sigma(X_m, ..., X_{m+n})$, according to Theorem 3.4, we have

$$p(A_{i_0,i_1,...,i_{T-1},i}, B_m, T = m|X_m = i) = p(A_{i_0,i_1,...,i_{T-1},i}, T = m|X_m = i)p(B_m|X_m = i)$$

Because of the homogeneity of the Markov chain,

$$p(B_m|X_m = i) = p_{ij_1}p_{j_1j_2}...p_{j_{n-1}j_n}.$$

$$p(A_{i_0,i_1,...,i_{T-1},i}, B, T = m, X_T = i) = p(A_{i_0,i_1,...,i_{T-1},i}, T = m|X_m = i)p_{ij_1}p_{j_1j_2}...p_{j_{n-1}j_n}p(X_m = i)$$

$$p(A_{i_0,i_1,...,i_{T-1},i}, B, T = m, X_m = i) = p(A_{i_0,i_1,...,i_{T-1},i}, T = m)p_{ij_1}p_{j_1j_2}...p_{j_{n-1}j_n}$$

Hence

$$p(A_{i_0,i_1,...,i_{T-1},i}, B, T < \infty, X_T = i) = \sum_{m=0}^{\infty} p(A_{i_0,i_1,...,i_{T-1},i}, T = m)p_{ij_1}p_{j_1j_2}...p_{j_{n-1}j_n}.$$

$$p(A_{i_0,i_1,...,i_{T-1},i}, B, T < \infty, X_T = i) = p(A_{i_0,i_1,...,i_{T-1},i}, T < \infty)p_{ij_1}p_{j_1j_2}...p_{j_{n-1}j_n}.$$

$$p(A_{i_0,i_1,...,i_{T-1},i}, B|T < \infty, X_T = i) = \frac{p(A_{i_0,i_1,...,i_{T-1},i}, T < \infty)p_{ij_1}p_{j_1j_2}...p_{j_{n-1}j_n}}{p(T < \infty, X_T = i)}$$

$$p(A_{i_0,i_1,...,i_{T-1},i}, B|T < \infty, X_T = i) = p(A_{i_0,i_1,...,i_{T-1},i}, T < \infty|T < \infty, X_T = i)p_{ij_1}p_{j_1j_2}...p_{j_{n-1}j_n}$$

$$p(B|T < \infty, X_T = i) = \sum_{i_0,i_1,...,i_{T-1}} p(A_{i_0,i_1,...,i_{T-1},i}, T < \infty|T < \infty, X_T = i)p_{ij_1}p_{j_1j_2}...p_{j_{n-1}j_n}$$

$$p(B|T < \infty, X_T = i) = p(\bigcup_{i_0,i_1,...,i_{T-1}} A_{i_0,i_1,...,i_{T-1},i}, T < \infty|T < \infty, X_T = i)p_{ij_1}p_{j_1j_2}...p_{j_{n-1}j_n}$$

Finally,

$$p(B|T < \infty, X_T = i) = p_{ij_1}p_{j_1j_2}...p_{j_{n-1}j_n}.$$

Theorem 3.11 (Other variant of the strong Markov property)
For any two events $A \subset \sigma(X_1, ..., X_T)$ and $B \subset \sigma(X_T, ..., X_{T+n})$, we have

$$p(A \bigcap B | X_T = i, T < \infty) = p(A | X_T = i, T < \infty) p(B | X_T = i, T < \infty).$$

Proof.
We were able to prove that for a constant T, 3.10 and 3.11 are equivalent. Using a similar reasoning, we can show this is also true if T is a Markov time.

Example 3.10 Let's consider an irreducible MC $(X_n)_{n\geq 0}$ on the finite state space I, P being its transition matrix. We suppose that $(X_n)_{n\geq 0}$ has no absorbing state. We define the following sequence:

$$\tau_0 := 0 \text{ and } \tau_{m+1} := inf\{n > \tau_m : X_n \neq X_{\tau_m}\}$$

for $m = 0, 1, 2 ...$ (τ_m) is the sequence of times at which the Markov chain $(X_n)_{n\geq 0}$ jumps into a new position. It is easy to see that τ_m is a Markov time because $\{\tau_m = k\}$ is entirely determined by $(X_0, ..., X_k)$. Since $(X_n)_{n\geq 0}$ has no absorbing state, $\forall m \geq 0$, p($\tau_m < \infty$) = 1.

We'll now define $Z_n := X_{\tau_n}$ for $n = 0, 1, ...$ It is, so to speak, the succession of all **new** values. We have because of Theorem 3.11,

$$\begin{aligned}
&p(Z_{n+1} = i_{n+1} | Z_0 = i_0, ..., Z_n = i_n) = p(X_{\tau_{n+1}} = i_{n+1} | X_{\tau_0} = i_0, ..., X_{\tau_n} = i_n) \\
&= \frac{p(X_{\tau_{n+1}} = i_{n+1}, X_{\tau_0} = i_0, ..., X_{\tau_n} = i_n)}{p(X_{\tau_0} = i_0, ..., X_{\tau_n} = i_n)} \\
&= \frac{p(X_{\tau_{n+1}} = i_{n+1}, X_{\tau_0} = i_0, ..., X_{\tau_n} = i_n, \tau_n < \infty)}{p(X_{\tau_0} = i_0, ..., X_{\tau_n} = i_n, \tau_n < \infty)}
\end{aligned}$$

because $p(\tau_n < \infty) = 1, \forall n > 0$. We have

$$\begin{aligned}
&p(X_{\tau_{n+1}} = i_{n+1}, X_{\tau_0} = i_0, ..., X_{\tau_n} = i_n, \tau_n < \infty) = \\
&p(X_{\tau_{n+1}} = i_{n+1}, X_{\tau_0} = i_0, ...X_{\tau_n} = i_n | X_{\tau_n} = i_n, \tau_n < \infty) p(X_{\tau_n} = i_n, \tau_n < \infty) \\
&= p(X_{\tau_{n+1}} = i_{n+1} | X_{\tau_n} = i_n, \tau_n < \infty) p(X_{\tau_0} = i_0, ...X_{\tau_n} = i_n | X_{\tau_n} = i_n, \tau_n < \infty) \\
&\times p(X_{\tau_n} = i_n, \tau_n < \infty) \\
&= p(X_{\tau_{n+1}} = i_{n+1} | X_{\tau_n} = i_n, \tau_n < \infty) p(X_{\tau_0} = i_0, ...X_{\tau_n} = i_n, \tau_n < \infty)
\end{aligned}$$

Because of the strong Markov property, if $\tau_n < \infty$ and $X_{\tau_n} = i_n$, $(X_{\tau_n+m})_{m\geq 0}$ is a (δ_{i_n}, P) Markov chain.

$$
\begin{aligned}
&p(Z_{n+1} = i_{n+1}|Z_0 = i_0, ..., Z_n = i_n) = \\
&= \frac{p(X_{\tau_{n+1}} = i_{n+1}|X_{\tau_n} = i_n, \tau_n < \infty)p(X_{\tau_0} = i_0, ...X_{\tau_n} = i_n, \tau_n < \infty)}{p(X_{\tau_0} = i_0, ..., X_{\tau_n} = i_n, \tau_n < \infty)} \\
&= p(X_{\tau_{n+1}} = i_{n+1}|X_{\tau_n} = i_n, \tau_n < \infty) \\
&= p(X_{\tau_{n+1}} = i_{n+1}|X_{\tau_n} = i_n) \\
&= p(Z_{n+1} = i_{n+1}|Z_n = i_n) \\
&= p(X_{\tau_1} = i_{n+1}|X_0 = i_n) = p(Z_1 = i_{n+1}|Z_0 = i_n)
\end{aligned}
$$

because $\{\tau_n < \infty\}$ is a sure event and because of the strong Markov property which stipulates that given $\tau_n < +\infty$ and $X_{\tau_n} = i_n$, $(X_{\tau_n+m})_{m\geq 0}$ is a (δ_{i_n}, P)-MC independent of $X_0, ..., X_{\tau_n}$. $(Z_n)_{n\geq 0}$ is thus an homogeneous Markov chain. Since $Z_0 = X_{\tau_0} = X_0$, $(Z_n)_{n\geq 0}$ has the same initial distribution as X_0.

We introduce the matrix $P^{\sim}$ such that

$$p(Z_{n+1} = i_{n+1}|Z_0 = i_0, ..., Z_n = i_n) = p(Z_{n+1} = i_{n+1}|Z_n = i_n) = p^{\sim}_{i_n i_{n+1}}$$

whereby $p^{\sim}_{ii} = 0$ and for $i \neq j$

$$p^{\sim}_{ij} = p(X_{\tau_{n+1}} = j|X_{\tau_n} = i) = p(X_{\tau_1} = j|X_0 = i)$$

which is the probability that we go from state i to state j knowing that the next step cannot be i (because of the definition of τ_{n+1}). We thus have for $i \neq j$

$$
\begin{aligned}
&p^{\sim}_{ij} = p(X_{\tau_1} = j|X_0 = i) = \frac{p(X_{\tau_1} = j, X_0 = i)}{p(X_0 = i)} \\
&= \frac{\sum_{n=1}^{\infty} p(X_n = j, X_0 = i, \tau_1 = n)}{p(X_0 = i)} = \frac{\sum_{n=1}^{\infty} p(X_0 = i, , ..., X_{n-1} = i, X_n = j)}{p(X_0 = i)} \\
&= \frac{\sum_{n=1}^{\infty} p(X_0 = i)p_{ii}^{n-1}p_{ij}}{p(X_0 = i)} = \sum_{n=1}^{\infty} p_{ii}^{n-1}p_{ij} \\
&= \frac{p_{ij}}{1 - p_{ii}} = \frac{p_{ij}}{\sum_{k\neq i} p_{ik}}
\end{aligned}
$$

We can easily check that $\sum_{j\in I} p^{\sim}_{ij} = 1$.
$(Z_n)_{n\geq 0}$ is thus a Markov chain with the transition matrix $P^{\sim} = (p^{\sim}_{ij})_{i,j\in I}$.

3.4 recurrence and transiency

Let $(X_n)_{n\geq 0}$ be an homogeneous Markov chain with a transition matrix P on the state space I. Let's consider the first return time in i defined above:

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

For $i, j \in I$, we define

$$f_{ii}^{(n)} := p_i(T_i = n) = p_i(X_n = i, X_k \neq i, 1 \leq k \leq n-1)$$

$$f_{ij}^{(n)} := p_i(T_j = n) = p_i(X_n = j, X_k \neq j, 1 \leq k \leq n-1)$$

and

$$f_{ij} := \sum_{n=1}^{\infty} f_{ij}^{(n)}.$$

We then have

$$p_i(T_j < \infty) = p_i\Big(\bigcup_{n=1}^{\infty}\{T_j = n\})\Big) = \sum_{n=1}^{\infty} p_i(T_j = n) = f_{ij}.$$

As a consequence, f_{ij} is the probability that the Markov chain starting in i will ever visit state j after the beginning.

Definition 3.10 (Transiency, recurrence)

A state i is called *recurrent* if and only if $f_{ii} = 1$.

A state i is called *transient* if and only if $f_{ii} < 1$.

A recurrent state i is called *positive recurrent* if and only if $E_i[t_i] < \infty$ and *null recurrent* otherwise.

The different situations can be visualised in Figure 3.8. From an intuitive standpoint, if $f_{ii} = 1$, state i is bound to reoccur an endless number of times. If $f_{ii} < 1$, state i is bound to reoccur only a finite number of times, because every time i is reached, it can fail to be reached again with a probability of $1 - f_{ii} > 0$. If $f_{ii} = 0$, state i can only be the starting point.

In what follows, we'll abbreviate *infinitely often* by *i.o.*

Lemma 3.6 For all $i, j \in I$, we have

$$p_i(X_n = j \quad i.o.) = \begin{cases} 0 & if \quad f_{jj} < 1 \\ f_{ij} & if \quad f_{jj} = 1 \end{cases} \tag{3.12}$$

Proof.

Let us define the arrival time N_1 and the **return times** $N_2, ..., N_k$ such that $1 \leq N_1 < ... < N_k$ and $X_{N_1} = X_{N_2} = ... = X_{N_k} = j$ and $X_m \neq j$ for $m \in \{1, 2, ..., N_k\}\backslash\{N_1, N_2, ..., N_k\}$. As we saw above, arrival and return times are Markov times.

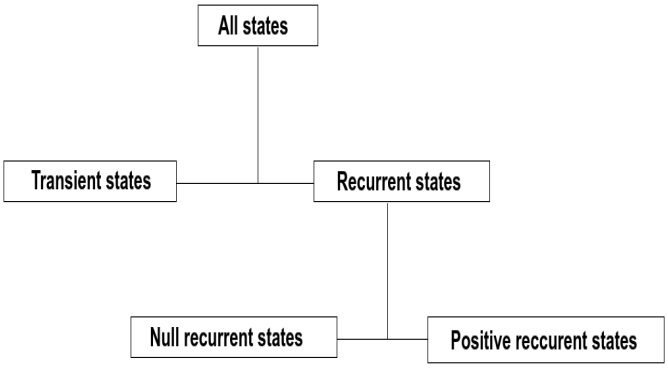


Figure 3.8: Classification of the states according to their asymptotic properties

Let us further define $A^{(n_1,...,n_k)} = \{N_1 = n_1, ..., N_k = n_k\}$. We want to prove recursively that

$$p_i(A^{(n_1,...,n_k)}) = f_{ij}^{(n_1)} f_{jj}^{(n_2-n_1)} ... f_{jj}^{(n_k-n_{k-1})}.$$

For $k = 1$, we obviously have $p_i(A^{(n_1)}) = f_{ij}^{(n_1)}$ because of the very definition of $f_{ij}^{(n_1)}$. For $i = j$, we have

$$p_j(A^{(n_1,...,n_k)}) = f_{jj}^{(n_1)} f_{jj}^{(n_2-n_1)} ... f_{jj}^{(n_k-n_{k-1})}.$$

We now want to show that the relationship holds for $k + 1$.

$$p_i(A^{(n_1,...,n_k,n_{k+1})}) = \frac{p(A^{(n_1,...,n_k,n_{k+1})})}{p(X_0 = i)}.$$

We can decompose $A^{(n_1,...,n_k,n_{k+1})}$ into

$$B = \{X_{n_1} = ... = X_{n_k} = j, X_n \neq j, \text{ for n in between}\} \in \sigma(X_0, X_1, ..., X_{n_k})$$

and

$$C = \{X_{n_k} = X_{n_{k+1}} = j, X_n \neq j, \ n_k < n < n_{k+1}\} \in \sigma(X_{n_k}, X_{n_k+1}, ..., X_{n_{k+1}}).$$

According to Theorem 3.4, we know that

$$p(A^{(n_1,...,n_k,n_{k+1})} | X_{n_k} = j) = p(B|X_{n_k} = j) p(C|X_{n_k} = j)$$

$$\frac{p(A^{(n_1,...,n_k,n_{k+1})})}{p(X_{n_k} = j)} = \frac{p(B)}{p(X_{n_k} = j)} p(C|X_{n_k} = j)$$

$$p(A^{(n_1,...,n_k,n_{k+1})}) = p(B) p(C|X_{n_k} = j)$$

$$p_i(A^{(n_1,...,n_k,n_{k+1})}) = \frac{p(B)}{p(X_0 = i)} p(C|X_{n_k} = j)$$

$$\frac{p(B)}{p(X_0 = i)} = p_i(A^{(n_1,...,n_k)}) = f_{ij}^{(n_1)} f_{jj}^{(n_2-n_1)} ... f_{jj}^{(n_k-n_{k-1})}$$

Given that $N_k = n_k < \infty$ and $X_{n_k} = j$, we know that $(X_{N_k+n})_{n\geq 0}$ is a (δ_j, P) Markov chain, P being the transition matrix of the MC $(X_n)_{n\geq 0}$. As a consequence, we have

$$p(C \mid X_{n_k} = j) = p(X_{n_{k+1}-n_k} = X_0 = j, X_n \neq j,\ 0 < n < n_{k+1} - n_k \mid X_0 = j)$$

$$p(C \mid X_{n_k} = j) = f_{jj}^{(n_{k+1}-n_k)}$$

Finally, we find that

$$p_i(A^{(n_1,...,n_k,n_{k+1})}) = f_{ij}^{(n_1)} f_{jj}^{(n_2-n_1)} ... f_{jj}^{(n_k-n_{k-1})} f_{jj}^{(n_{k+1}-n_k)}.$$

The relationship is thus valid $\forall k \geq 1$.

The probability that the state i is visited at least k times is equal to the sum of $p_i(A^{(n_1,...,n_k)})$ over all possible tuples $\{n_1, ..., n_k\}$.

$$p_i(X_n = j, \text{ at least } k \text{ times}) = \sum_{n_1 \geq 1} f_{ij}^{n_1} \sum_{n_2 \geq n_1} f_{jj}^{n_2-n_1} ... \sum_{n_k \geq n_{k-1}} f_{jj}^{n_k-n_{k-1}} = f_{ij} f_{jj}^{k-1}.$$

The last equality can be easily proven by considering the definition of f_{ij} and f_{jj} just above and performing a change of variable for each sum while beginning with the last one. For $k \to +\infty$, we obtain the result we've sought to prove.

For $j = i$, we see that the event $\{X_n = i \quad i.o.\}$ follows a binary probabilistic law:

$$p_i(X_n = i \quad i.o.) = \begin{cases} 0 & if \quad f_{ii} < 1 \\ 1 & if \quad f_{ii} = 1 \end{cases}$$

However, we must always keep in mind that the events $\{X_n = i\}$ themselves are not independent.

Theorem 3.12 (Equivalent characterisation of recurrence and transiency)

(i) The recurrence of i is equivalent to $p_i(X_n = i\ i.o.) = 1$ and $\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$.

(ii) The transiency of i is equivalent to $p_i(X_n = i\ i.o.) = 0$ and $\sum_{n=1}^{\infty} p_{ii}^{(n)} < \infty$.

Proof.

Let us first prove the first equivalency. Lemma 3.6 demonstrates that if i is recurrent ($f_{ii} = 1$), we then have $p_i(X_n = i \quad i.o.) = 1$. If i is transient ($f_{ii} < 1$), we have $p_i(X_n = i \quad i.o.) = 0$. On the other hand,

if $p_i(X_n = i \quad i.o.) = 0$, we cannot have $f_{ii} = 1$ but must have $f_{ii} < 1$. Likewise, if $p_i(X_n = i \quad i.o.) = 1$, we cannot have $f_{ii} < 1$ but must have $f_{ii} = 1$.

We must now show that $\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty \Leftrightarrow f_{ii} = 1$ Since $f_{ii} \in [0, 1]$, if the latter equivalency is true, $\sum_{n=1}^{\infty} p_{ii}^{(n)} < \infty$ implies that $f_{ii} < 1$. Similarly, if $f_{ii} < 1$, we necessarily have $\sum_{n=1}^{\infty} p_{ii}^{(n)} < \infty$.

We have

$$p_{ii}^{(n)} = \frac{p(X_n = i)}{p(X_0 = i)} = \frac{\sum_{k=0}^{n-1} p(X_n = i \cap T_i = n - k)}{p(X_0 = i)}$$

$$p_{ii}^{(n)} = \sum_{k=0}^{n-1} \frac{p(X_1 \neq i, ..., X_{n-k-1} \neq i, X_{n-k} = i, X_n = i)}{p(X_0 = i)}$$

$$p_{ii}^{(n)} = \sum_{k=0}^{n-1} p(X_n = i | X_1 \neq i, ..., X_{n-k-1} \neq i, X_{n-k} = i) p_i(X_1 \neq i, ..., X_{n-k-1} \neq i, X_{n-k} = i)$$

Thanks to the strong Markov property applied to the first return time and the homogeneity of the MC, we have

$$p_{ii}^{(n)} = \sum_{k=0}^{n-1} p_i(X_k = i) f_{ii}^{(n-k)} = \sum_{k=0}^{n-1} p_{ii}^{(k)} f_{ii}^{(n-k)}$$

Through a change of index $k' = n - k$, we find that

$$p_{ii}^{(n)} = \sum_{k=1}^{n} f_{ii}^{(k)} p_{ii}^{(n-k)}.$$

$$\sum_{n=1}^{\infty} p_{ii}^{(n)} = \sum_{n=1}^{\infty} \sum_{k=1}^{n} f_{ii}^{(k)} p_{ii}^{(n-k)}.$$

We know that $\{k \in [1, n], \quad n \in [1, +\infty]\} = \{k \in [1, +\infty], \quad n \in [k, +\infty]\}$. Hence

$$\sum_{n=1}^{\infty} p_{ii}^{(n)} = \sum_{k=1}^{\infty} \sum_{n=k}^{\infty} f_{ii}^{(k)} p_{ii}^{(n-k)}.$$

$$\sum_{n=1}^{\infty} p_{ii}^{(n)} = \sum_{k=1}^{\infty} f_{ii}^{(k)} \sum_{n=k}^{\infty} p_{ii}^{(n-k)} = \sum_{k=1}^{\infty} f_{ii}^{(k)} \sum_{n=0}^{\infty} p_{ii}^{(n)}$$

$$\sum_{n=1}^{\infty} p_{ii}^{(n)} = f_{ii} \sum_{n=0}^{\infty} p_{ii}^{(n)} = f_{ii} \left(1 + \sum_{n=1}^{\infty} p_{ii}^{(n)} \right)$$

If i is recurrent ($f_{ii} = 1$), we then have $\sum_{n=1}^{\infty} p_{ii}^{(n)} = 1 + \sum_{n=1}^{\infty} p_{ii}^{(n)}$. This is only possible if $\sum_{n=1}^{\infty} p_{ii}^{(n)} = +\infty$.

Let's now suppose that $\sum_{n=1}^{\infty} p_{ii}^{(n)} = +\infty$. Since $p_{ii}^{(n)} = \sum_{k=1}^{n} f_{ii}^{(k)} p_{ii}^{(n-k)}$,

$$\sum_{n=1}^{N} p_{ii}^{(n)} = \sum_{n=1}^{N} \sum_{k=1}^{n} f_{ii}^{(k)} p_{ii}^{(n-k)}$$

By reorganising the indices as we did before

$$\sum_{n=1}^{N} p_{ii}^{(n)} = \sum_{k=1}^{N} f_{ii}^{(k)} \sum_{n=k}^{N} p_{ii}^{(n-k)} \leq \sum_{k=1}^{N} f_{ii}^{(k)} \sum_{l=0}^{N} p_{ii}^{(l)},$$

Therefore

$$f_{ii} = \sum_{k=1}^{\infty} f_{ii}^{(k)} \geq \sum_{k=1}^{N} f_{ii}^{(k)} \geq \frac{\sum_{n=1}^{N} p_{ii}^{(n)}}{\sum_{n=0}^{N} p_{ii}^{(n)}}.$$

$$\frac{\sum_{n=1}^{N} p_{ii}^{(n)}}{\sum_{n=0}^{N} p_{ii}^{(n)}} = \frac{\sum_{n=1}^{N} p_{ii}^{(n)}}{\sum_{n=1}^{N} p_{ii}^{(n)} + p_{ii}(0)}$$

Finally

$$f_{ii} \geq \frac{\sum_{n=1}^{N} p_{ii}^{(n)}}{\sum_{n=1}^{N} p_{ii}^{(n)} + 1} \to 1, \text{ if N } \to +\infty.$$

Since this inequality is valid for all values of N, we can conclude that $\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty \Leftrightarrow f_{ii} = 1$.

Example 3.11 (Random walk on $\mathbb{Z}$)

Let us consider the independent identically distributed random variables $(\varepsilon_1, \varepsilon_2, ...)$ with

$$p(\varepsilon_1 = 1) = p \text{ and } p(\varepsilon_1 = -1) = q = 1 - p, p \in]0, 1[.$$

We're interested in the random walk $(X_n)_{n \geq 0}$ defined through

$$X_0 = 0 \text{ and } X_n = \sum_{i=1}^{n} \varepsilon_i \text{ for } n \geq 1.$$

Obviously, the random walk is irreducible because every state can be attained from any other state with a positive probability. For any $(a, b) \in \mathbb{Z}^2$, if $X_0 = a$ and $n = b - a > 0$,

$$p(X_n = b | X_0 = a) = p(X_0 = a, X_1 = a+1, X_2 = a+2, ..., X_n = b | X_0 = a) = p^n \geq 0.$$

Likewise, if $n = b - a < 0$,

$$p(X_n = b | X_0 = a) = p(X_0 = a, X_1 = a-1, X_2 = a-2, ..., X_n = b | X_0 = a) = (1-p)^n \geq 0.$$

As we shall see later on, transiency and recurrence are class properties. Thus, for an irreducible Markov chain, it suffices to prove that one single

state (such as "0") is recurrent or transient to show that the whole chain is recurrent or transient. To that end, we'll use theorem 3.12. For all $n \in \mathbb{N}$, $p_{00}^{(2n+1)} = 0$. Indeed, to return to the state 0, we would need k positive changes ($\varepsilon_i = 1$) and k negative changes ($\varepsilon_i = -1$) which is impossible for an odd number such as $2n+1$. Consequently, we only need to consider the transition probability at even time steps. The Stirling formula is given by

$$n! \sim \sqrt{2\pi n} n^n e^{-n} \text{ for } n \to \infty.$$

$a_n \sim b_n$ signifies that $lim_{n\to\infty} \frac{a_n}{b_n} = 1$.

For the Markov chain starting in 0, there are $\binom{2n}{n}$ possibilities to come back to 0. Consequently,

$$p_{00}^{(2n)} = \binom{2n}{n} p^n q^n = \frac{(2n)!}{(n!)^2} (pq)^n.$$

$$p_{00}^{(2n)} \sim \frac{\sqrt{4\pi n}(2n)^{2n} e^{-2n}}{2\pi n n^{2n} e^{-2n}} (pq)^n = \frac{1}{\sqrt{n\pi}} (4pq)^n = \frac{1}{\sqrt{n\pi}} (4p(1-p))^n.$$

The maximum of $(4pq)^n$ is obtained for $p = 1/2$ and is equal to 1. For $p \neq 1/2$, $(4pq)^n < 1$.

Thus, for $p = 1/2$,

$$\sum_{n=1}^{\infty} p_{00}^{(2n)} \sim \frac{1}{\sqrt{\pi}} \sum_{n=1}^{\infty} \frac{1}{n^{0.5}}.$$

We recognise an hyperharmonic series which diverges because $0.5 \leq 1$.

For $p \neq 1$,

$$\sum_{n=1}^{\infty} p_{00}^{(2n)} \sim \frac{1}{\sqrt{\pi}} \sum_{n=1}^{\infty} \frac{(4pq)^n}{n^{0.5}} \leq \frac{1}{\sqrt{\pi}} \sum_{n=1}^{\infty} (4pq)^n$$

We recognise a geometrical series which converges by virtue of the fact that $4pq < 1$.

To conclude, a symmetrical random walk on $\mathbb{Z}$ is recurrent. An asymmetrical random walk on $\mathbb{Z}$ is transient.

Theorem 3.13 (Theorem of Polya) A symmetrical random walk on $\mathbb{Z}^d$ is a Markov chain $X = (X_n)_{n\geq 0}$ whose transition probability is given by

$$p_{xy} = p(X_{n+1} = y | X_n = x) = \begin{cases} \frac{1}{2d} & if \quad y \in \{x \pm e_i : i = 1, ..., d\} \\ 0: & \text{otherwise} \end{cases}$$

whereby e_i is the i-th unit vector. The particle chooses thus one of its $2d$ neighbours and jumps there.

The theorem of Polya stipulates that a **symmetrical** random walk X is recurrent for $d \leq 2$ and transient for $d \geq 3$. In Example 3.11, we just

proved this is the case for $d = 1$. As we did before, we can easily prove that any random walk is irreducible as it is always possible to find one path from one point to the other by selecting the values of the jumps $\pm e_i$ in a suitable way. To show that the chain is recurrent or transient, we thus just need to consider 0^d. As for $d = 1$, $a_{2n+1}^d = p_{00}^{(2n+1)} = 0$ because the random walk needs to take the same number of steps along e_i than along $-e_i$ to get back to the origin.

We now want to prove that

$$\sum_{n=1}^{\infty} a_{2n}^{(d)} = \begin{cases} = \infty & : \quad d = 2 \\ < \infty & : \quad d = 3 \end{cases}$$

Example 3.12 (Proof for d = 2)

In order to return to the origin after $2n$ steps, the MC needs to take as many steps u to the west as to the east and as many steps $n - u$ to the north as to the south. For a given $u \in [0, n]$, the total number of trajectories leading back to the starting point is given by the multinomial coefficient

$$\binom{2n}{u, u, n-u, n-u} = \binom{u}{u}\binom{2u}{u}\binom{n+u}{n-u}\binom{2n}{n+u}.$$

$$\binom{2n}{u, u, n-u, n-u} = \frac{(2u)!}{u!u!}\frac{(n+u)!}{(n-u)!(2u)!}\frac{(2n)!}{(n+u)!(n-u)!}$$

$$\binom{2n}{u, u, n-u, n-u} = \frac{(2n)!}{u!u!(n-u)!(n-u)!}$$

We also have $\binom{2n}{n} = \frac{(2n)!}{n!n!}$, hence $(2n!) = \binom{2n}{n} n!n!$. This leads to

$$\binom{2n}{u, u, n-u, n-u} = \frac{(2n)!}{u!u!(n-u)!(n-u)!} = \binom{2n}{n}\binom{n}{u}\binom{n}{n-u}.$$

$a_{2n}^{(2)}$ is obtained by considering the combination of all possible values of $u \in [0, n]$. We finally get

$$a_{2n}^{(2)} = \sum_{u=0}^{n} \binom{2n}{n}\binom{n}{u}\binom{n}{n-u}\frac{1}{4^{2n}}.$$

$$a_{2n}^{(2)} = \binom{2n}{n}\frac{1}{4^{2n}}\sum_{u=0}^{n}\binom{n}{u}\binom{n}{n-u}$$

According to the combinatorial identity, we have

$$\sum_{u=0}^{n}\binom{n}{u}\binom{n}{n-u} = \binom{2n}{n}.$$

As a consequence,

$$a_{2n}^{(2)} = \binom{2n}{n}^2 \frac{1}{4^{2n}}.$$

Through the use of the Stirling formula, we can prove that $a_{2n}^{(2)} \sim \frac{1}{\pi n}$ so that $\sum_{n=1}^{\infty} a_{2n}^{(2)} = \infty$. This shows that the MC is recurrent.

Example 3.13 (Proof for d = 3)

In that case, in order to return to the origin after $2n$ steps, the random walk must take as many steps u to the west as to the east, as many steps v to the north as to the south and as many steps upwards w as downwards. For a given triplet (u, v, w), the total number of trajectories leading back to the starting point is given by the multinomial coefficient

$$\binom{2n}{u,u,v,v,w,w} = \binom{2u}{u}\binom{2u+v}{v}\binom{2u+2v}{v}\binom{2u+2v+w}{w}\binom{2n}{w}.$$

$$\binom{2n}{u,u,v,v,w,w} = \binom{2u}{u}\binom{2u+v}{v}\binom{2u+2v}{v}\binom{2u+2v+w}{w}\binom{2n}{w}.$$

$$\binom{2n}{u,u,v,v,w,w} = \frac{(2u)!(2u+v)!(2u+2v)!(2u+2v+w)!(2n)!}{u!u!v!(2u)!v!(2u+v)!w!(2u+2v)!w!(2n-w)!}$$

$$\binom{2n}{u,u,v,v,w,w} = \frac{(2n)!}{u!u!v!v!w!w!}$$

Like for $d = 2$, the probability that the chain returns to the origin after $2n$ steps involves all possible combinations of (u, v, w). Therefore,

$$p_{00}^{(2n)} = a_{2n}^{(3)} = \sum_{\substack{u,v,w\geq 0 \\ u+v+w=n}} \frac{(2n)!}{u!u!v!v!w!w!} \frac{1}{6^{2n}}$$

Since $(2n!) = \binom{2n}{n} n!n!$,

$$a_{2n}^{(3)} = \sum_{\substack{u,v,w\geq 0 \\ u+v+w=n}} \binom{2n}{n} \frac{n!n!}{u!u!v!v!w!w!} \frac{1}{6^{2n}}$$

$$a_{2n}^{(3)} = \binom{2n}{n} \frac{1}{2^{2n}} \frac{1}{3^n} \sum_{\substack{u,v,w\geq 0 \\ u+v+w=n}} \frac{n!n!}{u!u!v!v!w!w!} \frac{1}{3^n}$$

$$\binom{n}{u,v,w} = \binom{(u+v)!}{v!}\binom{n!}{w!} = \frac{(u+v)!n!}{v!u!w!(u+v)!} = \frac{n!}{v!u!w!} =$$

Finally, we find that

$$a_{2n}^{(3)} = \binom{2n}{n} \frac{1}{2^{2n}} \frac{1}{3^n} \sum_{\substack{u,v,w \geq 0 \\ u+v+w=n}} \binom{n}{u,v,w}^2 \frac{1}{3^n}$$

According to the multinomial theorem,

$$(1+1+1)^n = \sum_{\substack{u,v,w \geq 0 \\ u+v+w=n}} \binom{n}{u,v,w} 1^u 1^v 1^w$$

$$(3)^n = \sum_{\substack{u,v,w \geq 0 \\ u+v+w=n}} \binom{n}{u,v,w}$$

$$\sum_{\substack{u,v,w \geq 0 \\ u+v+w=n}} \binom{n}{u,v,w} \frac{1}{3^n} = 1$$

The latter formula can also be seen as the total sum of the multinomial distribution $\left(n, (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})\right)$.

Furthermore, for every $m \in \mathbb{N}$,

$$\binom{3m}{u,v,w} \leq \binom{3m}{m,m,m} \text{ for all } u,v,w \geq 0 \text{ with } u+v+w = 3m.$$

We have

$$\sum_{n=1}^{\infty} a_{2n}^{(3)} = \sum_{m=1}^{\infty} (a_{2*3m}^{(3)} + a_{2*(3m-1)}^{(3)} + a_{2*(3m-2)}^{(3)} = \sum_{m=1}^{\infty} (a_{6m}^{(3)} + a_{6m-2}^{(3)} + a_{6m-4}^{(3)}).$$

We also have

$$\sum_{\substack{u,v,w \geq 0 \\ u+v+w=3m}} \binom{3m}{u,v,w} \frac{1}{3^{3m}} = 1$$

and

$$a_{6m}^{(3)} = \binom{6m}{3m} \frac{1}{2^{6m}} \frac{1}{3^{3m}} \sum_{\substack{u,v,w \geq 0 \\ u+v+w=3m}} \binom{3m}{u,v,w}^2 \frac{1}{3^{3m}}$$

$$a_{6m}^{(3)} \leq \binom{6m}{3m} \frac{1}{2^{6m}} \frac{1}{3^{3m}} \sum_{\substack{u,v,w \geq 0 \\ u+v+w=3m}} \binom{3m}{m,m,m} \binom{3m}{u,v,w} \frac{1}{3^{3m}}$$

Thanks to the formula of Stirling, we can write that

$$a_{6m}^{(3)} \leq \binom{6m}{3m} \frac{1}{2^{6m}} \frac{1}{3^{3m}} \binom{3m}{m,m,m} \sim C \frac{1}{m^{3/2}}.$$

As a consequence $\sum_{m=1}^{\infty} a_{6m}^{(3)} < \infty$ because $3/2 > 1$.

Thanks to the Chapman-Komolgorov equation, we have

$$p_{00}^{(6m)} \geq p_{00}^{(2)} p_{00}^{(6m-2)} \text{ with } p_{00}^{(2)} = 6(1/6)^2 = 1/6$$

so that

$$p_{00}^{(6m)} \geq (1/6) p_{00}^{(6m-2)}$$

or in our notations

$$a_{6m}^{(3)} \geq (1/6) a_{6m-2}^{(3)}.$$

For the same reason, we also have

$$p_{00}^{(6m)} \geq p_{00}^{(4)} p_{00}^{(6m-4)} \text{ with } p_{00}^{(4)} > 0$$

, i.e.

$$a_{6m}^{(3)} \geq p_{00}^{(4)} a_{6m-4}^{(3)}.$$

Therefore, $\sum_{m=1}^{\infty} a_{6m-2}^{(3)} < \infty$ and $\sum_{m=1}^{\infty} a_{6m-4}^{(3)} < \infty$. This leads to

$$\sum_{m=1}^{\infty} \left(a_{2*3m}^{(3)} + a_{2*(3m-1)}^{(3)} + a_{2*(3m-2)}^{(3)} < \infty \right).$$

$$\sum_{n=1}^{\infty} a_{2n}^{(3)} < \infty.$$

The random walk is transient.

Theorem 3.14 Let $X = (X_n)_{n \geq 0}$ be an irreducible Markov chain. There are then only two alternatives:

(i) All states are transient, $p_i(\cup_{j \in I} \{X_n = j \ \ i.o.\}) = 0$ and $\sum_{n=1}^{\infty} p_{ij}^{(n)} < \infty$ for all $i, j \in I$.

(ii) All states are recurrent, $p_i(\cap_{j \in I} \{X_n = j \ \ i.o.\}) = 1$ and $\sum_{n=1}^{\infty} p_{ij}^{(n)} = \infty$ for all $i, j \in I$.

Consequently, transiency and recurrence are class properties.

Proof.

We'll first prove that if a state $i \in I$ is recurrent, every other state $j \in I$ is also recurrent and that if i is transient, j is transient as well.

Since the Markov chain is irreducible, all states of I are communicating. Consequently, there exists $n, m \in \mathbb{N}$ such that $p_{ij}^{(m)} > 0$ and $p_{ji}^{(n)} > 0$. Therefore, $\alpha = p_{ji}^{(n)} p_{ij}^{(m)} > 0$. Through an obvious generalisation of the Chapman-Komolgorov equation, we can write that for every $k \in \mathbb{N}_0$ we have

$$p_{ii}^{(m+k+n)} \geq p_{ij}^{(m)} p_{jj}^{(k)} p_{ji}^{(n)} = \alpha p_{jj}^{(k)}.$$

Likewise

$$p_{jj}^{(m+k+n)} \geq p_{ji}^{(n)} p_{ii}^{(k)} p_{ij}^{(m)} = \alpha p_{ii}^{(k)}.$$

For these reasons,

$$\sum_{k=1}^{\infty} \alpha p_{jj}^{(k)} \leq \sum_{k=1}^{\infty} p_{ii}^{(m+k+n)} = \sum_{k=1}^{\infty} p_{ii}^{(k)} - \sum_{k=1}^{m+n} p_{ii}^{(k)}.$$

If i is transient, $\sum_{k=1}^{\infty} p_{ii}^{(k)} < \infty$ so that j is transient as well.

We also have

$$\sum_{k=1}^{\infty} \alpha p_{ii}^{(k)} \leq \sum_{k=1}^{\infty} p_{jj}^{(m+k+n)} = \sum_{k=1}^{\infty} p_{jj}^{(k)} - \sum_{k=1}^{m+n} p_{jj}^{(k)}.$$

If i is recurrent ($\sum_{k=1}^{\infty} p_{ii}^{(k)} = \infty$), j must also be recurrent, i.e. non-transient. As a consequence, two states of an irreducible chain are either both transient or both recurrent.

Let us now consider the situation where the Markov chain is transient, i.e. $f_{jj} < 1$. According to Lemma 3.6, $p_i(X_n = j \ \ i.o.) = 0$ for all i and j so that $p_i(\cup_j \{X_n = j \ \ i.o.\}) = 0$. The first arrival time in state j is given by $S_j := inf\{n \geq 0 : X_n = j\}$.

$$p_{ij}^{(n)} = p(X_n = j | X_0 = i) = \frac{p(X_n = j, X_0 = i)}{p(X_0 = i)} = \frac{\sum_{k=0}^{n} p(X_n = j, S_j = k, X_0 = i)}{p(X_0 = i)}$$

Thanks to the Markov property, we can write that for $k \geq 1$ and $i \neq j$,

$$\begin{aligned}
p(X_0 = i, S_j = k, X_n = j) = & \sum_{i_1,...,i_{k-1} \neq j} \sum_{i_{k+1},...,i_{n-1} \in I} p(X_0 = i, X_1 = i_1, ..., X_{k-1} = i_{k-1}, \\
& X_k = j, X_{k+1} = i_{k+1}, ..., X_{n-1} = i_{n-1}, X_n = j) \\
= & \sum_{i_1,...,i_{k-1} \neq j} \sum_{i_{k+1},...,i_{n-1} \in I} \nu_i p_{ii_1} ... p_{i_{k-1}j} p_{ji_{k+1}} ... p_{i_{n-1}j} \\
= & \nu_i \sum_{i_1,...,i_{k-1} \neq j} p_{ii_1} ... p_{i_{k-1}j} \sum_{i_{k+1},...,i_{n-1} \in I} p_{ji_{k+1}} ... p_{i_{n-1}j} \\
= & \nu_i f_{ij}^{(k)} \sum_{i_{k+1},...,i_{n-1} \in I} \frac{p(X_k = j, X_{k+1} = i_{k+1}, ..., X_{n-1} = i_{n-1}, X_n = j)}{p(X_k = j)} \\
= & \nu_i f_{ij}^{(k)} \frac{p(X_k = j, X_n = j)}{p(X_k = j)} \\
= & \nu_i f_{ij}^{(k)} p(X_n = j | X_k = j) \\
= & p(X_0 = i) f_{ij}^{(k)} p_{jj}^{(n-k)} \\
\Rightarrow p_{ij}^{(n)} = & p(X_n = j | X_0 = i) = \sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)}
\end{aligned}$$

Consequently,

$$\sum_{n=1}^{\infty} p_{ij}^{(n)} = \sum_{n=1}^{\infty}\sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)} = \sum_{k=1}^{\infty}\sum_{n=k}^{\infty} f_{ij}^{(k)} p_{jj}^{(n-k)}$$
$$= \sum_{k=1}^{\infty} f_{ij}^{(k)} \sum_{n=k}^{\infty} p_{jj}^{(n-k)}$$
$$= \sum_{k=1}^{\infty} f_{ij}^{(k)} \sum_{n=0}^{\infty} p_{jj}^{(n)} = f_{ij} \sum_{n=0}^{\infty} p_{jj}^{(n)}$$
$$\leq \sum_{n=0}^{\infty} p_{jj}^{(n)}$$

If $i = j$, $p_{ij}^{(n)} = p_{jj}^{(n)}$ so that

$$\sum_{n=1}^{\infty} p_{ij}^{(n)} \leq \sum_{n=0}^{\infty} p_{jj}^{(n)}$$

is true under every circumstance.

If the Markov chain is transient, we have $\sum_{n=0}^{\infty} p_{jj}^{(n)} < \infty$ which implies that $\sum_{n=1}^{\infty} p_{ij}^{(n)} < \infty$.

We shall now consider the case of a recurrent Markov chain. We want to prove that $p_i(\cap_{j\in I}\{X_n = j \;\; i.o.\}) = 1$ and that $\sum_{n=1}^{\infty} p_{ij}^{(n)} = \infty$.

According to Theorem 3.12, $p_j(X_n = j \;\; i.o.) = 1$. Therefore, we can write that

$$p_{ji}^{(m)} = p_j(X_m = i) = p_j(X_m = i \cap \{X_n = j \;\; i.o.\})$$
$$= \sum_{n>m} p_j(X_m = i, X_{m+1} \neq j, ..., X_{n-1} \neq j, X_n = j \cap \{(X_{n+n2})_{n2>0} = j \;\; i.o.\})$$
$$\leq \sum_{n>m} p_j(X_m = i, X_{m+1} \neq j, ..., X_{n-1} \neq j, X_n = j)$$

$$p_j(X_m = i, X_{m+1} \neq j, ..., X_{n-1} \neq j, X_n = j) = \frac{p(X_m = i, X_{m+1} \neq j, ..., X_{n-1} \neq j, X_n = j)}{p(X_0 = j)}$$
$$= p(X_{m+1} \neq j, ..., X_{n-1} \neq j, X_n = j | X_m = i) p_j(X_m = i)$$
$$= p(X_1 \neq j, ..., X_{n-m-1} \neq j, X_{n-m} = j | X_0 = i) p_{ji}^{(m)}$$
$$= p_{ji}^{(m)} f_{ij}^{(n-m)}$$

As a consequence, we find that

$$p_{ji}^{(m)} \leq \sum_{n>m} p_{ji}^{(m)} f_{ij}^{(n-m)}$$
$$= p_{ji}^{(m)} f_{ij}$$
$$p_{ji}^{(m)} \leq p_{ji}^{(m)} f_{ij}.$$

Because of the irreducibility of the Markov chain, we know that there is at least one $m > 0$ such that $p_{ji}^{(m)} > 0$. Hence $f_{ij} \geq 1$. f_{ij} is the probability that a Markov chain starting in state i would ever visit state j. We thus have $f_{ij} = 1, \forall i, j \in I$. According to Lemma 3.6, for $f_{jj} = 1$, $p_i(X_n = j \ \ i.o.) = f_{ij} = 1$. Since this is valid for all $i, j \in I$, we have $p_i(\cap_{j \in I}\{X_n = j \ \ i.o.\}) = 1$.

Let us now suppose that $\sum_{n=1}^{\infty} p_{ij}^{(n)}$ converges in spite of the irreducibility of the chain. This means that

$$\sum_{n=1}^{\infty} p_{ij}^{(n)} = \sum_{n=1}^{\infty} p(X_n = j | X_0 = i) = \frac{\sum_{n=1}^{\infty} p(X_n = j, X_0 = i)}{p(X_0 = i)}$$

converges. According to the Lemma of Borel-Cantelli **??**, this entails that $p(\{X_n = j, X_0 = i\} \ \ i.o.) = 0)$. If we choose to use a Markov chain starting in state i ($p(X_0 = i) = 1$), this leads to $p(\{X_n = j\} \ \ i.o.) = 0$. This contradicts our hypothesis that the chain is recurrent.

Corollary 3.2 An irreducible Markov chain on a finite state space is recurrent.

Proof. Let $|I|$ be the number of states of the Markov chain. Let us suppose on the contrary that $\sum_{n=1}^{\infty} p_{ij}^{(n)} < \infty$ for all $i, j \in I$. As a consequence, for a fixed i there exists a j^* that maximises this sum, that is to say $\sum_{n=1}^{\infty} p_{ij}^{(n)} \leq \sum_{n=1}^{\infty} p_{ij^*}^{(n)}$. This leads to

$$\begin{aligned}\sum_{n=1}^{\infty} p_{ij}^{(n)} \leq \sum_{j \in I} \sum_{n=1}^{\infty} p_{ij}^{(n)} &\leq \sum_{j \in I} \sum_{n=1}^{\infty} p_{ij^*}^{(n)} \\ &\leq |I| \sum_{n=1}^{\infty} p_{ij^*}^{(n)} < \infty\end{aligned}$$

On the other hand, we also have $\sum_{j \in I} p_{ij}^{(n)} = 1$. This implies that

$$\sum_{j \in I} \sum_{n=1}^{\infty} p_{ij}^{(n)} = \sum_{n=1}^{\infty} \sum_{j \in I} p_{ij}^{(n)} = \infty.$$

This stands in contradiction to the implication of our hypothesis. As a consequence, we must conclude that $\sum_{n=1}^{\infty} p_{ij}^{(n)} = \infty$. Because the chain is irreducible, we can apply Theorem 3.14 which tells us that the Markov chain must be recurrent.

Remark 3.1 For all Markov chains on a finite state space, if state i is essential ($\forall j \in I, \exists m \in \mathbb{N}, \ p_{ji}^{(m)} > 0$), it is also recurrent. Likewise, an unessential state must be a transient state.

Example 3.14 For a simple symmetrical random walk on $\mathbb{Z}^d$, all states are essential. We saw that for $d \leq 2$, every state is visited an infinite number of times. However, for $d = 3$ the probability is equal to 0 which means that the sequence diverges to infinity. Remark 3.1 is thus only valid for finite spaces.

3.5 Invariant distributions

In this section, we'll consider invariant distributions and invariant measures of Markov chains. A measure on a countable state space I is a row vector with non-negative members. Prior to giving a formal definition, we shall go into a practical example.

Example 3.15 (Markov chain with two states) Let us consider a Markov chain on $I = \{1, 2\}$ whose transition matrix is given by

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

We suppose that $\alpha, \beta \in]0; 1[$ as the other cases aren't interesting. Let us consider the probability distribution

$$\pi = (\pi_1, \pi_2) = (\frac{1-\beta}{2-\alpha-\beta}, \frac{1-\alpha}{2-\alpha-\beta}).$$

We'll prove recurrently that

$$p(X_n = 1) = \pi_1 + (\alpha + \beta - 1)^n (p(X_0 = 1) - \pi_1)$$

and

$$p(X_n = 2) = 1 - p(X_n = 1) = \pi_2 + (\alpha + \beta - 1)^n (p(X_0 = 2) - \pi_2).$$

For $n = 0$, it is obviously true. Let us now suppose it is true for n.

$$\begin{aligned}
p(X_{n+1} = 1) &= p(X_{n+1} = 1|X_n = 1)p(X_n = 1) + p(X_{n+1} = 1|X_n = 2)p(X_n = 2) \\
&= \alpha\pi_1 + \alpha(\alpha + \beta - 1)^n (p(X_0 = 1) - \pi_1) + (1 - \beta)\pi_2 \\
&+ (1 - \beta)(\alpha + \beta - 1)^n (p(X_0 = 2) - \pi_2) \\
&= \pi_1 + (\alpha - 1)\pi_1 + (\alpha + \beta - 1)^{n+1} (p(X_0 = 1) - \pi_1) \\
&- (\beta - 1)(\alpha + \beta - 1)^n (p(X_0 = 1) - \pi_1) + (1 - \beta)\pi_2 \\
&+ (1 - \beta)(\alpha + \beta - 1)^n (p(X_0 = 2) - \pi_2) \\
&= \pi_1 + (\alpha + \beta - 1)^{n+1} (p(X_0 = 1) - \pi_1) + (\alpha - 1)\pi_1 \\
&- (\beta - 1)(\alpha + \beta - 1)^n (p(X_0 = 1) - \pi_1) \\
&+ (1 - \beta)\pi_2 + (1 - \beta)(\alpha + \beta - 1)^n (p(X_0 = 2) - \pi_2)
\end{aligned}$$

We must now prove that

$$A = (\alpha-1)\pi_1-(\beta-1)(\alpha+\beta-1)^n(p(X_0 = 1)-\pi_1)+(1-\beta)\pi_2+(1-\beta)(\alpha+\beta-1)^n(p(X_0 = 2)-\pi_2) = 0$$

$$\begin{aligned}
(\alpha - 1)\pi_1 + (1 - \beta)\pi_2 &= \\
&= \frac{(\alpha - 1)(1 - \beta)}{2 - \alpha - \beta} + \frac{(1 - \beta)(1 - \alpha)}{2 - \alpha - \beta} \\
&= 0
\end{aligned}$$

$$
\begin{aligned}
A &= -(\beta-1)(\alpha+\beta-1)^n(p(X_0=1)-\pi_1)+(1-\beta)(\alpha+\beta-1)^n(p(X_0=2)-\pi_2)\\
&= (1-\beta)(\alpha+\beta-1)^n(p(X_0=1)+p(X_0=2)-\pi_1-\pi_2)\\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
p(X_{n+1}=2) &= 1-p(X_{n+1}=1)\\
&= 1-\pi_1-(\alpha+\beta-1)^{n+1}(p(X_0=1)-\pi_1)\\
&= \pi_2-(\alpha+\beta-1)^{n+1}(1-p(X_0=2)-1+\pi_2)\\
&= \pi_2-(\alpha+\beta-1)^{n+1}(\pi_2-p(X_0=2))
\end{aligned}
$$

This has several important consequences:

- If the initial distribution is given by $(p(X_0=1),p(X_0=2))=\pi$, then for $n\to\infty$, then

$$\forall n>0,(p(X_n=1),p(X_n=2))=\pi \tag{3.13}$$

This means that the probability distribution of the two states remains always the same. Moreover,

$$
\begin{aligned}
\pi P &= \Big(\frac{(1-\beta)\alpha+(1-\alpha)(1-\beta)}{2-\alpha-\beta},\frac{(1-\beta)(1-\alpha)+(1-\alpha)\beta}{2-\alpha-\beta}\Big)\\
&= \Big(\frac{1-\beta}{2-\alpha-\beta},\frac{(1-\alpha)(1-\beta+\beta)}{2-\alpha-\beta}\Big)\\
&= \Big(\frac{1-\beta}{2-\alpha-\beta},\frac{(1-\alpha)}{2-\alpha-\beta}\Big)\\
&= \pi.
\end{aligned}
$$

This signifies that the Markov chain never changes the distribution π which is called **invariant**.

- If we start with any other distribution, it will surely converge exponentially fast towards π.

We can interpret π as an *equilibrium distribution*. (ii) tells us that any other initial distribution will approach exponentially fast the equilibrium distribution.

Definition 3.11 We say that a measure $(\pi_i)_{i\in I}$ is an invariant measure for the Markov chain with the transition matrix P if and only if

$$\pi P=\pi \tag{3.14}$$

If in addition to that $\sum_{i\in I}\pi_i=1$, then π is called an *invariant distribution*.

If π is an invariant distribution and if $X_0=\pi$, we have

$$p_\pi(X_n=i)=(\pi P^n)_i=(\pi PP^{n-1})_i=(\pi P^{n-1})_i=...=\pi_i=p_\pi(X_0=i).$$

Consequently, invariant distributions are also called stationary distributions. For all $m, n \in \mathbb{N}_0$,

$$(X_0, ..., X_m) \text{ and } (X_n, ..., X_{n+m})$$

have the same distribution (although they aren't, of course, independent).

Remark 3.2 Definition 3.11 implies that π is a left eigenvector of P whose eigenvalue is 1. As a consequence, we can use the means of linear algebra to find the invariant distribution. For that sake we must determine the left eigenvector of P whose eigenvalue is 1 and which has non-negative values with one of them being strictly positive. We can then normalise the eigenvector in such a way that we get a probability distribution.

In what follows, we'll need the Weierstrass M-test to deal with limits of sums over an infinite space.

Lemma 3.7 (Weierstrass M-test)
We consider a sequence $(x_{n,k})_{n,k=1,2...}$ such that

$$(i) \quad \forall k, \lim_{n\to\infty} x_{n,k} = x_k$$

$$(ii) \quad \forall n, k, \ |x_{n,k}| \leq M_k \text{ with } \sum_{k=1}^{\infty} M_k < \infty$$

We then know that $\sum_{k=1}^{\infty} x_k$ and $\sum_{k=1}^{\infty} x_{n,k}$ (for $n \in \mathbb{N}$) converge. Moreover

$$\lim_{n\to\infty} \sum_{k=1}^{\infty} x_{n,k} = \sum_{k=1}^{\infty} x_k. \tag{3.15}$$

Proof.
For each n, because of (ii) we know that

$$\sum_{k=1}^{\infty} |x_{n,k}| \leq \sum_{k=1}^{\infty} M_k < \infty$$

so that this series converges. Thanks to (i) and to the Lemma of Fatou [1], we have

$$\sum_{k=1}^{\infty} |x_k| = \sum_{k=1}^{\infty} |\lim_{n\to\infty} x_{n,k}| = \sum_{k=1}^{\infty} \liminf_{n\to\infty} |x_{n,k}|$$
$$\leq \liminf_{n\to\infty} \sum_{k=1}^{\infty} |x_{n,k}| \leq \liminf_{n\to\infty} \sum_{k=1}^{\infty} M_k = \sum_{k=1}^{\infty} M_k < \infty.$$

[1] See "Fatou's lemma" on Wikipedia.

We can conclude that $\sum_{k=1}^{\infty} |x_k|$ converges as well.

For all $k_0 > 1$,

$$\begin{aligned}
|\sum_{k=1}^{\infty} x_{n,k} - \sum_{k=1}^{\infty} x_k| &= |\sum_{k\le k_0} x_{n,k} - \sum_{k\le k_0} x_k + \sum_{k>k_0} x_{n,k} - \sum_{k>k_0} x_k| \\
&\le |\sum_{k\le k_0} x_{n,k} - \sum_{k\le k_0} x_k| + |\sum_{k>k_0} x_{n,k} - \sum_{k>k_0} x_k| \\
&= |\sum_{k\le k_0} (x_{n,k} - x_k)| + |\sum_{k>k_0} (x_{n,k} - x_k)| \\
&\le \sum_{k\le k_0} |x_{n,k} - x_k| + \sum_{k>k_0} |x_{n,k} - x_k| \\
&\le \sum_{k\le k_0} |x_{n,k} - x_k| + \sum_{k>k_0} \Big(|x_{n,k}| + |x_k|\Big)
\end{aligned}$$

$\forall n > 0, |x_{n,k}| \le M_k$

$$\begin{aligned}
&\Rightarrow x_k = \lim_{n\to\infty} |x_{n,k}| \le \lim_{n\to\infty} M_k = M_k \\
&\Rightarrow |x_{n,k}| + |x_k| \le 2M_k \\
&\Rightarrow |\sum_{k=1}^{\infty} x_{n,k} - \sum_{k=1}^{\infty} x_k| \le \sum_{k\le k_0} |x_{n,k} - x_k| + 2\sum_{k>k_0} M_k
\end{aligned}$$

On the other hand,

$$\lim_{k_0\to\infty} \sum_{k>k_0} M_k = \lim_{k_0\to\infty} \sum_{k=1}^{\infty} M_k - \lim_{k_0\to\infty} \sum_{k=1}^{k_0} M_k = 0.$$

As a consequence, $\forall \epsilon > 0$ and $\epsilon/3 > 0$,

$$\exists k_a, k_0 \ge k_a \Rightarrow |\sum_{k>k_0} M_k| \le \frac{\epsilon}{3}.$$

Since

$$\lim_{n\to\infty} x_{n,k} = x_k$$

$$\text{for } \epsilon_2 = \frac{\epsilon}{3k_a}, \exists n_1, n \ge n_1 \to |x_{n,k} - x_k| \le \epsilon_2 = \frac{\epsilon}{3k_a}.$$

As a consequence, $\forall \epsilon > 0, \exists\ k_a, n_1$ such that $n \ge n_1 \Rightarrow$

$$|\sum_{k=1}^{\infty} x_{n,k} - \sum_{k=1}^{\infty} x_k| \le \sum_{k\le k_a} |x_{n,k} - x_k| + 2\sum_{k>k_a} M_k \le \sum_{k\le k_a} \frac{\epsilon}{3k_a} + 2\frac{\epsilon}{3} = \epsilon.$$

We have thus just proven that

$$\lim_{n\to\infty} \sum_{k=1}^{\infty} x_{n,k} = \sum_{k=1}^{\infty} x_k.$$

Theorem 3.15 (recurrence of a Markov chain, uniqueness and positivity of the invariant distribution)

Let $(X_n)_{n\geq 0})$ be an irreducible and aperiodic Markov chain on the state space I with a transition matrix P (which may be of infinite dimension). The existence of an invariant distribution π such that $\pi_i \geq 0$, $\sum_{i\in I} \pi_i = 1$ and $\pi P = \pi$ has three consequences.

(i) The Markov chain is recurrent.

(ii) $\forall j \in I, \pi_j > 0$

(iii) $\forall i, j \in I, \ \lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$ whereby π is unique/unambiguous.

(iii) tells us that the Markov chain "forgets" any initial state (and hence also the initial probability distribution). Indeed, if $\nu = (\nu_i)_{i\in I}$ is a probability distribution on I, we have

$$p_\nu(X_n = j) = \sum_{i\in I} \nu_i p_{ij}^{(n)} \overset{n\to+\infty}{=} \sum_{i\in I} \nu_i \pi_j = \pi_j \sum_{i\in I} \nu_i = \pi_j.$$

This is the origin of the name *equilibrium distribution.*

Proof. We'll first prove that the Markov chain is recurrent. Since it is irreducible, if I is finite, Corollary 3.2 tells us it must be recurrent because of its irreducibility. We'll consider the case where I is countably infinite and we'll suppose that the Markov chain is transient instead. According to Theorem 3.14, $\sum_{n=1}^{\infty} p_{ij}^{(n)} < \infty$ for all $i, j \in I$. According to Theorem 2.10,

$$\lim_{n\to\infty} p_{ij}^{(n)} = 0.$$

Since π is an invariant distribution with respect to P, it is an invariant distribution with respect to P^n, $n > 0$ so that $\forall i$,

$$\pi_i = \sum_{j\in I} \pi_j p_{ji}^{(n)}.$$

On the other hand, we have

$$a_{ij} := \lim_{n\to\infty} \pi_j p_{ji}^{(n)} = 0, \quad \forall i, j \in I,$$

$\pi_j p_{ji}^{(n)} \leq \pi_j$ and $\sum_j \pi_j = 1$.

Therefore, according to the Weierstrass M-test (see 3.7),

$$\begin{aligned}\lim_{n\to\infty} \sum_{j\in I} \pi_j p_{ji}^{(n)} &= \lim_{n\to\infty} \pi_i = \pi_i \\ &= \sum_{j\in I} \lim_{n\to\infty} \pi_j p_{ji}^{(n)} = \sum_{j\in I} a_{ij} = \sum_{j\in I} 0 = 0.\end{aligned}$$

Since this is valid for all $i \in I$, we have a contradiction as π couldn't be a probability distribution. The existence of the invariant distribution allows us thus to conclude that the Markov chain is recurrent.

We now want to prove the uniqueness and the convergence of the Markov chain (iii). To achieve this, we shall use the *coupling technique*. We define the coupling $(Z_n, Y_n)_{n\geq 0}$ as a Markov chain on $I \times I$ with the following transition matrix which gathers the transition probabilities from state (i,j) to state (k,l):

$$\tilde{p}((i,j),(k,l)) = p_{ik}p_{jl} \tag{3.16}$$

While the index is two-dimensional, $\tilde{P} = (\tilde{p}((i,j),(k,l))_{(i,j),(k,l)\in I\times I})$ is a stochastic matrix. Obviously $\tilde{p}((i,j),(k,l)) = p_{ik}p_{jl} \geq 0$. $\forall i,j \in I$,

$$\sum_{k,l\in I} \tilde{p}((i,j),(k,l)) = \sum_{k\in I}\sum_{l\in I} p_{ik}p_{jl} = \sum_{k\in I} p_{ik} \sum_{l\in I} p_{jl} = 1 \times 1 = 1.$$

We must further show that $\tilde{\pi}(i,j) = \pi_i\pi_j$ is an invariant distribution of that Markov chain.

$$\begin{aligned}\left(\tilde{\pi}(i,j)\tilde{P}\right)_{i_2j_2} &= \sum_{(i,j)\in I} \tilde{\pi}(i,j)\tilde{p}((i,j),(i_2,j_2)) = \sum_{(i,j)\in I} \pi_i\pi_j p_{ii_2}p_{jj_2} \\ &= \sum_{i\in I} \pi_i p_{ii_2} \sum_{j\in I} \pi_j p_{jj_2} = \pi_{i_2}\pi_{j_2} = \tilde{\pi}(i_2,j_2)\end{aligned}$$

because π is an invariant distribution of Z_n and Y_n.

We now want to demonstrate that $(Z_n, Y_n)_{n\geq 0}$ is an **irreducible** Markov chain. We must demonstrate that

$$\forall j_0,l_0,j,l \in I,\ \exists m > 0, p(\{(Z_m,Y_m) = (j,l)\}|\{(Z_0,Y_0) = (j_0,l_0)\}) > 0.$$

$$\begin{aligned}&p(\{(Z_m,Y_m) = (j,l)\}|\{(Z_0,Y_0) = (j_0,l_0)\}) \\ &= p_{(j_0,l_0)}((Z_m,Y_m) = (j,l)) \\ &= \frac{p((Z_m,Y_m) = (j,l) \cap (Z_0,Y_0) = (j_0,l_0))}{p((Z_0,Y_0) = (j_0,l_0))} \\ &= \sum_{\substack{1\leq k<m \\ j_k,l_k\in I}} \frac{p((Z_m,Y_m) = (j,l) \cap (Z_0,Y_0) = (j_0,l_0) \cap (Z_k,Y_k) = (j_k,l_k))}{p((Z_0,Y_0) = (j_0,l_0))} \\ &= \sum_{\substack{1\leq k<m \\ j_k,l_k\in I}} p_{(j_0,l_0)}((Z_m,Y_m) = (j,l),(Z_k,Y_k) = (j_k,l_k))\end{aligned}$$

in virtue of the law of total probability.

Because of the Markov property applied to the coupled Markov chain,

$$p(\{(Z_m, Y_m) = (j,l)\}|\{(Z_0, Y_0) = (j_0, l_0)\}) =$$

$$\sum_{\substack{1\leq k<m \\ j_k, l_k \in I}} p_{(j_0,l_0)(j_1,l_1)} p_{(j_1,l_1)(j_2,l_2)} \cdots p_{(j_{m-1},l_{m-1})(j,l)}$$

$$= \sum_{\substack{1\leq k<m \\ j_k, l_k \in I}} (p_{j_0,j_1} p_{j_1,j_2} \cdots p_{j_{m-1}j})(p_{l_0,l_1} p_{l_1,l_2} \cdots p_{l_{m-1}l})$$

$$= \sum_{\substack{1\leq k<m \\ j_k \in I}} p_{j_0,j_1} p_{j_1,j_2} \cdots p_{j_{m-1}j} \sum_{\substack{1\leq k<m \\ l_k \in I}} p_{l_0,l_1} p_{l_1,l_2} \cdots p_{l_{m-1}l}$$

$$= \sum_{\substack{1\leq k<m \\ j_k \in I}} p_{j_0}(Z_1 = j_1, Z_2 = j_2, ..., Z_{m-1} = j_{m-1}, Z_m = j) \times$$

$$\sum_{\substack{1\leq k<m \\ l_k \in I}} p_{l_0}(Y_1 = l_1, Y_2 = l_2, ..., Y_{m-1} = l_{m-1}, Y_m = l)$$

$$\Rightarrow p_{(j_0,l_0)}((Z_m, Y_m) = (j,l)) = p_{j_0}(Z_m = j) p_{l_0}(Y_m = l)$$

by applying again the law of total probability.

$$\Rightarrow p^{(m)}_{(j_0,l_0)(j,l)} = p^{(m)}_{j_0 j} p^{(m)}_{l_0 l}$$

Since $(Z_n)_{n\geq 0}$ is an irreducible aperiodic Markov chain, according to Theorem 3.7,

$$\exists m_1 > 0, n \geq m_1 \Rightarrow p_{j_0}(Z_n = j) > 0.$$

Likewise, because $(Y_n)_{n\geq 0}$ is also an irreducible aperiodic Markov chain,

$$\exists m_2 > 0, n \geq m_2 \Rightarrow p_{l_0}(Y_n = l) > 0.$$

For $m = max(m_1, m_2), n \geq m \Rightarrow$

$$p_{(j_0,l_0)}((Z_n, Y_n) = (j,l)) = p_{j_0}(Z_n = j) p_{l_0}(Y_n = l) > 0.$$

As a consequence, the MC is irreducible. Since this is valid for all $(j,l) = (j_0, l_0)$, it is also aperiodic.

Since we know it also has an invariant distribution, we can apply the first part of the demonstration and draw the conclusion that (Z_n, Y_n) is recurrent. Following Theorem 3.14,

$$p_{(i,j)}((Z_n, Y_n) = (i_0, i_0) \ \ i.o.) = 1.$$

Let us now introduce the Markov time $\tau = inf\{n \geq 0 : (Z_n, Y_n) = (i_0, i_0)\}$. Let's consider the events

$$A = \{\tau < \infty\}$$

and

$$B = \{(Z_n, Y_n) = (i_0, i_0) \ \ i.o.\} \text{ with } p(B) = 1$$

$$\bar{A} = \{\cap_{n\geq 0}\ (Z_n, Y_n) \neq (i_0, i_0)\}$$

$$\bar{B} = \{(Z_n, Y_n) = (i_0, i_0) \text{ at most a limited number of times}\}$$

$$\bar{A} \subset \bar{B} \Rightarrow p(\bar{A}) \leq p(\bar{B}) \Rightarrow p(\bar{A}) \leq 1 - p(B) = 1 - 1 = 0$$

because the Markov chain is recurrent. Consequently, $p(\bar{A}) = 0 \Rightarrow p(A) = p(\tau < \infty) = 1$. Since $\{\tau < \infty\}$ is a sure event, it also always occurs when $(Z_0, Y_0) = (i, j), \forall i, j \in I$ so that

$$p_{i,j}(\tau < \infty) = 1.$$

Since τ is a Markov time, because of the strong Markov property, given that $\tau = n_0 \leq n < \infty$ we have $(Z_{n=n_0}, Y_{n=n_0}) = (i_0, i_0))$, (Z_n, Y_n) behaves probabilistically like (Z_{n-n_0}, Y_{n-n_0}) when the Markov chain starts in the state (i_0, i_0). This means that for all $n_0, k, i_0 \in I$,

$$p((Z_n, Y_n) = (k, k_2)|\tau = n_0 < \infty) = p((Z_{n-n_0}, Y_{n-n_0}) = (k, k_2)|(Z_0, Z_0) = (i_0, i_0)).$$

Thanks to the law of total probability,

$$\begin{aligned} p(Z_n = k|\tau = n_0 < \infty) &= \sum_{k_2 \in I} p(Z_n = k, Y_n = k_2|\tau = n_0 < \infty) \\ &= \sum_{k_2 \in I} p(Z_{n-n_0} = k, Y_{n-n_0} = k_2|(Z_0, Y_0) = (i_0, i_0)) \\ &= p(Z_{n-n_0} = k|(Z_0, Y_0) = (i_0, i_0)) \end{aligned}$$

because of the law of total probability. Likewise,

$$p(Y_n = k|\tau = n_0 < \infty) = p(Y_{n-n_0} = k|(Z_0, Y_0) = (i_0, i_0)).$$

Since Z_n and Y_n have the same probabilistic behaviour, we can conclude that

$$p(Y_{n-n_0} = k|(Z_0, Y_0) = (i_0, i_0)) = p(Z_{n-n_0} = k|(Z_0, Y_0) = (i_0, i_0))$$

so that

$$p(Z_n = k|\tau = n_0 < \infty) = p(Y_n = k|\tau = n_0 < \infty).$$

As a consequence,

$$
\begin{aligned}
&p_{(i,j)}(Z_n = k) = p_{(i,j)}(Z_n = k, \tau \leq n) + p_{(i,j)}(Z_n = k, \tau > n) \\
&p_{(i,j)}(Z_n = k, \tau > n) \leq p_{(i,j)}(\tau > n) \Rightarrow \\
&p_{(i,j)}(Z_n = k) \leq p_{(i,j)}(Z_n = k, \tau \leq n) + p_{(i,j)}(\tau > n) \\
&= p_{(i,j)}(Y_n = k, \tau \leq n) + p_{(i,j)}(\tau > n) \\
&\leq p_{(i,j)}(Y_n = k) + p_{(i,j)}(\tau > n)
\end{aligned}
$$

We saw before that $p^{(n)}_{(i,j)(k,l)} = p^{(n)}_{ik} p^{(n)}_{jl}$. On the other hand, we have

$$
\begin{aligned}
p_{(i,j)}(Z_n = k) &= p(Z_n = k | (Z_0, Y_0) = (i,j)) \\
&= \sum_{l \in I} \frac{p(Z_n = k, Y_n = l | (Z_0, Y_0) = (i,j))}{p((Z_0, Y_0) = (i,j))} \\
&= \sum_{l \in I} p^{(n)}_{(i,j)(k,l)} = \sum_{l \in I} p^{(n)}_{ik} p^{(n)}_{jl} \\
&= p^{(n)}_{ik} \sum_{l \in I} p^{(n)}_{jl} = p^{(n)}_{ik}
\end{aligned}
$$

Likewise, $p_{(i,j)}(Y_n = k) = p^{(n)}_{jk}$. Consequently,

$$
\begin{aligned}
&p^{(n)}_{ik} \leq p^{(n)}_{jk} + p_{(i,j)}(\tau > n) \\
&\Rightarrow \lim_{n \to +\infty} p^{(n)}_{ik} \leq \lim_{n \to +\infty} p^{(n)}_{jk} + \lim_{n \to +\infty} p_{(i,j)}(\tau > n) = \lim_{n \to +\infty} p^{(n)}_{jk}
\end{aligned}
$$

Since this is valid for all $i, j \in I$, we also have $\lim_{n \to +\infty} p^{(n)}_{jk} \leq \lim_{n \to +\infty} p^{(n)}_{ik}$, hence $\lim_{n \to +\infty} p^{(n)}_{ik} = \lim_{n \to +\infty} p^{(n)}_{jk}$. We now want to show that $p^{(n)}_{jk} = \pi_k$.

Since π is an invariant distribution, we have

$$
\pi_k - p^{(n)}_{jk} = \sum_i \pi_i p^{(n)}_{ik} - p^{(n)}_{jk} \sum_i \pi_i = \sum_i \pi_i (p^{(n)}_{ik} - p^{(n)}_{jk}).
$$

On the other hand, we know that $\forall i, j, k \in I, n \in \mathbb{N}, \pi_i (p^{(n)}_{ik} - p^{(n)}_{jk}) \leq \pi_i$, $\sum_i \pi_i = 1 \leq +\infty$ and $\lim_{n \to +\infty} \pi_i (p^{(n)}_{ik} - p^{(n)}_{jk}) = 0$. Therefore, we can apply Weierstrass' theorem so that

$$
\begin{aligned}
&\lim_{n \to +\infty} (\pi_k - p^{(n)}_{jk}) = \lim_{n \to +\infty} \sum_i \pi_i (p^{(n)}_{ik} - p^{(n)}_{jk}) = \sum_i \pi_i \lim_{n \to +\infty} (p^{(n)}_{ik} - p^{(n)}_{jk}) = \sum_i 0 = 0. \\
&\Rightarrow \\
&\lim_{n \to +\infty} (\pi_k - p^{(n)}_{jk}) = 0 \\
&\Rightarrow \\
&\lim_{n \to +\infty} p^{(n)}_{jk} = \pi_k
\end{aligned}
$$

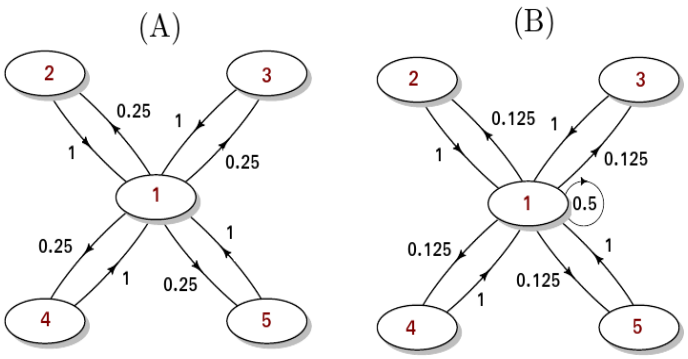


Figure 3.9: Markov chain with a period of 2 and its aperiodic counterpart

which is valid $\forall i, j, k \in I$. As a consequence, the invariant distribution is unambiguously determined by $\pi_k = \lim_{n \to +\infty} p_{jk}^{(n)}$.

Finally, we must demonstrate that π_k is always strictly positive. There must at least be one $i \in I$ such that $\pi_i > 0$ because otherwise π couldn't be a probability distribution. Since the Markov chain is irreducible, all states communicate with one another. Consequently, for every other state j

$$\exists k, l \in I, p_{ji}^{(k)} > 0 \text{ and } p_{ij}^{(l)} > 0.$$

According to the Markov-Chapman inequality,

$$\begin{aligned} & p_{jj}^{(k+n+l)} \geq p_{ji}^{(k)} p_{ii}^{(n)} p_{ij}^{(l)} \\ & \Rightarrow \lim_{n \to +\infty} p_{jj}^{(k+n+l)} \geq \lim_{n \to +\infty} p_{ji}^{(k)} p_{ii}^{(n)} p_{ij}^{(l)} \\ & \lim_{n \to +\infty} p_{jj}^{(n)} = \lim_{n \to +\infty} p_{jj}^{(k+n+l)} \geq p_{ji}^{(k)} p_{ij}^{(l)} \lim_{n \to +\infty} p_{ii}^{(n)} \end{aligned}$$

$$\pi_j \geq p_{ji}^{(k)} p_{ij}^{(l)} \pi_i > 0.$$

The existence of one $\pi_i > 0$ entails thus that all other $\pi_j > 0$.

In what follows, we'll establish a connection between the **average time** that a Markov chain spends in a state and the weight of that state in the invariant distribution. If we look at Figure 3.9, it makes intuitive sense to think it'd spend more time in state 1 than in the other states. We'll soon see why this is true in the case of *positive recurrent* Markov chains.

As we saw earlier, we defined the return time in k as

$$T_k := inf\{n \geq 1 : X_n = k\}.$$

For $k \in I$, we define the row vector $\gamma^k := (\gamma_i^k : i \in I)$ through

$$\gamma_i^k = E_k\Big[\sum_{n=0}^{T_k-1} 1_{\{X_n=i\}}\Big]$$
$$= \sum_{\substack{t_k\in[|1;+\infty[|\\ x_1,\ldots,x_{t_k-1}\in I}} \sum_{n=0}^{t_k-1} 1_{\{x_n=i\}} p_k(T_k = t_k, X_n = x_n, n \in [|1; t_k - 1|])$$

Theorem 3.16 For any irreducible and recurrent Markov chain $X = (X_n)_{n=0,1,\ldots}$ we have

(i) $\gamma_k^k = 1$

(ii) γ_k is an invariant measure, i.e. $\gamma^k P = \gamma^k$

(iii) $0 < \gamma_i^k < \infty,$ for all $i \in I$.

(i) follows from the very definition of γ_k^k given that the Markov chain always begins in state k.

Proof of (ii)

$$\gamma_j^k = E_k\Big[\sum_{n=0}^{T_k-1} 1_{\{X_n=j\}}\Big]$$
$$= E_k\Big[1_{\{X_0=j\}}\Big] + E_k\Big[\sum_{n=1}^{T_k-1} 1_{\{X_n=j\}}\Big]$$

whereby we pose that $E_k\Big[\sum_{n=1}^{T_k-1} 1_{\{X_n=j\}}\Big] = 0$ if $T_k = 1$. This leads to

$$\gamma_j^k = \delta_k(j) + E_k\Big[\sum_{n=1}^{T_k-1} 1_{\{X_n=j\}}\Big] = E_k\Big[1_{\{X_n=j\}}\Big] + E_k\Big[\sum_{n=1}^{T_k-1} 1_{\{X_n=j\}}\Big]$$
$$= E_k\Big[\sum_{n=1}^{T_k} 1_{\{X_n=j\}}\Big]$$
$$= E_k\Big[\sum_{n=1}^{\infty} 1_{\{X_n=j \text{ and } n\le T_k\}}\Big]$$
$$\gamma_j^k = \sum_{n=1}^{\infty} E_k\Big[1_{\{X_n=j \text{ and } n\le T_k\}}\Big]$$

$$
\begin{aligned}
&E_k\Big[1_{\{X_n=j \text{ and } n\leq T_k\}}\Big] \\
&= \sum_{t_k=n}^{\infty} \sum_{x_n \in I} 1_{\{x_n=j, t_k \geq n\}} p_k(X_n = x_n, T_k = t_k) \\
&= \sum_{t_k=n}^{\infty} 1_{\{t_k \geq n\}} p_k(X_n = j, T_k = t_k) \\
&= \sum_{t_k=n}^{\infty} p_k(X_n = j, T_k = t_k) \\
&= p_k(X_n = j, T_k \geq n) \\
&\textbf{(Law of total probability)}
\end{aligned}
$$

As a consequence,

$$
\begin{aligned}
\gamma_j^k &= \sum_{n=1}^{\infty} E_k\Big[1_{\{X_n=j\} \text{ and } n\leq T_k}\Big] \\
&= \sum_{n=1}^{\infty} p_k(X_n = j, T_k \geq n) \\
&= \sum_{n=1}^{\infty} \sum_{i \in I} p_k(X_n = j, X_{n-1} = i, T_k \geq n) \\
&= \sum_{i \in I} \sum_{n=1}^{\infty} p_k(X_n = j, X_{n-1} = i, T_k \geq n)
\end{aligned}
$$

According to Theorem 3.4, for any two events $A \in \sigma(X_0, ..., X_m)$ and $B \in \sigma(X_m, ..., X_{m+n})$, we have

$$p(A \bigcap B | X_m = i) = p(A|X_m = i)p(B|X_m = i)$$

.

We know that $\{T_k \geq n, X_{n-1} = i, X_0 = k\} \in \sigma(X_0, ..., X_{n-1})$ and that $\{X_n = j\} \in \sigma(X_{n-1}, X_n)$. Therefore,

$$
\begin{aligned}
&p(X_n = j, X_{n-1} = i, X_0 = k, T_k \geq n | X_{n-1} = i) \\
&= p(X_n = j | X_{n-1} = i) p(T_k \geq n, X_{n-1} = i, X_0 = k | X_{n-1} = i) \\
&\Rightarrow \\
&\frac{p(X_n = j, X_{n-1} = i, X_0 = k, T_k \geq n)}{p(X_{n-1} = i)} = p_{ij} \frac{p(T_k \geq n, X_{n-1} = i, X_0 = k)}{p(X_{n-1} = i)} \\
&p(X_n = j, X_{n-1} = i, X_0 = k, T_k \geq n) = p_{ij} p(T_k \geq n, X_{n-1} = i, X_0 = k)
\end{aligned}
$$

On the other hand, we have

$$p_k(X_n = j, X_{n-1} = i, T_k \geq n) = \frac{p_k(X_n = j, X_{n-1} = i, X_0 = k, T_k \geq n)}{p(X_0 = k)}$$
$$= \frac{p_{ij} p(T_k \geq n, X_{n-1} = i, X_0 = k)}{p(X_0 = k)}$$
$$= p_{ij} p_k(T_k \geq n, X_{n-1} = i).$$

Consequently,

$$\begin{aligned}
\gamma_j^k &= \sum_{i \in I} \sum_{n=1}^{\infty} p_k(X_n = j, X_{n-1} = i, T_k \geq n) \\
&= \sum_{i \in I} \sum_{n=1}^{\infty} p_{ij} p_k(T_k \geq n, X_{n-1} = i) \\
&= \sum_{i \in I} p_{ij} \sum_{n=1}^{\infty} p_k(T_k \geq n, X_{n-1} = i) \\
&= \sum_{i \in I} p_{ij} \sum_{m=0}^{\infty} p_k(T_k \geq m+1, X_m = i) \\
&= \sum_{i \in I} p_{ij} \sum_{m=0}^{\infty} p_k(X_m = i, m \leq T_k - 1) \\
&= \sum_{i \in I} p_{ij} \sum_{m=0}^{\infty} E_k(1_{\{X_m = i, m \leq T_k - 1\}}) \\
&= \sum_{i \in I} p_{ij} E_k\Big[\sum_{m=0}^{\infty} 1_{\{X_m = i, m \leq T_k - 1\}}\Big] \\
&= \sum_{i \in I} p_{ij} E_k\Big[\sum_{m=0}^{T_k - 1} 1_{\{X_m = i\}}\Big] \\
\gamma_j^k &= \sum_{i \in I} p_{ij} \gamma_i^k = \sum_{i \in I} \gamma_i^k p_{ij}
\end{aligned}$$

We now know that γ^k is an invariant measure, that is $\gamma^k P = \gamma^k$.

Proof of (iii)

Because of the irreducibility of the Markov chain, for any $i \in I$ there exists $n, m > 0$ such that $p_{ik}^{(n)}, p_{ki}^{(m)} > 0$. By applying (ii), we know that $\gamma^k P = \gamma^k$. It is easy to show recurrently that $\gamma^k P^{(N)} = \gamma^k, \forall n \in \mathbb{N}_0$.

Therefore,

$$\gamma_i^k = \sum_{j \in I} \gamma_j^k p_{ji}^{(m)} \geq \gamma_k^k p_{ki}^{(m)} > 0.$$

This also leads to

$$\gamma_i^k p_{ik}^{(n)} \leq \sum_{j \in I} \gamma_j^k p_{jk}^{(n)} = \gamma_k^k = 1.$$

Since $p_{ik}^{(n)} > 0$,

$$\gamma_i^k p_{ik}^{(n)} \leq 1 \Rightarrow \gamma_i^k \leq \frac{1}{p_{ik}^{(n)}} < \infty.$$

Finally, we find that

$$0 < \gamma_i^k < \infty.$$

Theorem 3.17 If P is the matrix of an irreducible MC and if λ is an invariant measure for P such that $\lambda_k = 1$, then we have $\lambda \geq \gamma^k$ componentwise. What is more, if the MC is recurrent we have $\lambda = \gamma^k$.

Proof.
For each $i \in I$, we have

$$\begin{aligned}
\lambda_j &= \sum_{i_1 \in I} \lambda_{i_1} p_{i_1 j} = \sum_{i_1 \neq k} \lambda_{i_1} p_{i_1 j} + p_{kj} \\
&= \sum_{i_1 \neq k} \sum_{i_2 \neq k} (\lambda_{i_2} p_{i_2 i_1} + p_{k i_1}) p_{i_1 j} + p_{kj} \\
&= \sum_{i_1 \neq k} (\sum_{i_2 \neq k} \lambda_{i_2} p_{i_2 i_1} p_{i_1 j}) + p_{k i_1} p_{i_1 j} + p_{kj} \\
&= \sum_{i_1, i_2 \neq k} \lambda_{i_2} p_{i_2 i_1} p_{i_1 j} + \sum_{i_1 \neq k} p_{k i_1} p_{i_1 j} + p_{kj} \\
&\vdots \\
&= \sum_{i_1, \ldots, i_n \neq k} \lambda_{i_n} p_{i_n i_{n-1}} \ldots p_{i_1 j} + \Big(p_{kj} + \sum_{i_1 \neq k} p_{k i_1} p_{i_1 j} \\
&+ \ldots + \sum_{i_1, \ldots, i_{n-1} \neq k} p_{k i_n - 1} \ldots p_{i_2 i_1} p_{i_1 j} \Big).
\end{aligned}$$

As a consequence, $\forall n > 0$,

$$\lambda_j \geq p_{kj} + \sum_{i_1 \neq k} p_{k i_1} p_{i_1 j} + \ldots + \sum_{i_1, \ldots, i_{n-1} \neq k} p_{k i_n - 1} \ldots p_{i_2 i_1} p_{i_1 j}.$$

$$p_k(X_1 = j, T_k \geq 1) = \frac{p(X_1 = j, T_k \geq 1, X_0 = k)}{p(X_0 = k)}$$
$$= \frac{p(T_k \geq 1 | X_1 = j, X_0 = k) p(X_1 = j, X_0 = k)}{p(X_0 = k)}$$
$$p_k(X_1 = j, T_k \geq 1) = p(T_k \geq 1 | X_1 = j, X_0 = k) p_{kj}$$

Since $T_k \geq 1$ is a sure event, we have

$$p_k(X_1 = j, T_k \geq 1) = p_{kj}.$$

Similarly,

$$p_k(X_2 = j, T_k \geq 2) = p_k(X_2 = j, X_1 \neq k) = \sum_{i_1 \neq k} p_k(X_1 = i_1, X_2 = j)$$
$$p_k(X_2 = j, T_k \geq 2) = \sum_{i_1 \neq k} p_{k i_1} p_{i_1 j}$$

$$p_k(X_n = j, T_k \geq n) = p_k(X_n = j, X_i \neq k, 0 < i < n)$$
$$= \sum_{i_1, ..., i_{n-1} \neq k} p_k(X_1 = i_1, X_2 = i_2, ..., X_{n-1} = i_{n-1}, X_n = j)$$
$$p_k(X_n = j, T_k \geq n) = \sum_{i_1, ..., i_{n-1} \neq k} p_{k i_{n-1}} ... p_{i_2 i_1} p_{i_1 j}$$

Therefore,

$$\Big(p_{kj} + \sum_{i_1 \neq k} p_{k i_1} p_{i_1 j} + ... + \sum_{i_1, ..., i_{n-1} \neq k} p_{k i_{n-1}} ... p_{i_2 i_1} p_{i_1 j}\Big) =$$
$$p_k(X_1 = j, T_k \geq 1) + p_k(X_2 = j, T_k \geq 2) + ... + p_k(X_n = j, T_k \geq n)$$
$$= E_k[1_{\{X_1 = j, T_k \geq 1\}}] + E_k[1_{\{X_2 = j, T_k \geq 2\}}] + ... + E_k[1_{\{X_n = j, T_k \geq n\}}]$$
$$= E_k\Big[\sum_{n'=1}^{n} 1_{\{X'_n = j, T_k \geq n'\}}\Big] \overset{n \to +\infty}{=} E_k\Big[\sum_{n=1}^{\infty} 1_{\{X_n = j, T_k \geq n\}}\Big]$$
$$= E_k[\sum_{n=0}^{T_k - 1} 1_{\{X_n = i\}}]$$
$$= \gamma_j^k.$$

Since $\forall n > 0$,

$$\lambda_j \geq p_{kj} + \sum_{i_1 \neq k} p_{ki_1} p_{i_1 j} + ... + \sum_{i_1,...,i_{n-1} \neq k} p_{ki_{n-1}} ... p_{i_2 i_1} p_{i_1 j}$$

, we have $\lambda_j \geq \gamma_j^k$ which is valid for all $j \in I$.

If P is recurrent, Theorem 3.16 tells us that γ^k is an invariant measure. It is easy to show that $\mu = \lambda - \gamma^k$ is an invariant measure as well. Since P is irreducible, for each $i \in I$ there exists $n > 0$ such that $p_{ik}^{(n)} > 0$. Consequently,

$$\lambda_k - \gamma_k^k = 1 - 1 = 0 = \mu_k = \sum_{j \in i} \mu_j p_{jk}^{(n)} \geq \mu_i p_{ik}^{(n)}.$$

Since $p_{ik}^{(n)} > 0$, this entails that

$$\mu_i = \lambda_i - \gamma_i^k = 0, \forall i \in I.$$

In what follows, m_i designates the expected return time in state i, that is $m_i = E_i[T_i]$. According to Definition 3.8, i is positive recurrent if $m_i < \infty$ and null recurrent if $m_i = \infty$ and $p_i(T_i < \infty) = f_{ii} = 1$.

Theorem 3.18 Let $X = (X_n)_{n=0,1,...}$ be an irreducible Markov chain. The three following statements are equivalent:

(i) every state is positive recurrent

(ii) one state is positive recurrent

(iii) X has an invariant distribution.

(vi) What is more, this invariant distribution is given by

$$\pi_i = \frac{1}{m_i}, i \in I.$$

It is obvious that $(i) \Rightarrow (ii)$.

Proof of $(ii) \Rightarrow (iii)$. If i is a positive recurrent state, it is also recurrent. Since recurrence is a class property, all other states of the irreducible Markov chain are recurrent too and so is P as a whole. Following Theorem 3.16, γ^i is an invariant measure. Furthermore,

$$\sum_{j \in I} \gamma_j^i = \sum_{j \in I} E_i\Big[\sum_{n=0}^{T_i - 1} 1_{\{X_n = j\}}\Big] = E_i\Big[\sum_{j \in I} \sum_{n=0}^{T_i - 1} 1_{\{X_n = j\}}\Big] = E_i\Big[\sum_{n=0}^{T_i - 1} \sum_{j \in I} 1_{\{X_n = j\}}\Big]$$

$$= E_i\Big[\sum_{n=0}^{T_i - 1} 1_{\{X_n \in I\}}\Big] = E_i[T_i] = m_i.$$

We now want to demonstrate that $\pi = \frac{\gamma^i}{m_i}$ is an invariant distribution. It is quite clear that

$$\forall i, j \in I, 0 \leq = \frac{\gamma_j^i}{\sum_{j \in I} \gamma_j^i} \leq 1$$

and

$$\sum_{j \in I} \pi_j = \sum_{j \in I} \frac{\gamma_j^i}{\sum_{j \in I} \gamma_j^i} = 1.$$

$$\pi P = \frac{\pi \gamma^i}{m_i} = \frac{\gamma^i}{m_i} = \pi.$$

Proof of $(iii) \Rightarrow (i)$. Let's consider an **arbitrary** $k \in I$. $\pi = \frac{\gamma^{i_2}}{m_{i_2}}$ is an invariant distribution so that there must at least be a $j \in I$ with $\pi_j > 0$. Because of the irreducibility of the Markov chain, there exists a $n > 0$ such that $p_{jk}^{(n)} > 0$.

$$\pi_k = \sum_{i \in I} \pi_i p_{ik}^{(n)} \geq \pi_j p_{jk}^{(n)} > 0.$$

We can thus define $\lambda_i = \pi_i / \pi_k$. Since k and π_k are constants, $\lambda_i = \pi_i / \pi_k$ is an invariant measure of the Markov chain with $\lambda_k = 1$. Given the irreducibility of the MC, Theorem 3.17 tells us that $\lambda \geq \gamma^k$ so that

$$m_k = \sum_{i \in I} \gamma_i^k \leq \sum_{i \in I} \lambda_i = \frac{1}{\pi_k} < \infty.$$

This shows that every state k is positive recurrent so that the Markov chain and P are recurrent.

It remains to be shown that $\pi_i = 1/m_i$.

Proof of vi)

The Markov chain is irreducible and recurrent. λ is an invariant distribution with $\lambda_k = \pi_k / \pi_k = 1$. Therefore, by applying the last part of Theorem 3.17, we have $\forall i, k \in I$,

$$\begin{aligned}
&\lambda_i = \gamma_i^k \\
&\Rightarrow \\
&\sum_{i \in I} \gamma_i^k = m_k = \sum_{i \in I} \lambda_i = \frac{1}{\pi_k} < \infty \\
&\Rightarrow \\
&\pi_k = \frac{1}{m_k}
\end{aligned}$$

Theorem 3.19 If an aperiodic, irreducible Markov chain on I hasn't any stationary distribution, then we have $\forall i, j \in I$

$$\lim_{n\to\infty} p_{ij}^{(n)} = 0.$$

Proof.

As we saw in the long demonstration of Theorem 3.15, the coupled Markov chain (Y_n, Z_n) is irreducible. Let's first suppose it is transient. According to Theorem 3.14,

$$\begin{aligned}
&\forall i_1, i_2, j_1, j_2 \in I, \sum_{n=0}^{\infty} \tilde{p}((i_1, i_2), (j_1, j_2))^{(n)} < \infty \\
&\Rightarrow \\
&\sum_{n=0}^{\infty} \tilde{p}((i, i), (j, j))^{(n)} = \sum_{n=0}^{\infty} \left(p_{ij}^{(n)}\right)^2 < \infty
\end{aligned}$$

Following Theorem 2.10, we can conclude that

$$\lim_{n\to\infty} \left(p_{ij}^{(n)}\right)^2 = 0 \Rightarrow \lim_{n\to\infty} p_{ij}^{(n)} = 0.$$

Let us now assume that the coupled Markov chain is recurrent. As we saw during the demonstration of Theorem 3.15,

$$\forall i, j, k \in I,\ |p_{ik}^{(n)} - p_{jk}^{(n)}| \overset{n+\infty}{\to} 0.$$

Let's now suppose that there exists $i_1, k_1 \in I$ with $p_{i_1 k_1}^{(n)} \overset{n+\infty}{\nrightarrow} 0$. As we realised during the demonstration of Theorem 2.10, this means that there exists an infinite subsequence $n' \subset \mathbb{N}$ and $\delta > 0$ such that $p_{i_1 k_1}^{(n')} \in [\delta, 1]$ everywhere (because otherwise the whole sequence would go to 0). According to the famous Bolzano-Weierstrass theorem [2], there exists a subsequence $n'' \subset n'$ such that $p_{i_1 k_1}^{(n'')}$ is convergent, i.e.

$$\lim_{n''\to\infty} p_{i_1 k_1}^{(n'')} = \alpha_{i_1, k_1} \text{ with } \alpha_{i_1, k_1} \in]0; 1]$$

We have

$$\forall i \in I, |p_{i_1 k_1}^{(n'')} - p_{i k_1}^{(n'')}| \overset{n''+\infty}{\to} 0.$$

This is only possible if

$$\lim_{n''\to\infty} p_{i k_1}^{(n'')} = \lim_{n''\to\infty} p_{i_1 k_1}^{(n'')} = \alpha_{i_1 k_1} = \alpha_{k_1} > 0, \forall i \in I.$$

[2] `https://en.wikipedia.org/wiki/Bolzano-Weierstrass_theorem`

Let's now consider $k \neq k_1$. Because of the irreducibility and aperiodicity of the Markov chain, $\exists m_1 > 0, n \geq m_1 \Rightarrow p_{kk_1}^{(n)} > 0$ and $\exists m_2 > 0, n \geq m_2 \Rightarrow p_{k_1k}^{(n)} > 0$ so that for $m = max(m_1, m_2), n \geq m \Rightarrow p_{kk_1}^{(n)} > 0$ and $p_{k_1k}^{(n)} > 0$.

For $n'' > m$, because of the Chapman-Kolmogorov equation (see 3.5) we have

$$p_{ik_1}^{(n'')} \geq p_{ik}^{(n''-m)} p_{kk_1}^{(m)}$$

Let's suppose that there exists $k \in I$ such that $p_{ik}^{(n'')}$ doesn't converge but keeps fluctuating instead. This means that

$$\nexists \alpha_k \in [0;1], \lim_{n'' \to +\infty} p_{ik}^{(n'')} = \lim_{n'' \to +\infty} p_{ik}^{(n''-m)} = \alpha_k.$$

Let's introduce

$$a_{ik_1}^{(n'')} = |p_{ik_1}^{(n''+1)} - p_{ik_1}^{(n'')}| \overset{n'' \to +\infty}{\to} 0$$

and

$$a_{ik}^{(n'')} = |p_{ik}^{(n''+1)} - p_{ik}^{(n'')}| \overset{n'' \to +\infty}{\not\to} 0.$$

As a consequence, there exists an infinite subsequence $n''' \subset n''$ and $\epsilon > 0$ such that

$$a_{ik}^{(n''')} > \epsilon \Leftrightarrow |p_{ik}^{(n'''+1)} - p_{ik}^{(n''')}| > \epsilon.$$

Through the Chapman-Kolmogorov equation, we know that for $n''' > m$

$$p_{ik_1}^{(n'''+m)} \geq p_{ik}^{(n''')} p_{kk_1}^{(m)}$$

and

$$p_{ik_1}^{(n'''+m+1)} \geq p_{ik}^{(n'''+1)} p_{kk_1}^{(m)}.$$

This leads to

$$p_{ik_1}^{(n'''+m+1)} - p_{ik_1}^{(n'''+m)} \geq p_{kk_1}^{(m)} (p_{ik}^{(n'''+1)} - p_{ik}^{(n''')}).$$

Likewise, it is possible to show that

$$p_{ik}^{(n'''+1)} - p_{ik}^{(n''')} \geq p_{k_1k}^{(m)} (p_{ik_1}^{(n'''-m+1)} - p_{ik_1}^{(n'''-m)})$$

$$\Rightarrow p_{ik_1}^{(n'''-m+1)} - p_{ik_1}^{(n'''-m)} \leq \frac{p_{ik}^{(n'''+1)} - p_{ik}^{(n''')}}{p_{k_1k}^{(m)}}.$$

We must now distinguish between two possibilities: either $p_{ik}^{(n'''+1)} - p_{ik}^{(n''')} > 0$ or $p_{ik}^{(n'''+1)} - p_{ik}^{(n''')} < 0$ because of the very definition of the subsequence n'''. If $p_{ik}^{(n'''+1)} - p_{ik}^{(n''')} > 0$, $p_{ik}^{(n'''+1)} - p_{ik}^{(n''')} > \epsilon$ so that

$$p_{ik_1}^{(n'''+m+1)} - p_{ik_1}^{(n'''+m)} \geq p_{kk_1}^{(m)} (p_{ik}^{(n'''+1)} - p_{ik}^{(n''')}) > p_{kk_1}^{(m)} \epsilon > 0.$$

If $p_{ik}^{(n'''+1)} - p_{ik}^{(n''')} < 0$, $p_{ik}^{(n'''+1)} - p_{ik}^{(n''')} < -\epsilon$ so that

$$p_{ik_1}^{(n'''-m+1)} - p_{ik_1}^{(n'''-m)} \leq \frac{p_{ik}^{(n'''+1)} - p_{ik}^{(n''')}}{p_{k_1k}^{(m)}} < -\frac{\epsilon}{p_{k_1k}^{(m)}} < 0.$$

For all $n''' > m$ we have thus either

$$|p_{ik_1}^{(n'''+m+1)} - p_{ik_1}^{(n'''+m)}| > p_{kk_1}^{(m)}\epsilon$$

or

$$|p_{ik_1}^{(n'''-m+1)} - p_{ik_1}^{(n'''-m)}| > \frac{\epsilon}{p_{k_1k}^{(m)}}.$$

In this way, it is possible to construct a subsequence $n^{(4)} \subset n''' \subset \mathbb{N}$ such that

$$|p_{ik_1}^{n^{(4)}+1} - p_{ik_1}^{n^{(4)}}| > min(p_{kk_1}^{(m)}\epsilon, \frac{\epsilon}{p_{k_1k}^{(m)}}) > 0.$$

This would show that $a_{ik_1}^{(n'')} = p_{ik_1}^{(n''+1)} - p_{ik_1}^{(n'')}|$ doesn't go to zero so that $p_{ik_1}^{(n'')}$ doesn't converge towards any limit. This contradiction proves that $p_{ik}^{(n'')}$ must be convergent as well $\forall k, i \in I$. We can thus write that

$$\lim_{n''\to} p_{ik}^{(n'')} = \alpha_k \in [0;1], \forall i, k \in I.$$

We know that $\alpha_{k_1} > 0$.

Let us now consider a **finite** subset M of I. We have

$$\sum_{k\in M} \alpha_k = \lim_{n''\to\infty} \sum_{k\in M} p_{ik}^{(n'')} \leq 1$$

so that for $M \uparrow I$,

$$0 < \alpha := \sum_{k\in I} \alpha_k \leq 1 \text{ because } \alpha_{k_1} > 0.$$

Thanks to the Chapman-Komolgorov equation, we know that

$$\sum_{k\in M} p_{ik}^{(n'')} p_{kj} \leq p_{ij}^{(n''+1)} = \sum_{k\in I} p_{ik} p_{kj}^{(n'')}.$$

Since M is a finite subset of I,

$$\lim_{n''\to\infty} \sum_{k\in M} p_{ik}^{(n'')} p_{kj} = \sum_{k\in M} \lim_{n''\to\infty} p_{ik}^{(n'')} p_{kj} = \sum_{k\in M} \alpha_k p_{kj}.$$

Moreover, we have $p_{ik}p_{kj}^{(n'')} \leq p_{ik}$ with

$$\sum_{k\in I} p_{ik} = 1 < \infty \text{ and } \lim_{n\to\infty} p_{ik}p_{kj}^{(n'')} = \alpha_j p_{ik}.$$

As a consequence, we can apply Weierstrass' M-test to the sum over I so that

$$\lim_{n\to\infty}\sum_{k\in I} p_{ik}p_{kj}^{(n'')} = \sum_{k\in I}\lim_{n\to\infty} p_{ik}p_{kj}^{(n'')} = \sum_{k\in I}\alpha_j p_{ik}.$$

We have

$$\sum_{k\in M} p_{ik}^{(n'')}p_{kj} \leq \sum_{k\in I} p_{ik}p_{kj}^{(n'')}.$$

Hence

$$\lim_{n''\to\infty}\sum_{k\in M} p_{ik}^{(n'')}p_{kj} \leq \lim_{n''\to\infty}\sum_{k\in I} p_{ik}p_{kj}^{(n'')}.$$

$$\sum_{k\in M}\alpha_k p_{kj} \leq \sum_{k\in I}\alpha_j p_{ik} = \alpha_j.$$

If we consider the limit $M \uparrow I$, we find that

$$\sum_{k\in I}\alpha_k p_{kj} \leq \alpha_j.$$

Let's first consider that

$$\exists j \in I, \sum_{k\in I}\alpha_k p_{kj} < \alpha_j.$$

This would lead straight to

$$\alpha = \sum_{k\in I}\alpha_k = \sum_{k\in I}\alpha_k\sum_{j\in I}p_{kj} = \sum_{k\in I}\sum_{j\in I}\alpha_k p_{kj} = \sum_{j\in I}\sum_{k\in I}\alpha_k p_{kj} < \sum_{j\in I}\alpha_j,$$

which is self-contradictory. Consequently, $\sum_{k\in I}\alpha_k p_{kj} = \alpha_j$ for all $j \in I$. Let us define $\pi_j = \alpha_j/\alpha$. It's easy to verify that

$$\forall j \in I, \pi_j \geq 0, \sum_j \pi_j = 1 \text{ and } \sum_{k\in I}\pi_k p_{kj} = \pi_j.$$

This would be an invariant distribution, which would contradict one fundamental hypothesis of the theorem. Therefore, we must conclude that

$$\forall i, j \in I, \lim_{n\to\infty} p_{ij}^n = 0.$$

The following theorem summarises the main results of this section.

Theorem 3.20 For any irreducible and aperiodic Markov chain, there are only three alternatives:

(i) The Markov chain is transient. It has no invariant distribution. For all $i, j \in I$, we have $\lim_{n\to\infty} p_{ij}^{(n)} = 0$ with a convergence rate such that $\sum_n p_{ij}^{(n)} < \infty$ and $p_i(T_j < \infty) < 1$.

(ii) The Markov chain is null recurrent and there is no invariant distribution. For all $i, j \in I$, we have $\lim_{n\to\infty} p_{ij}^{(n)} = 0$ with a convergence rate such that $\sum_n p_{ij}^{(n)} = \infty$ and $m_j = E_j[T_j] = \infty$ and $p_i(T_j < \infty) = 1$.

(iii) The Markov chain is positive recurrent. There is one single invariant distribution π. For all $i, j \in I$, $p_i(T_j < \infty) = 1$, $\sum_n p_{ij}^{(n)} = \infty$ and

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j > 0 \text{ and } m_j = E_j[T_j] = \frac{1}{\pi_j}.$$

Proof of (i).

Because of the definition of transiency, $p_i(T_j < \infty) < 1$ and $m_j = E_j[T_j] = \infty$. According to Theorem 3.18, since the MC is not positive recurrent, it cannot have any invariant distribution. Following Theorem 3.19, this entails that $\lim_{n\to\infty} p_{ij}^{(n)} = 0$. Finally, Theorem 3.12 tells us that $\sum_n p_{ij}^{(n)} < \infty$.

Proof of (ii).

Because of the definition of null-recurrence, $p_i(T_j < \infty) = 1$ and $m_j = E_j[T_j] = \infty$. According to Theorem 3.18, since the MC is not positive recurrent, it cannot have any invariant distribution. Like before, Theorem 3.19 implies that $\lim_{n\to\infty} p_{ij}^{(n)} = 0$. Theorem 3.12 leads to $\sum_n p_{ij}^{(n)} = \infty$.

Proof of (iii).

Because of the definition of positive-recurrence, $p_i(T_j < \infty) = 1$ and $m_j = E_j[T_j] < \infty$. Theorem 3.12 tells us that $\sum_n p_{ij}^{(n)} = \infty$. Finally, Theorem 3.18 entails that the MC has an invariant distribution given by

$$\pi_i = \frac{1}{m_i}, i \in I.$$

Since $m_i = E_i[T_i]$ is unambiguously defined, π is the unique invariant distribution of the Markov chain.

Theorem 3.21 (Markov chains on finite state spaces)

Let $X = (X_n)_{n=0,1,\ldots}$ be a Markov chain on a finite state space with a transition matrix P. The following statements hold:

(i) If X is irreducible and aperiodic, the Markov chain converges towards the unambiguous invariant distribution:

$$\forall\ i, j \in I, \lim_{n\to\infty} p_{ij}^{(n)} = \pi_j > 0.$$

(ii) If the MC is irreducible and periodic with a period equal to d, for $m > 0$, we consider a uniformly distributed random integer $N_m \in [|1; m|]$. There exists a unique invariant distribution π such that

$$\lim_{m\to\infty} p(X_{N_m} = j | X_0 = i) = \pi_j, \forall i, j \in I.$$

(iii) If X is irreducible, X is also positive recurrent and has exactly one invariant distribution π.

(vi) In the general case, X always has at least one invariant distribution.

Proof of (i).

Let's suppose that our irreducible and aperiodic Markov chain is either transient or null-recurrent. Based on Theorem 3.20, this would mean that

$$\lim_{n\to\infty} p_{ij}^{(n)} = 0.$$

Consequently,

$$\sum_{j\in I} \lim_{n\to\infty} p_{ij}^{(n)} = 0.$$

On the other hand, since I is finite,

$$\sum_{j\in I} \lim_{n\to\infty} p_{ij}^{(n)} = \lim_{n\to\infty} \sum_{j\in I} p_{ij}^{(n)} = 0.$$

For, say, $\epsilon = 1/2 > 0$, $\exists n_0$ such that $n \geq n_0$ implies that $\sum_{j\in I} p_{ij}^{(n)} < 1/2$ which violates the axioms of probability theory. Therefore, any irreducible and aperiodic Markov chain on a finite state space must be positive recurrent. Theorem 3.20 tells us that it converges towards an unambiguous invariant distribution:

$$\forall\ i, j \in I, \lim_{n\to\infty} p_{ij}^{(n)} = \frac{1}{m_j} = \pi_j > 0.$$

Proof of (ii).

We'll show first that for any irreducible MC,

$$p(X_{N_m} = j | X_0 = i) = \frac{1}{m} \sum_{n=1}^{m} p_{ij}^{(n)}.$$

$$p(X_{N_m} = j | X_0 = i) = \sum_{n=1}^{m} p(X_{N_m} = j, N_m = n | X_0 = i) = \sum_{n=1}^{m} \frac{p(X_n = j, N_m = n, X_0 = i)}{p(X_0 = i)}$$
$$= \sum_{n=1}^{m} \frac{p(X_n = j, X_0 = i | N_m = n) p(N_m = n)}{p(X_0 = i)} = \sum_{n=1}^{m} \frac{p(X_n = j, X_0 = i | N_m = n)}{p(X_0 = i)} p(N_m = n)$$
$$= \sum_{n=1}^{m} \frac{p(X_n = j, X_0 = i)}{p(X_0 = i)} p(N_m = n) = \frac{1}{m} \sum_{n=1}^{m} p_{ij}^{(n)}$$

Let us now consider an irreducible and aperiodic Markov chain.

$$\forall i, j \in I, \lim_{n\to\infty} p_{ij}^{(n)} = \pi_j.$$

Hence,

$$\forall \epsilon > 0, \exists n_0 > 0, n \geq n_0 \Rightarrow \pi_j - \frac{\epsilon}{2} < p_{ij}^{(n)} < \pi_j + \frac{\epsilon}{2}.$$

For a sufficiently large m, we have

$$p(X_{N_m} = j | X_0 = i) = \frac{1}{m} \sum_{n=1}^{m} p_{ij}^{(n)} = \frac{1}{m} \sum_{n=1}^{n_0-1} p_{ij}^{(n)} + \frac{1}{m} \sum_{n=n_0}^{m} p_{ij}^{(n)}$$

$$\text{Since } \lim_{m\to\infty} \frac{1}{m} \sum_{n=1}^{n_0-1} p_{ij}^{(n)} = 0,$$

For $\frac{\epsilon}{2} > 0$,

$$\exists n_1 > 0, m \geq n_1 \Rightarrow 0 < \frac{1}{m} \sum_{n=1}^{n_0-1} p_{ij}^{(n)} < \frac{\epsilon}{2}.$$

Finally, we find that for $m \geq n_1$

$$\frac{1}{m} \sum_{n=n_0}^{m} (\pi_j - \frac{\epsilon}{2}) < \frac{1}{m} \sum_{n=1}^{n_0-1} p_{ij}^{(n)} + \frac{1}{m} \sum_{n=n_0}^{m} p_{ij}^{(n)} < \frac{\epsilon}{2} + \frac{1}{m} \sum_{n=n_0}^{m} (\pi_j + \frac{\epsilon}{2})$$

$$-\frac{m-n_0+1}{m}\frac{\epsilon}{2} + \frac{1}{m} \sum_{n=n_0}^{m} \pi_j < \frac{1}{m} \sum_{n=1}^{m} p_{ij}^{(n)} < \frac{\epsilon}{2} + \frac{m-n_0+1}{m}\frac{\epsilon}{2} + \frac{1}{m} \sum_{n=n_0}^{m} \pi_j$$

$$\pi_j - \frac{m-n_0+1}{m}\frac{\epsilon}{2} - \frac{1}{m} \sum_{n=1}^{n_0-1} \pi_j < \frac{1}{m} \sum_{n=1}^{m} p_{ij}^{(n)} < \pi_j + \frac{\epsilon}{2} + \frac{m-n_0+1}{m}\frac{\epsilon}{2} + \frac{1}{m} \sum_{n=1}^{n_0-1} \pi_j$$

Since

$$\lim_{m\to\infty} \frac{1}{m} \sum_{n=1}^{n_0-1} \pi_j = 0$$

and since $\epsilon > 0$ can be chosen arbitrarily small, we must have

$$\lim_{m\to\infty} p(X_{N_m} = j | X_0 = i) = \lim_{m\to\infty} \frac{1}{m} \sum_{n=1}^{m} p_{ij}^{(n)} = \pi_j$$

for all irreducible **aperiodic** chains.

We'll now prove that for any irreducible periodic Markov chain,

$$\lim_{m\to\infty} p(X_{N_m} = j | X_0 = i) = \pi_j, \forall i, j \in I.$$

Two cases can be distinguished.

1) i and j belong to the same class K_i. In that case, $Y_n = (X_{dn})_{n\geq 0}$ is an irreducible aperiodic MC whose transition matrix is given by P^d which must be irreducible in order for P itself to be irreducible. As a consequence, there exists a unique invariant distribution π' such that

$$\lim_{m\to\infty} p(Y_{N_m} = j|Y_0 = i) = \pi'_j, \forall i, j \in I.$$

We have

$$\begin{aligned} p(X_{N_m} = j|X_0 = i) &= \frac{1}{m}\sum_{n=1}^{m} p_{ij}^{(n)} = \frac{1}{m}\sum_{n'=1}^{\lfloor \frac{m}{d} \rfloor} p_{ij}^{(dn')} \\ &= \frac{\lfloor \frac{m}{d} \rfloor}{m}\frac{1}{\lfloor \frac{m}{d} \rfloor}\sum_{n'=1}^{\lfloor \frac{m}{d} \rfloor} p_{ij}^{(dn')} \\ &= \frac{\lfloor \frac{m}{d} \rfloor}{m} p(Y_{N_{\lfloor \frac{m}{d} \rfloor}} = j|Y_0 = i) \end{aligned}$$

This entails that

$$\begin{aligned} &\lim_{m\to\infty} p(X_{N_m} = j|X_0 = i) = \lim_{m\to\infty} \frac{\lfloor \frac{m}{d} \rfloor}{m} \lim_{\lfloor \frac{m}{d} \rfloor \to\infty} p(Y_{N_{\lfloor \frac{m}{d} \rfloor}} = j|Y_0 = i) \\ &= \frac{1}{d}\pi'_j \text{ which is unambiguous.} \end{aligned}$$

2) i and j belong to two distinct periodic classes K_{l_1} and K_{l_2}. We have (if m is large enough and if we pose that $l_2 - l_1 = 1$ if $l_1 = d-1$ and $l_2 = 1$.)

$$\begin{aligned} &p(X_{N_m} = j|X_0 = i) = \frac{1}{m}\sum_{n=1}^{m} p_{ij}^{(n)} = \frac{1}{m}\sum_{n=1}^{l_2-l_1} p_{ij}^{(n)} + \frac{1}{m}\sum_{n=l_2-l_1+1}^{m} p_{ij}^{(n)} \\ &\Rightarrow \lim_{m\to\infty} p(X_{N_m} = j|X_0 = i) = \lim_{m\to\infty} \frac{1}{m}\sum_{n=l_2-l_1+1}^{m} p_{ij}^{(n)} \end{aligned}$$

On the other hand, we have

$$p_{ij}^{(n)} = p(X_n = j|X_0 = i) = \sum_{i_2 \in K_{l_2}} p(X_n = j, X_{l_2-l_1} = i_2|X_0 = i)$$
$$= \sum_{i_2 \in K_{l_2}} \frac{p(X_n = j, X_{l_2-l_1} = i_2, X_0 = i)}{p(X_0 = i)}$$
$$= \sum_{i_2 \in K_{l_2}} p(X_n = j|X_{l_2-l_1} = i_2, X_0 = i)p(X_{l_2-l_1} = i_2|X_0 = i)$$
$$= \sum_{i_2 \in K_{l_2}} p_{i_2 j}^{(n-l_2+l_1)} p_{i i_2}^{(l_2-l_1)}$$

Consequently,

$$\lim_{m\to\infty} p(X_{N_m} = j|X_0 = i) = \sum_{i_2 \in K_{l_2}} p_{i i_2}^{(l_2-l_1)} \lim_{m\to\infty} \frac{1}{m} \sum_{n=l_2-l_1+1}^{m} p_{i_2 j}^{(n-l_2+l_1)}$$

$$\frac{1}{m} \sum_{n=l_2-l_1+1}^{m} p_{i_2 j}^{(n-l_2+l_1)} = \frac{1}{m} \sum_{n=1}^{m-l_2+l_1} p_{i_2 j}^{(n)}$$
$$= \frac{m-l_2+l_1}{m} \frac{1}{m-l_2+l_1} \sum_{n=1}^{m-l_2+l_1} p_{i_2 j}^{(n)}$$

$$\lim_{m\to\infty} \frac{1}{m-l_2+l_1} \sum_{n=1}^{m-l_2+l_1} p_{i_2 j}^{(n)} = \lim_{m\to\infty} \frac{1}{m} \sum_{n=1}^{m} p_{i_2 j}^{(n)}$$

Since both i_2 and j belongs to the periodic class K_{l_2}, we are in the same situation as in 1):

$$\lim_{m\to\infty} \frac{1}{m} \sum_{n=l_2-l_1+1}^{m} p_{i_2 j}^{(n-l_2+l_1)} = \lim_{m\to\infty} \sum_{n=1}^{m} p_{i_2 j}^{(n)} = \frac{1}{d}\pi'_j.$$

$$\lim_{m\to\infty} p(X_{N_m} = j|X_0 = i) = \sum_{i_2 \in K_{l_2}} p_{i i_2}^{(l_2-l_1)} \lim_{m\to\infty} \frac{1}{m} \sum_{n=l_2-l_1+1}^{m} p_{i_2 j}^{(n-l_2+l_1)}$$
$$= \sum_{i_2 \in K_{l_2}} p_{i i_2}^{(l_2-l_1)} \frac{1}{d}\pi'_j = \frac{1}{d}\pi'_j.$$

In all cases, we thus have

$$\lim_{m\to\infty} p(X_{N_m} = j|X_0 = i) = \frac{1}{d}\pi'_j = \pi_j.$$

It remains to be shown that the unambiguous limit distribution π is also an invariant distribution of P. We have

$$p(X_{N_m} = j|X_0 = i) = \frac{1}{m}\sum_{n=1}^{m} p_{ij}^{(n)}$$

and

$$\lim_{m\to\infty} p(X_{N_m} = j|X_0 = i) = \frac{1}{d}\pi'_j = \pi_j.$$

$$\begin{aligned}
(\pi P)_j &= \sum_{i2\in I} \pi_{i_2} p_{i_2 j} = \sum_{i2\in I} \lim_{m\to\infty} p(X_{N_m} = i_2|X_0 = i) p_{i_2 j} \\
&= \lim_{m\to\infty} \sum_{i2\in I} \frac{1}{m}\sum_{n=1}^{m} p_{ii_2}^{(n)} p_{i_2 j} = \lim_{m\to\infty} \frac{1}{m}\sum_{n=1}^{m}\sum_{i2\in I} p_{ii_2}^{(n)} p_{i_2 j} \\
&= \lim_{m\to\infty} \frac{1}{m}\sum_{n=1}^{m} p_{ij}^{(n+1)} \\
&= \lim_{m\to\infty} \frac{1}{m}\sum_{n=1}^{m} p_{ij}^{(n)} = \lim_{m\to\infty} p(X_{N_m} = j|X_0 = i) = \pi_j.
\end{aligned}$$

π is thus the unique invariant distribution of the irreducible aperiodic Markov chain.

Proof of (iii). Following Corollary 3.2, the irreducible MC must be recurrent because I is finite. If X is an irreducible Markov chain on a finite state space, it is either periodic or aperiodic. In both cases, (i) and (ii) allow us to state it has at least one invariant distribution. Theorem 3.18 tells us then that the Markov chain is positive recurrent and that the invariant distribution is uniquely defined by $\pi_i = \frac{1}{m_i} = \frac{1}{T_i}$.

Proof of (iv). In the general case, there is at least one **irreducible communicating** class $C \subset I$ because all states can't be transient owing to the finiteness of I. Let's suppose we start off in state $i \in C$. The random chain X on I is then equivalent to a random chain X_C restricted to C as state space and with a transition matrix P_C restricted to C. X_C and P_C are irreducible on the finite space C. We can thus apply (i) which tells us that there's an unambiguous distribution π_C such that $\pi_C P_C = \pi_C$. Let us define

$$\pi_i = \begin{cases} \pi_{C,i} \text{ if } i \in C \\ 0 \text{ otherwise.} \end{cases}.$$

$$\text{For } j \in C, (\pi P)_j = \sum_{i \in I} \pi_i p_{ij} = \sum_{i \in C} \pi_{C,i} p_{C,ij}$$
$$= \pi_{C,j} = \pi_j.$$

$$\text{For } j \in I \setminus C, (\pi P)_j = \sum_{i \in I} \pi_i p_{ij} = \sum_{i \in C} \pi_{C,i} p_{ij}$$
$$= 0 \text{ because } p_{ij} = 0 = \pi_j \text{ if } i \in C, j \in I \setminus C$$

Consequently, π is an invariant distribution of X.

Actually, a reducible Markov chain on a finite state space has as many invariant distributions as it has irreducible communicating classes.

3.6 Time reversal and reversible Markov chains

In this section, we shall take a look at reversible Markov chains. These are Markov chains where the evolution of the Markov chain going forward in time is the same as the evolution of the Markov chain going backward in time. We'll see that this requires very specific conditions involving both the transition matrix and the probability distribution at time n.

Theorem 3.22 Let P be an irreducible transition matrix which has an invariant distribution. Let's consider $(X_n)_{0 \leq n \leq N}$ which is a projection of a (π, P)- Markov chain and its **time-reversal** $Y_n = X_{N-n}$. In that case, $(Y_n)_{0 \leq n \leq N}$ is a $(\pi, \widehat{P})$- Markov chain with $\widehat{P} = (\hat{p}_{ij})_{i,j \in I}$ and

$$\widehat{p}_{ji} = \frac{\pi_i}{\pi_j} p_{ij} \text{ for all } i, j.$$

What is more, $\widehat{P}$ is irreducible and π is its invariant distribution.

Proof. Since π is invariant for P, we have for all $j \in I$

$$\sum_{i \in I} \widehat{p}_{ji} = \frac{1}{\pi_j} \sum_{i \in I} \pi_i p_{ij} = 1$$

which implies that $\widehat{P}$ is a stochastic matrix.

Moreover, owing to the very definition of $\widehat{p}_{ji}$

$$\sum_{j \in I} \pi_j \widehat{p}_{ji} = \sum_{j \in I} \pi_i p_{ij} = \pi_i \sum_{j \in I} p_{ij} = \pi_i.$$

This shows that π is an invariant distribution for $\widehat{P}$. Furthermore,

$$\begin{aligned}
&p(Y_0 = i_0, Y_1 = i_1, ..., Y_N = i_N) = p(X_0 = i_N, X_1 = i_{N-1}, ..., X_N = i_0) \\
&= \pi_{i_N} p_{i_N i_{N-1}} ... p_{i_1 i_0} \\
&= \pi_{i_N} \frac{\pi_{i_{N-1}}}{\pi_{i_N}} \widehat{p}_{i_{N-1} i_N} ... \frac{\pi_{i_0}}{\pi_{i_1}} \widehat{p}_{i_0 i_1} = \pi_{i_0} \widehat{p}_{i_0 i_1} ... \widehat{p}_{i_{N-1} i_N}.
\end{aligned}$$

By integration, we also find that for $n < N$,

$$p(Y_0 = i_0, Y_1 = i_1, ..., Y_n = i_n) = \pi_{i_0} \widehat{p}_{i_0 i_1} ... \widehat{p}_{i_{n-1} i_n}.$$

Corollary 3.1 allows us to state that $\widehat{P}$ is the transition matrix of the Markov chain $(Y_n)_{0 \leq n \leq N}$ [3]. We still have to prove that $\widehat{P}$ is irreducible. Since P is irreducible, $\forall i, j \in I, \exists i_1, i_2, ..., i_n \in I$ so that $i_1 = i, i_n = j$ and $p_{i_1 i_2} ... p_{i_{n-1} i_n} > 0$. As a consequence, we have

$$\widehat{p}_{i_n i_{n-1}} ... \widehat{p}_{i_2 i_1} = \frac{\pi_1}{\pi_n} p_{i_1 i_2} ... p_{i_{n-1} i_n} > 0,$$

which proves that $\widehat{P}$ is irreducible too.

Definition 3.12 A measure μ is called *reversible* with respect to a transition matrix P if the detailed balance conditions are satisfied:,

$$\forall i, j \in I, \mu_i p_{ij} = \mu_j p_{ji} \tag{3.17}$$

If we consider a large population of individuals that *independently* evolve following the same Markov process, the detailed balance conditions can be interpreted as follows: if μ is the probability distribution of the Markov chain values at time n, between time n and time $n+1$ there are as many individuals in state i switching to state j as there are individuals in state j switching to state i. In the case of physical and chemical processes, this is called a **dynamic equilibrium**. For example, a chemical reaction $A \rightleftharpoons B$ is in equilibrium if and only if there are (on average) as many molecules of A turning into B during dt as there are molecules of B turning into A: $\mu_A p_{AB} = \mu_B p_{BA}$. As a consequence, the following lemma makes a lot of sense intuitively:

Lemma 3.8 If μ is reversible with respect to P, then μ is invariant with respect to P.

Proof.

We have

$$(\mu P)_i = \sum_{j \in I} \mu_j p_{ji} = \sum_{j \in I} \mu_i p_{ij} = \mu_i.$$

[3] Actually, we're using a finite version of the theorem we won't further explore.

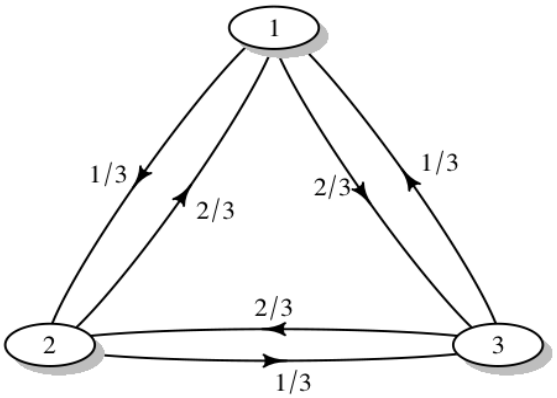


Figure 3.10: Example of a non-reversible Markov chain

Remark 3.3 It is only meaningful to define the transition matrix of the time reversal of a Markov chain $\widehat{P}$ with respect to an invariant distribution π. Otherwise, the Markov property in Theorem 3.22 wouldn't be valid.

Lemma 3.9 If μ is reversible with respect to P and if $(X_n)_{n\geq 0}$ is a (μ, P)-MC (a Markov chain starting with the probability distribution μ) then the time reversal of $(X_n)_{n\geq 0}$ is also a (μ, P)-MC, which means that $\widehat{P} = P$.

Proof.

Since μ is reversible with respect to P, we have $\mu_i p_{ij} = \mu_j p_{ji}$ and $\sum_{j\in I} \mu_j p_{ji} = \mu_i$. Moreover,

$$\widehat{p}_{ji} = \frac{\mu_i}{\mu_j} p_{ij} = \frac{\mu_j p_{ji}}{\mu_j}$$
$$\widehat{p}_{ji} = p_{ji}, \forall i, j \in I.$$

As a consequence,

$$\mu_i p_{ij} = \mu_j p_{ji} \Leftrightarrow \mu_i \widehat{p}_{ij} = \mu_j \widehat{p}_{ji}.$$

μ is thus the invariant distribution of the time-reversal Markov Chain whereas P is its transition matrix.

Example 3.16 (Non-reversible Markov chain) We consider the Markov chain shown in Figure 3.10. The transition matrix is the irreducible doubly stochastic matrix

$$P = \begin{pmatrix} 0 & 1/3 & 2/3 \\ 2/3 & 0 & 1/3 \\ 1/3 & 2/3 & 0 \end{pmatrix}$$

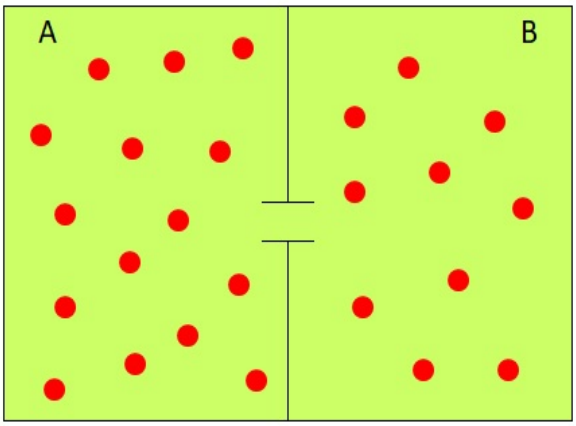


Figure 3.11: Ehrenfest model

and $\pi = (1/3, 1/3, 1/3)$ is the invariant distribution. It follows from the definition of $\widehat{P}$ that $\widehat{P} = P^T$. If the Markov chain were reversible, the detailed balance conditions would be satisfied so that

$$\pi_i p_{ij} = \pi_j p_{ji}.$$

Since the invariant distribution is colinear with $(1, 1, 1)$, the matrix would have to be symmetrical. This is clearly not the case, therefore the Markov chain is not reversible. The physical meaning of this can be easily recognised: the Markov chain is intuitively two times more likely to go clockwise than counterclockwise. This is the example of an irreversible process. If all transition probabilities had been equal to $1/2$, the Markov chain and the physical process would have been reversible so that the past and the future would have been indistinguishable.

Example 3.17 (Model of Ehernfest)

Let us consider a fluid contained in two containers separated by a membrane as can be seen in Figure 3.11. Let us suppose we choose to describe the content of the two containers through the total number of molecules N instead of using the concentrations. Let X_n be the number of molecules in the right container at time n. If we assume that the probability that a molecule goes from one container to the other is uniform and does not vary in time and that there can only be one molecule moving by time unit, we (logically) have

$$\forall n > 0,\ p(X_{n+1} = i + 1 | X_n = i) = \frac{N - i}{N},\ p(X_{n+1} = i - 1 | X_n = i) = \frac{i}{N}$$

Let us consider the probability distribution π with $\pi_j = 2^{-N}\binom{N}{j}$. We have

$$\sum_{i=1}^{N} \pi_i = 2^{-N} \sum_{i=1}^{N} \binom{N}{i} = 2^{-N} 2^N = 1$$

and for $0 < j < N$

$$\begin{aligned}
&(\pi P)_j = \sum_{i=0}^{N} \pi_i p_{i,j} = \pi_{j-1} p_{j-1,j} + \pi_{j+1} p_{j+1,j} \\
&= 2^{-N} \binom{N}{j-1} \frac{N-j+1}{N} + 2^{-N} \binom{N}{j+1} \frac{j+1}{N} \\
&= 2^{-N} \left(\frac{N!}{(j-1)!(N-j+1)!} \frac{N-j+1}{N} + \frac{N!}{(j+1)!(N-j-1)!} \frac{j+1}{N} \right) \\
&= \frac{2^{-N} N!}{N} \left(\frac{1}{(j-1)!(N-j)!} + \frac{1}{(j)!(N-j-1)!} \right) \\
&= \frac{2^{-N} N!}{N} \left(\frac{j}{j!(N-j)!} + \frac{N-j}{(j)!(N-j)!} \right) \\
&= \frac{2^{-N} N!}{j!(N-j)!} = \pi_j.
\end{aligned}$$

We also have

$$\begin{aligned}
(\pi P)_0 &= \sum_{i=0}^{N} \pi_i p_{i,0} \\
&= \pi_1 p_{1,0} = 2^{-N} \binom{N}{1} \frac{1}{N} = 2^{-N} = \pi_0
\end{aligned}$$

and

$$\begin{aligned}
(\pi P)_N &= \sum_{i=0}^{N} \pi_i p_{i,N} \\
&= \pi_{N-1} p_{N-1,N} = 2^{-N} \binom{N}{N-1} \frac{1}{N} = 2^{-N} = \pi_N
\end{aligned}$$

We thus always have $(\pi P)_j = \pi_j$. π is therefore the invariant distribution of X.

We also have $\forall i < N$

$$\pi_i p_{i,i+1} = 2^{-N}\binom{N}{i}\frac{N-i}{N} = 2^{-N}\frac{N!}{i!(N-i-1)!}\frac{1}{N}$$
$$= 2^{-N}\frac{N!}{(i+1)!(N-(i+1))!}\frac{i+1}{N} = 2^{-N}\binom{N}{i+1}\frac{i+1}{N}$$
$$= \pi_{i+1}p_{i+1,i}$$

For $|j-i|>1$, $p_{i,j} = p_{j,i} = 0$. Therefore, we always have

$$\pi_i p_{i,j} = \pi_j p_{j,i}.$$

The Markov chain is thus reversible with respect to π.

There are two physical interpretations of the invariant distribution (also called equilibrium distribution). According to the ensemble interpretation, we consider a very large number of containers such that $\forall i \in [1;n]$ the percentage of containers where $X_n = i$ is given by π_i. At time $n+1$, the percentage of containers where $X_n = i$ will still be equal to π_i $\forall i \in [1;n]$ because there are as many containers where the right half loses a particle as there are containers where it is the left half which loses a particle (dynamic equilibrium).

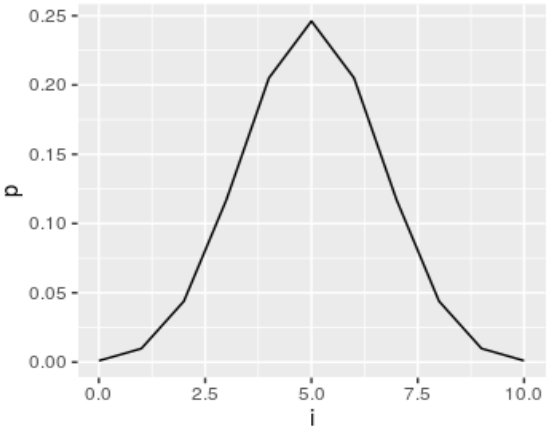
According to the time-average interpretation, if we consider a very long period of time T, $\pi_i T$ is the amount of time where $X_n = i$. The fundamental concept which allows us to link the two interpretations is the **ergodic theorem** we'll explore later on. The probability distribution at the equilibrium is shown in 3.12. We can see that the distribution quickly becomes a peak centred at N as N increases. In the case of a real liquid or gas in two containers separated by a membrane (which involves millions and millions of molecules), this means that the ratio of the number of molecules in the right container to the total number of molecules will be extremely close to 0.5 so that *at equilibrium*, the concentrations in the two containers will be equal in every practical sense (π_i for $i \neq N$ is extremely small).

Example 3.18 (Random walk on a graph).

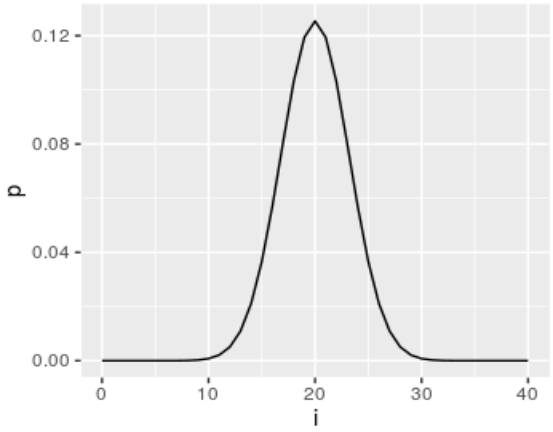
A graph is a set $G = (V, K)$ where V is the countable set of the *vertices* and K is the set of all edges connecting the vertices and thus a subset of $\{(V_i, V_j), i \neq j\}$. All points are either directly or indirectly connected. The set K can be described through an adjacency matrix $A = (a_{ij})_{i,j\in V}$:

$$a_{ij} = \begin{cases} 1 : (v_i, v_j) \in K \\ 0 \text{ otherwise} \end{cases}$$

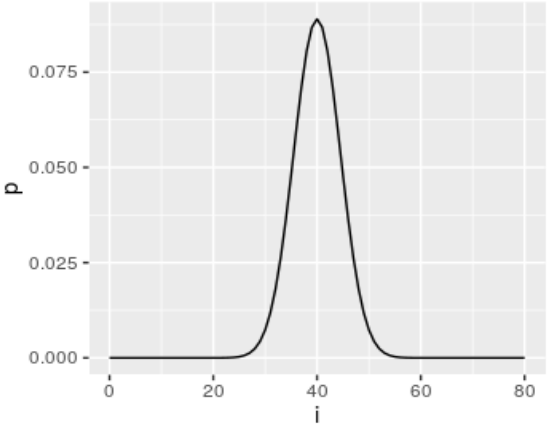
We'll only consider here undirected graphs without loops, i.e. $a_{ij} = a_{ji}$ and $a_{ii} = 0$. While $\{v_i,\ i \in V\}$ might be countably infinite, we assume that the



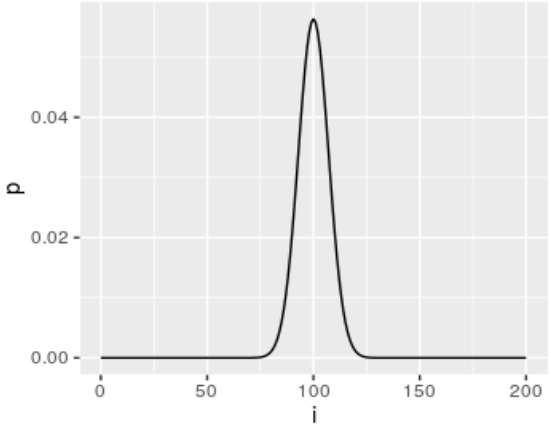
(a) N = 10



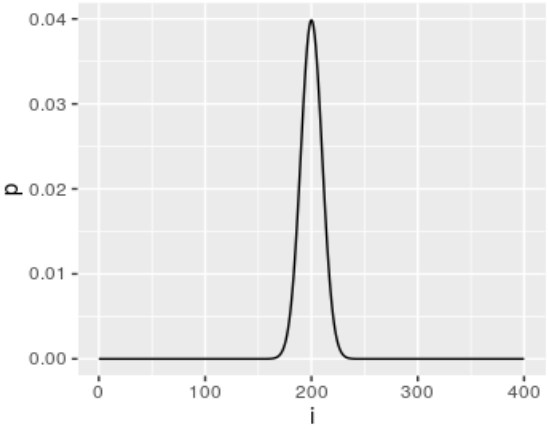
(b) N = 40



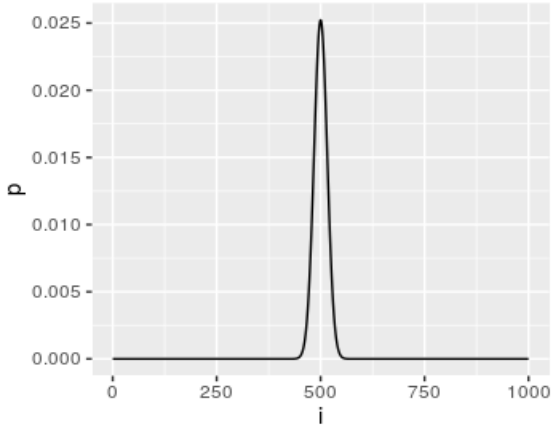
(c) N = 80



(d) N = 200

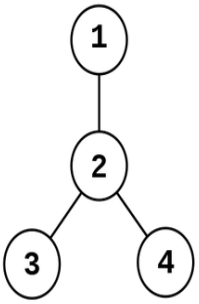


(e) N = 400



(f) N = 1000

Figure 3.12: Equilibrium distribution (Ehrenfest model)



	1	2	3	4
1	0	1	0	0
2	1	0	1	1
3	0	1	0	0
4	0	1	0	0

Figure 3.13: A graph with its adjacency matrix

subset of all edges around a vertex i is finite:

$$\forall i \in V, 0 < \mu_i = \sum_{j \in V} a_{ij} < \infty$$

and we pose that

$$p_{ij} = \frac{a_{ij}}{\mu_i}.$$

P is then the transition matrix of a random walk on the graph G. If the Markov chain is in state i at time n, the next state at $n+1$ is randomly and uniformly chosen out of all neighbours of i. An example of graph along its adjacency matrix is shown in Figure 3.13. Based on the definition of p_{ij}, it is clear that

$$\mu_i p_{ij} = a_{ij} = a_{ji} = \mu_j p_{ji}.$$

As a consequence, μ is a reversible measure with respect to P. If V is countably finite, $\sum_{i \in V} \mu_i < \infty$ and π with

$$\pi_i = \frac{\mu_i}{\sum_{i \in V} \mu_i}$$

is a reversible probability distribution. In the case of the graph in Figure 3.13, we have $\pi_1 = \pi_3 = \pi_4 = 1/6$ and $\pi_2 = 1/2 = 3\pi_1$.

Instead of using binary values for the a_{ij}, we could also use real positive values so that

$$a_{ij} = a_{ji} \geq 0,\ 0 < \mu_i = \sum_{j \in V} a_{ij} < \infty \text{ and } p_{ij} = \frac{a_{ij}}{\mu_i}.$$

The next vertex is then chosen according to its relative weight.

Definition 3.13 (Scalar product)

Let P be the transition matrix of an aperiodic, irreducible and positive recurrent Markov chain on a state space I. We then know that for all $i \in I$, $(p_{i,j}^{(n)})_{j \in I}$ converges towards the unambiguous invariant distribution $(\pi_j)_{j \in I}$.

For any probability distribution π on I with $\pi > 0$, $i \in I$, we define the scalar product as

$$< f, g >_\pi = \sum_{i \in I} f_i g_i \pi_i \tag{3.18}$$

$$\text{for } f, g \in L^2(\pi) = \{h : I \to \mathbb{R} : \|h\|_\pi^2 < \infty, \text{ with } \|h\|_\pi^2 = \sum_{i \in I} (h_i)^2 \pi_i\}$$

Definition 3.14 The matrix $\widehat{P}$ with

$$\widehat{p}_{i,j} = \frac{p_{j,i}\pi_j}{\pi_i}, \quad i, j \in I$$

is called the ($\pi-$) adjoint matrix of P. If $P = \widehat{P}$, the matrix is symmetrical *with respect to* π. In that case, P satisfies the detailed-balance conditions

$$\pi_i p_{i,j} = \pi_j p_{j,i} \quad i, j \in I$$

so that the Markov chain is reversible with respect to π.

Lemma 3.10 If π is symmetrical with respect to P, π is symmetrical with respect to $P^{(k)}$, $\forall k > 0$.

Proof.

For $k = 1$ this is true. Let's suppose it is true for $k' \leq k$, i.e.

$$\pi_i p_{i,j}^{(k')} = \pi_j p_{j,i}^{(k')}.$$

$$\begin{aligned}
\pi_i p_{i,j}^{(k+1)} &= \sum_{i2 \in I} \pi_i p_{i,i_2}^{(k)} p_{i_2,j} \\
&= \sum_{i2 \in I} \pi_{i_2} p_{i_2,i}^{(k)} p_{i_2,j} = \sum_{i2 \in I} \pi_{i_2} p_{i_2,j} p_{i_2,i}^{(k)} \\
&= \sum_{i2 \in I} \pi_j p_{j,i_2} p_{i_2,i}^{(k)} = \pi_j p_{j,i}^{(k+1)}
\end{aligned}$$

It is thus also true for $k + 1$ and hence for all $k > 1$.

If π is reversible with respect to P, P is symmetrical with respect to π and we immediately see that the matrices P and $\widehat{P}$ are normal so that

$$P\widehat{P} = \widehat{P}P.$$

Theorem 3.23 The following relationship holds:

$$< Pf, g >_\pi = < f, \widehat{P}g >_\pi, f, g \in L^2(\pi),$$

with $Pf_i = \sum_{j \in I} p_{i,j} f_j$.

Proof.

$$
\begin{aligned}
&< Pf, g >_\pi = \sum_{i \in I} (Pf)_i g_i \pi_i \\
&= \sum_{i \in I} \sum_{k \in I} p_{ik} f_k g_i \pi_i \\
&= \sum_{i \in I} \sum_{k \in I} \widehat{p}_{k,i} \pi_k f_k g_i \\
&= \sum_{k \in I} \pi_k f_k \sum_{i \in I} \widehat{p}_{k,i} g_i \\
&= \sum_{k \in I} \pi_k f_k (\widehat{P} g)_k \\
&= \sum_{k \in I} < f, \widehat{P} g >_\pi
\end{aligned}
$$

Example 3.19 (Torus of length N)

Let $I = \{0, 1, ..., N-1\}$. Let $q(k), k \in I$ be a probability distribution. The transition probabilities are defined in such a way that $\forall i, j \in I, p_{i,j} = q((i-j) mod N)$. An example of torus can be seen in Example 3.16. Is this Markov chain irreducible? Not always, because if $q(2) = 1$ and if we start off in state i, we will never be able to reach state $i+1$. Let π be the uniform distribution on I, i.e. $\pi_i = \frac{1}{N}, \forall i \in I$. We have

$$
\begin{aligned}
(\pi P)_j &= \sum_{i \in I} \pi_i p_{i,j} = \frac{1}{N} \sum_{i \in I} p_{i,j} \\
&= \frac{1}{N} \sum_{i \in I} q\big((i-j) mod N\big) \\
&= \frac{1}{N} \sum_{i \in I} q(i) = \frac{1}{N} = \pi_j.
\end{aligned}
$$

π is thus invariant with respect to P. By definition, we have

$$
\widehat{p}_{i,j} = \frac{p_{ji} \pi_j}{\pi_i} = p_{ji}
$$

and

$$
\begin{aligned}
(\widehat{P}P)_{ij} &= \sum_{k\in I} \widehat{p}_{i,k}p_{k,j} \\
&= \sum_{k\in I} p_{k,i}p_{k,j} \\
&= \sum_{k\in I} q\big((k-i)modN\big)q\big((k-j)modN\big) \\
&= \sum_{k\in I} q\big((k-i)modN\big)q\big((k-i+i-j)modN\big) \\
&= \sum_{k\in I} q\big((k-i)modN\big)q\big([(k-i)modN+(i-j)modN]modN\big) \\
&= \sum_{K\in I} q(K)q\big([K+(i-j)modN]modN\big) \\
&= \sum_{K\in I} q(K)q\big([K+a]modN\big) \\
&\text{with } a=(i-j)modN \in I.
\end{aligned}
$$

On the other hand, we have

$$
\begin{aligned}
(P\widehat{P})_{ij} &= \sum_{k'\in I} p_{i,k'}\widehat{p}_{k',j} \\
&= \sum_{k'\in I} p_{i,k'}p_{j,k'} \\
&= \sum_{k'\in I} q\big((i-k')modN\big)q\big((j-k')modN\big) \\
&= \sum_{k'\in I} q\big((i-k')modN\big)q\big((j-i+i-k')modN\big) \\
&= \sum_{k'\in I} q\big((i-k')modN\big)q\big([(j-i)modN+(i-k')modN]modN\big) \\
&= \sum_{K'\in I} q(K')q\big([K'-a]modN\big)
\end{aligned}
$$

In order to prove that

$$(\widehat{P}P)_{ij} = \sum_{K\in I} q(K)q\big([K+a]modN\big) = (P\widehat{P})_{ij} = \sum_{K'\in I} q(K')q\big([K'-a]modN\big)$$

we can show that

$$\forall K\in I, \exists K'\in I \text{ with } q(K')q\big([K'-a]modN\big) = q(K)q\big([K+a]modN\big).$$

For $K' = [K + a] mod N$, we have

$$
\begin{aligned}
q(K')q([K' - a] mod N) &= q\Big([K + a] mod N\Big) q\Big(([K + a] mod N - a) mod N\Big) \\
&= q\Big([K + a] mod N\Big) q\Big(([K + a] mod N - a mod N) mod N\Big) \text{ (because } a \in I \text{)} \\
&= q\Big([K + a] mod N\Big) q\Big([K + a - a] mod N\Big) \\
&= q\Big(K mod N\Big) q\Big([K + a] mod N\Big) = q(K)q((K + a) mod N)
\end{aligned}
$$

We know thus that $(\widehat{P}P)_{ij} = (P\widehat{P})_{ij}$ so that P and $\widehat{P}$ are normal. However, as we saw in Example 3.16, there is no guarantee that the uniform distribution π is reversible with respect to P so that the detailed balance conditions are satisfied. $\widehat{P}P = P\widehat{P}$ is therefore a necessary but not a sufficient condition for reversibility.

Example 3.20 Let $I = \mathbb{Z}$ and let the stochastic matrix P be defined through $p_{i,j} = p_{-j,-i} > 0, |i - j| \leq 1$ and $p_{i,j} = 0, |i - j| > 1$. The corresponding Markov chain is thus irreducible and aperiodic and it either jumps to one of its direct neighbours or it stays at the same position. We know that if the coefficients of P are such that the Markov chain is positive recurrent, according to Theorem 3.20 there is then an unambiguous invariant distribution that is positive. We can further show that P is symmetric with respect to π, i.e. the detailed balance conditions are satisfied and the Markov chain starting in π is reversible (Exercise!).

3.7 Convergence of Markov chains on finite state spaces

In this section, we limit ourselves to aperiodic irreducible Markov chains on finite state spaces. As we saw before, they are necessarily positive definite and must thus converge towards an unambiguous invariant distribution. We want to determine their convergence rate, i.e. *how fast* they converge towards that distribution. For that sake, we need to define the **distance** between two probability distributions. If μ and ν are two probability distributions on a discrete or continuous state space Ω, the *total variation distance* is such that

$$\|\mu - \nu\|_{TV} = max_{A \in \Omega} |\mu(A) - \nu(A)|.$$

As a consequence, a sequence of probability distributions converges towards a stable probability distribution if and only if the probability of each state converges towards the stable probability (this is a necessary condition implied by the definition). In what follows, we'll call the discrete state space I.

Lemma 3.11

$$\|\mu - \nu\|_{TV} = max_{A \subset I} |\mu_A - \nu_A|$$
$$= \frac{1}{2} \Sigma_{i \in I} |\mu_i - \nu_i|$$

Proof.
Let $B = \{i \in I : \mu_i \geq \nu_i\}$ and A be any event. We have

$$\mu(A) - \nu(A) = \mu(A \bigcap B) - \nu(A \bigcap B) + \mu(A \bigcap B^c) - \nu(A \bigcap B^c).$$

$\forall i \in A \bigcap B^c, \mu_i - \nu_i \leq 0$. As a consequence, $\mu(A \bigcap B^c) - \nu(A \bigcap B^c) \leq 0$. It follows that

$$\mu(A) - \nu(A) \leq \mu(A \bigcap B) - \nu(A \bigcap B).$$

$$\mu(A \bigcap B) - \nu(A \bigcap B) = \sum_{i \in A \bigcap B} (\mu_i - \nu_i) \leq \sum_{i \in B} (\mu_i - \nu_i) = \mu(B) - \nu(B)$$

because $\mu_i - \nu_i \geq 0$ for $i \in B$ and $A \bigcap B \subset B$. Therefore, we have

$$\mu(A) - \nu(A) \leq \mu(A \bigcap B) - \nu(A \bigcap B) \leq \mu(B) - \nu(B).$$

Likewise, by switching the roles of μ and ν, we can show that

$$\nu(A) - \mu(A) \leq \nu(A \bigcap B^c) - \mu(A \bigcap B^c) \leq \nu(B^c) - \mu(B^c).$$

$$\nu(B^c) - \mu(B^c) = 1 - \nu(B) - 1 + \mu(B) = \mu(B) - \nu(B)$$

$$\mu(A) - \nu(A) \geq -(\nu(B^c) - \mu(B^c)) = -(\mu(B) - \nu(B))$$

This leads to

$$-(\mu(B) - \nu(B)) \leq \mu(A) - \nu(A) \leq \mu(B) - \nu(B).$$

Consequently, since this is valid $\forall A \subset I$,

$$\|\mu - \nu\|_{TV} = max_{A \subset I} |\mu_A - \nu_A| \leq \mu(B) - \nu(B).$$

This bound is realised for $A = B$ or $A = B^c$. Therefore, we have

$$\|\mu - \nu\|_{TV} = \mu(B) - \nu(B) = \nu(B^c) - \mu(B^c)$$

$$\|\mu - \nu\|_{TV} = \sum_{i, \mu_i \geq \nu_i} (\mu_i - \nu_i) = \sum_{i, \mu_i \leq \nu_i} (\nu_i - \mu_i).$$

$$\frac{1}{2}\sum_{i,\mu_i\geq\nu_i}(\mu_i-\nu_i)=\frac{1}{2}\sum_{i,\mu_i\geq\nu_i}|\mu_i-\nu_i|=\frac{1}{2}\|\mu-\nu\|_{TV}$$

$$\frac{1}{2}\sum_{i,\mu_i\leq\nu_i}(\nu_i-\mu_i)=\frac{1}{2}\sum_{i,\mu_i\leq\nu_i}|\mu_i-\nu_i|=\frac{1}{2}\|\mu-\nu\|_{TV}$$

Finally, we get

$$\|\mu-\nu\|_{TV}=\frac{1}{2}(\sum_{i,\mu_i\geq\nu_i}|\mu_i-\nu_i|+\sum_{i,\mu_i\leq\nu_i}|\mu_i-\nu_i|)$$

$$\|\mu-\nu\|_{TV}=\frac{1}{2}\sum_{i\in I}|\mu_i-\nu_i|.$$

Definition 3.15 (Distance to the stationary distribution)

Let us now consider the initial distribution of an irreducible Markov chain on a finite state space $(\nu_i=\delta_i)_{i\in I}$. We want to examine the convergence of such a Markov chain. We therefore take a look at the probability distribution of the MC at time n: $\nu_j^{(n)}=p(X_n=j|X_0=i)=p_{ij}^{(n)},\forall j\in I$. We want to know how far $(\nu_j^{(n)})_{j\in I}$ is from the stationary distribution $(\pi_j)_{j\in I}$ for any possible initial state $i\in I$. This leads to the definition of the **distance** to the stationary distribution:

$$d(n):=max_{i\in I}\|P_{i,\cdot}^{(n)}-\pi\|_{TV}.$$

In other words, we consider the worst-case scenario.

When and in which sense does a Markov chain converge towards its stationary distribution? As we saw in the previous sections, for a finite state space I, the Markov chain must be both irreducible (so that every state can be reached) and aperiodic (so that it can converge towards an unambiguous distribution instead of constantly fluctuating between periodic classes). We'll soon see that the speed of convergence must be exponential.

Lemma 3.12 Let π be a probability distribution on I and Π be the $|I|\times|I|$ matrix with identical rows equal to π. Then for the $|I|\times|I|$ matrices M and M', the following statements hold:

(i) If M is stochastic, we have $M\Pi=\Pi$.

(ii) For any matrix M with $\pi M=\pi$ we have $\Pi M=\Pi$.

(iii) If M and M' are stochastic, so is MM'.

Proof: $\forall i,j\in I$, we have

(i)

$$(M\Pi)_{ij}=\sum_{k\in I}M_{ik}\Pi_{kj}=\sum_{k\in I}M_{ik}\pi_j=\pi_j=\Pi_{ij}.$$

(ii)

$$(\Pi M)_{ij} = \sum_{k\in I} \Pi_{ik} M_{kj} = \sum_{k\in I} \pi_k M_{kj} = \pi_j = \Pi_{ij}.$$

(iii)

$$\sum_{j\in I} (MM')_{ij} = \sum_{j\in I} \sum_{k\in I} M_{ik} M'_{kj} = \sum_{k\in I} M_{ik} \sum_{j\in I} M'_{kj} = 1.$$

Theorem 3.24 (Convergence theorem)

Let P be the transition matrix of an irreducible and aperiodic Markov chain $X = (X_0, X_1, ...)$ on a finite state space whose invariant distribution is given by π. Then there exists $\alpha \in (0; 1)$ and $C > 0$ such that

$$d(n) := max_{i\in I} \| P^{(n)}_{i,.} - \pi \|_{TV} \leq C\alpha^n.$$

Proof. Since P is irreducible and aperiodic on a finite state space, $\exists r > 0$ such that $p^{(r)}_{ij} > 0, \forall i, j \in I$. Let Π be the $|I| \times |I|$ matrix with identical rows equal to π. Since $\pi_j > 0$ (by virtue of the positive recurrence) and $p^{(r)}_{ij} > 0$, it is possible to find a sufficiently small $\delta \in (0; 1)$ such that

$$p^{(r)}_{ij} \geq \delta\pi_j, \forall i, j \in I.$$

We then define $\theta := 1 - \delta$. We further introduce the matrix

$$Q := \theta^{-1} P^r + (1 - \theta^{-1})\Pi.$$

This means that

$$\begin{aligned} \theta Q &= P^r + \theta(1 - \theta^{-1})\Pi \\ \theta Q &= P^r + (\theta - 1)\Pi \\ P^r &= (1 - \theta)\Pi + \theta Q \end{aligned}$$

We have $\forall i, j \in I$

$$\begin{aligned} Q_{ij} &= \theta^{-1} p^{(r)}_{ij} + (1 - \theta^{-1})\pi_{ij} \\ &= \theta^{-1} p^{(r)}_{ij} + \theta^{-1}(\theta - 1)\pi_{ij} \\ &= \theta^{-1} p^{(r)}_{ij} - \theta^{-1}\delta\pi_{ij} \\ &= \theta^{-1}(p^{(r)}_{ij} - \delta\pi_{ij}) \geq 0. \end{aligned}$$

Moreover,

$$\sum_{j\in I} Q_{ij} = \theta^{-1}(\sum_{j\in I} p_{ij}^{(r)} - \sum_{j\in I} \delta\pi_{ij})$$
$$= \theta^{-1}(1-\delta) = 1.$$

Q is thus a stochastic matrix. We want to prove inductively that

$$P^{kr} = (1-\theta^k)\Pi + \theta^k Q^k.$$

Let' suppose this is true for all $k' \leq k$. We saw just above that this is the case for $k = 1$.

$$P^{r(k+1)} = P^{rk}P^r = [(1-\theta^k)\Pi + \theta^k Q^k]P^r$$
$$= (1-\theta^k)\Pi P^r + \theta^k Q^k P^r$$

Because of Lemma 3.12, since P is a stochastic matrix, we have $\Pi P^r = \Pi$.

On the other hand,

$$\theta^k Q^k P^r = \theta^k Q^k[(1-\theta)\Pi + \theta Q]$$
$$= \theta^k Q^k(1-\theta)\Pi + \theta^k Q^k \theta Q$$

Since Q is a stochastic matrix, so is Q^k and $Q^k\Pi = \Pi$. Hence

$$\theta^k Q^k P^r = \theta^k(1-\theta)\Pi + \theta^{k+1}Q^{k+1}.$$

This leads to

$$P^{r(k+1)} = (1-\theta^k)\Pi + \theta^k(1-\theta)\Pi + \theta^{k+1}Q^{k+1}$$
$$= (1-\theta^k+\theta^k-\theta^{k+1})\Pi + \theta^{k+1}Q^{k+1}$$
$$= (1-\theta^{k+1})\Pi + \theta^{k+1}Q^{k+1}$$

We have thus shown that $\forall k \geq 1$,

$$p^{rk} = (1-\theta^k)\Pi + \theta^k Q^k.$$

Let's now consider $k_2 \in \mathbb{N}$. We have

$$P^{rk+k_2} = P^{rk}P^{k_2} = (1-\theta^k)\Pi P^{k_2} + \theta^k Q^k P^{k_2}$$
$$= (1-\theta^k)\Pi + \theta^k Q^k P^{k_2}$$
$$= \Pi - \theta^k\Pi + \theta^k Q^k P^{k_2}$$

Consequently,

$$P^{rk+k_2} - \Pi = \theta^k (Q^k P^{k_2} - \Pi).$$

Now, we're almost finished. Let us consider a given initial state $i \in I$ and the total variation distance related to $P_{i,\cdot}^{rk+k_2}$ and π. By using Lemma 3.12 , since Q and P are stochastic matrix, so is $Q^k P^{k_2}$.

Because of Lemma 3.15 , the total variation distance is given by

$$\begin{aligned}
&\|P_{i,\cdot}^{(rk+k_2)} - \Pi\|_{TV} = \frac{1}{2} \sum_{j \in I} |P_{i,j}^{rk+k_2} - \pi_j| \\
&= \frac{1}{2} \sum_{j \in I} |P_{i,j}^{rk+k_2} - \Pi_{ij}| \\
&= \theta^k \frac{1}{2} \sum_{j \in I} |(Q^k P^{k_2})_{ij} - \Pi_{ij}| \\
&= \theta^k \|(Q^k P^{k_2})_{i,\cdot} - \Pi\|_{TV} \leq \theta^k \\
&\text{because } \|(Q^k P^{k_2})_{i,\cdot} - \Pi\|_{TV} \leq 1.
\end{aligned}$$

We have thus $\|P_{i,\cdot}^{(rk+k_2)} - \Pi\|_{TV} \leq \theta^k$.

Let us now consider step $n > r$. $\exists k \in \mathbb{N}, m \in \{0, 1, ..., r-1\}$ such that $n = rk + m$. Therefore,

$$\theta^k = \theta^{\frac{n-m}{r}} = (\theta^{\frac{1}{r}})^n (\frac{1}{\theta})^{\frac{m}{r}}.$$

$$\|P_{i,\cdot}^{(n)} - \Pi\|_{TV} = \|P_{i,\cdot}^{(rk+m)} - \Pi\|_{TV} \leq \theta^k = \alpha^n C^{\frac{m}{r}}$$

Let's define $\alpha = \theta^{\frac{1}{r}}$ and $C = \theta^{-1}$.

Since $0 < \theta < 1$ and $n > r > 1$ so that $\frac{1}{r} > 0$, $0 < \alpha < 1$ and $C > 1$. Since $m < r$, $\frac{m}{r} < 1$ and $C^{\frac{m}{r}} < C$. As a consequence,

$$\theta^k < \alpha^n C.$$

This leads to the desired result:

$$\|P_{i,\cdot}^{(n)} - \Pi\|_{TV} \leq \theta^k < \alpha^n C.$$

3.8 Ergodic theorem

In this section, we'll examine the long-term probabilistic behaviour of an irreducible Markov chain. We'll first mention important related theorems without trying to prove them.

Theorem 3.25 (Weak law of large numbers)

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables with finite expected values ($E(|X_n|) < \infty, \forall n \in \mathbb{N}$) that satisfy the following conditions:

- the X_n are pairwise uncorrelated, which means that

$$cov(X_i, X_j) = 0 \text{ for } i \neq j.$$

- $$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} var(X_i) = 0.$$

In that case, $(X_n)_{n\in\mathbb{N}}$ fulfils the weak law of large numbers. This means that the **centred mean**

$$\overline{X_n} := \frac{1}{n} \sum_{i=1}^{n} (X_i - E(X_i))$$

converges in probability towards 0, i.e.

$$\lim_{n\to\infty} p(|\overline{X_n}| \geq \epsilon) = 0$$

for all $\epsilon > 0$.

This means that

$$\forall \epsilon > 0, \forall \epsilon' > 0, \exists n_1, n \geq n_1$$
$$\Rightarrow p(|\overline{X_n}| > \epsilon) < \epsilon'.$$

Theorem 3.26 (Strong law of large numbers)

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of identically distributed and pairwise independent random variables with $E(|X_n|) < \infty, \forall n \in \mathbb{N}$. Then

$$\overline{X_n} := \frac{1}{n} \sum_{i=1}^{n} (X_i - E(X_i))$$

converges almost certainly towards 0, i.e.

$$p(\lim_{n\to\infty} \overline{X_n} = 0) = 1.$$

This means that for any infinite sequence $(X_n)_{n\in\mathbb{N}}$,

$$\forall \epsilon > 0, \exists N_1, n \geq N_1$$
$$\Rightarrow |\overline{X_n}| < \epsilon.$$

Figure 3.14: Dice rolling

(N_1 is a random variable).

The strong law of large numbers logically implies the weak law of large numbers whereas the reverse isn't true.

Another way to express the law of large numbers is that we have *almost certainly*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k = \mu = E(X_1) = E(X_n).$$

We're now interested in the limit theorem describing the behaviour of a Markov chain in the long run where the X_i are no longer independent of one another. This will lead us to the notion of ergodicity. We already saw an example earlier on regarding the Ehrenfest model (see 3.17). If we consider a large number of elements *at equilibrium* following the same probabilistic law, the ergodic theorem tells us that the proportion of elements in a given state is equal to the proportion of time that a single element would be in that state after a very long period of time. Several other examples follow now.

Example 3.21 (Dice rolling)

We're interested in examining the behaviour of a large number of dice in a casino (see Fig. 3.21). To approximate the discrete probability distribution describing the die behaviour, we have two options. We can either throw a very large number of identical dice (e.g. 10,000) at the same time and measure the relative frequencies of the different outcomes (approximation of the ensemble average). Or we can throw the same die 10,000 times and measure the corresponding frequencies. This is possible because the die is an ergodic system.

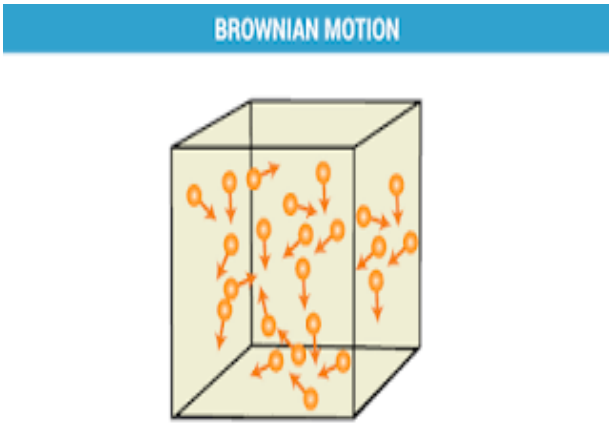


Figure 3.15: Brownian motion

Example 3.22 (Brownion motion)

Another example is the Brownion motion which is the random motion of particles in a container (see Figure 3.15). Contrarily to the topic of this chapter, it is a Markov process in continuous time and with a continuous state space (namely the positions of the particles in 3D). Nevertheless, the most fundamental theorems governing the behaviour of Markov chains can be extended to that case. If we wait a long period of time, if we consider a large number of particles, the percentage of particles in a volume V of the container (i.e. the relative density) will be approximately equal to the proportion of time **one single particle** spend in V. This has important practical consequences. If we want to determine the particulate density within the container, all we have to do is following a single particle over a long period of time. This can considerably simplify the task and reduce all related costs. An example of application can be found in [1].

Example 3.23 (Stock Market)

The stock market is not an ergodic system. After a very long-time, the average value of a large number of shares ("actions boursiaires" in French) will not generally be equal to the time-average of a single share.

In what follows, we'll call $V_i(n)$ the number of visits of the MC in state i until time n. $V_i(n)$ is also called the local time in i. It is defined as:

$$V_i(n) := \sum_{k=0}^{n-1} 1_{\{X_k=i\}} \leq n.$$

We then define recursively the return times in i:

$$T_i^{(0)} = 0, \text{ and } T_i^{(r+1)} := inf\{n \geq T_i^{(r)} + 1 : X_n = i\}, r = 0, 1, 2, ...$$

Figure 3.16: Stock Market

$T_i^{(0)} = 0$ is the starting time even if i isn't the first value of the Markov chain. We further introduce the waiting times

$$S_i^{(r)} = \begin{cases} T_i^{(r)} - T_i^{(r-1)} : \text{ if } T_i^{(r-1)} < \infty \\ 0 \text{ otherwise} \end{cases}$$

Lemma 3.13 For $r = 2, 3, ...$, the following statement holds: given that $T_i^{(r-1)} < \infty$, $S_i^{(r)}$ is independent of $\{X_m : m \leq T_i^{(r-1)}\}$ and

$$p(S_i^{(r)} = n | T_i^{(r-1)} < \infty) = p_i(T_i^{(1)} = n).$$

Proof.

$T_i^{(r-1)}$ is a Markov time because the event $T_i^{(r-1)} = n$ only depends on $X_0, X_1, X_2, ..., X_n$ but not on future values of the Markov chain. Let us pose $T = T_i^{(r-1)}$. Because of the very definition of $T_i^{(r-1)}$, $X_T = i$ for $T < \infty$. If we apply the strong Markov property on the Markov time T, given that $T < \infty$, $(X_{T+n})_{n \geq 0}$ is a (δ_i, P)-MC which is independent of $X_0, ..., X_T$. We further have

$$S_i^{(r)} = inf\{n \geq 1 : X_{T+n} = i\},$$

that is $S_i^{(r)}$ is the first return time of $(X_{T+n})_{n \geq 0}$ into state i. Given the fact that we consider all Markov chains in this lecture to be homogeneous in time, this leads to

$$p(S_i^{(r)} = n | T_i^{(r-1)} < \infty) = p_i(T_i^{(1)} = n).$$

Theorem 3.27 (Ergodic theorem applied to Markov chains)

Let $(X_n)_{n \geq 0}$ be an irreducible Markov chain whose transition matrix is P and whose initial distribution is λ. Let $m_i = E_i[T_i]$ be the expected return

time in state i. m_i can be infinite in the case of Markov chains that aren't positive-recurrent.

We then have

$$p\Big(\frac{V_i(n)}{n} \overset{n+\infty}{\to} \frac{1}{m_i}\Big) = 1. \tag{3.19}$$

If the MC is positive recurrent, we have for any bound function $f : I \to \mathbb{R}$

$$p\Big(\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \overset{n+\infty}{\to} < f, \pi > \Big) = 1, \tag{3.20}$$

where π is the unambiguous invariant distribution of the Markov chain (with $\pi_i = \frac{1}{m_i}$) and

$$< f, \pi >= \sum_{i \in I} f(i)\pi_i = E_\pi(f).$$

To interpret this theorem, let us consider that the Markov chain describes the evolution of the characteristics of an object and that we follow the evolution of **a very large number of such objects** M. Equation 3.19 tells us that *for a positive recurrent MC*, the proportion of time the Markov chain spends in state i converges towards π_i, which is the probability of finding one object in state i if we completely randomly choose it within the large population after a long period of time.

Let us suppose we're now interested in the average temperature of an object (from a very large population) after a long period of time N. That quantity is given by $\overline{T} = \sum_{i \in I} T(X_N = i)\pi_i$. According to the law of large numbers, it can be approximated as

$$\overline{T} \approx \frac{1}{M}\sum_{k=0}^{M-1} T(X_{N,k})$$

where $X_{N,k}$ corresponds to the k-th object at time N. If we don't want to examine all objects, we can instead follow a **single object** over a very long period of time through the use of Eq. 3.20. We then have

$$\overline{T} \approx \frac{1}{n}\sum_{k'=0}^{n-1} T(X'_k)$$

if n is large enough.

Proof of Equation 3.19.

Let's first consider the transient case where $m_i = +\infty$. If $V_i = \sum_{k=0}^{\infty} 1_{\{X_k = i\}}$, we have $p(V_i = \infty) = 0$ so that $V_i < \infty$ with almost certainty. Hence, we have almost certainly

$$\frac{V_i(n)}{n} \leq \frac{V_i}{n} \overset{n+\infty}{\to} 0 = \frac{1}{m_i}.$$

We'll now be dealing with the much more complex recurrent case. We'll show that Eq. 3.19 holds for any $i \in I$. Since state i is recurrent, we know that $p(T = T_i < \infty) = 1$. Given the fact that T is a Markov time, $(X_{T+n})_{n\geq 0}$ is a (δ_i, P) Markov chain independent of $X_0, ..., X_T$. As we saw above, the relative number of visits in state i is defined as

$$\frac{V_i(n)}{n} = \frac{1}{n}\sum_{k=0}^{n-1} 1_{\{X_k=i\}} \leq 1.$$

Consequently, it has the same probabilistic behaviour for $(X_{T+n})_{n\geq 0}$ as for $(X_n)_{n\geq 0}$ if the Markov chain starts off in state i. Let's suppose that the Markov chain starts off in a state $j \neq i$. Let us consider a very large $N >> T$ [4]. We have

$$\begin{aligned}
\frac{V_i(N)}{N} &= \frac{1}{N}\sum_{k=0}^{N-1} 1_{\{X_k=i\}} = \frac{1}{N}\sum_{k=0}^{T-1} 1_{\{X_k=i\}} + \frac{1}{N}\sum_{k=T}^{N} 1_{\{X_k=i\}} \\
&\approx \frac{1}{N}\sum_{k=T}^{N} 1_{\{X_k=i\}} \\
&= \frac{1}{N}\sum_{k=0}^{N-T} 1_{\{X_{T+k}=i\}} \\
&= \frac{1}{N}\sum_{k=0}^{N} 1_{\{X_{T+k}=i\}} - \frac{1}{N}\sum_{k=N-T+1}^{N} 1_{\{X_{T+k}=i\}} \qquad \approx \frac{1}{N}\sum_{k=0}^{N} 1_{\{X_{T+k}=i\}}
\end{aligned}$$

In a similar way, we can show that for any initial probability distribution of the Markov chain λ, in the long run $\frac{V_i(n)}{n}$ obtained with $(X_n)_{n\geq 0}$ will be very close to $\frac{V_i(n)}{n}$ obtained with $(X_{T+n})_{n\geq 0}$ (which logically starts off in i). As a consequence, we only need to consider the case where the initial distribution is given by $\lambda = \delta_i$.

According to Lemma 3.13, $S_i^{(r)}$ is independent of $\{X_m : m \leq T_i^{(r-1)}\}$ and

$$p(S_i^{(r)} = n | T_i^{(r-1)} < \infty) = p_i(T_i^{(1)} = n).$$

Given that i is recurrent, all return times $T_i^{(r)}$ are finite and so are the $S_i^{(r)}$. $S_i^{(r)}$ is also independent of $T_i^{(0)}, T_i^{(1)}, T_i^{(2)}, ..., T_i^{(r-1)}$ since they only depend on the $\{X_m : m \leq T_i^{(r-1)}\}$. Since $S_i^{(r-1)} = T_i^{(r-1)} - T_i^{(r-2)}$, $S_i^{(r)}$ is independent of $S_i^{(r-1)}$, $S_i^{(r-2)}$, ..., $S_i^{(1)}$. We can therefore conclude that the $S_i^{(r)}, r > 0$ are independent identically distributed random variables.

$$E(S_i^{(r)}) = E(S_i^{(r-1)}) = E(T_i) = m_i$$

[4] which is a non-rigorous way of considering the limit.

We have

$$S_i^{(1)} + ... + S_i^{(V_i(n)-1)} = T_i^{(V_i(n)-1)} - T_i^{(V_i(n)-2)} + T_i^{(V_i(n)-2)} - T_i^{(V_i(n)-3)} + ... - T_i^{(1)} + T_i^{(1)} - 0$$
$$S_i^{(1)} + ... + S_i^{(V_i(n)-1)} = T_i^{(V_i(n)-1)}$$

As we saw above, the return times are defined as

$$T_i^{(0)} = 0, \text{ and } T_i^{(r+1)} := inf\{n \geq T_i^{(r)} + 1 : X_n = i\}, r = 0, 1, 2, ...$$

$V_i(n)$ is the number of times $\{X_k = i\}$ for k between 0 and $n-1$. Between $k = 1$ and $k = n-1$, we have $(V_i(n) - 1)$ times $\{X_k = i\}$. Hence, $T_i^{(V_i(n)-1)} = V_i(n) - 1 \leq n - 1$. $T_i^{(V_i(n))}$ is the time until which there are $V_i(n)$ times $\{X_k = i\}$ for $k \geq 1$. At time $n-1$, there are only $(V_i(n) - 1)$ times $\{X_k = i\}$ for $k \geq 1$. Therefore, $T_i^{(V_i(n))} \geq n$. Consequently, we have

$$T_i^{(V_i(n)-1)} \leq n \leq T_i^{(V_i(n))}.$$

This allows us to write that

$$S_i^{(1)} + S_i^{(2)} + ... + S_i^{(V_i(n)-1)} \leq n \leq S_i^{(1)} + S_i^{(2)} + ... + S_i^{(V_i(n))} \tag{3.21}$$
$$\frac{S_i^{(1)} + S_i^{(2)} + ... + S_i^{(V_i(n)-1)}}{V_i(n)} \leq \frac{n}{V_i(n)} \leq \frac{S_i^{(1)} + S_i^{(2)} + ... + S_i^{(V_i(n))}}{V_i(n)} \tag{3.22}$$
$$\tag{3.23}$$

As we showed above, the $S_i^{(r)}, r > 0$ are independent identically distributed random variables. Consequently, by applying the strong law of large numbers, we have

$$p\Big(\frac{S_i^{(1)} + ... + S_i^{(n)}}{n} \overset{n+\infty}{\to} m_i\Big) = 1$$

What is more, since i is recurrent, it is visited an infinite number of times so that $p(V_i(n) \overset{n+\infty}{\to} \infty) = 1$ and V_i visits all positive integers. As a consequence, we almost certainly have

$$\frac{S_i^{(1)} + ... + S_i^{(V_i(n))}}{V_i(n)} \overset{n+\infty}{\to} \frac{S_i^{(1)} + ... + S_i^{(n)}}{n} \overset{n+\infty}{\to} m_i$$

Moreover,

$$\frac{S_i^{(1)} + ... + S_i^{(V_i(n)-1)}}{V_i(n)} = \frac{S_i^{(1)} + ... + S_i^{(V_i(n)-1)}}{V_i(n) - 1} \frac{V_i(n) - 1}{V_i(n)}.$$

$$\frac{V_i(n) - 1}{V_i(n)} = 1 - \frac{1_{\{X_n = i\}}}{V_i(n)} \overset{n+\infty}{\to} 1$$

Therefore, we almost certainly have

$$\frac{S_i^{(1)} + ... + S_i^{(V_i(n)-1)}}{V_i(n)} \overset{n+\infty}{\to} m_i.$$

It follows that $\frac{n}{V_i(n)}$ almost certainly converges towards m_i. As a consequence,

$$p\Big(\frac{V_i(n)}{n} \overset{n+\infty}{\to} \frac{1}{m_i} = \pi_i\Big) = 1.$$

Proof of Equation 3.20.

Let $(X_n)_{n\geq 0}$ be a positive recurrent Markov chain whose invariant distribution is π and let $f : I \to \mathbb{R}$ be a bound function. Without loss of generality, we can assume that $|f| < 1$ (otherwise, we just have to consider $f/max_{i\in I}|f(i)|$ and remultiply the function by $max_{i\in I}|f(i)|$ in the result

we'll obtain). For any subset $J \subset I$, we have

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) - < f, \pi > \right| = \left|\frac{1}{n}\sum_{i\in I} V_i(n) f(i) - < f, \pi > \right| \\
&= \left|\sum_{i\in I}\Big(\frac{V_i(n)}{n} - \pi_i\Big) f(i)\right| \\
&\leq \left|\sum_{i\in I} \frac{V_i(n)}{n} - \pi_i\right| \\
&\leq \sum_{i\in I}\left|\frac{V_i(n)}{n} - \pi_i\right| \\
&= \sum_{i\in J}\left|\frac{V_i(n)}{n} - \pi_i\right| + \sum_{i\notin J}\left|\frac{V_i(n)}{n} - \pi_i\right| \\
&\leq \sum_{i\in J}\left|\frac{V_i(n)}{n} - \pi_i\right| + \sum_{i\notin J}\left(\frac{V_i(n)}{n} + \pi_i\right) \\
&= \sum_{i\in J}\left|\frac{V_i(n)}{n} - \pi_i\right| + \sum_{i\notin J}\left(\frac{V_i(n)}{n} + \pi_i\right) + 1 - 1 \\
&= \sum_{i\in J}\left|\frac{V_i(n)}{n} - \pi_i\right| + \sum_{i\notin J}\left(\frac{V_i(n)}{n} + \pi_i\right) + \sum_{i\in I}\left(\pi_i - \frac{V_i(n)}{n}\right) \\
&= \sum_{i\in J}\left|\frac{V_i(n)}{n} - \pi_i\right| + \sum_{i\notin J}\left(\frac{V_i(n)}{n} + \pi_i\right) + \sum_{i\in J}\left(\pi_i - \frac{V_i(n)}{n}\right) + \sum_{i\notin J}\left(\pi_i - \frac{V_i(n)}{n}\right) \\
&= \sum_{i\in J}\left|\frac{V_i(n)}{n} - \pi_i\right| + \sum_{i\notin J} 2\pi_i + \sum_{i\in J}\left(\pi_i - \frac{V_i(n)}{n}\right) \\
&\leq 2\sum_{i\in J}\left|\frac{V_i(n)}{n} - \pi_i\right| + \sum_{i\notin J} 2\pi_i
\end{aligned}$$

According to Theorem 3.20, we have $\pi_i = \frac{1}{m_i}$ and because of Equation 3.19,

$$p\Big(\frac{V_i(n)}{n} \overset{n+\infty}{\to} \frac{1}{m_i}, i \in I\Big) = 1.$$

Let us now consider $\epsilon > 0$. It is always possible to choose a subset J (including the set I) such that

$$\sum_{i\notin J} \pi_i < \frac{\epsilon}{4}.$$

Let $\omega \in \Omega$ be the event leading to the Markov chain. There must exist a $N(\omega)$ such that $n \geq N(\omega)$ implies that

$$\sum_{i \in J} \left| \frac{V_i(n)}{n} - \pi_i \right| \leq \frac{\epsilon}{4}$$

because otherwise $\frac{V_i(n)}{n}$ couldn't converge towards $\pi_i, \forall i \in I$. We finally find that for $n \geq N(\omega)$,

$$\left| \frac{1}{n} \sum_{k=0}^{n-1} f(X_k) - < f, \pi > \right| < \epsilon.$$

Since this is valid $\forall \epsilon > 0$, we've shown that Equation 3.20 holds.

Chapter 4

Continuous Markov chains in discrete time

In this short section, we consider Markov chains taking values in a continuous state space. If the state space is the interval $[a; b] \subset \mathbb{R}$ (allowing for infinite values), we could first consider a discrete Markov chain taking values in $I = \{a + k\frac{b-a}{n}, k \in (0, n)\}$. When $n \to \infty$, the Markov chain taking values in the discrete state space transitions into a Markov chain taking values in the continuous state space $S = [a; b]$. This allows us to extrapolate the results and theorems of Chapter 3. We shall write them here without any proof.

Definition 4.1 (Transition kernel)

In a continuous state space, the transition probability between two precise values x and y is always equal to 0. We need thus to introduce a **transition kernel** $P(x, A)$ which is the probability that we'll arrive in the subset $A \subset S$ given that we are in state $x \in S$.

Definition 4.2 (Markov property)

If $(X_n)_{n\geq 0}$ is a Markov chain taking values in the continuous state space S, we have $\forall n \geq 0, x_n, x_{n-1}, ..., x_0 \in S$, $\forall A \subset S$

$$P(X_{n+1} \in A | X_n = x_n, X_{n-1} = x_{n-1}, ..., X_0 = x_0) = P(X_{n+1} \in A | X_n = x_n).$$

Definition 4.3 (Transition kernel density)

For all $x \in S$, $p(x, y)$ is defined as a nonnegative function such that

$$P(x, A) = \int_{y\in A} p(x, y)dy.$$

For a given x, $p(x, .)$ is a probability density function so that

$$p(x, S) = \int_{y\in S} p(x, y)dy = 1.$$

Definition 4.4 (Transition probability between two subsets)

Let f be the current probability density of the Markov chain. $\forall A, B \subset S$, the transition probability is given by

$$p(B|A)) = p(A, B) = \frac{\int_{x \in A} p(x, B) f(x) dx}{\int_{x \in A} f(x) dx}.$$

Definition 4.5 (Product of two transition kernels)

For any $x \in S$ and any subset $A \subset S$, the product of two transition kernels P and Q is defined as

$$PQ(x, A) = \int_{y \in S} P(x, dy) Q(y, A) |dy|$$

where dy is a small interval centred at y.

In this way, we can compute

$$P^k(x, A) = P(X_k \in A | X_0 = x).$$

Definition 4.6 (ϕ-irreducibility)

We consider that ϕ is a measure of S (which would be a length for $\mathbb{R}$, an area for $\mathbb{R}^2$ and a volume for $\mathbb{R}^3$). The Markov chain $(X_n)_{n \geq 0}$ is called ϕ-irreducible if and only if for every $x \in S$ and every $A \subset S$ with $\phi(A) > 0$, there exists an integer $n > 0$ such that $P^n(x, A) > 0$.

Alternatively, $(X_n)_{n \geq 0}$ is called ϕ-irreducible if and only if for any two subsets $A, B \subset S$ with $\phi(A) > 0$ and $\phi(B) > 0$, $\exists n > 0$ such that $P^n(A, B) = p(X^n \in A | X_0 \in B) > 0$.

Definition 4.7 (Harris recurrence)

"i.o." stands for *infinitely often*. $(X_n)_{n \geq 0}$ is Harris-recurrent if and only if for every measurable set B with $\phi(B) > 0$ and $\forall x_0 \in S$ we have

$$P(X_i \in B \ i.o. | X_0 = x_0) = 1.$$

Definition 4.8 (Stationary distribution)

A probability distribution π is called a *stationary distribution* of the Markov chain with transition density p (and transition kernel P) if and only if

$$\pi(y) = \int_{x \in S} p(x, y) \pi(x) dx,$$

for all $y \in S$ or equivalently,

$$\pi(A) = \int_{x \in S} P(x, A) \pi(x) dx$$

for all measurable sets $A \subset S$. In that case, as in the discrete case we write that $\pi P = \pi$.

Definition 4.9 (Reversible distribution)

π is a reversible distribution if and only if $\forall x, y \in S$,

$$\pi(x)p(x,y) = \pi(y)p(y,x)$$

or $\forall A, B \subset S$

$$\pi(A)P(A,B) = \pi(B)p(B,A)$$

Theorem 4.1 Any reversible distribution π is an invariant distribution. The reverse isn't always true.

Definition 4.10 (Periodicity and aperiodicity)

A Markov chain $(X_n)_{n\geq 0}$ is periodic if and only if there exists $n \geq 2$ and a sequence of non-empty disjoint measurable sets $A_1, A_2, ..., A_n$ such that $\forall x \in A_j$, if $j < n$ we have $P(x, A_{j+1}) = 1$ and $P(x, A_1) = 1$ $\forall x \in A_n$.

An aperiodic Markov chain is a Markov chain that isn't periodic.

Theorem 4.2 (Sufficient condition for aperiodicity)

If $p(x, .)$ is the transition kernel of a Markov chain, if there exists $x \in S$ such that $p(x, .)$ is strictly positive in the neighbourhood of x, the Markov chain $(X_n)_{n\geq 0}$ must be aperiodic because it can remain for an arbitrarily long period arbitrarily close to x.

Theorem 4.3 (Convergence theorem)

Let $(X_n)_{n\geq 0}$ be a ϕ-irreducible Markov chain with a transition kernel P and an invariant distribution π. π is then the **unambiguous** (unique) invariant distribution of the Markov chain.

If P is aperiodic, we have $\forall x \in S$

$$\|P^n(x, A) - \pi(A)\|_{TV} \overset{n+\infty}{\to} 0.$$

We recall that the total variation norm is defined as

$$\|\pi_1 - \pi_2\|_{TV} = sup_A|\pi_1(A) - \pi_2(A)|$$

.

In other words, we can start off in pretty much any state x, run the Markov chain for a long time and the probability distribution we observe then is a good approximation of π. How much we need to wait is an open question that cannot be answered in advance. The time we need to reach the invariant distribution is known as the **burn-in time**.

Theorem 4.4 (Markov chain central limit theorem)

For independent and identically distributed random variables X_i, the central limit theorem tells us that for any real-valued measurable function g such that $\mu = E(g(X_i)) < \infty$ and $\sigma^2 = var(g(X_i)) < \infty$ if $\widehat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} g(X_k)$, we have

$$\hat{\mu}_n \approx \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right) \text{ for } n \to \infty.$$

If $X_1, X_2, ..., X_n$ are instead elements of a Markov chain whose probability distributions are all equal to the unique invariant distribution of the Markov chain, for

$$\mu = \mathrm{E}(g(X_1)) < \infty,$$
$$\sigma^2 = \mathrm{var}(g(X_1)) + 2\sum_{k=1}^{\infty} \mathrm{cov}(g(X_1), g(X_{1+k})) < \infty,$$

and

$$\widehat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} g(X_k)$$

we have

$$\hat{\mu}_n \approx \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right) \text{ for } n \to \infty.$$

The central limit theorem can be used to estimate confidence intervals for $\mu = \mathrm{E}(g(X_1))$.

Chapter 5

Markov-Chain Monte-Carlo method

5.1 Introduction

In Bayesian reasoning, we use experimental data to update one or several prior probability distributions into posterior probability distributions that are supposed to reflect all knowledge at our disposal [1]. Let's consider a problem of chemical kinetics we mentioned in the lecture "Frequentist and Bayesian approaches to uncertainty". We are interested in the elementary reaction $\mathrm{R} + \mathrm{R} \rightarrow \mathrm{P}_2$ whose speed is given by

$$r = k[R]^2 \text{ with } k \in S = [5E+04; 5E+09].$$

We measured the evolution of the concentration of species R as a function of time. The posterior probability distribution of the values of k is given by

$$f(k|data) = \frac{L(data|k)f_0(k)}{\int_{k\in S} L(data|k)f_0(k)dk}$$

Based on a prior probability distribution $f_0(k) \neq 0, k \in S$, we can easily compute the value of the likelihood and $L(data|k)f_0(k)$ for every value of k. To calculate

$$p(data) = \int_{k\in S} L(data|k)f_0(k)dk,$$

we can either use a deterministic integration or a **Monte-Carlo integration**. The latter goes as follows:

$$\int_{k\in S} L(data|k)f_0(k)dk = E_k(L(data|k)) \approx \frac{1}{n}\sum_{i=1}^{n} L(data|K_i)$$

[1] As we saw during the first lecture, we must use SEVERAL priors to represent our total lack of knowledge.

where $K_i, i \in (1;n)$ is a random variable taking values in S and whose probability density is given by $f_0(k)$.

But let us now consider a more realistic problem of chemical kinetics where the concentration of R depends on 6 different reactions and thus on the reaction rate coefficients $k = (k_1, k_2, k_3, k_4, k_5, k_6) \in S$. The posterior probability distribution is still given by

$$f(k|data) = \frac{L(data|k)f_0(k)}{\int_{k\in S} L(data|k)f_0(k)dk}$$

This time however,

$$p(data) = \int_{k_1,k_2,k_3,k_4,k_5,k_6\in S} L(data|k_1,k_2,k_3,k_4,k_5,k_6)f_0(k_1)$$
$$f_0(k_2)f_0(k_3)f_0(k_4)f_0(k_5)f_0(k_6)dk_1dk_2dk_3dk_4dk_5dk_6$$

Due to a numerical phenomenon known as the **curse of dimensionality**, the time needed to calculate this integral increases exponentially with the number of dimensions if we use either direct numerical integration or the straightforward Monte-Carlo method. The computation of the integral above could take us very long. And if there are 100 or more parameters, it clearly becomes completely impossible.

It is worth noting that while we cannot get the integral directly, we have no problem computing the posterior ratio

$$\frac{f(k_1|data)}{f(k_2|data)} = \frac{L(data|k_1)f_0(k_1)}{L(data|k_2)f_0(k_2)}$$

for any $k_1, k_2 \in S$.

We can create a Markov chain based on the values of $\frac{f(k_1|data)}{f(k_2|data)}$ whose invariant distribution is the posterior probability distribution $f(k|data)$. This is the very principle of the Markov-chain Monte-Carlo method (MCMC). In what follows, we shall discover three algorithms to perform this: the simple Metropolis algorithm, the Metropolis-Hasting algorithm and the Gibbs sampling algorithm.

An important assumption is that both the prior and the likelihood are **strictly positive everywhere in the entire state space**. It follows from this that the posterior is strictly positive everywhere too. In case where the likelihood is equal to zero for certain values of the parameters, i.e. the data at our disposal allow us to rule out some ranges of parameter values, we can simply reformulate the problem by redefining the sample space S in such a way that the likelihood is strictly positive everywhere. It follows from this that the posterior distribution is always strictly positive:

$$\forall k \in S, f(k|data) > 0.$$

We'll also always make the assumption that the parameter space S is **finite**.

5.2 Metropolis algorithm

5.2.1 Algorithm and convergence

We want to determine the posterior probability distribution

$$f(x|data) = \frac{L(data|x)f_0(x)}{\int_{x\in S} L(data|x)f_0(x)dx}.$$

Let $g(x) = L(data|x)f_0(x)$. The metropolis algorithm goes as follows [2]:

1. Initialisation: Choose an arbitrary point $X_0 = x_0$ to be the first sample and choose an arbitrary probability density $Q(y|x)$ that suggests a candidate for the next sample value X_{t+1}, given the previous sample value $X_t = x$. For the Metropolis algorithm, Q must be symmetrical, in other words, it must satisfy $Q(y|x) = Q(x|y)$. A usual choice is to let $Q(y|x)$ be a **Gaussian distribution** centred at x, so that points closer to x are more likely to be visited next, thereby making the sequence of samples into a **random walk**. The function Q is referred to as the "proposal density" or "jumping distribution".

2. For each iteration t:

   - *Generate* a candidate y for the next sample by picking from the jumping distribution $Q(y|x_t)$.
   - *Calculate* the *acceptance ratio*

     $$\alpha(x_t, y) = min\Big(1, \frac{g(y)}{g(x_t)}\Big) = min\Big(1, \frac{f(y|data)}{f(x_t|data)}\Big),$$

     which will be used to decide whether to accept or reject the candidate.
   - Accept y with a probability of $\alpha(x_t, y)$. Otherwise reject y.
     - Generate a uniform random number $u \in [0, 1]$.
     - If $u \leq \alpha$, then "accept" the candidate by setting $x_{t+1} = y$,
     - If $u > \alpha$, then "reject" the candidate and set $x_{t+1} = x_t$ instead.

This algorithm proceeds by randomly attempting to move about the sample space, sometimes accepting the moves and sometimes remaining in place. Note that the acceptance ratio α indicates how probable the new proposed sample is with respect to the current sample, according to the posterior distribution $f(x|data)$. If we attempt to move to a point that is more probable than the existing point (i.e. a point in a higher-density

[2] We'll use more rigorous notations in the next subsection.

region of $f(x|data)$), we will always accept the move. However, if we attempt to move to a less probable point, we will sometimes reject the move, and the higher the relative drop in probability, the more likely we are to reject the new point. Thus, we will tend to stay in (and return large numbers of samples from) high-density regions of $f(x|data)$, while only occasionally visiting low-density regions. Intuitively, this is why this algorithm works and returns samples that follow the desired distribution $f(x|data)$.

5.3 Metropolis-Hasting algorithm

Since we're dealing with continuous variables, for any $a \in S$ in what follows we call $\delta_{a,\phi} \subset S$ an extremely small part of S containing a with $\phi(\delta_{a,\phi}) = \phi > 0$.

The Metropolis-Hasting algorithm is very similar to the Metropolis algorithm, the only difference being that $Q(y|x)$ is no longer necessarily symmetrical; which can change the definition of the acceptance probability.

It consists of the following steps:

1. Initialisation: Choose an arbitrary point $X_0 = x_0$ to be the first sample and choose an arbitrary probability density $Q(y|x)$ that suggests a candidate for the next sample value X_{t+1}, given the previous sample value $X_t = x$. The function Q is referred to as the "proposal density" or "jumping distribution".

   From a rigorous **mathematical** standpoint, we write that $X_0 \in \delta_{x_0,\phi}$.

2. For each iteration t:

   - *Generate* a candidate y for the next sample by picking from the distribution $Q(y|x_t)$.
     Mathematically, we have [3]

     $$q(y|x_t) = p(y \in \delta_{y,\phi}|x_t \in \delta_{x_t,\phi})$$

   - *Calculate* the *acceptance ratio*

     $$\alpha(x_t, y) = min\Big(1, \frac{g(y)q(x_t|y)}{g(x_t)q(y|x_t)}\Big) = min\Big(1, \frac{f(y|data)q(x_t|y)}{f(x_t|data)q(y|x_t)}\Big),$$

     which will be used to decide whether to accept or reject the candidate.

[3] To be entirely rigorous, we should write

$$q(y|x_t) = \lim_{\phi \to 0} p(y \in \delta_{y,\phi}|x_t \in \delta_{x_t,\phi}).$$

- Accept y with a probability of $\alpha(x_t, y)$. Otherwise reject y.
  - Generate a uniform random number $u \in [0, 1]$.
  - If $u \leq \alpha$, then "accept" the candidate by setting $x_{t+1} = y$.
  - If $u > \alpha$, then "reject" the candidate and set $x_{t+1} = x_t$ instead.

    From a **mathematical** point of view, if $\delta_{x_t,\phi} \cap \delta_{y,\phi} = \varnothing$,

    $$p(X_{t+1} \in \delta_{y,\phi} | X_t \in \delta_{x_t,\phi}, Y \in \delta_{y,\phi}) = \alpha(x_t, y)$$

    and

    $$p(X_{t+1} \in \delta_{x_t,\phi} | X_t \in \delta_{x_t,\phi}, Y \in \delta_{y,\phi}) = 1 - \alpha(x_t, y).$$

    Otherwise,

    $$p(X_{t+1} \in \delta_{y,\phi} | X_t \in \delta_{x_t,\phi}, Y \in \delta_{y,\phi}) \geq \alpha(x_t, y)$$

    because if y is not accepted, there is still the possibility that $X_{t+1} \in \delta_{x_t,\phi} \cap \delta_{y,\phi}$. Likewise,

    $$p(X_{t+1} \in \delta_{x_t,\phi} | X_t \in \delta_{x_t,\phi}, Y \in \delta_{y,\phi}) \geq 1 - \alpha(x_t, y).$$

    The last two inequalities are thus true under every circumstance.

Theorem 5.1 Any Markov chain produced in this way converges towards its invariant distribution $f(x|data)$.

Proof.

It is obvious that the sequence defined in this way is a Markov chain as X_{t+1} only depends on X_t.

We'll begin by proving that the Markov chain is ϕ-irreducible (see Def. 4.6).

For any $\delta_{x_t,\phi} \subset S$ and $\delta_{x_{t+1},\phi} \subset S$, the probability of the transition from $\delta_{x_t,\phi}$ to $\delta_{x_{t+1},\phi}$ is given by

$$\begin{aligned}
P(X_{t+1} \in \delta_{x_{t+1},\phi} | X_t \in \delta_{x_t,\phi}) &= \frac{P(X_{t+1} \in \delta_{x_{t+1},\phi}, X_t \in \delta_{x_t,\phi})}{p(X_t \in \delta_{x_t,\phi}))} \\
&= \frac{P(X_{t+1} \in \delta_{x_{t+1},\phi}, X_t \in \delta_{x_t,\phi}, Y \in S)}{p(X_t \in \delta_{x,\phi}))} \\
&= \frac{P(X_{t+1} \in \delta_{x_{t+1},\phi}, X_t \in \delta_{x_t,\phi}, Y \in \delta_{x_{t+1},\phi})}{p(X_t \in \delta_{x,\phi}))} \\
&+ \frac{P(X_{t+1} \in \delta_{x_{t+1},\phi}, X_t \in \delta_{x_t,\phi}, Y \notin \delta_{x_{t+1},\phi})}{p(X_t \in \delta_{x,\phi}))}
\end{aligned}$$

We further have

$$\frac{P(X_{t+1} \in \delta_{x_{t+1},\phi}, X_t \in \delta_{x_t,\phi}, Y \in \delta_{x_{t+1},\phi})}{p(X_t \in \delta_{x,\phi}))} =$$
$$\frac{P(X_{t+1} \in \delta_{x_{t+1},\phi} | X_t \in \delta_{x_t,\phi}, Y \in \delta_{x_{t+1},\phi}) p(X_t \in \delta_{x_t,\phi}, Y \in \delta_{x_{t+1},\phi})}{p(X_t \in \delta_{x,\phi}))}$$
$$= P(X_{t+1} \in \delta_{x_{t+1},\phi} | X_t \in \delta_{x_t,\phi}, Y \in \delta_{x_{t+1},\phi}) p(Y \in \delta_{x_{t+1},\phi} | X_t \in \delta_{x_t,\phi})$$
$$= P(X_{t+1} \in \delta_{x_{t+1},\phi} | X_t \in \delta_{x_t,\phi}, Y \in \delta_{x_{t+1},\phi}) q(x_{t+1}|x_t)$$
$$\geq \alpha(x_t, x_{t+1}) q(x_{t+1}|x_t)$$

This entails that

$$P(X_{t+1} \in \delta_{x_{t+1},\phi} | X_t \in \delta_{x_t,\phi}) \geq \alpha(x_t, y) q(x_{t+1}|x_t).$$

We always have $q(x_{t+1}|x_t) > 0, \forall x_{t+1}, x_t \in S$. On the other hand,

$$\alpha(x_t, y) = min\Big(1, \frac{f(y|data)q(x_t|y)}{f(x_t|data)q(y|x_t)}\Big).$$

As we saw at the end of Section 5.1, our assumptions imply that the posterior is strictly positive everywhere. As a consequence, $\alpha(x_t, y) > 0$ and

$$P(X_{t+1} \in \delta_{x_{t+1},\phi} | X_t \in \delta_{x_t,\phi}) > 0.$$

The Markov chain is thus ϕ-irreducible. What is more, this is also valid if $\delta_{x_{t+1},\phi} = \delta_{x_t,\phi}$ for an arbitrarily small $\phi > 0$. Consequently, the MC is aperiodic as well based on Theorem 4.2. Since the Markov chain is both irreducible and aperiodic, according to Theorem 4.3, if we find an invariant distribution π, we know it is the unique invariant distribution the Markov chain inevitably converges towards. If we are able to show that the Markov chain is reversible with respect to $\pi = f(.|data)$, Theorem 4.1 tells us it must be the invariant / stationary distribution of the MC.

We want to prove that for any disjoint $\delta_{x_{t+1},\phi}$ and $\delta_{x_t,\phi}$,

$$\pi(\delta_{x_t,\phi}) P(\delta_{x_t,\phi}, \delta_{x_{t+1},\phi}) = \pi(\delta_{x_{t+1},\phi}) P(\delta_{x_{t+1},\phi}, \delta_{x_t,\phi})$$

which is equivalent to

$$\pi(\delta_{x_t,\phi}) P(X_{t+1} \in \delta_{x_{t+1},\phi} | X_t \in \delta_{x_t,\phi}) = \pi(\delta_{x_{t+1},\phi}) P(X_{t+1} \in \delta_{x_{t,\phi}} | X_t \in \delta_{x_{t+1},\phi})$$

As we saw above

$$P(X_{t+1} \in \delta_{x_{t+1},\phi} | X_t \in \delta_{x_t,\phi}) = \frac{P(X_{t+1} \in \delta_{x_{t+1},\phi}, X_t \in \delta_{x_t,\phi})}{p(X_t \in \delta_{x_t,\phi}))}$$
$$= \frac{P(X_{t+1} \in \delta_{x_{t+1},\phi}, X_t \in \delta_{x_t,\phi}, Y \in S)}{p(X_t \in \delta_{x,\phi}))}$$
$$= \frac{P(X_{t+1} \in \delta_{x_{t+1},\phi}, X_t \in \delta_{x_t,\phi}, Y \in \delta_{x_{t+1},\phi})}{p(X_t \in \delta_{x,\phi}))}$$
$$+ \frac{P(X_{t+1} \in \delta_{x_{t+1},\phi}, X_t \in \delta_{x_t,\phi}, Y \notin \delta_{x_{t+1},\phi})}{p(X_t \in \delta_{x,\phi}))}$$

Since $\delta_{x_{t+1},\phi}$ and $\delta_{x_t,\phi}$ are disjoint and that $Y \notin \delta_{x_{t+1},\phi}$,

$$\frac{P(X_{t+1} \in \delta_{x_{t+1},\phi}, X_t \in \delta_{x_t,\phi}, Y \notin \delta_{x_{t+1},\phi})}{p(X_t \in \delta_{x,\phi}))} = 0.$$

This implies that

$$P(X_{t+1} \in \delta_{x_{t+1},\phi} | X_t \in \delta_{x_t,\phi}) = \frac{P(X_{t+1} \in \delta_{x_{t+1},\phi}, X_t \in \delta_{x_t,\phi}, Y \in \delta_{x_{t+1},\phi})}{p(X_t \in \delta_{x,\phi}))}$$
$$= P(X_{t+1} \in \delta_{x_{t+1},\phi} | X_t \in \delta_{x_t,\phi}, Y \in \delta_{x_{t+1},\phi}) q(x_{t+1}|x_t) \Rightarrow$$
$$P(X_{t+1} \in \delta_{x_{t+1},\phi} | X_t \in \delta_{x_t,\phi}) = \alpha(x_t, x_{t+1}) q(x_{t+1}|x_t).$$

We have

$$\alpha(x_t, x_{t+1}) = min\Big(1, \frac{f(x_{t+1}|data) q(x_t|x_{t+1})}{f(x_t|data) q(x_{t+1}|x_t)}\Big)$$

and

$$\alpha(x_{t+1}, x_t) = min\Big(1, \frac{f(x_t|data) q(x_{t+1}|x_t)}{f(x_{t+1}|data) q(x_t|x_{t+1})}\Big).$$

We must make a distinction between three cases:

- $f(x_{t+1}|data) q(x_t|x_{t+1}) < f(x_t|data) q(x_{t+1}|x_t)$.
  In that case, $\alpha(x_{t+1}, x_t) = 1$ and

$$\alpha(x_t, x_{t+1}) = \frac{f(x_{t+1}|data) q(x_t|x_{t+1})}{f(x_t|data) q(x_{t+1}|x_t)}.$$

$$
\begin{aligned}
&f(x_t|data)P(X_{t+1} \in \delta_{x_{t+1},\phi}|X_t \in \delta_{x_t,\phi}) = f(x_t|data)\alpha(x_t, x_{t+1})q(x_{t+1}|x_t) \\
&= f(x_t|data)q(x_{t+1}|x_t)\frac{f(x_{t+1}|data)q(x_t|x_{t+1})}{f(x_t|data)q(x_{t+1}|x_t)} \\
&= f(x_{t+1}|data)q(x_t|x_{t+1}) \\
&= f(x_{t+1}|data)\alpha(x_{t+1}, x_t)q(x_t|x_{t+1}) \\
&= f(x_{t+1}|data)P(X_{t+1} \in \delta_{x_t,\phi}|X_t \in \delta_{x_{t+1},\phi})
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\frac{f(x_t|data)}{f(x_{t+1}|data)} &= \frac{P(X_{t+1} \in \delta_{x_t,\phi}|X_t \in \delta_{x_{t+1},\phi})}{P(X_{t+1} \in \delta_{x_{t+1},\phi}|X_t \in \delta_{x_t,\phi})} \\
\frac{\pi(\delta_{x_t,\phi})}{\pi(\delta_{x_{t+1},\phi})} &= \frac{P(X_{t+1} \in \delta_{x_t,\phi}|X_t \in \delta_{x_{t+1},\phi})}{P(X_{t+1} \in \delta_{x_{t+1},\phi}|X_t \in \delta_{x_t,\phi})}
\end{aligned}
$$

Finally, we find that

$$\pi(\delta_{x_t,\phi})P(X_{t+1} \in \delta_{x_{t+1},\phi}|X_t \in \delta_{x_t,\phi}) = \pi(\delta_{x_{t+1},\phi})P(X_{t+1} \in \delta_{x_t,\phi}|X_t \in \delta_{x_{t+1},\phi}).$$

- $f(x_{t+1}|data)q(x_t|x_{t+1}) > f(x_t|data)q(x_{t+1}|x_t)$.
  In that case, $\alpha(x_t, x_{t+1}) = 1$ and

$$\alpha(x_{t+1}, x_t) = \frac{f(x_t|data)q(x_{t+1}|x_t)}{f(x_{t+1}|data)q(x_t|x_{t+1})}.$$

$$
\begin{aligned}
&f(x_t|data)P(X_{t+1} \in \delta_{x_{t+1},\phi}|X_t \in \delta_{x_t,\phi}) = f(x_t|data)\alpha(x_t, x_{t+1})q(x_{t+1}|x_t) \\
&= f(x_t|data)q(x_{t+1}|x_t) \\
&= \alpha(x_{t+1}, x_t)f(x_{t+1}|data)q(x_t|x_{t+1}) \\
&= f(x_{t+1}|data)\alpha(x_{t+1}, x_t)q(x_t|x_{t+1}) \\
&= f(x_{t+1}|data)P(X_{t+1} \in \delta_{x_t,\phi}|X_t \in \delta_{x_{t+1},\phi})
\end{aligned}
$$

Like before, by forming the ratio $\frac{f(x_t|data)}{f(x_{t+1}|data)}$, we can show that

$$\pi(\delta_{x_t,\phi})P(X_{t+1} \in \delta_{x_{t+1},\phi}|X_t \in \delta_{x_t,\phi}) = \pi(\delta_{x_{t+1},\phi})P(X_{t+1} \in \delta_{x_t,\phi}|X_t \in \delta_{x_{t+1},\phi}).$$

- $f(x_{t+1}|data)q(x_t|x_{t+1}) = f(x_t|data)q(x_{t+1}|x_t)$.
  In that case, $\alpha(x_{t+1}, x_t) = 1$ and $\alpha(x_t, x_{t+1}) = 1$.

$$
\begin{aligned}
& f(x_t|data)P(X_{t+1} \in \delta_{x_{t+1},\phi}|X_t \in \delta_{x_t,\phi}) = f(x_t|data)\alpha(x_t, x_{t+1})q(x_{t+1}|x_t) \\
& = \frac{f(x_{t+1}|data)q(x_t|x_{t+1})}{q(x_{t+1}|x_t)}q(x_{t+1}|x_t) \\
& = f(x_{t+1}|data)q(x_t|x_{t+1})\alpha(x_{t+1}, x_t) \\
& = f(x_{t+1}|data)P(X_{t+1} \in \delta_{x_t,\phi}|X_t \in \delta_{x_{t+1},\phi})
\end{aligned}
$$

Similarly, by forming the ratio $\frac{f(x_t|data)}{f(x_{t+1}|data)}$, we can show that

$$\pi(\delta_{x_t,\phi})P(X_{t+1} \in \delta_{x_{t+1},\phi}|X_t \in \delta_{x_t,\phi}) = \pi(\delta_{x_{t+1},\phi})P(X_{t+1} \in \delta_{x_t,\phi}|X_t \in \delta_{x_{t+1},\phi}).$$

We have thus shown that the detailed-balance conditions are always satisfied. $f(.|data)$ is thus the unique invariant distribution of the irreducible and aperiodic Markov chain and Theorem 4.3 tells us that the Markov chain converges probabilistically towards $f(.|data)$.

5.3.1 Autocorrelation and optimal choice of the proposal distribution

Definition 5.1 (Autocorrelation)
If $(X_t)_{t\geq 0}$ is a Markov chain, the autocorrelation at time t_1 and t_2 is defined as

$$\gamma(t_1, t_2) = \mathrm{Cov}(X_{t_1}, X_{t_2}) = \mathrm{E}[(X_{t_1} - \mu_{t_1})(X_{t_2} - \mu_{t_2})].$$

For a Markov chain where all variables are independent (e.g. the successive tossing of a coin), $\gamma(t_1, t_2) = 0$. In general, $\gamma(t_1, t_2) \neq 0$ but it gets smaller when $\tau = t_2 - t_1$ increases. Autocorrelation is problematic for the Markov-Chain Monte-Carlo algorithms. To approximate the posterior distribution $f(x|data)$ (or any other invariant distribution), we must be able to sample all parts of the state space depending on how probable they are. A high autocorrelation reduces the freedom of the Markov chain to explore new areas.

A good proposal distribution generates candidate parameter values that cover all regions of the invariant distribution after a reasonable number of iterations (called the *burn-in*) and it brings about candidate values that are not accepted or rejected too often:

- If the proposal distribution is too diffuse with respect to the target distribution, i.e. it has a strong variance, the candidate values will be dismissed frequently and thus the chain will necessitate many iterations to adequately explore the space of the invariant distribution.

- If the proposal distribution is too focused, i.e. it has too small a variance, then the chain will stay in one small region of the target distribution for many iterations while other regions of the target distribution won't be explored, thereby resulting in a strong autocorrelation of the successive values of the random sequence.

The choice of the variance of the proposal distribution is thus always a compromise for which there unluckily doesn't exist any general rules.

5.3.2 Independence Chains

We can choose a proposal distribution such that the newly proposed parameter value is completely independent of the current parameter value: $q(y|x_t) = q(y)$. This leads to a so-called **independence chain** where the choice of the new candidate does not depend on the past, which leads to the following acceptance ratio:

$$R(x_t, y) = \frac{f(y|data)q(x_t)}{f(x_t|data)q(y)}.$$

The Markov chain generated in this way is irreducible and aperiodic if $q(x) > 0$ whenever $f(x|data) > 0$. The proposal distribution should resemble the desired invariant distribution and especially covers its tails.

In the context of Bayesian inference, a straightforward strategy consists of using the **prior distribution** as proposal distribution for the Markov chain. We thus have $Q(y|x_t) = f_0(y), \forall x_t, y \in S$. The acceptance ratio is

$$\begin{aligned}
\alpha(x_t, y) &= min\Big(1, \frac{f(y|data)q(x_t|y)}{f(x_t|data)q(y|x_t)}\Big) \\
&= min\Big(1, \frac{f(y|data)f_0(x_t)}{f(x_t|data)f_0(y)}\Big) \\
&= min\Big(1, \frac{f_0(y)L(data|y)f_0(x_t)}{f_0(x_t)L(data|x_t)f_0(y)}\Big) \\
\alpha(x_t, y) &= min\Big(1, \frac{L(data|y)}{L(data|x_t)}\Big)
\end{aligned}$$

In other words, the acceptance probability is equal to the likelihood ratio. Likelier values are always accepted. The less likely a value is, the less likely it is to be accepted.

Example 5.1 (Mixture distribution)

Let us suppose we have observed N independent and identically distributed data y_i which follow the mixture distribution:

$$f(y) = \delta N(7, 0.5^2) + (1 - \delta)N(10, 0.5^2).$$

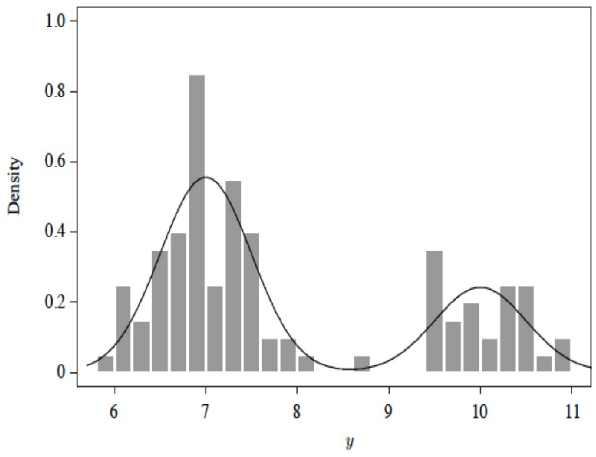


Figure 5.1: Mixture distribution

While the expected values and variances are known to us, we don't know δ. Based on the choice of the data, we want to use the data to determine the posterior probability distribution of δ. Each time, the prior is also the proposal distribution. The beta distribution $Beta(p, q)$ is defined by

$$Beta(p, q, x) = \frac{1}{B(p, q)} x^{p-1}(1-x)^{q-1},$$

whereby the normalisation constant is given by

$$B(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} = \int_0^1 u^{p-1}(1-u)^{q-1} du.$$

We choose to use two different proposal distributions in order to generate two Markov chains: $f_1(\delta) = Beta(1, 1)$ and $f_2(\delta) = Beta(2, 10)$ which are shown in Figure 5.2. The first prior is uniform whereas the second one is skewed to the right with a mean equal to approximately 0.167. Values of δ higher than 0.7 are very unlikely to be chosen by the second distribution whereas all values have the same probability to be chosen by the first uniform one. The samplings based on the two priors / proposal distributions are shown in Fig. 5.3 for 10,000 iterations. A **path** is a sequence of realisations of δ_t represented as a function of time. It is very useful for understanding the behaviour of the Markov chain.

The upper plot displays a Markov chain that moves quickly away from its starting value and can readily sample values from all parts of the parameter space supported by the posterior of δ. Such behaviour is called good mixing.

The lower plot corresponds to the chain based on $Beta(2, 10)$ as proposal density. The resulting chain moves slowly from its starting value and does a poor job of exploring the region where the posterior has non-negligible values (i.e., poor mixing). This chain has clearly not converged to its stationary

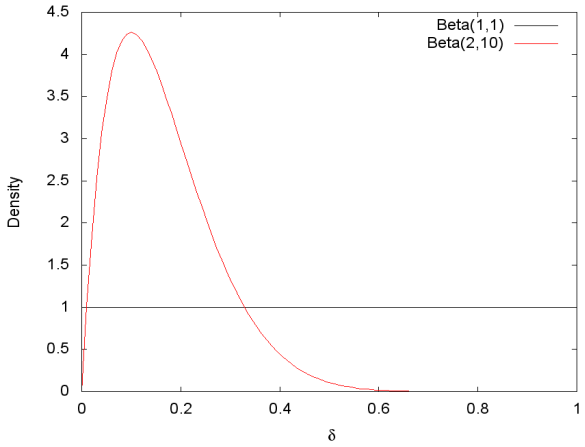


Figure 5.2: Prior / Proposal distributions

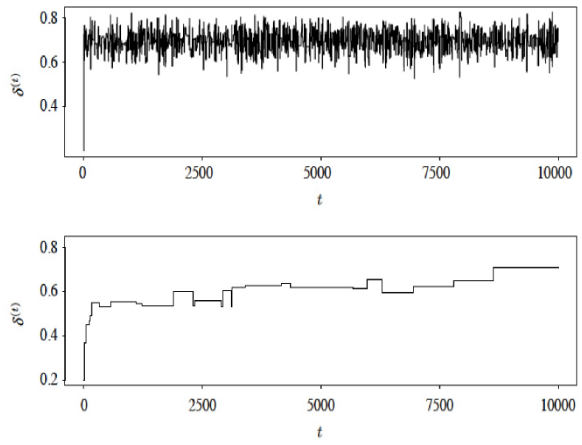


Figure 5.3: Sampling of δ based on $Beta(1,1)$ (top) and $Beta(2,10)$ (bottom)

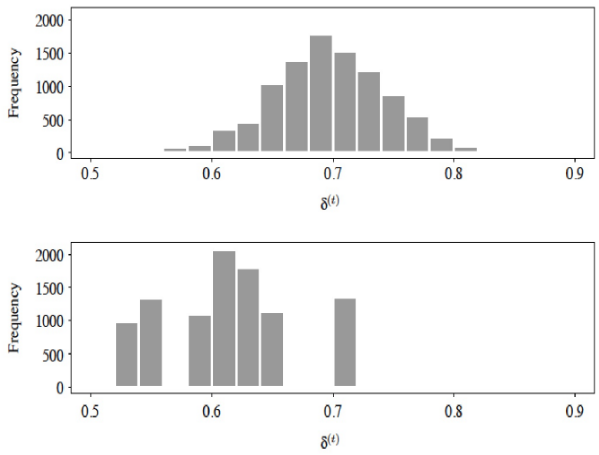


Figure 5.4: Histograms based on the two Markov chains

distribution since a systematic drift is still apparent. Such a plot should lead the MCMC user to reconsider the proposal density. The histograms obtained with the two Markov chains can be visualised in Figure 5.4. Unsurprisingly, the second distribution is completely unrealistic because the second Markov chain had huge difficulties sampling values higher than 0.70.

5.3.3 Random Walk Chains

A random walk chain is another simple variant of the Metropolis-Hasting algorithm. Let y be generated by first choosing $\epsilon \sim h(\epsilon)$ such that $y = x_t + \epsilon$. This brings into being a **random walk**. In this case, the proposal distribution is given by $q(y|x_t) = h(y - x_t)$. Usually, h may be a uniform distribution in a ball centred at x_t, a scaled standard normal distribution or a scaled Student's t distribution.

If the support region of $f(x|data)$ [4] is compact and if h is positive in a neighbourhood of 0, the resulting chain is irreducible and aperiodic. If h is symmetrical, i.e. $h(-\epsilon) = h(\epsilon)$, the acceptance ratio becomes

$$\begin{aligned}
&\alpha(x_t, y) = min\Big(1, \frac{f(y|data)q(x_t|y)}{f(x_t|data)q(y|x_t)}\Big) \\
&= min\Big(1, \frac{f(y|data)h(y - x_t)}{f(x_t|data)h(y - x_t)}\Big) \\
&= min\Big(1, \frac{f(y|data)}{f(x_t|data)}\Big) = min\Big(1, \frac{g(y)}{g(x_t)}\Big) = min\Big(1, \frac{L(data|y)f_0(y)}{L(data|x_t)f_0(x_t)}\Big)
\end{aligned}$$

[4] The region where it has non-negligible values.

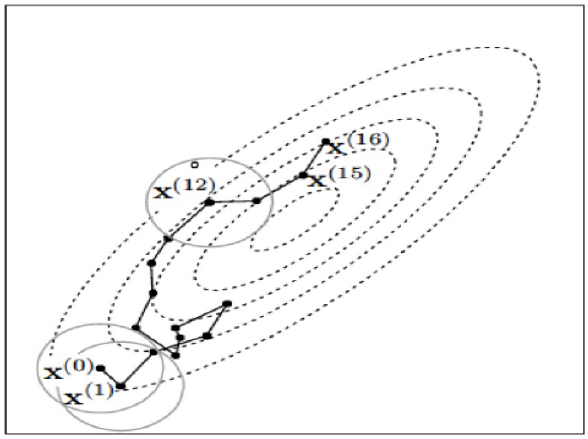


Figure 5.5: Path based on a random walk chain

In Figure 5.5, the goal is to sample values from a two-dimensional target distribution (dotted contours) by using proposed increments sampled uniformly from a disk centred around the current value of the MC.

Example 5.2 Let's suppose we want to build up a random walk Metropolis-Hasting sampler to produce a sample of 100,000 observations from the Laplace distribution

$$f(x|data) = f(x) = \frac{1}{2}e^{-|x|}, -\infty < x < +\infty.$$

We choose to define our random-walk based on $\epsilon \sim N(0, \sigma^2)$ to generate a proposal $y = x_t + \epsilon$. The results for two different values of σ are shown in Figure 5.6. It can be clearly recognised that for a small jump ($\sigma = 0.1$), the samples stemming from the random walk are more strongly auto-correlated so that the most extreme values of the posterior distribution aren't explored at all. On the contrary, for a larger jump ($\sigma = 10$) the autocorrelation is much smaller and the whole parameter space of the posterior is evenly explored. In Figure 5.7, the histogram found with the second proposal distribution is shown alongside the real distribution.

5.4 Gibbs sampling

So far, we've dealt with X_t without paying any attention to its dimensionality. The Gibbs sampler has been specifically designed for handling **multidimensional target distributions**. The goal is to construct a Markov chain whose stationary distribution - or some marginalisation thereof - is equal to the target distribution $f(x|data)$. The principle of the Gibbs sampler is to sequentially produce samples from univariate conditional distributions, which are fortunately often available in closed form.

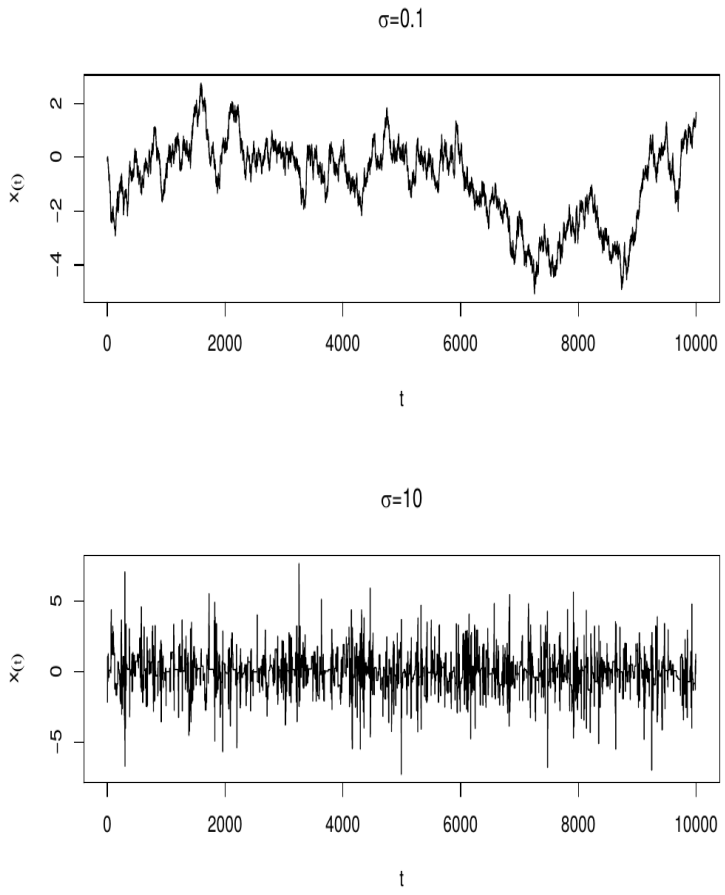


Figure 5.6: Trajectories of the two Markov chains

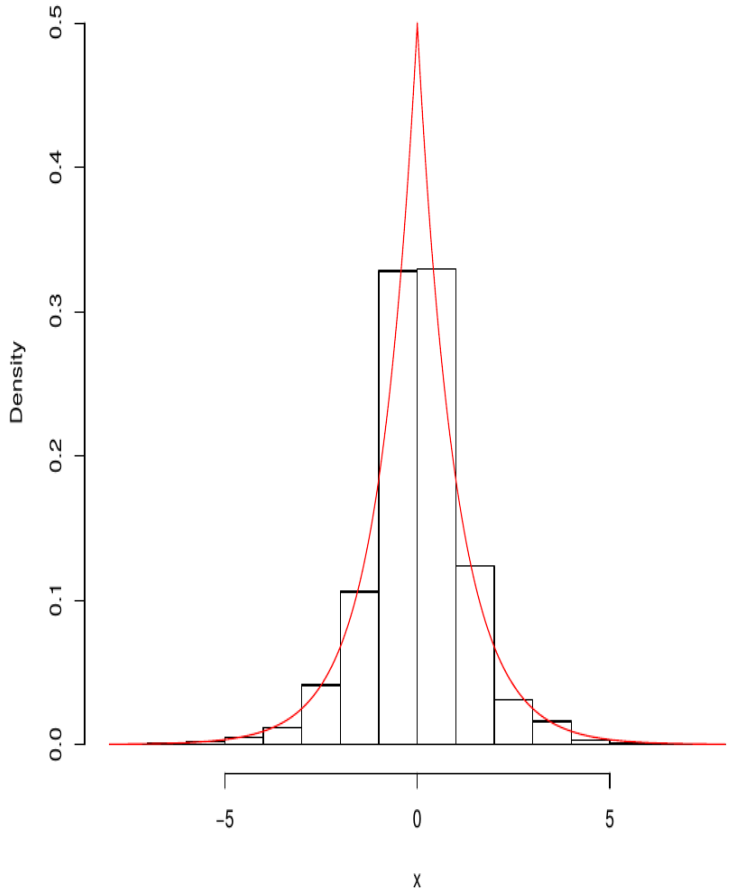


Figure 5.7: Histogram grounded on simulated values from $t = 200$ to $t = 10,000$ obtained with $\sigma = 10$.

5.4.1 Basic Gibbs sampler

This time, the parameter viewed as a Bayesian random variable is given by $X = (X_1, ..., X_p)^T$ and

$$X_{-i} = (X_1, ..., X_{i-1}, X_{i+1}, ..., X_p)^T.$$

Let's assume that the univariate conditional density of $X_i|X_{-i} = x_{-i}, data$ which we call $f(x_i|x_{-i}, data)$, is easily sampled for $i = 1, ..., p$.

The algorithm of Gibbs in its most general form can be written as follows [5]:

1. Select the starting values x^0 and set $t = 0$.

2. Produce, in turn

$$x_1^{t+1}|. \sim f(x_1|x_2^{(t)}, ..., x_p^{(t)}, data)$$
$$x_2^{t+1}|. \sim f(x_2|x_1^{(t+1)}, x_3^{(t)}, x_p^{(t)}, data)$$

.

.

.

$$x_{p-1}^{t+1}|. \sim f(x_{p-1}|x_1^{(t+1)}, x_2^{(t+1)}, ..., x_{p-2}^{(t+1)}, x_p^{(t)}, data)$$

$$x_p^{t+1}|. \sim f(x_p|x_1^{(t+1)}, x_2^{(t+1)}, ..., x_{p-2}^{(t+1)}, x_{p-1}^{(t+1)}, data)$$

   The operator $|.|$ means that we're conditioning on the most recent updates of all other elements of X.

3. We increment t and return to step t.

The Gibbs sampler is especially fruitful for Bayesian applications where the goal is to make inference based on the posterior distribution of multiple parameters. In the case of Bayesian hierarchical models (such as *belief networks*), most parameters only depend on a small number of other ones to a considerable extent, thereby greatly facilitating the expression of the conditional probability distributions Gibbs sampling relies upon.

As we saw several times during the course of this series of lecture, the purpose of Bayesian inference is to determine or to approximate the posterior probability distribution that can always be written as $f(x|data) =$

[5] As we saw in 5.3, the formulation given below is a simplification that doesn't completely consider the continuous nature of the random variable.

$\frac{f_0(x)L(data|x)}{c}$, where $c = p(data) = \int\limits_{x \in S} f_0(x)L(data|x)dx$ is an unknown integration constant.

When the required univariate conditional densities are easily computable, the Gibbs sampler can be applied and does not require the evaluation of the constant c. In this case, the i-th step in a cycle of the Gibbs sampler at iteration t consists of

$$x_i^{t+1}|(x_{-i}^t, data) \sim f(x_i|x_{-i}^{(t)}, data),$$

whereby

$$x_{-i}^t = (x_1^{(t+1)}, ..., x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, ..., x_p^{(t)}).$$

Theorem 5.2 Let us consider a posterior $f(x|data)$ which is such that any conditional distribution thereof is strictly positive everywhere. This means that

$$\forall x \in S, \forall i \in (1; p), \forall y \in S_i, , f(y|x_{-i}, data) > 0.$$

Any sequence of random variables generated with the Gibbs algorithm is a Markov chain converging towards its unique invariant distribution $f(x|data)$.

Proof.

Let $\pi(x) = f(x|data), x \in S$ be the posterior distribution we want to simulate.

In order to generate x^{t+1} iteratively, we only need to know the values of x^t and don't require the knowledge of any earlier sample [6]. $(X_n)_{n \geq 0}$ is thus a Markov chain.

What is more, it is irreducible as starting from $x^t \in S$, any $x \in S$ can be reached component-wise after one single step comprising p iterations. It is aperiodic as well since x can be arbitrarily close to x^t (see Theorem 4.2).

We must first compute the transition kernel. Given $x = (x_1, ..., x_p)$, y_1 only depends on $x_2, ..., x_p$, y_2 only depends on $y_1, x_3, ..., x_p$,... and y_p only depends on $y_1, ..., y_{d-1}$. We can define a random sequence $(Z_k)_{1 \leq k \leq d}$ of finite dimension such that $Z_k = (Y_1, Y_2, ..., Y_{k-1}, Y_k)$. $(Z_k)_{1 \leq k \leq d}$ is a finite Markov chain since $Z_2 = (Y_1, Y_2)$ only depends on $Z_1 = Y_1$, $Z_3 = (Y_1, Y_2, Y_3)$ only depends on Z_2 and Z_p only depends on Z_{d-1}. As a consequence, we can apply the Markov property which tells us that [7]

$$P(x, dy) = P(x, dz) = f(z_1|x)f(z_2|x, z_1)...f(z_p|x, z_{d-1})dy_1...dy_p$$
$$= f(y_1|x)f(y_2|x, y_1)...f(y_p|x, y_1, ..., y_{d-1})dy_1...dy_p.$$

[6] The fact that we don't use all components of x^t doesn't change anything to the situation.

[7] f simply denotes a generic probability density.

Given what we know about the relationship between the y_k and the $x_{k'}$, we find that

$$\begin{aligned} P(x, dy) &= \pi(y_1|x_2, ..., x_p)\pi(y_2|y_1, x_3, ..., x_p)...\pi(y_p|y_1, ..., y_{d-1})dy_1...dy_p \\ &= \Pi_{k=1}^{p}\pi(y_k|y_1, ..., y_{k-1}, x_{k+1}, ..., x_p)dy_k \end{aligned}$$

with the convention that all x_k and y_k in the expressions are ignored if $k > p$ or $k < 1$.

We shall now prove that π is the invariant distribution so that

$$\int_{x\in S} P(x, dy)\pi(x)dx = \pi(y)dy.$$

We have

$$\int_{x\in S} P(x, dy)\pi(x)dx = \int_{x\in S} \Pi_{k=1}^{p}\pi(y_k|y_1, ..., y_{k-1}, x_{k+1}, ..., x_p)dy_k\pi(x_1, ..., x_p)dx_1...dx_p$$

Moreover,

$$\begin{aligned} &\pi(y_k|y_1, ..., y_{k-1}, x_{k+1}, ..., x_p) = \frac{\pi(y_k, y_1, ..., y_{k-1}, x_{k+1}, ..., x_p)}{\pi(y_1, ..., y_{k-1}, x_{k+1}, ..., x_p)} \\ &= \frac{\pi(x_{k+1}, ..., x_p|y_1, ..., y_{k-1}, y_k)\pi(y_1, ..., y_{k-1}, y_k)}{\pi(x_{k+1}, ..., x_p|y_1, ..., y_{k-1})\pi(y_1, ..., y_{k-1})} \\ &= \frac{\pi(y_k|y_1, ..., y_{k-1})\pi(x_{k+1}, ..., x_p|y_1, ..., y_{k-1}, y_k)}{\pi(x_{k+1}, ..., x_p|y_1, ..., y_{k-1})} \end{aligned}$$

Consequently,

$$\begin{aligned} &\int_{x\in S} P(x, dy)\pi(x)dx = \\ &\int_{x\in S} \Pi_{k=1}^{p} \frac{\pi(y_k|y_1, ..., y_{k-1})\pi(x_{k+1}, ..., x_p|y_1, ..., y_{k-1}, y_k)}{\pi(x_{k+1}, ..., x_p|y_1, ..., y_{k-1})} dy_k\pi(x_1|x_2, ..., x_p)\pi(x_2, ..., x_p)dx_1...dx_p \\ &= \Pi_{k=1}^{p}\pi(y_k|y_1, ..., y_{k-1})dy_k \int_{x\in S} \frac{\pi(x_{k+1}, ..., x_p|y_1, ..., y_{k-1}, y_k)}{\pi(x_{k+1}, ..., x_p|y_1, ..., y_{k-1})}\pi(x_1|x_2, ..., x_p)\pi(x_2, ..., x_p)dx_1...dx_p \\ &= \Pi_{k=1}^{p}\pi(y_k|y_1, ..., y_{k-1})dy_k \\ &\times \Pi_{k=1}^{p} \int_{x\in S} \frac{\pi(x_{k+1}, ..., x_p|y_1, ..., y_{k-1}, y_k)}{\pi(x_{k+1}, ..., x_p|y_1, ..., y_{k-1})}\pi(x_1|x_2, ..., x_p)\pi(x_2, ..., x_p)dx_1...dx_p \end{aligned}$$

Based on the same kind of reasoning we used while introducing the Z_k, we can prove that

$$\Pi_{k=1}^{p}\pi(y_k|y_1, ..., y_{k-1})dy_k = \pi(y)dy.$$

We must now show that

$$A = \frac{\int_{x\in S} P(x,dy)\pi(x)dx}{\Pi_{k=1}^{p}\pi(y_k|y_1,...,y_{k-1})dy_k} = 1.$$

$$\begin{aligned}
A &= \int_{x\in S} \Pi_{k=1}^{p} \frac{\pi(x_{k+1},...,x_p|y_1,...,y_{k-1},y_k)}{\pi(x_{k+1},...,x_p|y_1,...,y_{k-1})}\pi(x_1|x_2,...,x_p)\pi(x_2,...,x_p)dx_1...dx_p \\
&= \int_{x\in S} \frac{\pi(x_2,...,x_p|y_1)}{\pi(x_2,...,x_p)}\frac{\pi(x_3,...,x_p|y_1,y_2)}{\pi(x_3,...,x_p|y_1)}\frac{\pi(x_4,...,x_p|y_1,...,y_3)}{\pi(x_4,...,x_p|y_1,y_2)}...\frac{\pi(x_p|y_1,...,y_p)}{\pi(x_p|y_1,...,y_{d-1})} \\
&\times \pi(x_1|x_2,...,x_p)\pi(x_2,...,x_p)dx_1...dx_p \\
&= \int_{x\in S} \Pi_{k=1}^{p} \frac{\pi(x_{k+1},...,x_p|y_1,...,y_{k-1},y_k)}{\pi(x_{k+2},...,x_p|y_1,...,y_k)}\pi(x_1|x_2,...,x_p)dx_1...dx_p \\
&= \int_{x\in S} \Pi_{k=1}^{p} \frac{\pi(x_{k+1},...,x_p,y_1,...,y_{k-1},y_k)\pi(y_1,...,y_k)}{\pi(x_{k+2},...,x_p,y_1,...,y_k)\pi(y_1,...,y_{k-1},y_k)}\pi(x_1|x_2,...,x_p)dx_1...dx_p \\
&= \int_{x\in S} \Pi_{k=1}^{p} \pi(x_{k+1},...,x_p,y_1,...,y_k|x_{k+2},...,x_p,y_1,...,y_k)\pi(x_1|x_2,...,x_p)dx_1...dx_p \\
&= \int_{x\in S} \Pi_{k=1}^{p} \pi(x_{k+1}|x_{k+2},...,x_p,y_1,...,y_k)\pi(x_1|x_2,...,x_p)dx_1...dx_p \\
&= \int_{x\in S} \Pi_{k=1}^{p} \pi(x_k|x_{k+1},...,x_p,y_1,...,y_{k-1})dx_1...dx_p
\end{aligned}$$

We can decompose this integral according to each x_k. x_1 only appears in $\pi(x_1|x_2,...,x_p)$ and $\int_{x_1\in S_1}\pi(x_1|x_2,...,x_p) = 1$. Likewise, x_2 only shows up in $\int_{x_2\in S_2}\pi(x_2|x_3,...,x_p,y_1) = 1$. In this way, we can demonstrate that $A = 1$. Finally, we find that

$$\int_{x\in S} P(x,dy)\pi(x)dx = \pi(y)dy.$$

A Markov chain based on Gibbs' sampling converges thus always towards the posterior probability distribution we're interested in.

Example 5.3 Let the *data* be made of the independent and identically distributed random variables $Y_1,...,Y_n \sim N(\mu, h^{-1})$, where $h = 1/\sigma^2$ is the precision. $h \to +\infty$ means that the only possible value is μ whereas $h \approx 0$ means that Y_i can take on values over a very large interval. Let us further assume that we decide to express our prior knowledge about μ and h by stipulating that they are initially independent and that $\mu \sim N(\mu_0, h_0^{-1})$ and $h \sim G(\alpha_0/2, \delta_0/2)$, where G designates the Gamma distribution, so that

$$f(h) = G(h) \propto h^{\alpha_0/2-1}exp(-\delta_0 h/2).$$

The posterior probability distribution is proportional to the product of the likelihood and of the two independent priors. Hence,

$$f(\mu, h|y) \propto L(y|\mu, h)f(\mu)f(h) = h^{n/2}exp\Big(-\frac{h}{2}\sum_{i=1}^{n}(y_i-\mu)^2\Big)\times$$
$$exp\Big(-\frac{h_0}{2}(\mu-\mu_0)^2\Big)h^{\alpha_0/2-1}exp(-\frac{\delta_0 h}{2}).$$

We can now calculate the conditional posterior distribution of h as

$$f(h|\mu, y) = \frac{f(h, \mu, y)}{f(\mu, y)}$$
$$\propto \frac{L(y|\mu, h)f(\mu)f(h)}{f(y|\mu)f(\mu)} = \frac{L(y|\mu, h)f(h)}{f(y|\mu)}$$

On the other hand, we have

$$L(y|\mu, h) = \frac{f(y, \mu, h)}{f(\mu, h)} = \frac{f(y, \mu, h)}{f(\mu)f(h)}$$
$$\Rightarrow f(y|\mu) = \frac{f(y, \mu)}{f(\mu)} = \frac{\int_{h=0}^{+\infty} f(y, \mu, h)dh}{f(\mu)} = \frac{\int_{h=0}^{+\infty} L(y|\mu, h)f(\mu)f(h)dh}{f(\mu)}$$
$$f(y|\mu) = \int_{h=0}^{+\infty} L(y|\mu, h)f(h)dh$$

Finally, we find that

$$f(h|\mu, y) = \frac{L(y|\mu, h)f(h)}{\int_{h=0}^{+\infty} L(y|\mu, h)f(h)dh}.$$

We have

$$L(y|\mu, h)f(h) \propto h^{n/2}exp\Big(-\frac{h}{2}\sum_{i=1}^{n}(y_i-\mu)^2\Big)h^{\alpha_0/2-1}exp(-\delta_0 h/2)$$
$$= h^{(\alpha_0+n)/2-1}exp\Big\{-\frac{h}{2}\Big(\delta_0+\sum_{i=1}^{n}(y_i-\mu)^2\Big)\Big\}$$

This leads to

$$f(h|\mu, y) \propto \frac{h^{(\alpha_0+n)/2-1}exp\Big\{-\frac{h}{2}\Big(\delta_0+\sum_{i=1}^{n}(y_i-\mu)^2\Big)\Big\}}{\int_{h=0}^{+\infty} h^{(\alpha_0+n)/2-1}exp\Big\{-\frac{h}{2}\Big(\delta_0+\sum_{i=1}^{n}(y_i-\mu)^2\Big)\Big\}dh}.$$

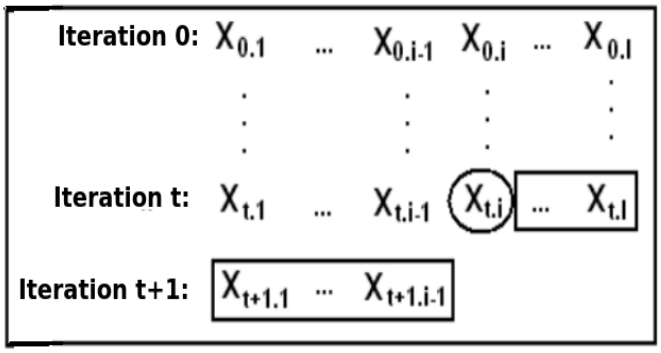


Figure 5.8: Single Component Metropolis-Hastings algorithm

Let us suppose we have μ^{t+1} and we want to sample the next h^{t+1} based on $f(h^{t+1}|\mu^{t+1}, y)$. For that sake, if we fail to find an analytical expression of the integral on the denominator $\int_{h=0}^{+\infty} L(y|\mu^{t+1}, h)f(h)dh$, we can compute it either deterministically or with the simple Monte-Carlo method [8]

Similarly, by reversing the roles of h and μ, we directly find the conditional posterior distribution of μ:

$$f(\mu|h, y) = \frac{L(y|h, \mu)f(\mu)}{\int_{\mu=0}^{+\infty} L(y|h, \mu)f(\mu)d\mu}.$$

Let us suppose we have h^t and we want to sample the next μ^{t+1} based on $f(\mu^{t+1}|h^t, y)$. Like before, the integral $\int_{\mu=0}^{+\infty} L(y|h^t, \mu)f(\mu)d\mu$ can be computed numerically if one fails to find an analytical expression for it.

5.4.2 Single Component Metropolis-Hastings

Gibbs' sampling is a specific case of the *single component Metropolis-Hastings*. The principle can be visualised in Figure 5.8. Let's introduce $x = (x_1, x_2, ..., x_d)$, $x_{.-i} = (x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_d)$ and $x_{t.-i} = (x_{t+1.1}, x_{t+1.2}, ..., x_{t+1.i-1}, x_{t.i+1}, ..., x_{t.d})$ To simplify things, we'll write the posterior probability distribution as $f(x|data) = f(x)$. This algorithm works as follows:

1. Select the starting values x^0 and set $t = 0$.
2. For $i = 1$ to d
   (a) Produce y_i according to the proposal distribution $q_i(y_i|x_{t.i}, x_{t.-i})$.

[8] This isn't much of an issue since it is unidimensional.

(b) Accept i with the probability

$$\alpha(x_{t.-i}, x_{t.i}, y_i) = min(1, \frac{f(y_i|x_{t.-i})q_i(x_{t.i}|y_i, x_{t.-i})}{f(x_{t.i}|x_{t.-i})q_i(y_i|x_{t.i}, x_{t.-i})}).$$

Reject it otherwise.

3. We increment t and return to the first step.

$f(.|x_{.-i})$ is the full conditional distribution which is given by

$$f(x_i|x_{.-i}) = \frac{f(x_i)}{\int_{x'_i \in S_i} f(x'_i, x_{.-i})dx'_i}.$$

Theorem 5.3 If the full conditional distributions and proposal distributions are strictly positive everywhere, the Markov chain produced by the Single Component Metropolis-Hastings algorithm converges towards the posterior probability distribution $f(x) = f(x|data)$ which is its invariant distribution.

Proof.

If the full conditional distributions are strictly positive everywhere, so is the full target distribution. We express the overall acceptance probability of y given x_t and show that it must be strictly positive everywhere as well if the $\alpha(x_{t.-i}, x_{t.i}, y_i)$ are strictly positive everywhere. We can then apply Theorem 5.3 to prove that the MC must converge towards its invariant distribution $f(x)$.

5.4.3 Implementation of the MCMC

All the MCMC methods we've explored so far converge towards the stationary distribution which interests us (generally a posterior probability distribution). From a practical standpoint, we must always make sure that we have sampled enough values of the Markov chain to accurately represent the target distribution. Unluckily, the convergence of MCMC methods can sometimes be excruciatingly slow, especially while dealing with a high dimensional parameter space. Moreover, widespread criteria used to deem whether a MCMC has successfully converged can be quite misleading. In this last part of the lecture, we want to focus on (mutually dependent) questions related to the long-run behaviour of the Markov chain such as

- Have we drawn samples of the MC for long enough?
- Has the Markov chain gone through all regions of the parameter space where the probability distribution takes on non-negligible values (the so-called "support" of the distribution)?
- How good is the approximation of the invariant distribution?

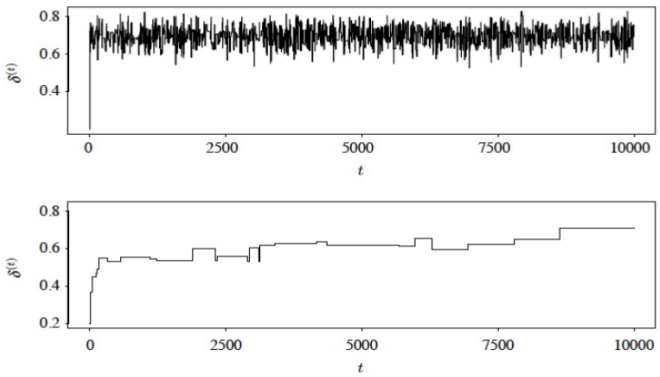


Figure 5.9: Sample paths for the example in Fig. 5.3. The proposal densities are Beta(1, 1) (top) and Beta(2, 10) (bottom).

- How can we use the output of our tedious computations to estimate quantities of interest to us and assess the reliability and precision of the approximated values we get?

5.4.4 Ensuring Good Mixing and Convergence

There are two main problems we must deal with:

- Mixing: how fast does the MC forget its starting value? How quickly does the Markov chain explore the support of the target distribution? How far apart must several sampled values be in order to be seen as approximately independent?

- Convergence: Has the Markov chain approximately reached its stationary distribution?

  These two questions and the methods and tools used to answer them are inextricably linked. It is always advisable to employ several verification techniques to increase our confidence in the outcome.

We'll begin by examining two simple **graphical diagnostics**.

The *sample path* is a plot of $X(t)$ as a function of the iteration number t. A good mixing goes hand in hand with quick and strong oscillations covering the whole region of support. When the mixing is poor, the MC will remain near the same value for a long time without any strong fluctuations. The empirical distribution will be very far from the invariant distribution. The two situations are shown in Figure 5.9.

The *autocorrelation plot* sums up the autocorrelation of the Markov chain $X(t)$ as a function of the **lag**, i.e. the difference $t_2 - t_1$ between two time steps t_1 and t_2. The autocorrelation at lag k is the correlation between $X(0)$ and $X(k)$. An example is shown in Figure 5.10. In the first case,

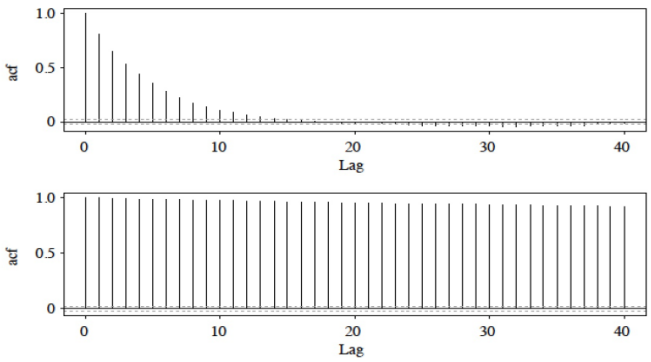


Figure 5.10: Autocorrelation plot for the example in Fig. 5.3. The proposal densities are Beta(1, 1) (top) and Beta(2, 10) (bottom).

the autocorrelation swiftly decreases, which shows that the members of the Markov chain quickly become independent. In the second case, it stays constant and very close to 1, which shows that the entire Markov chain is strongly determined by the first value.

The problem of autocorrelation naturally leads to the notion of **burn-in period**. As we saw before, the burn-in period is the time the Markov chain needs to become independent of its starting value. It is only after that period that the Markov chain will begin to yield a good approximation of the target distribution. As a consequence, we dismiss the D first values of the MC corresponding to that phase. Generally, $100 < D < 10000$ but the optimal value is always context-dependent. The burn-in period is extremely short (practically non-existent) if we start off in a region of high density of the target distribution (provided it is unimodal).

Another important aspect of a MCMC algorithm is the **choice of the proposal distribution** Q. A good mixing strongly depends on the properties of the proposal distribution and especially on its *spread* (i.e. variance). In the case of the Metropolis-Hastings algorithm, it is highly desirable to have a jumping distribution that approximates well the target distribution, so that new candidates within the high-density region will most often be accepted. It is even more important to make sure that Q has non-negligible values in the *tails* (low-density regions) of $f(x|data)$ so that the MC won't ignore them altogether.

Practically speaking, it is possible to fine-tune the variance of the proposal distribution **iteratively**. After having launched the Markov chain, we compute the acceptance rate and adjust the spread (variance) of the proposal distribution in such a way that the acceptance rate becomes close to a desirable value (which is generally comprised between 25 % and 50 %. In order to make this work, both the target and the proposal distribution must

be roughly normally distributed or at the very least unimodal. Multimodal distributions (such as the mixture distribution shown in Figure 5.1) are particularly problematic for MCMC algorithms. Indeed, if the variance of the proposal distribution is too small, the Markov chain can easily get stuck in one mode and fail to explore the other one. In such a situation, while the acceptance rate may be high, the probability of jumping from one mode to the other is very low. A foreknowledge about the target distribution is always very useful, even though it is often not available.

The notion of **effective sample size** is of utmost importance to the successful use of the MCMC method. If the realisations of the Markov chain are highly correlated, the information produced by the whole MC will be much smaller than that suggested by its total length. The effective sample size is the size of an independent identically distributed random sequence that contains the same amount of information as the real MC. To estimate this quantity, we first calculate the autocorrelation time defined as

$$\tau = 1 + 2\sum_{k=1}^{\infty} \rho(k),$$

where $\rho(k) = \frac{cov(X(0),X(k))}{\sqrt{Var(X(0))Var(X(k))}}$. A widespread approach to estimating τ is to truncate the sum whenever $\hat{\rho(k)} < 0.1$. The effective sample size of a MC of length L is then approximately given by $L/\hat{\tau}$. The effective sample size can be employed to compare the relative efficiency of different MCMC algorithms. For instance, if a Gibbs sampler has a much higher effective sample size than a regular Metropolis-Hastings algorithm with uncoupled variables, it means it is much more efficient to approximate the target distribution.

We'll finally take a look at how to summarise the results of a MCMC simulation. We're most often interested in quantities such as the mean and the variance of the target distribution and variables derived thereof. The most commonly used estimator of the mean $\mu = E\{h(X)\}$ is the empirical mean without considering the burn-in phase:

$$\hat{\mu} = \frac{1}{L}\sum_{t=D}^{D+L-1} h(X(t))$$

where L is the length of the MC without the burn-in period. This estimator is consistent even though the $X(t)$ are correlated.

To estimate how far the realisations of the estimators deviate from the true values stemming from the target distribution, we can use the **simulation standard error** (sse). It shows the variability of the estimator if the MC were to be run an infinite number of times. The naive estimate of the standard error for an estimator like μ is the sample standard deviation

of the L realisations after the burn-in phase divided by L. However, as we saw above, realisations belonging to the same Markov chain are typically correlated, which can lead to an underestimation of the standard error. To overcome this problem, we can resort to the *batch means method* which boils down to dividing the single Markov chain we've simulated into a series of Markov chains that can be considered to be roughly independent. We begin by examining the autocorrelations of a concrete Markov chain to determine a lag k_0 such that the autocorrelation is small enough to be ignored, e.g. $\rho(\hat{k}_0) \leq 0.05$. We then divide the L observations into $B \approx L/k_0$ batches. Let $\hat{\mu_b}$ be the mean of $h(X(t))$ in batch b. The sample variance of the means is then given by

$$S^2 = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\mu_b} - \hat{\mu})^2$$

and the estimated standard error is

$$s\hat{s}e(\hat{\mu}) = \sqrt{S^2/B}.$$

Chapter 6

Sources

The following materials have been used to create this lecture:

Thierry Denoeux: "Computational statistics: Markov Chain Monte Carlo methods"

Andrej Depperschmidt: "Markovketten", in German.

Anna Mikusheva, course materials for 14.384 Time Series Analysis, Fall 2007. MIT OpenCourseWare (http://ocw.mit.edu),Massachusetts Institute of Technology

Bregt Savat: "Monte Carlo Markov Chain methoden", in Dutch.

Bibliography

[1] S. Chakravarty, M. Fischer, P. García-Triñanes, D. Parker, O. Le Bihan, M. Morgeneyer, Powder Technology (2017)