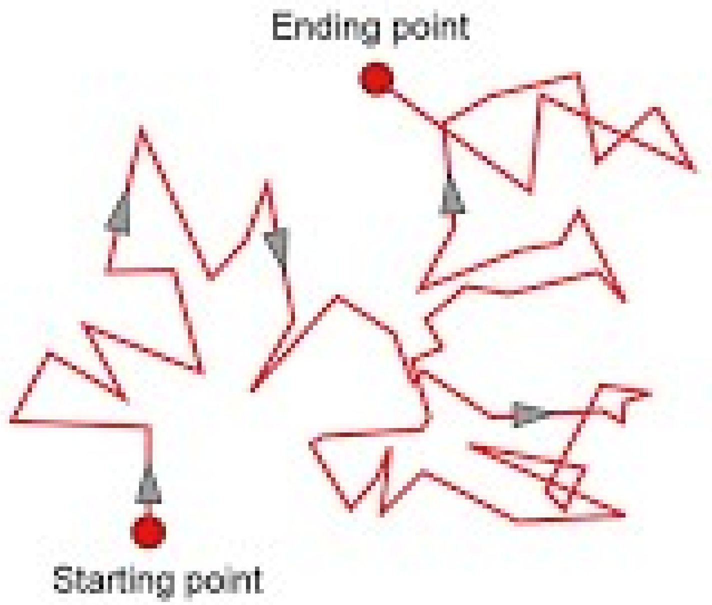
Markov chains
Markov Chain Monte-Carlo Method (MCMC)

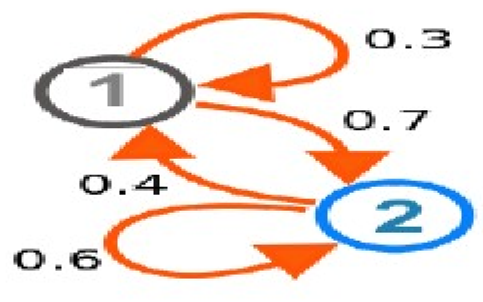
Dr. Marc Fischer (Assistant professor)
marc.fischer@emse.fr

Mines Saint-Etienne


Ending point
Starting point

1 Introduction

Markov chains (MC) are stochastic processes where the future only depends on the present and not on the past.



$$p(X_{i+1} = 1 | X_i = 1) = 0.3$$
$$p(X_{i+1} = 2 | X_i = 1) = 0.7$$
$$p(X_{i+1} = 2 | X_i = 2) = 0.6$$
$$p(X_{i+1} = 1 | X_i = 2) = 0.4$$

(Simple Markov chain)

In biology, the Galton-Watson process models the population of asexual beings.

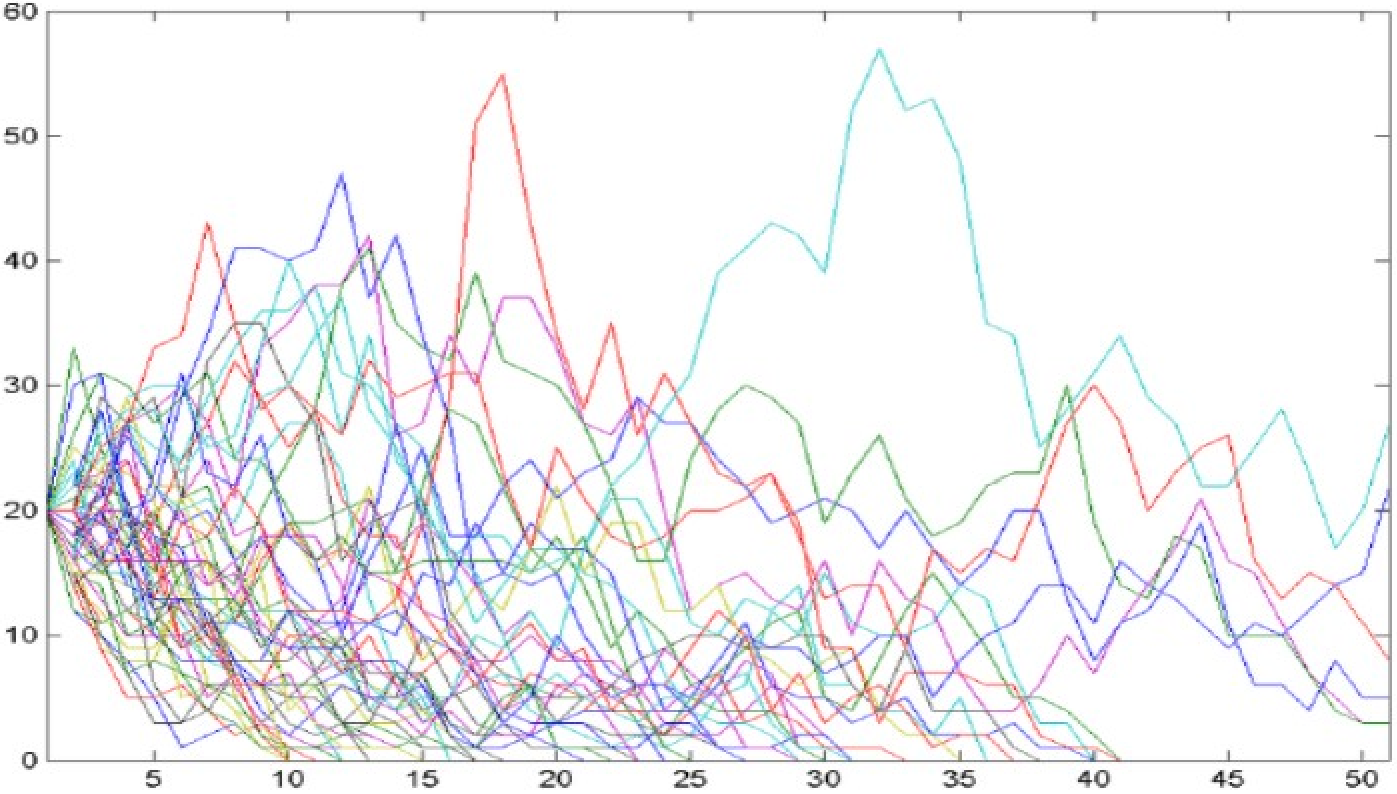
The number of newborn individuals at time *i*, Z_i follows a Poisson-distribution:

$$p(Z_i = k) = P_\lambda(k) = \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda}.$$

With $X_0 = a$ as the initial population, we have $X_{i+1} = X_i + \sum_{j=1}^{X_i} Z_j$.

Given the current population X_i, X_{i+1} is determined by the number of births coming from each of the X_i individual.

We don't need to know X_{i-1}, X_{i-2}, X_{i-3} etc. to determine the probability distribution of X_{i+1}.

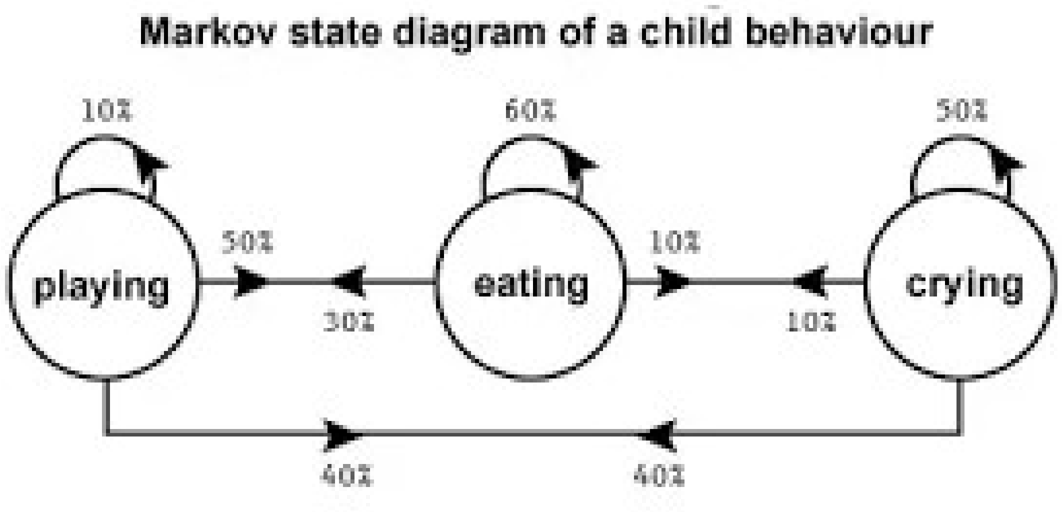


Galton process ($\lambda = 0.95$)

Only 6 out of 50 populations survived.

Galton process $(\lambda = 1)$

24 out of 50 populations survived.



Interpret this diagram!

Will this kid be thin or fat?

2 Markov chains in a finite state space

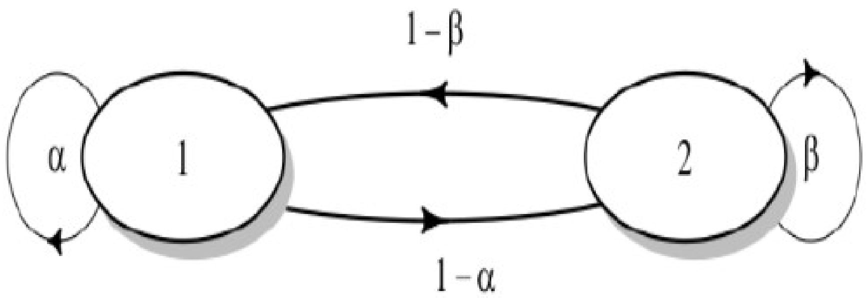
2.1 Basics

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

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

I is a finite state space such as N.

Here, $I = \{1,2\}$.



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

Definition 3.2 (Markov chain in discrete time)

Let $v = (v(i))_{i \in I}$ 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 v, 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

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

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. $\boxed{3.3}$ 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.

It is always possible to transform an <u>inhomogeneous</u> Markov chain into an homogeneous one.

Therefore, we can limit our study to homogeneous MC.

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 $i_0, ..., i_n \in I$

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

Any probability distribution υ and stochastic matrix *P* define a 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.

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 $X_0, ..., X_{m-1}$.

The MC restarts thus at time *i*.

The past is irrelevant / unimportant.

1, 3, -2, 4, 5, 9, -5, 2, **1, -8, 3, -4, 11, 9, ….**

At i = 9, with X_9 = 1, the MC behaves probabilistically as if it had began at i = 1 with X_1 = 1.

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 <u>given the present</u>, the past and the future are <u>independent</u>.

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.

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

Given a Markov chain such that ν is the probability distribution of X_0, (νP^n) is the probability distribution of X_n.

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

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 $k \in I$,

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

2.2 Communicating classes and periods

Let us consider the following random walk.

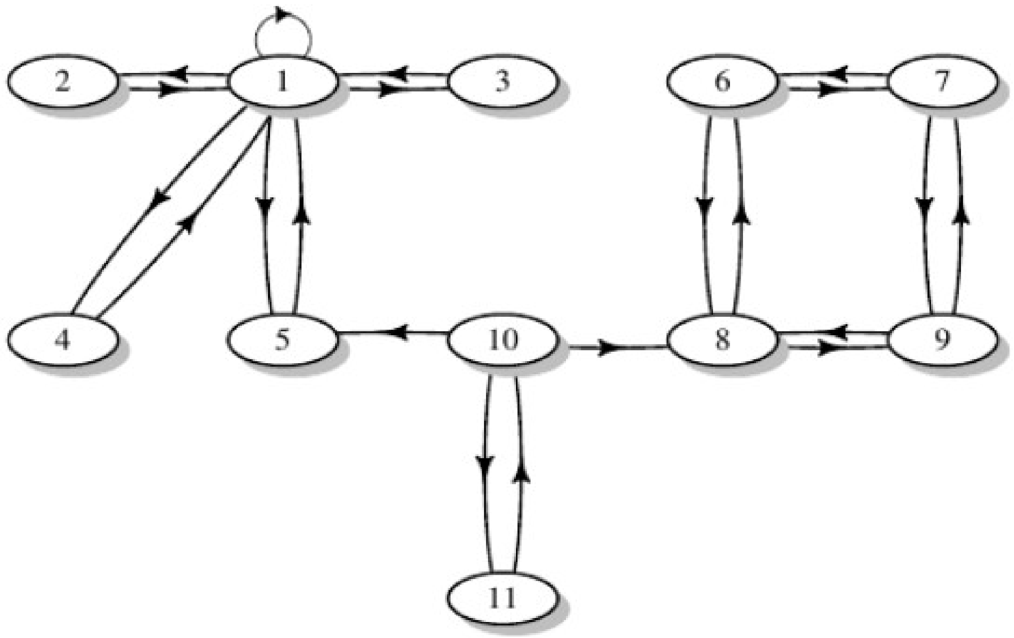


Figure 3.4: Random walk on a graph

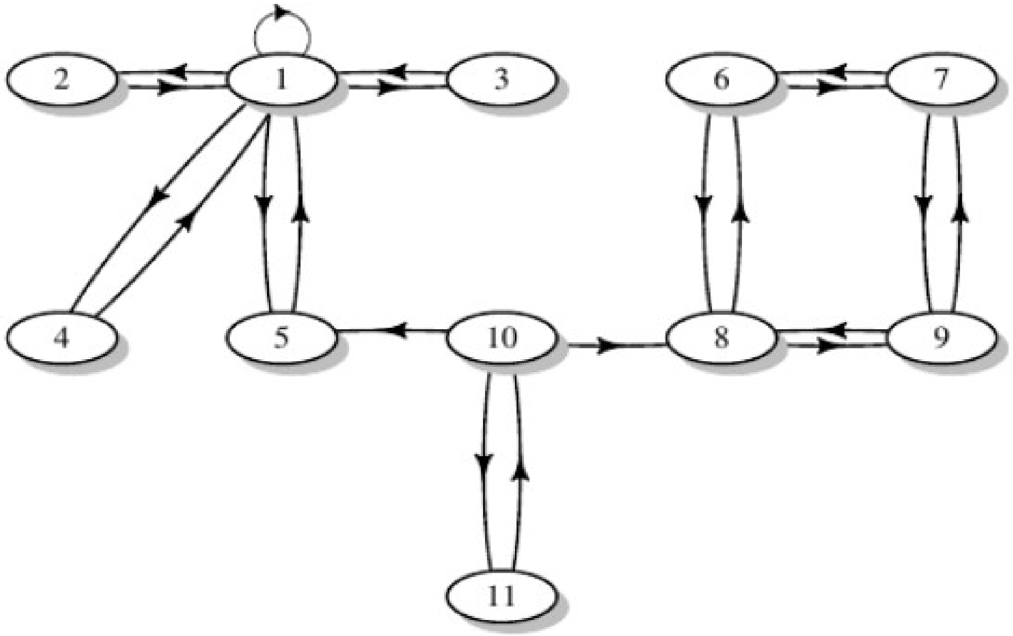


Figure 3.4: Random walk on a graph

Can the Markov chain spend an infinite amount of time in **11**?

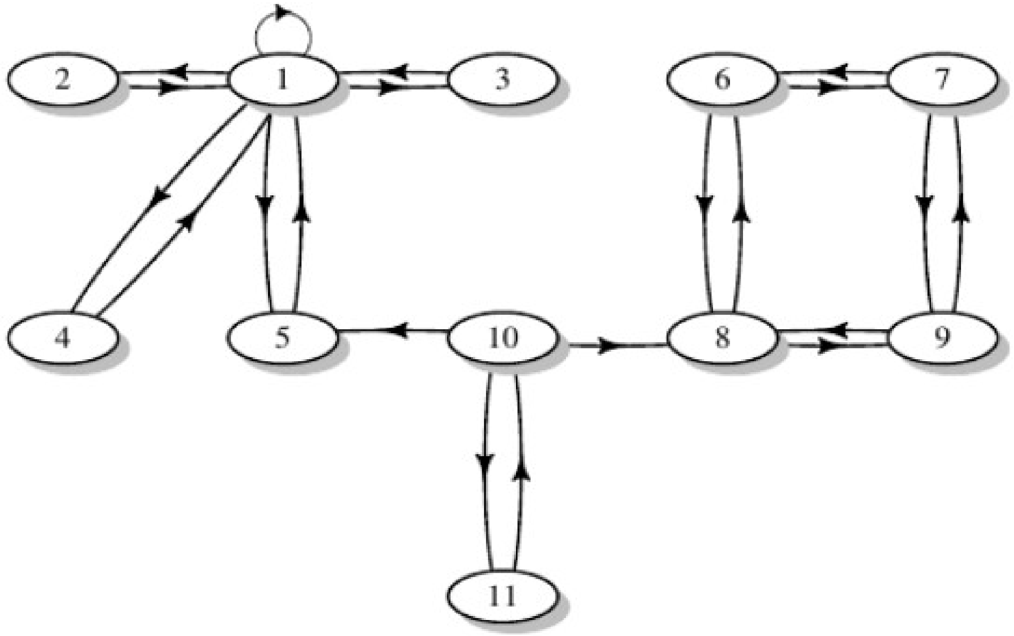


Figure 3.4: Random walk on a graph

Can the Markov chain spend an infinite amount of time in **11**?

No, it will sooner or later land in **5** or **8** from which it cannot return to **11**.

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**.

A **communicating class** is an ensemble of states that communicate with one another.

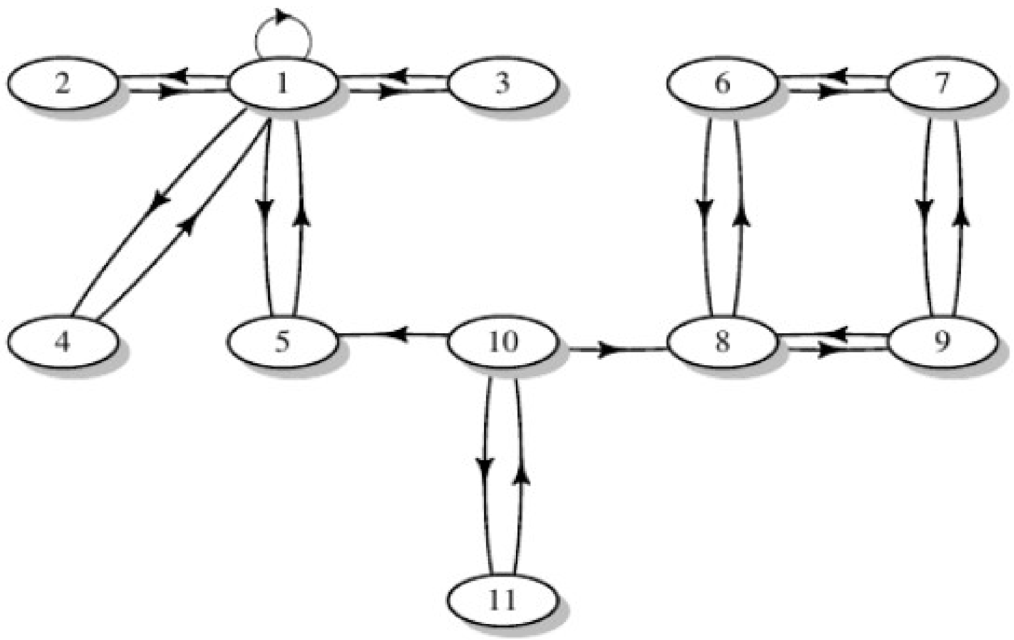


Figure 3.4: Random walk on a graph

Which states are essential and inessential?

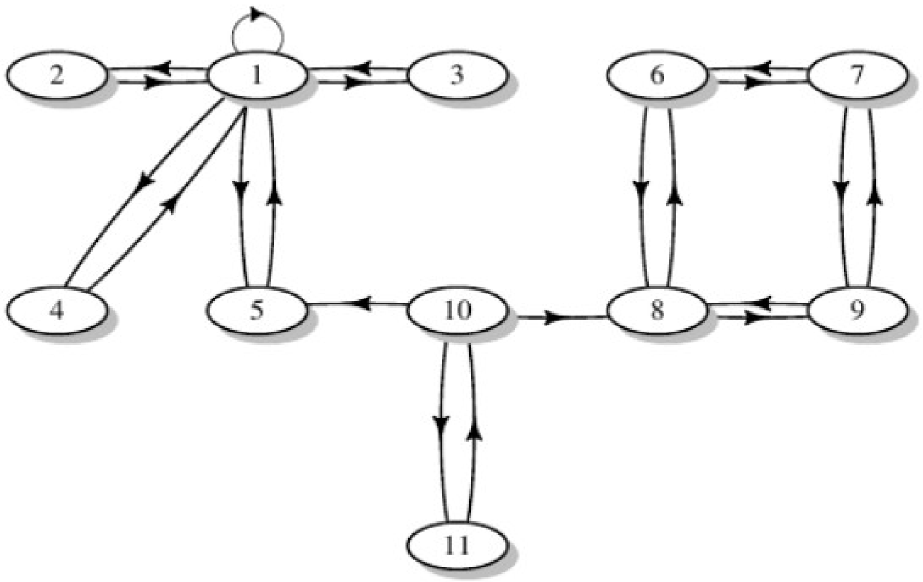


Figure 3.4: Random walk on a graph

Which states are essential and inessential?

10 and **11** are inessential. All other states are essential.

Only essential states are important for the asymptotic behaviour of the MC.

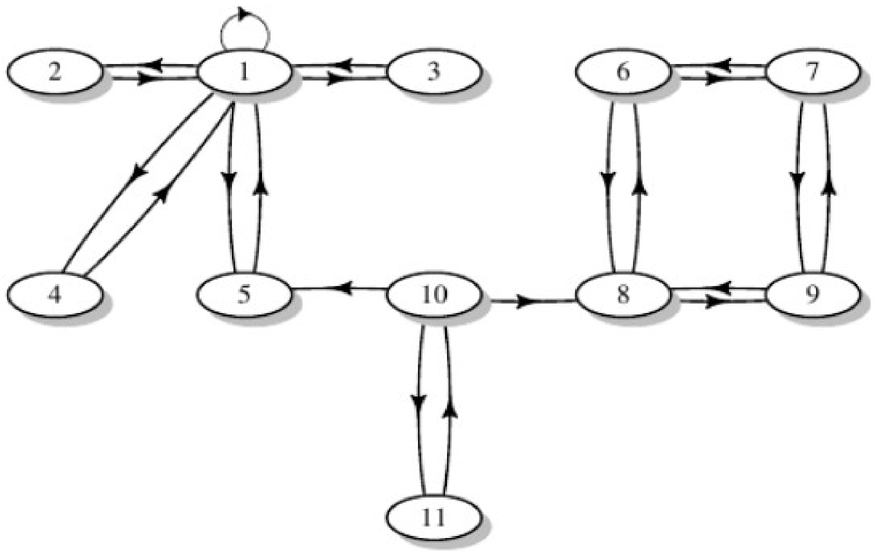


Figure 3.4: Random walk on a graph

How many communicating classes do you recognise?

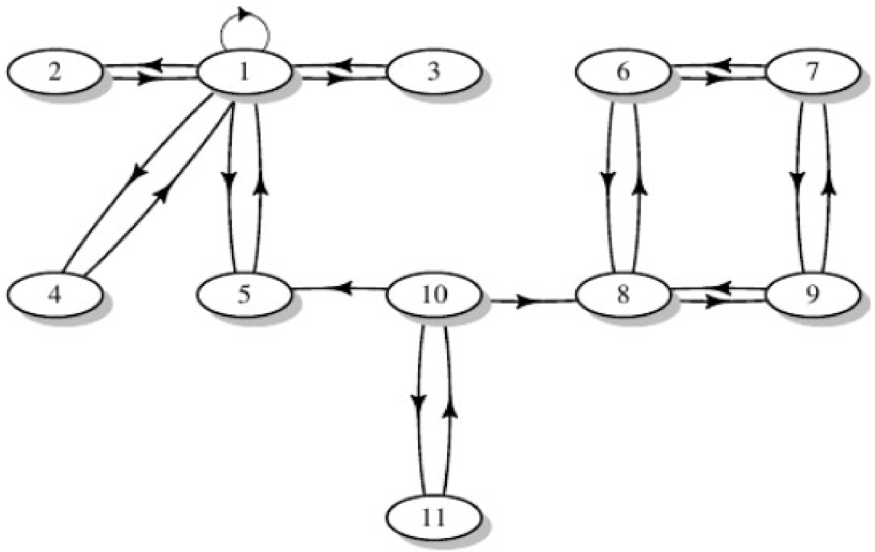


Figure 3.4: Random walk on a graph

How many <u>communicating classes</u> do you recognise?

{1,2,3,4,5} and **{6,7,8,9}** are communicating classes.

{10,11} are inessential states.

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 $\boxed{2.20}$, this is equivalent to the irreducibility of the transition matrix P.

(See “Markov_Chain_Lecture.pdf”)

The state space *I* is irreducible if and only if it forms a single communicating class.

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

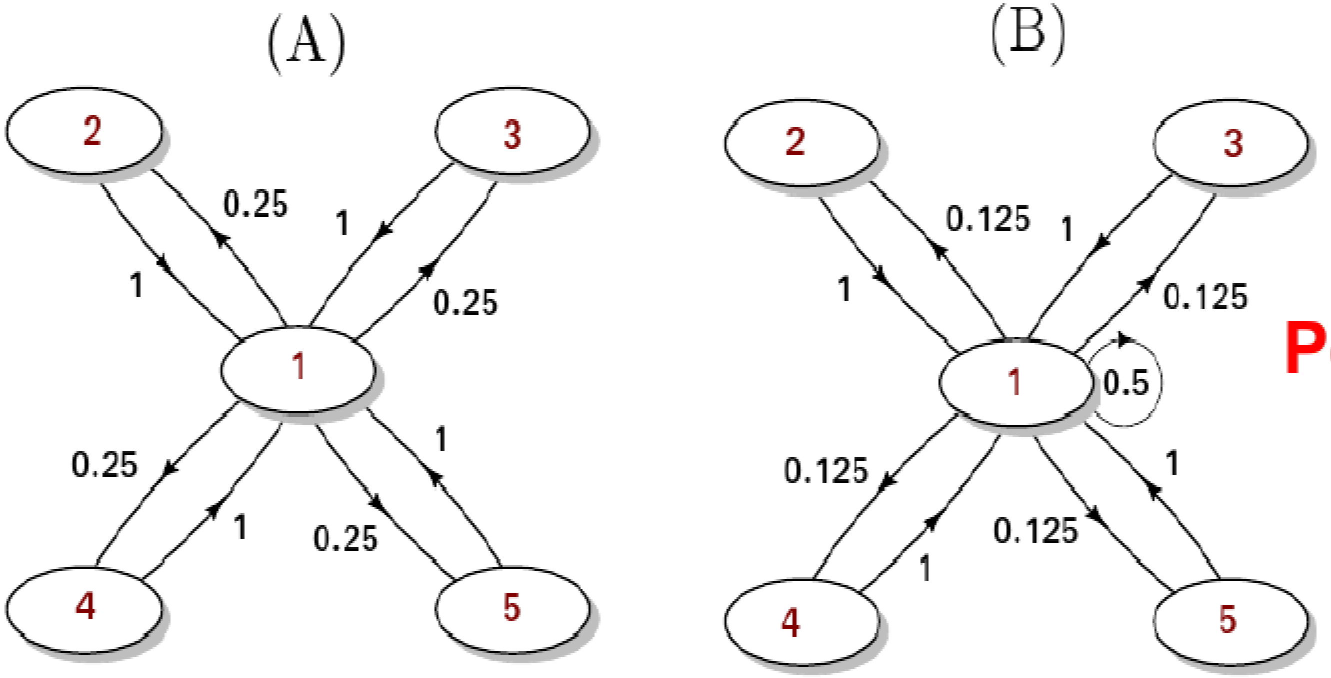
$$d_i = gcd\{n \geq 1 : p_{ii}^{(n)} > 0\}$$

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

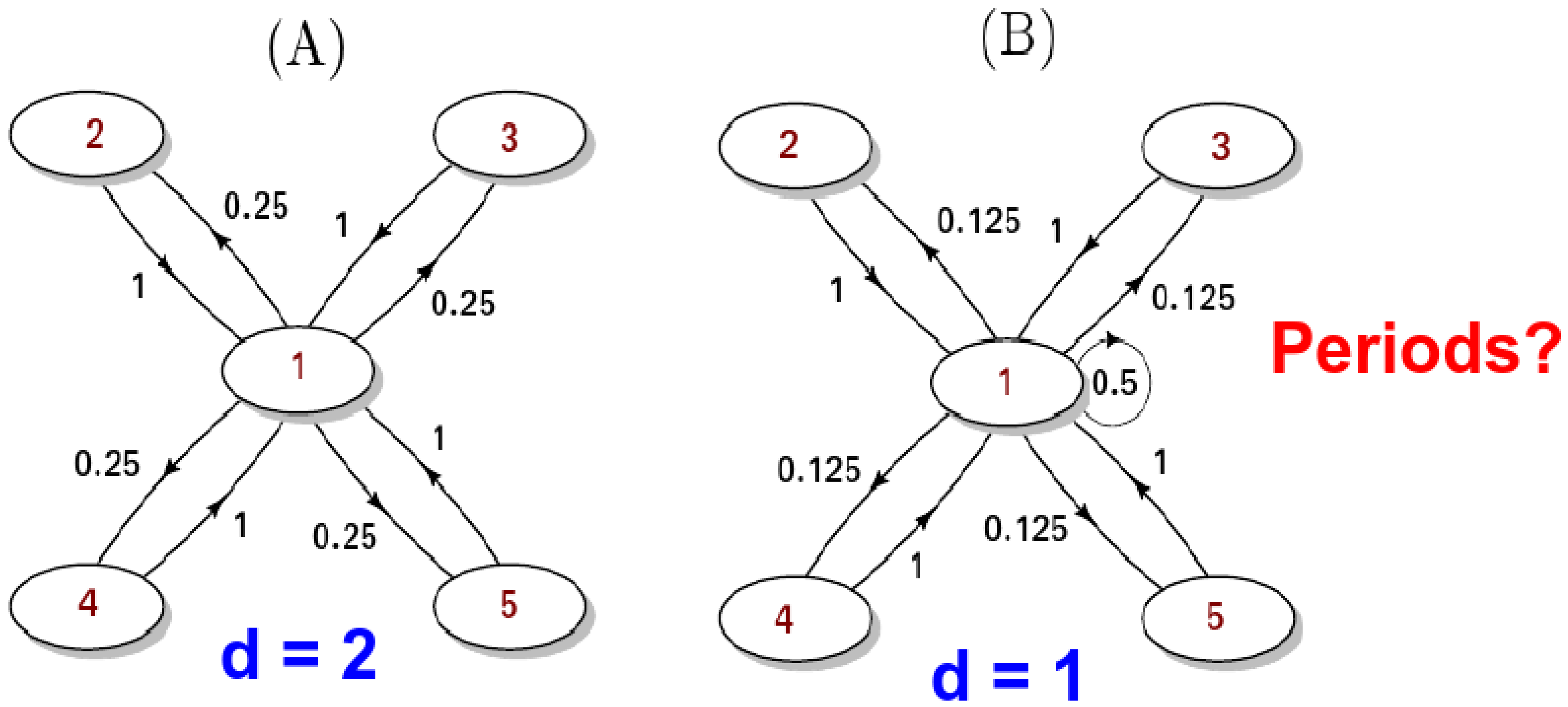
gcd: greatest common denominator.

The period is a property of a **communicating class**.

All members of a communicating class have the same period.


(A)
2
3
1
4
5
0.25
1
1
0.25
0.25
1
1
0.25
(B)
2
3
1
4
5
0.125
1
1
0.125
0.5
0.125
1
1
0.125

Periods?



B) is not periodic because the MC can remain in **1** during several steps.

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 $n \geq n_0$.

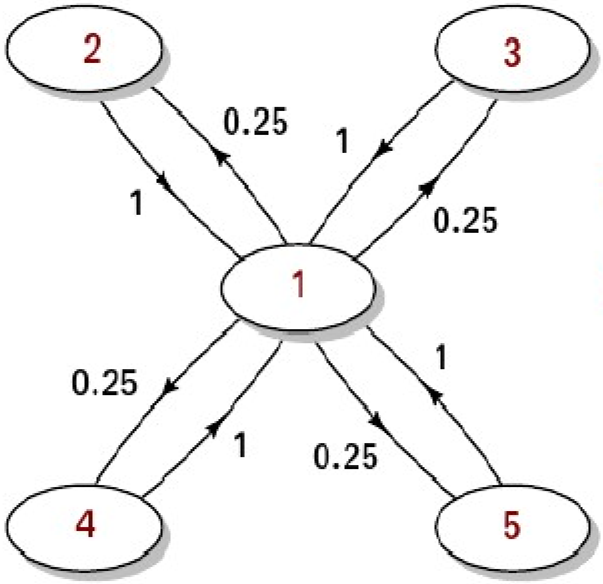
Definition 3.7 If C is a communicating class, then we call $d(C)$ the (common) period of the elements of that class. We say that the class is *aperiodic* if $d(C) = 1$.

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 $l \in \{0, 1, ..., d-1\}$,

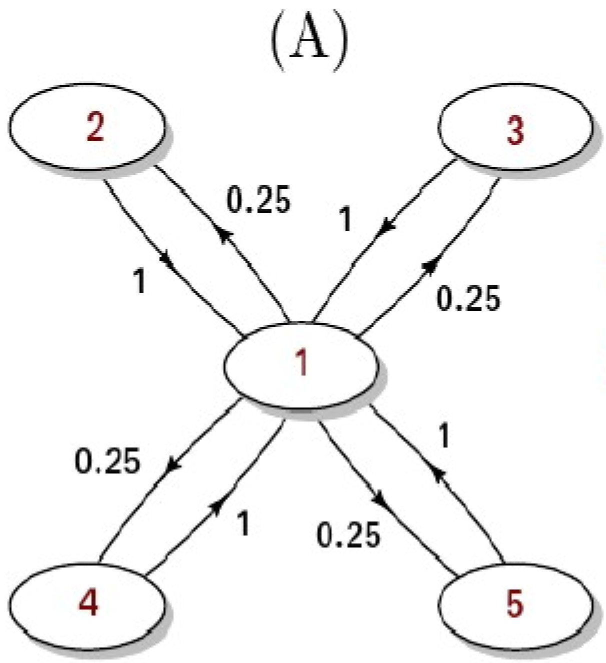
$$\sum_{j \in K_{l+1}} p_{ij} = 1,$$

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.

(A)



What are the cyclical subclasses?



What are the cyclical subclasses?

{1} and {2,3,4,5}.

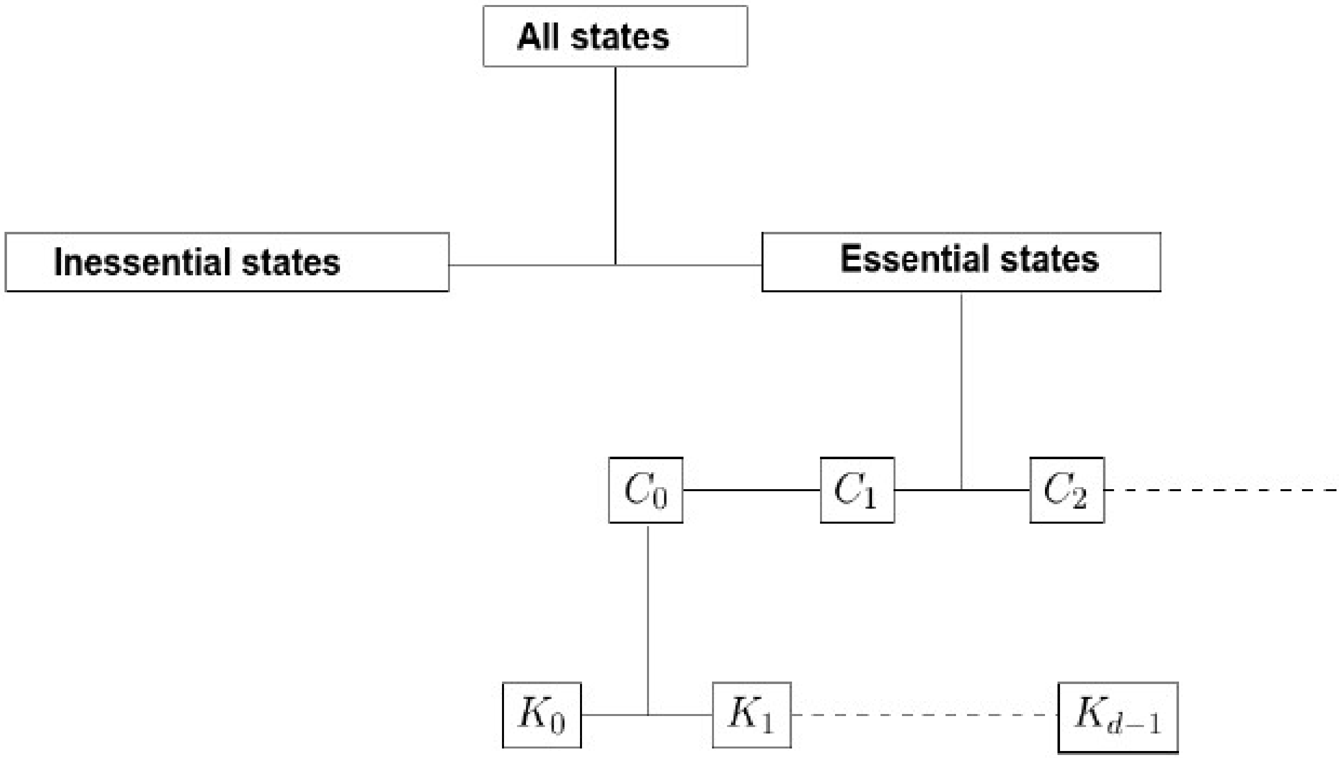
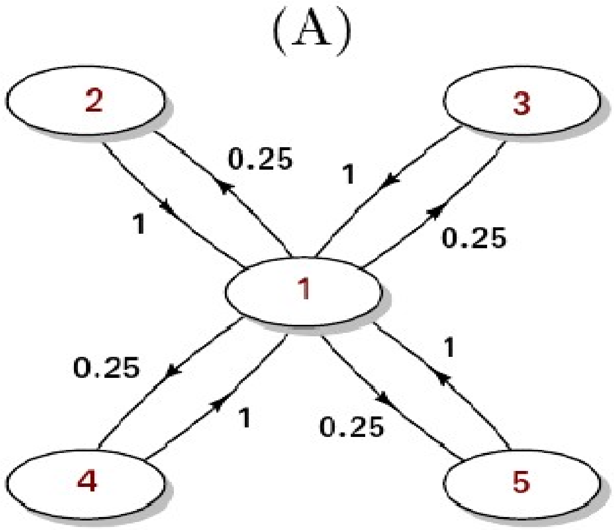


Figure 3.7: Classification of all states. $C_0, C_1, \ldots$ are communicating classes. $K_0, \ldots, K_{d-1}$ are cyclical subclasses of C_0 with $d = d(C_0)$



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

After 2 steps, the MC remains in the same cyclical subclass.

For a periodic Markov chain with d = 5, we have

$$P = \begin{pmatrix} 0 & P_{12} & 0 & 0 & 0 \\ 0 & 0 & P_{23} & 0 & 0 \\ 0 & 0 & 0 & P_{34} & 0 \\ 0 & 0 & 0 & 0 & P_{45} \\ P_{51} & 0 & 0 & 0 & 0 \end{pmatrix}. \qquad P^5 = \begin{pmatrix} E_{11} & 0 & 0 & 0 & 0 \\ 0 & E_{22} & 0 & 0 & 0 \\ 0 & 0 & E_{33} & 0 & 0 \\ 0 & 0 & 0 & E_{44} & 0 \\ 0 & 0 & 0 & 0 & E_{55} \end{pmatrix}$$

After d = 5 steps, we are back in the same cyclical subclass.

2.3 Markov times and the strong Markov property

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.

Example of event: $\{X_0 = 3, -2 < X_1 < 4, X_2 = 5, X_3 > 32\} \in F_3$.

Definition 3.9 (Markov time)
An application $T : \Omega \to \mathbb{N}_0 \bigcup \{+\infty\}$ is a **Markov time** of $(X_n)_{n>0}$ if and only if for each $n \in \mathbb{N}_0$, we have $\{T = n\} \in F_n$.

That sounds rather abstract, doesn't it? ;-)

The event $\{T = n\}$ can only depend on $X_0, X_1, ..., X_n$ and not on X_{n+1} or any subsequent value of the Markov chain.

Now, let us see some examples!

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

2. $T_A := inf\{n \geq 1 : X_n \in A\}$, the first return time in A is a Markov time.

$S_A := inf\{n \geq 0 : X_n \in A\}$, the first arrival time in A is a Markov time.

Both T_A and S_A are Markov times because $\{T_A = n\}$ and $\{S_A = n\}$ entirely depend on $F_n := \sigma(X_m, m \leq n)$.

3) Is the exit time from A $L^A := sup\{n \geq 0 : X_n \in A\}$, a Markov time?

3) Is the exit time from A $L^A := sup\{n \geq 0 : X_n \in A\}$, a Markov time?

No, because $\{L^A = n\}$ depends on all subsequent values of (X_k) until ∞!

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 $X_0, ..., X_T$.

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

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

2.4 Recurrence and transiency

Return time in *i*: $T_i := inf\{n \geq 1 : X_n = i\}$.

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

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

f_{ij} is the probability that $\boldsymbol{j}$ will be visited by a MC starting in $\boldsymbol{i}$ <u>in a finite time</u>.

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.

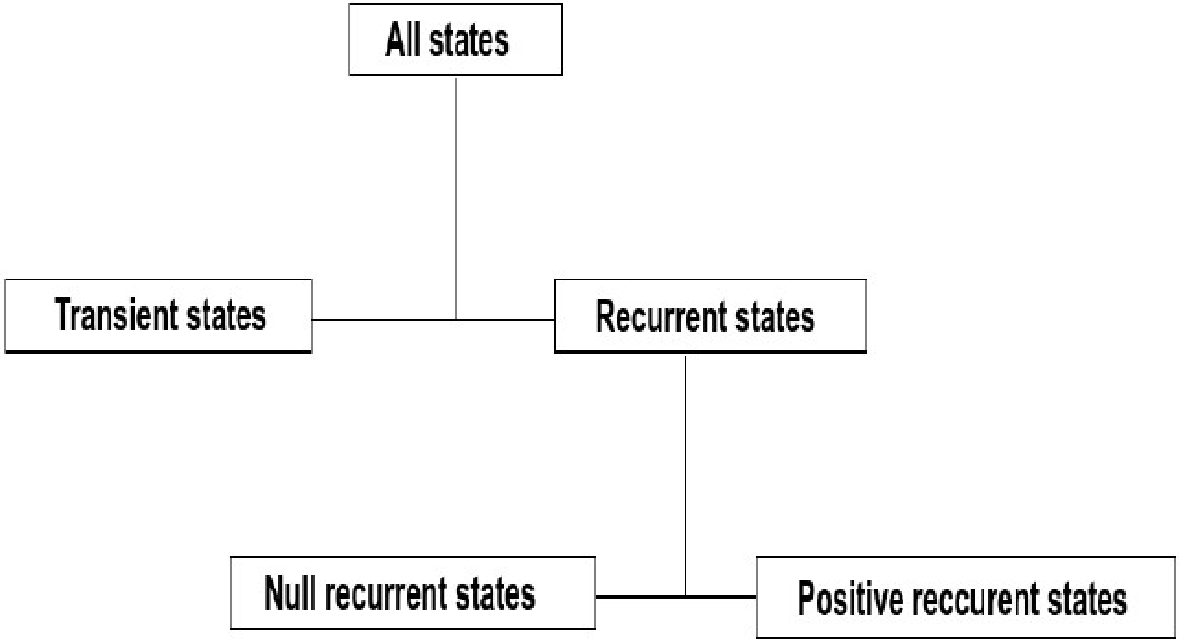


Figure 3.8: Classification of the states according to their asymptotic properties

Transiency and recurrence are class properties.

All elements of an irreducible MC are either transient or recurrent.

i.o. : abbreviation for **infinitely often**.

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

$$\boxed{p_i(T_i < \infty) = f_{ii} = 1}$$

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

$$\boxed{p_i(T_i < \infty) = f_{ii} < 1}$$

Logically enough, a recurrent state is visited an infinite number of times.

A transient state is only visited a finite number of times.

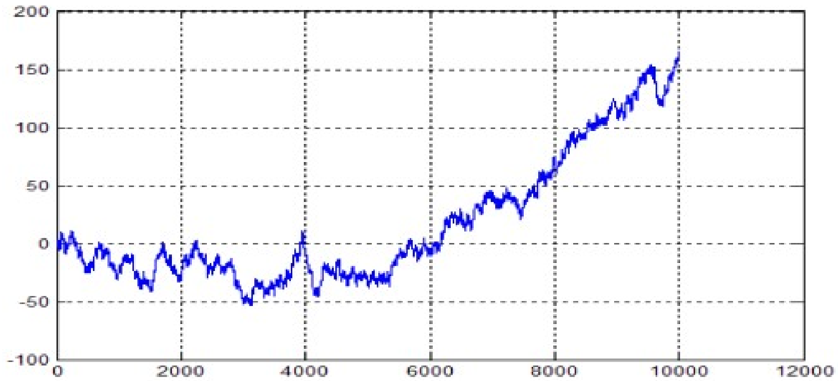
Example 3.11 (Random walk on $\mathbb{Z}$)

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

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



Is the random walk on ***Z*** recurrent or transient?

What do you **intuitively** think?

Through mathematical reasoning (see "**Markov_Chain_Lecture.pdf**") we can show that the Markov chain on ***Z*** is only recurrent if $p = q = ½$.

Otherwise, it is biased and hence transient.

For instance, if $p = 0.70$ the Markov Chain will tend to take on larger values as time goes by.

Likewise, if $p = 0.20$ the MC will progress towards $-\infty$.

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

Intuitively, this seems very plausible.

For all *(i,j)*, $\exists n \geq 0, p_{ij}^{(n)} > 0$ and the state space is finite.

Sooner or later, if the MC is in state ***i***, state ***j*** is bound to be visited.

Since state ***i*** will be revisited for the same reason, this will happen an endless number of times.

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.

2.5 Invariant distribution

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 have $0 < \alpha, \beta < 1$.

Let us consider the probability distribution

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

We can prove recurrently that

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

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

Since $\alpha + \beta - 1 < 1$, $\lim_{n \to \infty} p(X_n = 1) = \pi_1$ and $\lim_{n \to \infty} p(X_n = 2) = \pi_2$.

If the probability distribution of X_0 is (π_1, π_2), it NEVER CHANGES!

This is why it is called the **invariant distribution** of the Markov chain.

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

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

Now comes some harder stuff! :-)

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

The demonstration is not trivial (see “***Markov_Chain_Lecture.pdf***“).

Theorem 3.18 Let $X = (X_n)_{n=0,1,\ldots}$ 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

$$\pi_i = \frac{1}{m_i}, i \in I \text{ with } m_i = E_i[T_i].$$

Positive recurrence is thus directly related to the existence of a <u>unique</u> invariant distribution.

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

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

$\forall j \in I$**, the MC will only visit it a finite number of times.**

Obviously, I must be countably infinite as we saw above.

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

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j > 0 \text{ and } m_j = E_j[T_j] = \frac{1}{\pi_j}.$$

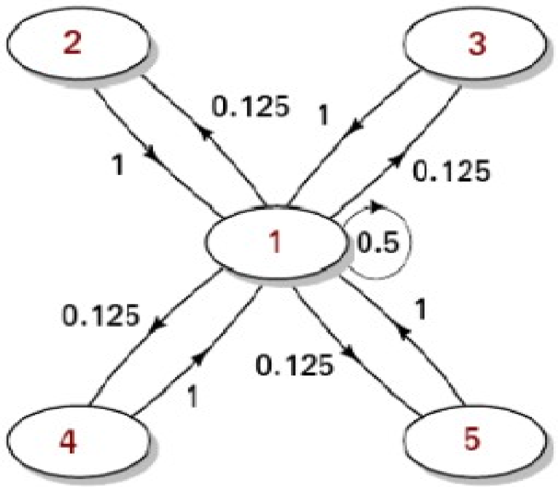
Theorem 3.21 (Markov chains on finite state spaces)

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:

$$\forall\ i, j \in I, \lim_{n\to\infty} p_{ij}^{(n)} = \pi_j > 0.$$

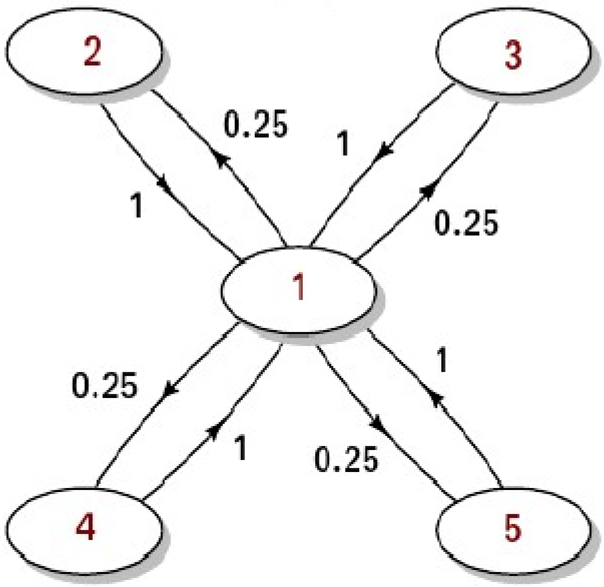
(B)



(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

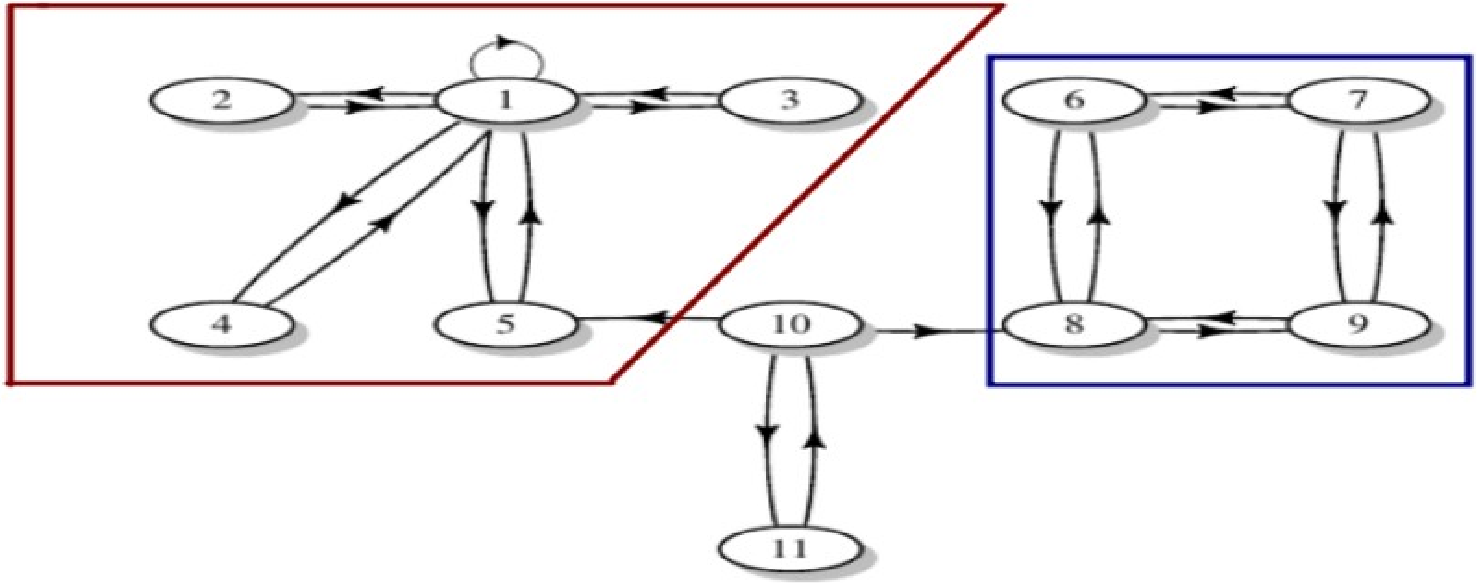
$$\lim_{m \to \infty} p(X_{N_m} = j | X_0 = i) = \pi_j, \forall i, j \in I.$$

(A)



(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.



A MC starting in **{1,2,3,4,5}** or **{6,7,8,9}** is irreducible.

2.6 Time reversal and reversible Markov chains

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 $\quad \widehat{p}_{ji} = \dfrac{\pi_i}{\pi_j} p_{ij}$ for all i, j.

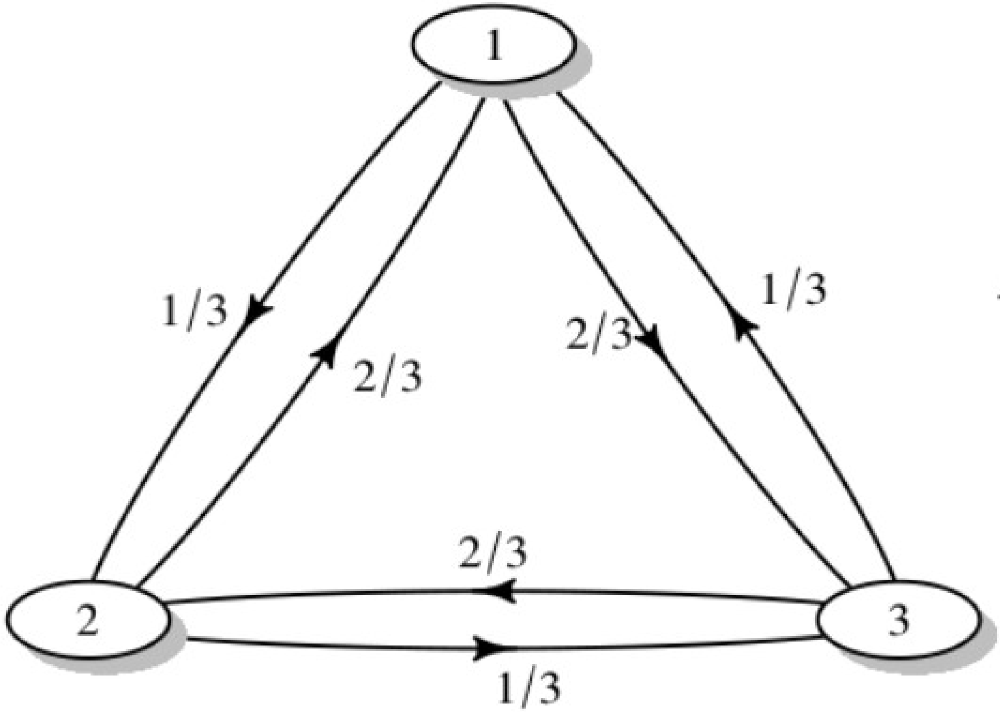
What is more, $\widehat{P}$ is irreducible and π is its invariant distribution.

Definition 3.12 A measure μ is called *reversible* with respect to a transition matrix P if the detailed balance conditions are satisfied:,

$$\forall i, j \in I, \mu_i p_{ij} = \mu_j p_{ji} \tag{3.17}$$

An irreducible Markov chain is reversible if and only if $\pi_i p_{ij} = \pi_j p_{ji}$.

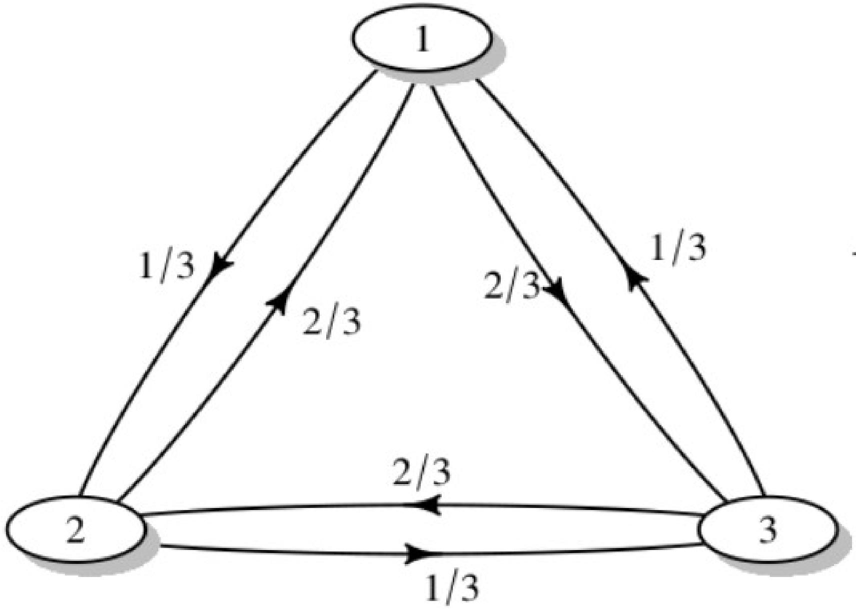
Example 3.16 We consider the Markov chain shown in Figure 3.10.
The transition matrix is the irreducible doubly stochastic matrix



$$P = \begin{pmatrix} 0 & 1/3 & 2/3 \\ 2/3 & 0 & 1/3 \\ 1/3 & 2/3 & 0 \end{pmatrix}$$

Intuitively, is the Markov chain reversible?

Example 3.16 We consider the Markov chain shown in Figure 3.10. The transition matrix is the irreducible doubly stochastic matrix

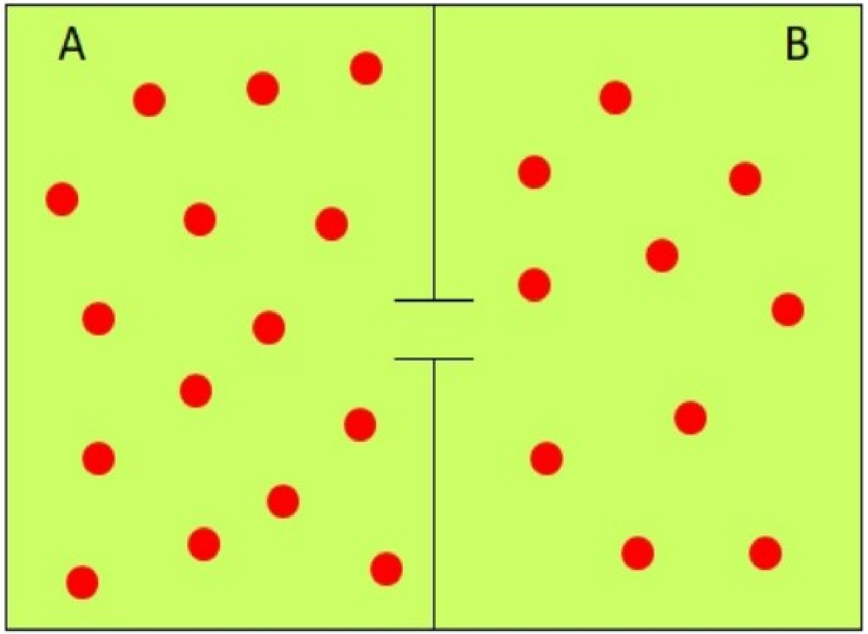


$$P = \begin{pmatrix} 0 & 1/3 & 2/3 \\ 2/3 & 0 & 1/3 \\ 1/3 & 2/3 & 0 \end{pmatrix}$$

Intuitively, is the Markov chain <u>reversible</u>?

If the MC were reversible, it would be as likely to go clockwise as to go counterclockwise. But that's clearly not the state: 2/3 = 2*1/3.

Example 3.17 (Model of Ehernfest)



We have two containers separated by a membrane.

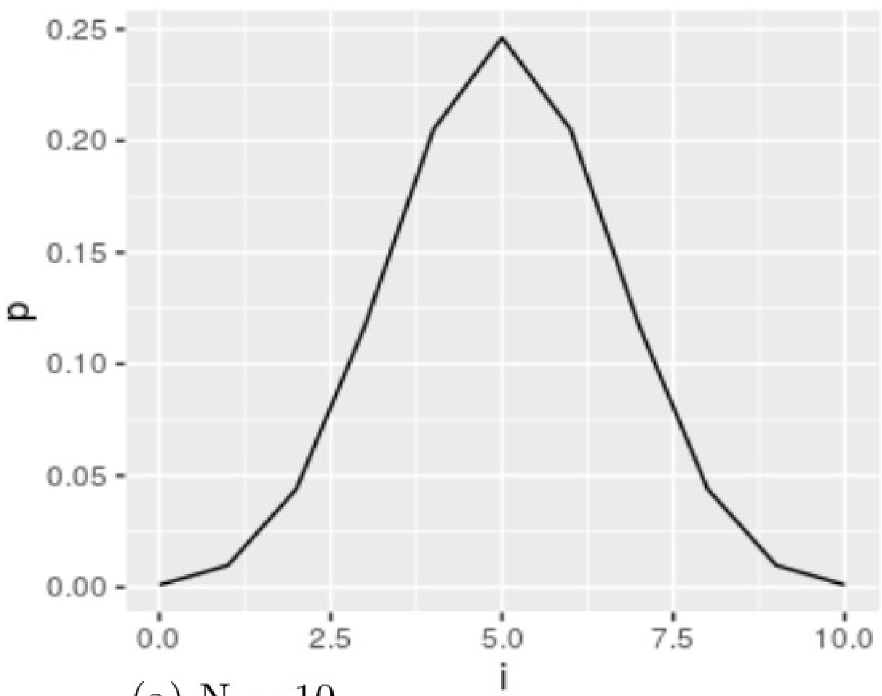
Let X_n be the number of molecules in the right container at time n.
The probability that one molecule goes to the other container is **uniform**.

Let *N* be the total number of molecules in the two containers.

$$\forall n > 0,\ p(X_{n+1} = i+1 | X_n = i) = \frac{N-i}{N},\ p(X_{n+1} = i-1 | X_n = i) = \frac{i}{N}$$

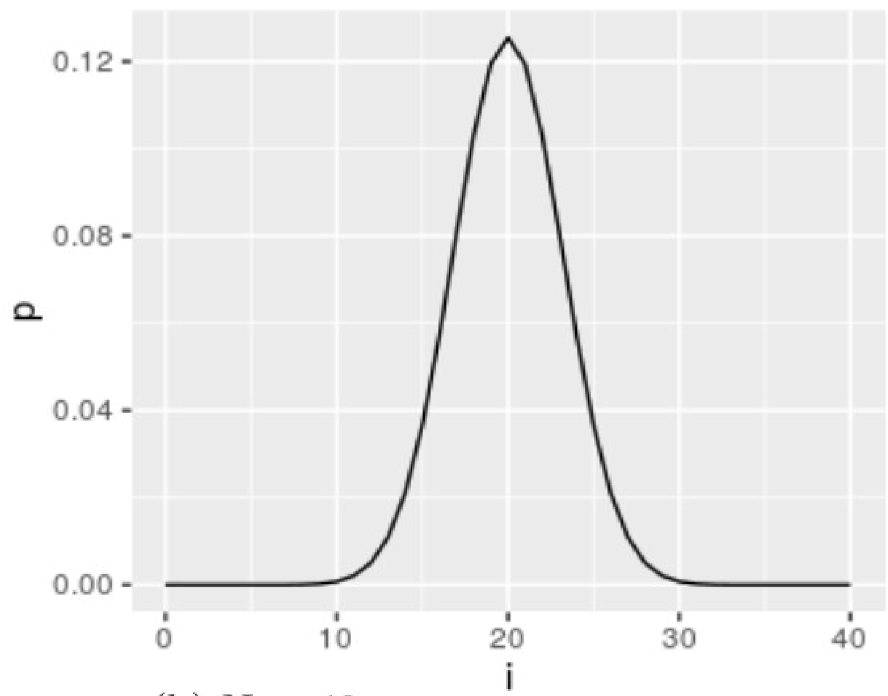
π defined by $\pi_j = 2^{-N}\binom{N}{j}$ is the **invariant and irreversible** distribution.

Let's plot the equilibrium distribution for different values of *N*.



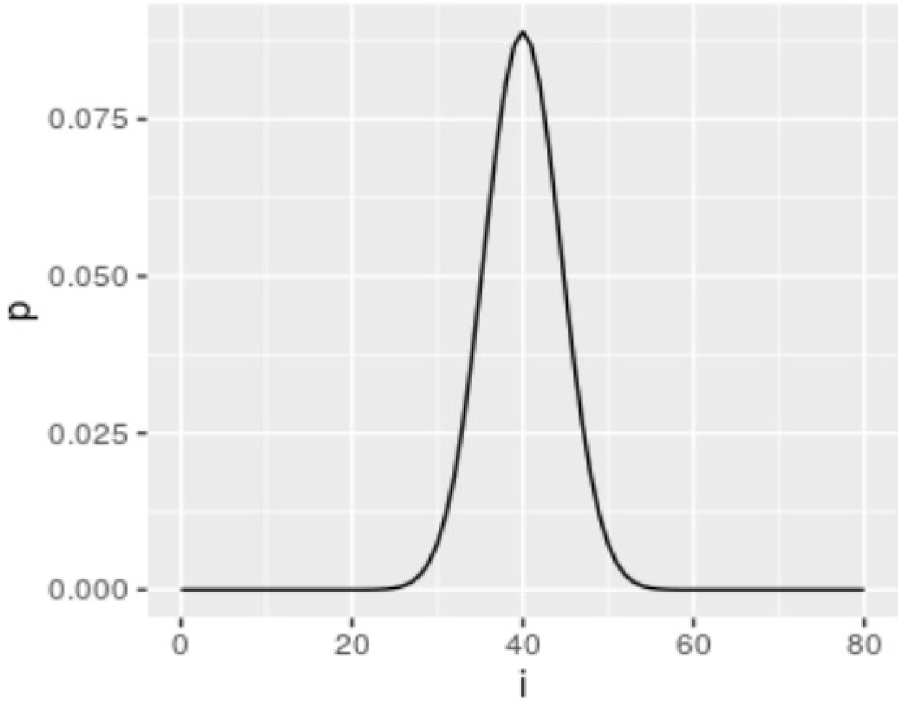
0.25
0.20
0.15
0.10
0.05
0.00
p
0.0
2.5
5.0
7.5
10.0
i

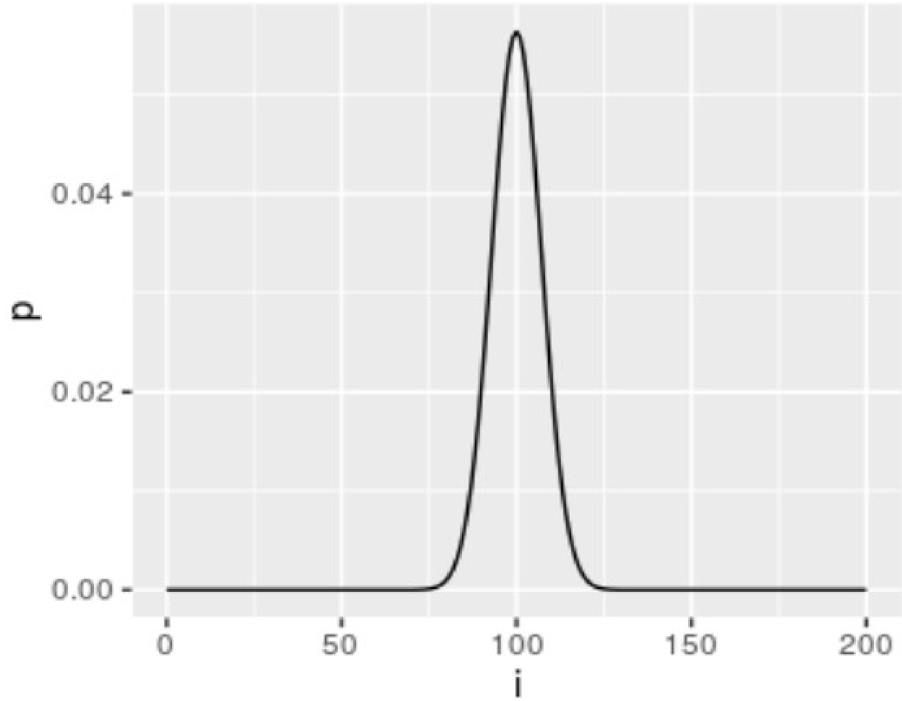
(a) N = 10

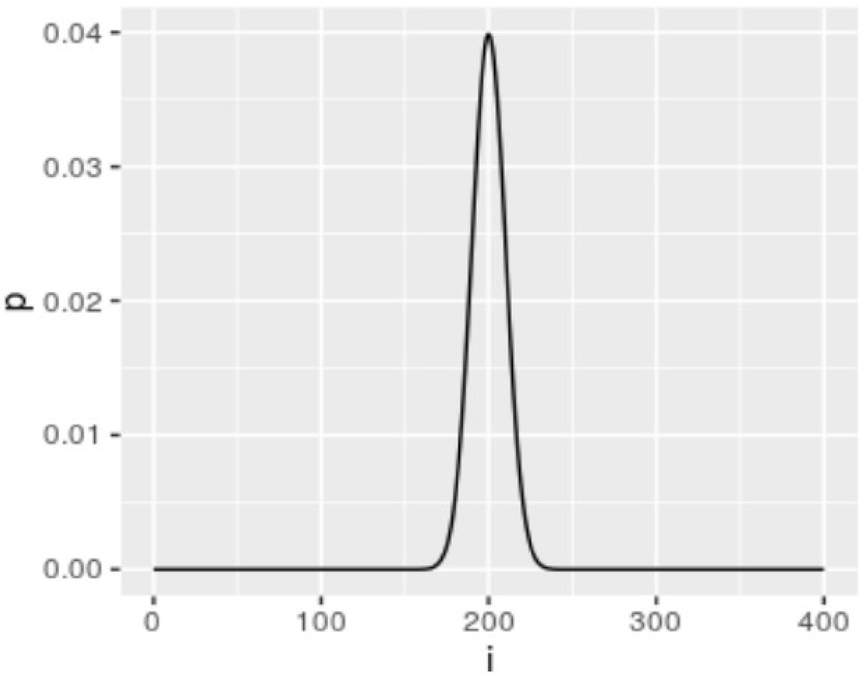


0.12
0.08
0.04
0.00
p
0
10
20
30
40
i

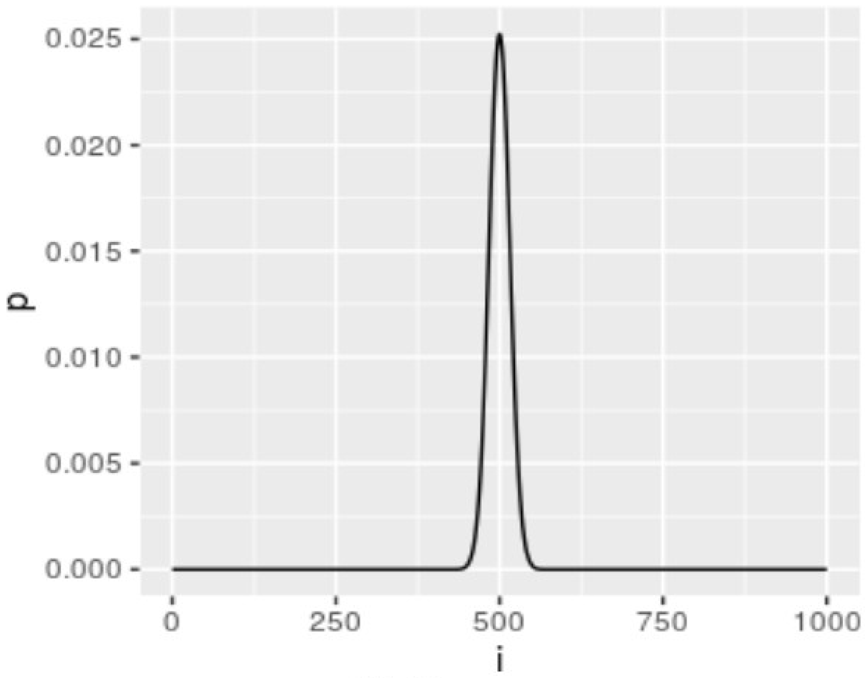
(b) N = 40


0.075
0.050
0.025
0.000
p
0
20
40
60
80
i
(c) N = 80


0.04
0.02
0.00
p
0
50
100
150
200
i
(d) N = 200



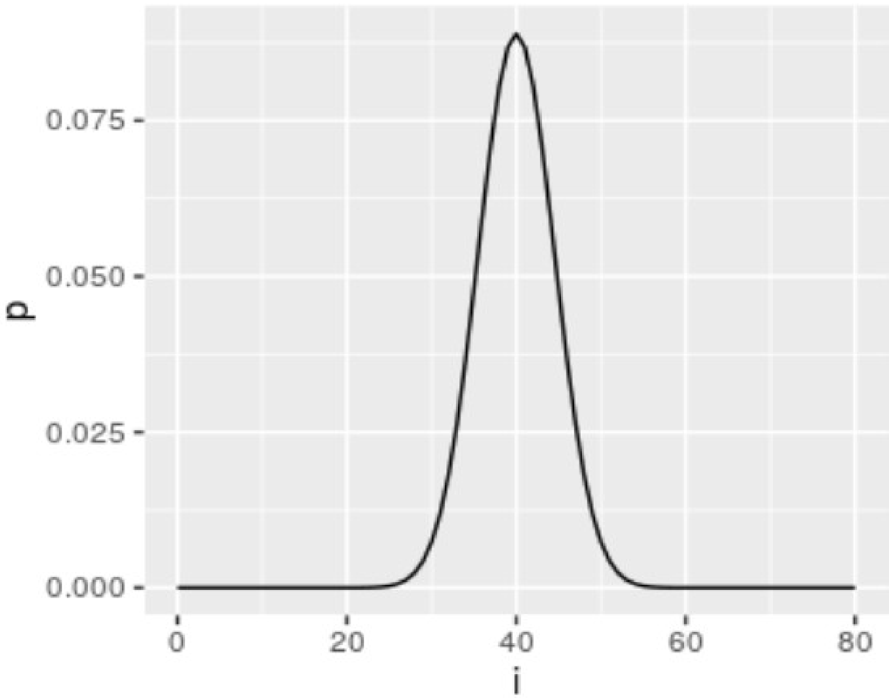
(e) N = 400



(f) N = 1000

For ***N* = 1E+06**, we would have a peak-shaped distribution: the number of molecules in both containers is roughly the same.

Let us reconsider



(c) $N = 80$

There are two ways to interpret that probability distribution:

Ensemble interpretation: if we consider a very large number M of systems, in roughly 8% of them there would be $n = 40$ particles in the right container.

Time-averaged interpretation: If we consider <u>only one system</u>, in **8% of the time** we would have $n = 40$ particles in the right container.

2.7 Convergence of Markov chains on finite state

Total variation distance:

$$\|\mu - \nu\|_{TV} = max_{A \subset I} |\mu_A - \nu_A| \ = \frac{1}{2} \Sigma_{i \in I} |\mu_i - \nu_i|$$

Distance to the invariant distribution:

$d(n) := max_{i \in I} \| P_{i,.}^{(n)} - \pi \|_{TV}$, *i* being the initial state.

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

$$d(n) := max_{i \in I} \|P_{i,.}^{(n)} - \pi\|_{TV} \leq C\alpha^n.$$

2.8 Ergodic theorem

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

$$\begin{aligned}&\forall\epsilon > 0, \forall\epsilon' > 0, \exists n_1, n \geq n_1\\&\Rightarrow p(|\overline{X_n}| > \epsilon) < \epsilon'.\end{aligned}$$

.

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 $(X_n)_{n\in\mathbb{N}}$,

$$\forall \epsilon > 0, \exists N_1, n \geq \boxed{N_1} \text{ random integer}$$
$$\Rightarrow |\overline{X_n}| < \epsilon.$$

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:

$$V_i(n) := \sum_{k=0}^{n-1} 1_{\{X_k=i\}} \leq n.$$

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

$$p\Big(\frac{V_i(n)}{n} \overset{n+\infty}{\rightarrow} \frac{1}{m_i}\Big) = 1. \tag{3.19}$$

If the MC is positive recurrent, we have for any bound function $f : I \to \mathbb{R}$

$$p\Big(\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \overset{n+\infty}{\rightarrow} < f, \pi > \Big) = 1, \tag{3.20}$$

where π is the unambiguous invariant distribution of the Markov chain (with $\pi_i = \frac{1}{m_i}$) and

$$< f, \pi > = \sum_{i \in I} f(i)\pi_i = E_\pi(f).$$

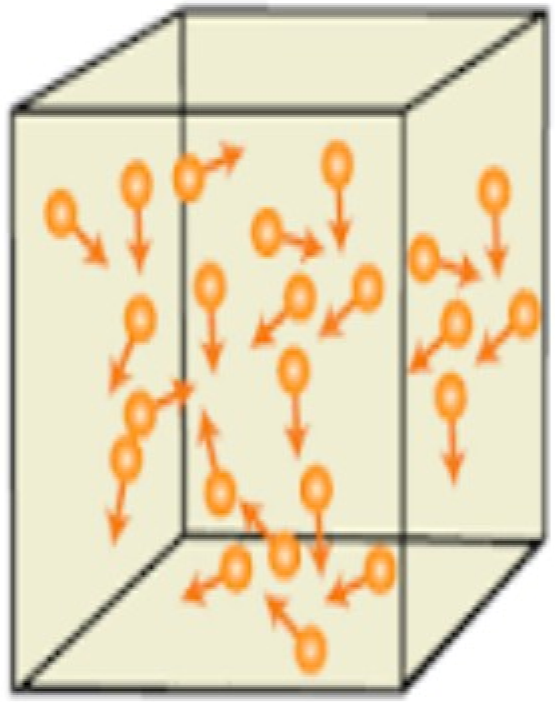
Example 3.21 (Dice rolling)

We throw 10,000 dices at the same time (**Ensemble interpretation**).

We throw 10,000 times the same die (**Time-averaged interpretation**).

In both cases, if the die is unbiased $f(1) \approx f(2) \approx f(3) \approx f(4) \approx f(5) \approx f(6) \approx 1/6$.

Brownian motion



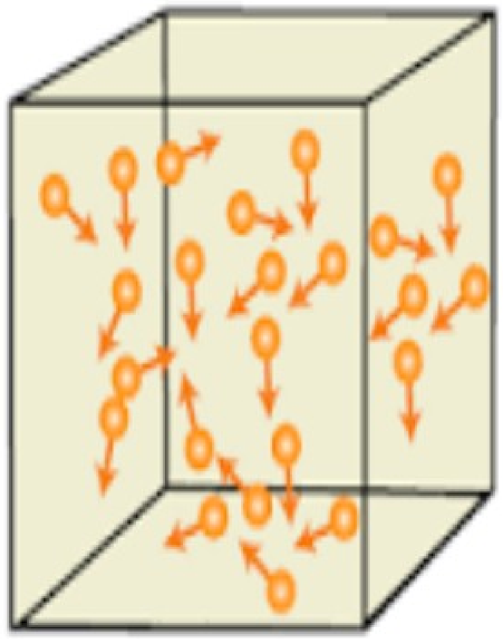
A very large number of particles move around in a container.

We want to determine the population densities in the different regions.

The system is in steady state.

Ensemble interpretation: we count the number of particles in every region $\rightarrow$ very expensive and tedious.

Brownian motion



A very large number of particles move around in a container.

We want to determine the population densities in the different regions.

The system is in steady state.

Time-averaged interpretation: we follow a single particle for a very large period of time and determine how often it is in a given region

→ **ergodic system!**

3 Continuous Markov chains in discrete time

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 $\forall x \in S$

$$\|P^n(x, A) - \pi(A)\|_{TV} \overset{n+\infty}{\rightarrow} 0.$$

We recall that the total variation norm is defined as

$$\|\pi_1 - \pi_2\|_{TV} = sup_A|\pi_1(A) - \pi_2(A)|$$

4 Markov-chain Monte-Carlo method

In the first lecture, we studied the chemical reaction $\mathrm{R} + \mathrm{R} \to \mathrm{P}_2$ with $r = k[R]^2$ with $k \in S = [5E+04; 5E+09]$.

The posterior density is $$f(k|data) = \frac{L(data|k) f_0(k)}{\int_{k \in S} L(data|k) f_0(k) dk}$$

with $$p(data) = \int_{k \in S} L(data|k) f_0(k) dk$$

We can directly compute the integral deterministically or through a Monte-Carlo simulation:

$$\int_{k\in S} L(data|k)f_0(k)dk = E_k(L(data|k)) \approx \frac{1}{n}\sum_{i=1}^{n} L(data|K_i)$$

where $K_i, i \in (1;n)$ is a random variable taking values in S and whose probability density is given by $f_0(k)$.

Let's now consider $k = (k_1, k_2, k_3, k_4, k_5, k_6) \in S$.

We still have $f(k|data) = \dfrac{L(data|k)f_0(k)}{\int_{k\in S} L(data|k)f_0(k)dk}$.

This time however,

$$p(data) = \int_{k_1,k_2,k_3,k_4,k_5,k_6 \in S} L(data|k_1,k_2,k_3,k_4,k_5,k_6) f_0(k_1)$$
$$f_0(k_2) f_0(k_3) f_0(k_4) f_0(k_5) f_0(k_6) dk_1 dk_2 dk_3 dk_4 dk_5 dk_6$$

Integrating *p(data)* is much harder.

With 100 parameters or more, it becomes nearly impossible.

But we have no problem computing $\dfrac{f(k_1|data)}{f(k_2|data)} = \dfrac{L(data|k_1) f_0(k_1)}{L(data|k_2) f_0(k_2)}$.

4.1 Metropolis-Hasting algorithm

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 $g(x) = L(data|x)f_0(x)$.

We must define the state space such that ***g(x) > 0*** everywhere.

The Metropolis-Hasting algorithm goes as follows:

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 $\phi(\delta_{a,\phi}) = \phi > 0$.

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 $X_0 \in \delta_{x_0,\phi}$.

Mathematically, we have

$$q(y|x_t) = p(y \in \delta_{y,\phi}|x_t \in \delta_{x_t,\phi})$$

- *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.

We must keep in mind that $\frac{g(y)}{g(x_t)} = \frac{f(y|data)}{f(x_t|data)} = \frac{L(data|y)f_0(y)}{L(data|x_t)f_0(x_t)}$

is **the posterior ratio** that can be easily computed.

- 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.

Theorem 5.1 Any Markov chain produced in this way converges towards its invariant distribution $f(x|data)$.

4.2 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

$$\gamma(t_1, t_2) = \mathrm{Cov}(X_{t_1}, X_{t_2}) = \mathrm{E}[(X_{t_1} - \mu_{t_1})(X_{t_2} - \mu_{t_2})].$$

If all variables are independent (e.g. coin tossing), $\gamma(t_1, t_2) = 0$.

Normally, $\gamma(t_1, t_2) \neq 0$.

Autocorrelation is problematic because it prevents us from exploring the state space.

If the proposal distribution is too focused (its variance is too small), the auto-correlation will be high
→ only a small part of the parameter space will be explored.

If the variance of the proposal distribution is too large, most moves will be rejected.

A **compromise** needs to be made.

4.3 Independence chains

Independence chains are characterised by $q(y|x_t) = q(y)$.

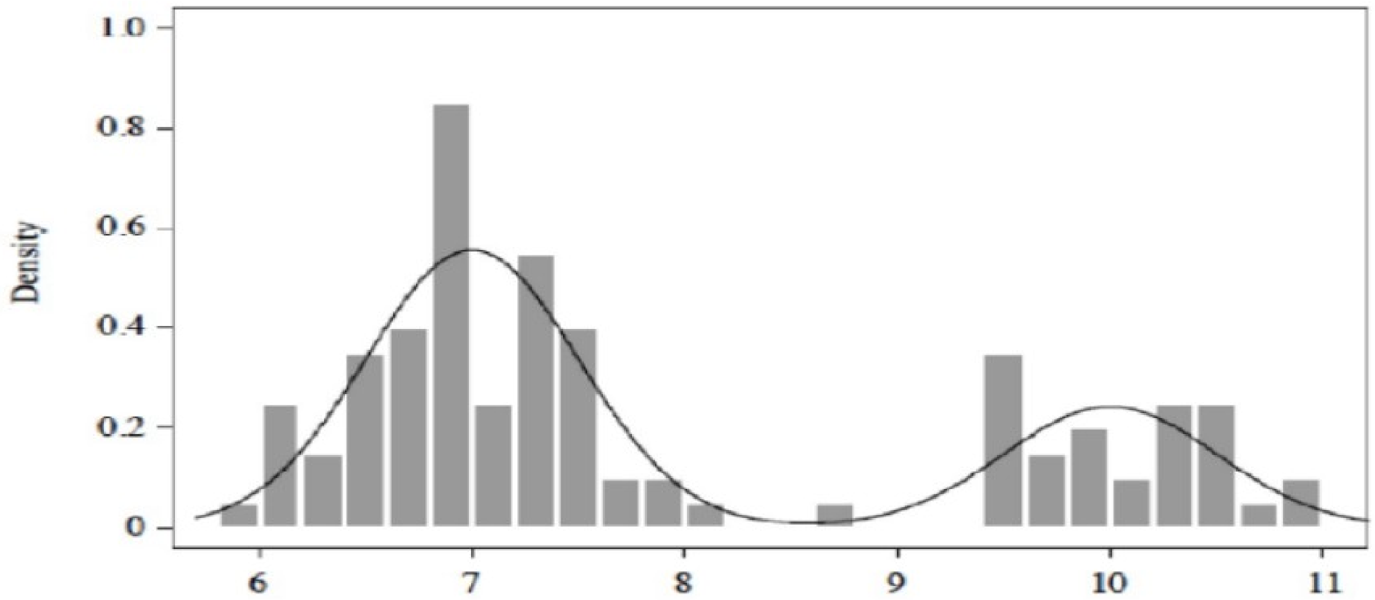
Therefore, x_{t+1} is completely independent of x_t.

Acceptance ratio $\alpha(x_t, y) = min\Big(1, \dfrac{L(data|y)}{L(data|x_t)}\Big)$.

The likelier parameter values are automatically accepted.

Example 5.1 (Mixture distribution)

$$f(y) = \delta N(7, 0.5^2) + (1-\delta)N(10, 0.5^2)$$



Y is the observed random variable.

δ is the unknown parameter. We want to determine its posterior density.

We decide that the prior and proposal distribution will be <u>the same</u>.

$$Beta(p,q,x) = \frac{1}{B(p,q)} x^{p-1}(1-x)^{q-1}$$ with the normalisation constant

$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} = \int_0^1 u^{p-1}(1-u)^{q-1}du$$.

We decide to use two beta distributions to generate two Markov chains:
$f_1(\delta) = Beta(1,1)$ and $f_2(\delta) = Beta(2,10)$.

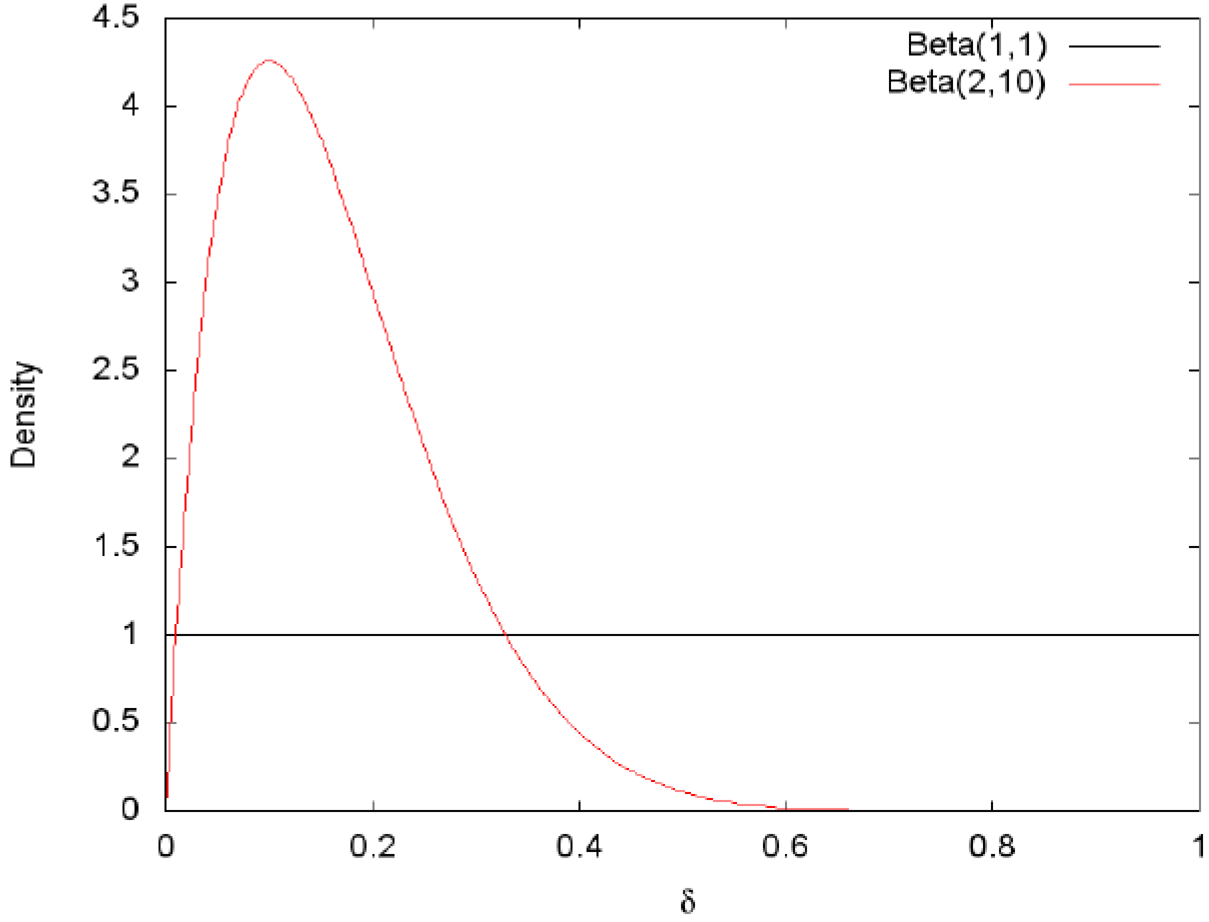
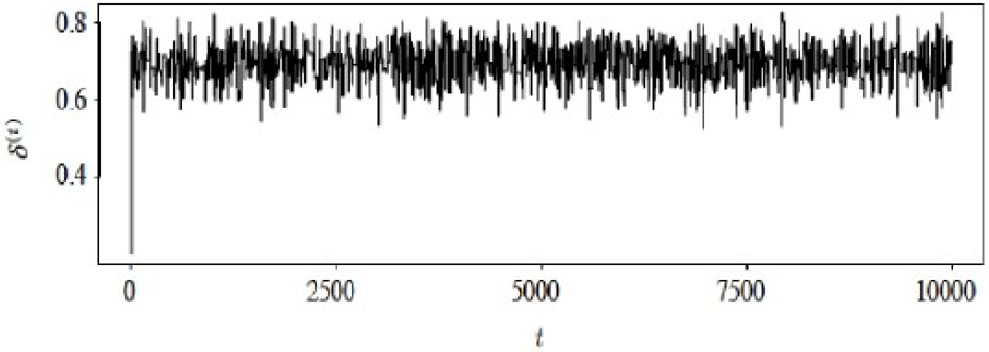


Figure 5.2: Prior / Proposal distributions

A **path** is a sequence of realisations of the Markov chain.

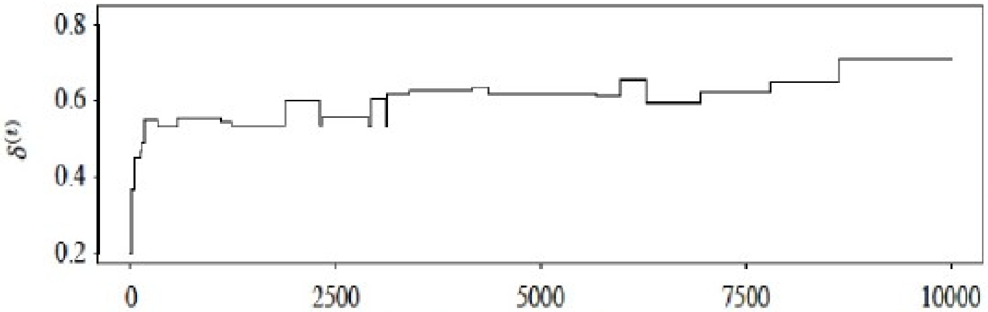
$$f_1(\delta) = Beta(1,1)$$



Strong oscillations

Good convergence

$$f_2(\delta) = Beta(2,10)$$



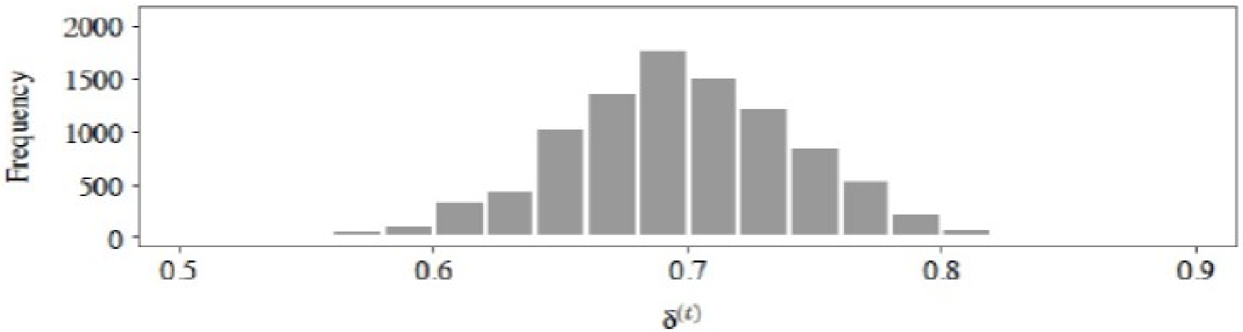
Weak oscillations

Bad convergence

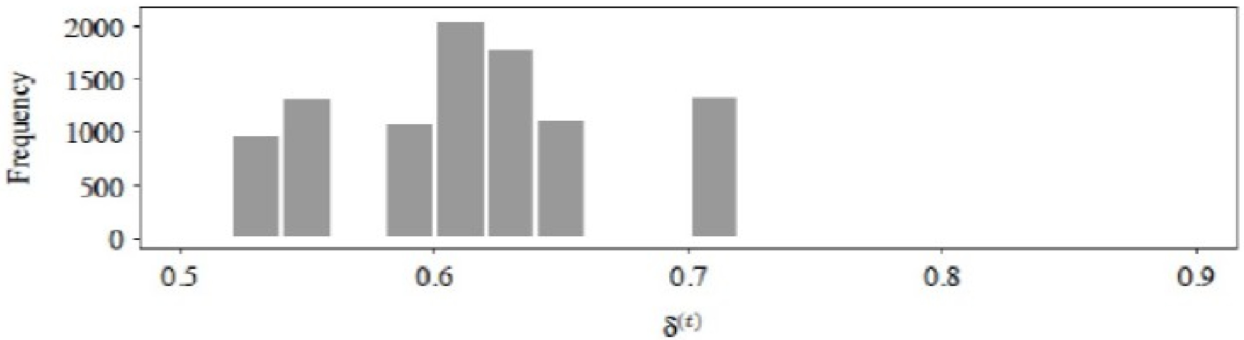
10,000 iterations

This results in two histograms:

$$f_1(\delta) = Beta(1,1)$$



$$f_2(\delta) = Beta(2,10)$$



4.4 Random walk chains

y is generated by first choosing the symmetrical function $\epsilon \sim h(\epsilon)$.

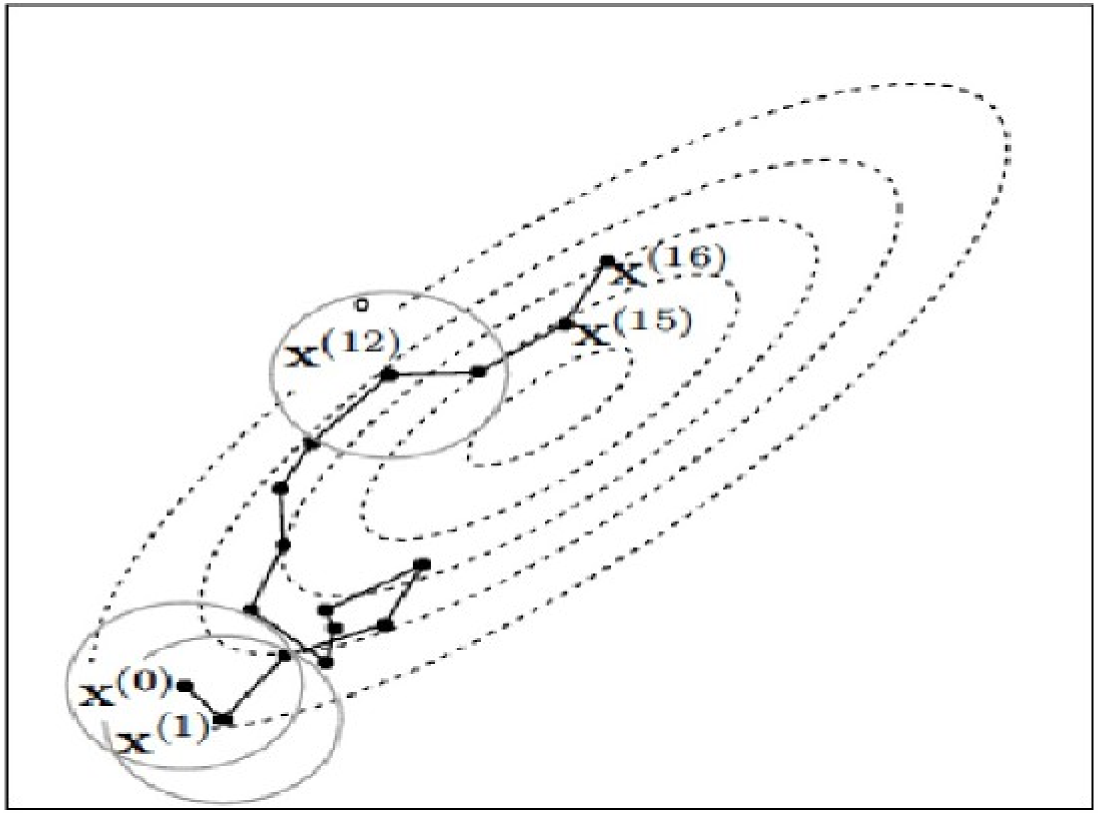
h might be e.g. a uniform or Gaussian distribution centred at x_t.

We then set $y = x_t + \epsilon$: $q(y|x_t) = h(y - x_t)$.

The acceptance rate is given by

$$\alpha(x_t, y) = min\left(1, \frac{f(y|data)}{f(x_t|data)}\right) = min\left(1, \frac{g(y)}{g(x_t)}\right) = min\left(1, \frac{L(data|y) f_0(y)}{L(data|x_t) f_0(x_t)}\right) \ .$$

Here is an example of a random walk:



Example

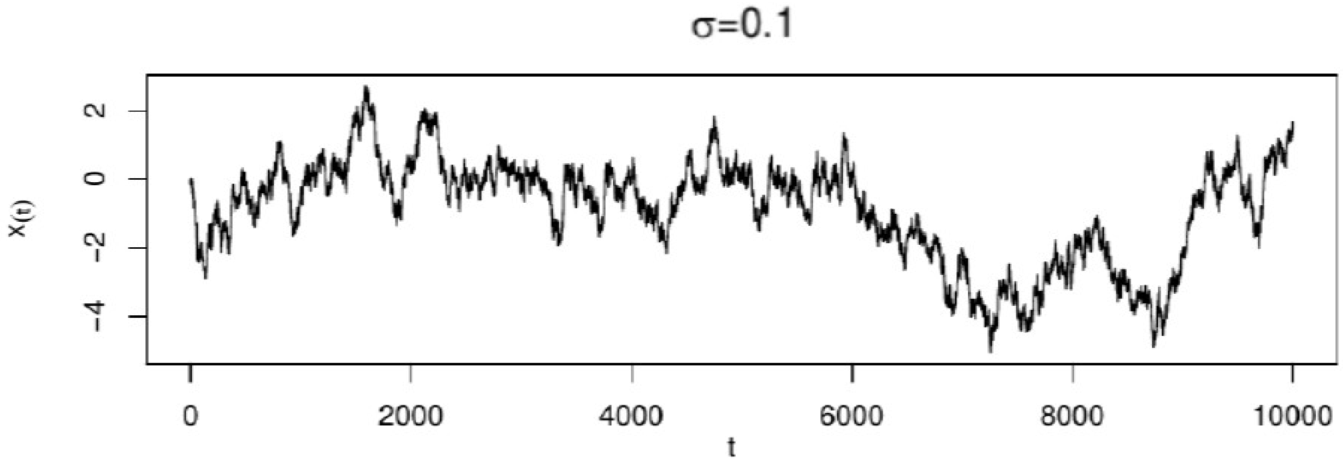
Let us suppose that the posterior probability distribution we want to approximate is the Laplace distribution:

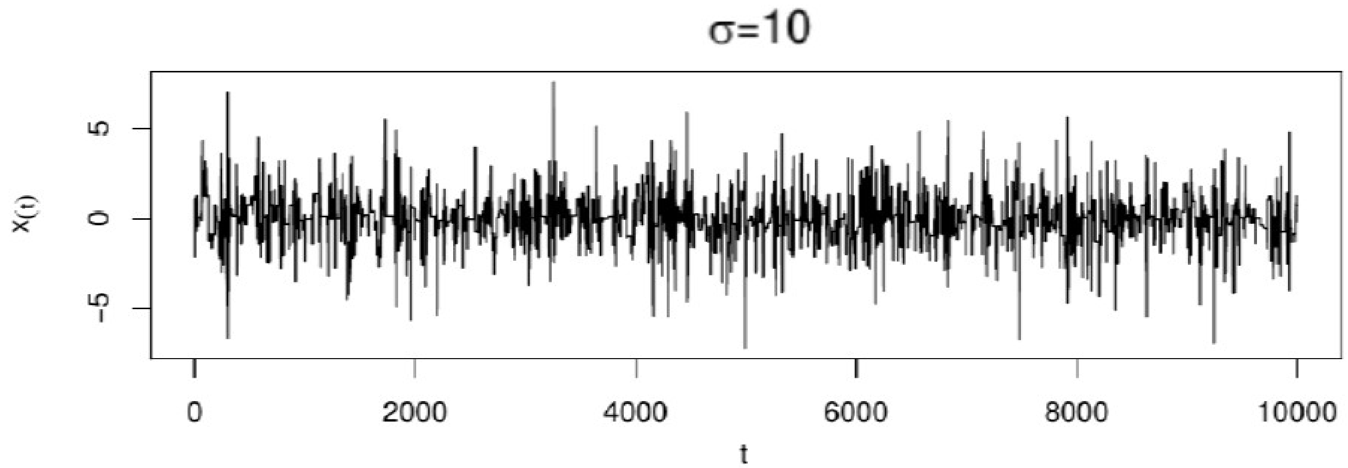
$$f(x|data) = f(x) = \frac{1}{2}e^{-|x|}, -\infty < x < +\infty$$

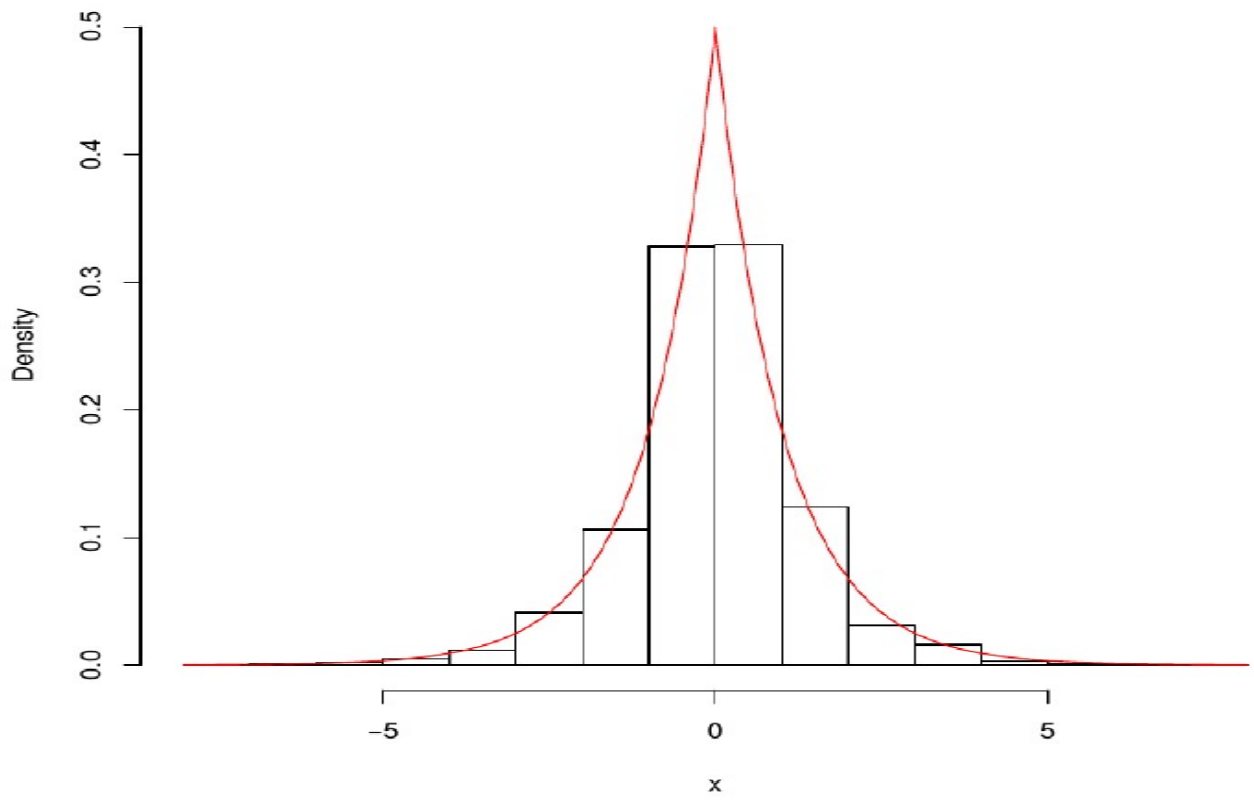
The random walk is defined by $\epsilon \sim N(0, \sigma^2)$ and $y = x_t + \epsilon$.

We consider two MC with $\sigma = 0.1$ and $\sigma = 10$.

As we can see on the next page, $\sigma = 0.1$ leads to a much stronger auto-correlation.


σ=0.1
2
0
−2
−4
x(t)
0
2000
4000
6000
8000
10000
t


σ=10
5
0
−5
x(t)
0
2000
4000
6000
8000
10000
t



$$\sigma = 10$$

4.5 Basic Gibbs sampler

We consider a multidimensional random variable $X = (X_1, ..., X_p)^T$.

Let's define $X_{-i} = (X_1, ..., X_{i-1}, X_{i+1}, ..., X_p)^T$, i.e. X without X_i.

The algorithm looks as follows:

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 very useful for **hierarchical models** where one variable only depends on few other ones (e.g. *belief networks*).

As always,

$$f(x|data) = \frac{f_0(x)L(data|x)}{c},$$

where $c = p(data) = \int_{x\in S} f_0(x)L(data|x)dx$.

In general, we have

$$x_i^{t+1}|(x_{-i}^t, data) \sim f(x_i|x_{-i}^{(t)}, data),$$

whereby

$$x_{-i}^t = (x_1^{(t+1)}, ..., x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, ..., x_p^{(t)}).$$

Example

Let us consider the random variables $Y_1, ..., Y_n \sim N(\mu, h^{-1})$.

μ (mean) and h (precision) are parameters and Bayesian random variables.

$h \to +\infty$ (maximum precision) means that every $Y_i \approx \mu$.

$h \approx 0$ (minimum precision) means that the Y_i can take on almost every value.

We define two **prior distributions**: $\mu \sim N(\mu_0, h_0^{-1})$ and $h \sim G(\alpha_0/2, \delta_0/2)$
with $f(h) = G(h) \propto h^{\alpha_0/2-1} exp(-\delta_0 h/2)$.

We have $$f(h|\mu,y) \propto \frac{L(y|\mu,h)f(h)}{\int\limits_{h=0}^{+\infty} L(y|\mu,h)f(h)dh}$$

$$L(y|\mu,h)f(h) \propto \; h^{(\alpha_0+n)/2-1} exp\Big\{-\frac{h}{2}\Big(\delta_0 + \sum_{i=1}^{n}(y_i-\mu)^2\Big)\Big\}$$

$$f(h|\mu,y) \propto \frac{h^{(\alpha_0+n)/2-1} exp\Big\{-\frac{h}{2}\Big(\delta_0 + \sum_{i=1}^{n}(y_i-\mu)^2\Big)\Big\}}{\int\limits_{h=0}^{+\infty} h^{(\alpha_0+n)/2-1} exp\Big\{-\frac{h}{2}\Big(\delta_0 + \sum_{i=1}^{n}(y_i-\mu)^2\Big)\Big\}dh}$$

$$f(\mu|h,y) = \frac{L(y|h,\mu)f(\mu)}{\int\limits_{\mu=0}^{+\infty} L(y|h,\mu)f(\mu)d\mu}$$

1) We set initial values μ^0, h^0.

2) $\mu^{t+1} \sim f(\mu|h^t, y)$ is generated.

3) $h^{t+1} \sim f(h|\mu^{t+1}, y)$ is generated.

4) $t = t+1$

5) Back to 2)

4.6 Single Component Metropolis-Hastings

Gibbs' sampling is a specific case of the single component Metropolis-Hastings.

Let's introduce $x = (x_1, x_2, ..., x_d)$, $x_{.-i} = (x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_d)$

$x_{t.-i} = (x_{t+1.1}, x_{t+1.2}, ..., x_{t+1.i-1}, x_{t.i+1}, ..., x_{t.d})$.

Let's define $f(x) = f(x|data)$. The 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 $q_i(y_i|x_{t.i}, x_{t.-i})$.

   (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

$$f(x_i|x_{.-i}) = \frac{f(x_i)}{\int_{x'_i \in S_i} f(x'_i, x_{.-i})dx'_i}.$$

The $single\ component\ Metropolis\text{-}Hastings$ converges towards the invariant distribution $f(x) = f(x|data)$.

4.7 Implementation of the MCMC

Several questions must be asked:

- 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?
- 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?

4.8 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?

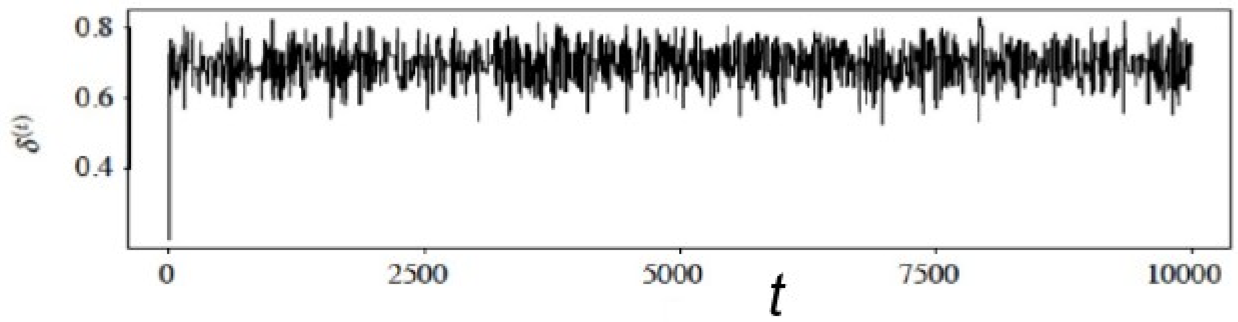
- Convergence: Has the Markov chain approximately reached its stationary distribution?

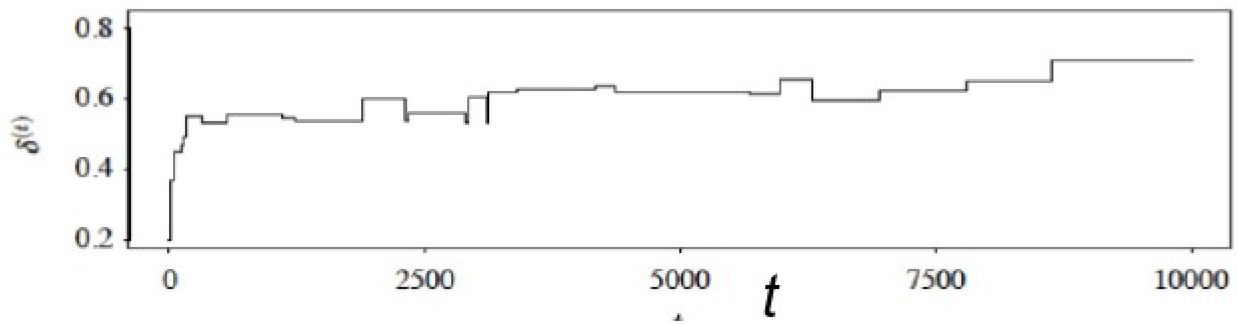
  These two questions and the methods and tools used to answer them are inextricably linked. It is always advisable to employ several verification techniques to increase our confidence in the outcome.

The two questions are inextricably linked.

We'll begin by examining two simple **graphical diagnostics**.

1) Autocorrelation plots

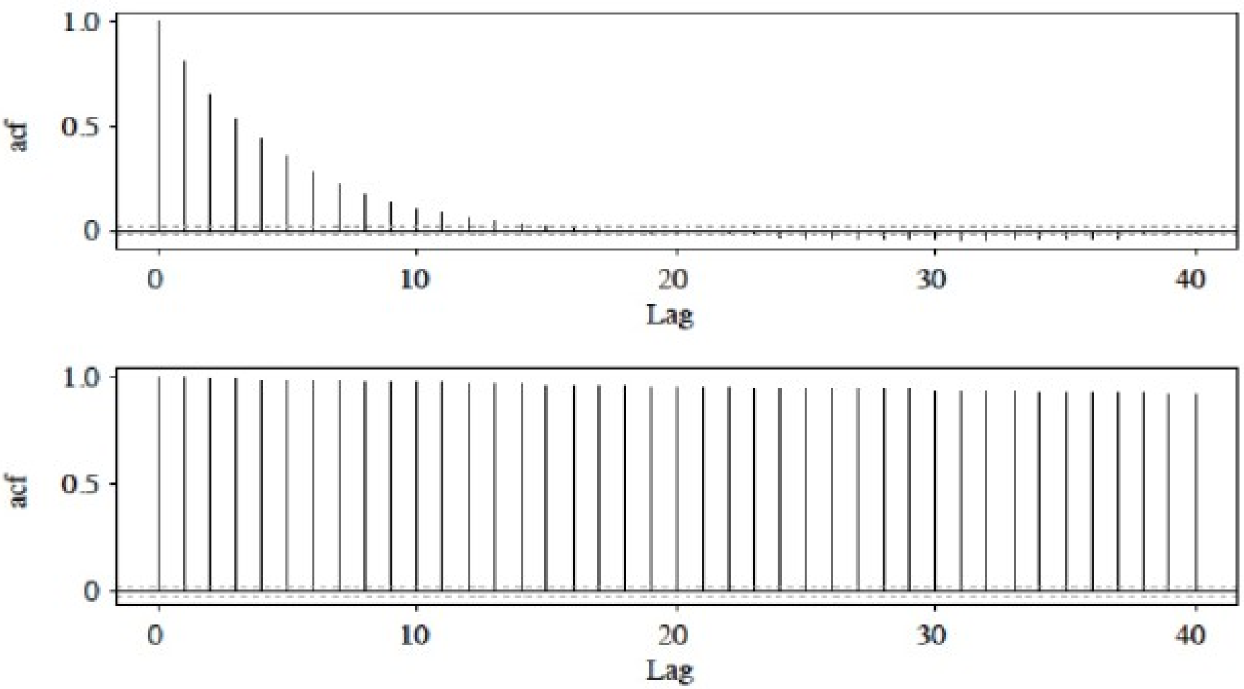




Beta(1, 1) (top) and Beta(2, 10) (bottom) for the mixed model (see above).

2) The autocorrelation plot

We have $lag = t_2 - t_1$.



Beta(1, 1) (top) and Beta(2, 10) (bottom) for the mixed model (see above).

For $\mathrm{Beta}(1,\ 1)$ (uniform distribution), the MC strongly oscillates and the autocorrelation quickly sinks → good convergence.

For $\mathrm{Beta}(2,\ 10)$, the MC very weakly oscillates and the autocorrelation remains constant → (very) bad convergence.

The choice of the proposal distribution Q is often vital.

It is important that *Q* resembles $f(x) = f(x|data)$.

Unfortunately, we seldom have any idea how $f(x|data)$ looks like.

Solution: iterative adaptation of the variance of Q

$\rightarrow$ goal: the acceptance rate oscillates between 25% and 50% (or some other values).

Multimodal invariant distributions are very problematic because the MC can be stuck in one mode.

The **autocorrelation time** is defined as

$\tau = 1 + 2\sum_{k=1}^{\infty} \rho(k)$ with $\rho(k) = \frac{cov(X(0),X(k))}{\sqrt{Var(X(0))Var(X(k))}}$.

The **effective sample size** of the MC is the size we'd need to get the same amount of information if the elements of the MC were independent.

It can be approximated by $L/\hat{\tau}$.

The **burn-in** period is the time the Markov chain needs to become independent of its starting value.

After this period, the MC is a good approximation of the posterior distribution.

We dismiss the D first values of the MC corresponding to that phase.

$100 < D < 10{,}000$ often (but not always) works.

4.9 Important estimators

Let $X \sim f(x|data)$ be the Markov chain, and *h* an arbitrary function.

We want to estimate the mean $\mu = E\{h(X)\}$.

If *L* is large enough, a good estimator is $\hat{\mu} = \frac{1}{L} \sum_{t=D}^{D+L-1} h(X(t))$.

This estimator is consistent even though the *X(t)* are correlated.

We now need the simulation standard error (sse) of $\hat{\mu}$ / μ .

If the elements of the MC were uncorrelated, we would have

$$\hat{sse}(\hat{\mu}) = \frac{\sqrt{\frac{1}{L-D-1}\sum_{t=D}^{L}(x_t - \overline{x})^2}}{\sqrt{L-D-1}}$$

Because of the correlations, $\hat{sse}(\hat{\mu})$ could be a strong underestimation.

Batch means method: we divide the MC into a series of approximately independent MC.

We first examine the autocorrelation of the MC until we find a lag k_0 such that $\rho(\hat{k}_0)$ is small enough, e.g. $\rho(\hat{k}_0) \leq 0.05$.

We then divide the L observations into $B \approx L/k_0$ batches.

The sample variance of the means is then given by

$$S^2 = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\mu_b} - \hat{\mu})^2$$

The estimated standard error is $s\hat{s}e(\hat{\mu}) = \sqrt{S^2/B}$.