UNIVERSITÀ DI PADOVA FACOLTÀ DI INGEGNERIA

Markov chains and Schrödinger bridges

Laureando DE BATTISTI RICCARDO

Relatore PAVON MICHELE

Corso di laurea Triennale in INGEGNERIA INFORMATICA

Anno Accademico 2009-2010

Autorizzo consultazione e prestito tesi

Firma

Introduction

The study of sequences of dependent random variables arose at the beginning of the twentieth century. In 1906 the Russian mathematician Andrei Andreyevich Markov (1856-1922), a Chebyshev's pupil, introduced some mathematical models to this end. His focus was where the present is a sufficient statistis of the past to predict the future. These sequence have been named *Markov chains*.

Even if it was a significant step in probability theory history, these models were not immediately considered by the scientific community. They were really appreciated only a few years later. Indeed, the study of Markov chains hugely spread only from the 1950s on. Nowadays, on the other hand, they are utilized in a variety of applications ranging from biology to psychology, from genetics to electrical engineering.

It is interesting to study how such models evolve over time and if they converge to a stationary situation, namely there is a limiting probability distribution. The convergence of a Markov chain, however is not always guaranteed, and it is not known a priori how much time takes to converge. These facts make the use of a Markov chain model more complex than its relatively simple theory.

Suppose we only deal with Markov chains whose convergence is guaranteed. In many applications, it is desirable to *control* the Markov chain by changing its transition mechanisms so as to achieve minimum cost, minimum queue length, etc. Another goal may be to drive the chain to a desired distribution at a given final time. This is achieve by the theory of *Schrödinger bridges*.

The purpose of this thesis is to describe the recently developed theory of Schrödinger bridges for Markov chains, and to investigate its effectiveness by simulation on various examples.

Of particular interest to us are chains that converges slowly to the equilibrium distribution such as those that arise from random geometric graphs.

Schrödinger bridges allow in principle the possibility of controlling a chain to its invariant distribution in finite time.

The outline of this thesis is as follows.

In Chapters $1-3$, we collect some basics material on *probability*, *combinatorics* and *random variables*;

In Chapters $4-7$, we introduce *Markov chains,* their properties and classificate them. We then give some examples distinguishing a Markov chain according its state space, namely finite or countable. Finally, we analyze the most important examples previously given.

In Chapters $8-9$, first we deal with general *maximum entropy problems*, then we focus on the theory of *Schrödinger bridges*.

In Chapters $10-11$, after introducing the *average consensus problem*, we discuss the importance of *random geoemtric graphs.* Finally we explain the algorithm of simulation for Schrödinger bridges giving a *time analysis* of its execution.

Contents

Chapter 1

Probability

1.1 Sample space and events

Let us consider experiments whose outcome cannot be predicted for certain. Although the outcome of the experiment is not known a priori, we suppose that the set of all the possible outcomes is finite and denote its cardinality by n.

Definition 1.1.1. The set of all the possible outcomes Ω is called *sample space*, and denoted with $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$.

Every subset E of the sample space, namely $\mathcal{P}(\Omega)$, is called an *event*. If the outcome of an experiment is contained in E, we say that E occurred. The event $\varnothing$ *(empty set)* is called impossible event as it never occurs. The event Ω is called certain as it always occurs.

Definition 1.1.2. Given two events E and F of the sample space Ω, we define $E \cup F$ as the *union* of these events, that means the set of all the possible outcomes which belong to E, F or both of them. On the other hand $E \cap F$, called *intersection* of E and F, represents all the outcomes contained both in E and F.

If $E \cap F = \varnothing$, E and F are *incompatible*. Given two events E and F, if all the outcomes of E are also of F, we say that E is contained in F ($E \subseteq F$) or with the same meaning F contains E ($F \supseteq E$).

Remark 1.1.3. If it is true that $E \subseteq F$ and $F \subseteq E$ at the same time, we say E and F coincide and write $E = F$.

Finally, every event E has its *complement* E^c which contains all the outcomes of the sample space that E does not.

Proposition 1.1.4. Operations above introduced, such as union, intersection and complementation, enjoy some properties:

1. Commutative
   (a) $E \cup F = F \cup E$;
   (b) $E \cap F = F \cap E$;
2. Associative
   (a) $(E \cup F) \cup G = E \cup (F \cup G)$;
   (b) $(E \cap F) \cap G = E \cap (F \cap G)$;
3. Distributive
   (a) $(E \cup F) \cap G = (E \cap G) \cup (F \cap G)$;
   (b) $(E \cap F) \cup G = (E \cup G) \cap (F \cup G)$.

There are also two relations, called *De Morgan's laws,* which connect union, intersection and complementation operation

$$\left(\bigcup_{i=1}^{n} E_i\right)^c = \bigcap_{i=1}^{n} E_i^c \tag{1.1}$$

$$\left(\bigcap_{i=1}^{n} E_i\right)^c = \bigcup_{i=1}^{n} E_i^c \tag{1.2}$$

1.2 Probability axioms

Let us consider a function

$$p : \Omega \to [0, 1], \tag{1.3}$$

such that

$$\sum_{i=1}^{n} p(\omega_i) = 1. \tag{1.4}$$

We call $p(\omega_i)$ the *probability* of outcome ω_i. For any event E its probability define

$$P(E) := \sum_{\omega_i \in E} p(\omega_i), \; P(\emptyset) = 0. \tag{1.5}$$

Such a map P is called a *probability measure* on $\mathcal{P}(\Omega)$. It must satisfy the following axioms:

1. $P(\Omega) = 1$;
2. $P\left(\bigcup_{i=1}^{m} E_i\right) = \sum_{i=1}^{m} P(E_i)$ whenever E_i, E_j are pairwise disjoint sets, namely $E_i \cap E_j \neq \emptyset$, for each couple (i, j) $i \neq j$.

Example 1.2.1. Let us toss a die. Supposing it is unbiased, each face is equiprobable. Hence we get $P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = \frac{1}{6}$. The probability of getting an even number is

$$P(\{2, 4, 6\}) = P(\{2\}) + P(\{4\}) + P(\{6\}) = \frac{1}{2}.$$

Due to our definition of probability of an event, we have some more properties:

1. E and E^c are always disjoint. Moreover $E \cup E^c = \Omega \Rightarrow P(E^c) = 1 - P(E)$;
2. $E \subset F \Rightarrow P(E) \leq P(F)$;
3. $P(E \cup F) = P(E) + P(F) - P(E \cap F) \leq P(E) + P(F)$.

Example 1.2.2. Now suppose to toss two distinguishable dice. Let E and F be the event of the first and second die respectively to get an even number. We get

$$P(E \cup F) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}.$$

1.3 Conditional probability and multiplication of probabilities

Conditional probability is the probability of some event E, given the occurrence of some other event F and is written $P(E|F)$.

Definition 1.3.1. For $P(F) > 0$, the conditional probability is defined by

$$P(E|F) := \frac{P(E \cap F)}{P(F)} \tag{1.6}$$

and let us now consider some basic properties:

1. $P(E|E) = 1$;
2. $P(\emptyset|E) = 0$;
3. $E \subseteq F \Rightarrow P(F|E) = 1$;
4. $F \subseteq E \Rightarrow P(F|E) = \frac{P(F)}{P(E)}$.

Using the conditional probability definition we calculate the probability of the intersection of two events i.e. E and F

$$P(E \cap F) = P(E|F) \cdot P(F) \tag{1.7}$$

namely the *multiplication of probabilities* formula.

Assuming $P(E_1 \cap E_2 \cdots \cap E_n) > 0$, the latter can be generalized by induction to n events

$$P(E_1 \cap E_2 \cdots \cap E_n) = P(E_1) \cdot P(E_2|E_1) \cdots P(E_n|E_1 \cap E_2 \cdots \cap E_{n-1}) \tag{1.8}$$

As before conditional probability enjoys the same three axioms:

1. $0 \leq P(E|F) \leq 1$;
2. $P(\Omega|F) = 1$;
3. $P\left(\bigcup_{j=1}^{\infty} E_i|F\right) = \sum_{j=1}^{\infty} P(E_i|F)$ whenever E_i, $i = 1, 2, \ldots$ are pairwise disjoint sets.

Example 1.3.2. This time, a die is tossed once. Let $E = \{5, 6\}$ and F be the event of getting an odd number at first toss, $F = \{1, 3, 5\}$. The conditional probability of E given F is

$$P(E \mid F) = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}.$$

1.4 Bayes' rule

By the multiplication of probabilities formula (1.7), in the case of $P(E) > 0$ and $P(F) > 0$, we get *Bayes' rule*

$$P(F|E) = \frac{P(E|F)\,P(F)}{P(E)}. \tag{1.9}$$

Definition 1.4.1. Let $F_1 \cdots F_n$ be pairwise disjoint sets with positive probabilities so that

$$\bigcup_{i=1}^{m} F_i = \Omega,$$

namely$\{F_i\}_{i=1}^{m}$ provide a *partition* of Ω.

For any event E, we have $E = \bigcup_{i=1}^{n} E \cap F_i$ composed by pairwise disjoint sets. We get the *law of total probability*

$$P(E) = \sum_{i=1}^{n} P(E|F_i) \cdot P(F_i), \tag{1.10}$$

that is the weighted average of conditional probabilities $P(E|F_i)$.

Since $\{F, F^c\}$ constitute a partition of Ω we get an important special case of (1.10)

$$P(E) = P(E|F)\,P(F) + P(E|F^c)\,P(F^c). \tag{1.11}$$

Then combining (1.9) with (1.10), we get a more general form of Bayes' formula:

$$P(F_j|E) := \frac{P(E|F_j) \cdot P(F_j)}{\sum_{i=1}^{n} P(E|F_i) \cdot P(F_i)}. \tag{1.12}$$

By (1.11) the latter has the special case

$$P(F|E) = \frac{P(E|F) \cdot P(F)}{P(E|F)\,P(F) + P(E|F^c)\,P(F^c)}. \tag{1.13}$$

Remark 1.4.2. In statistical applications, F_j are called *hypotheses*. The $P(F_j)$ are called *prior probabilities* because they do not take into account any information of E and the $P(F_j \mid E)$, the conditional probability of F_j, given E are called *posteriori probabilities* because they are derived from or depends upon the specified value of E.

Example 1.4.3. An urn contains two coins C_1 and C_2. The first is a fair coin, while C_2 has probability $\frac{1}{3}$ of getting head. Suppose a coins is drawn at random and toss: we get head. Hence we want to know the probability that it is the unbiased coin C_1.

We have a sample space $\Omega = \{C_1H, C_1T, C_2H, C_2T\}$. The probabilities are

$$\begin{aligned} P(C_1) = P(C_2) &= \frac{1}{2} \\ P(H|C_1) = P(T|C_1) &= \frac{1}{2} \\ P(H|C_2) &= \frac{1}{3}. \end{aligned}$$

By (1.13) we then get

$$P(C_1|H) = \frac{P(C_1)\,P(H|C_1)}{P(H|C_1)\,P(C_1) + P(H|C_2)\,P(C_2)} = \frac{\frac{1}{2}\cdot\frac{1}{2}}{\frac{1}{2}\cdot\frac{1}{2} + \frac{1}{3}\cdot\frac{1}{2}} = \frac{3}{5}.$$

1.5 Independent events

Definition 1.5.1. Given two events i.e. E and F we say they are *independent* if

$$P(E \cap F) := P(E) \cdot P(F) \tag{1.14}$$

assuming $P(F)$ and $P(E)$ are both positive. They are called *dependent* otherwise. This means the occurrence of F doesn't change the probability that E could occur.

Remark 1.5.2. In this case, the conditional probability

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = P(E) \tag{1.15}$$

Let E and F be two independent events. Then also E and F^c are independent. Moreover let E, F and G be three events, if they are independent they are linked by these relations:

1. $P(E \cap F \cap G) = P(E) \cdot P(F) \cdot P(G)$;
2. $P(E \cap F) = P(E) \cdot P(F)$;
3. $P(E \cap G) = P(E) \cdot P(G)$;
4. $P(F \cap G) = P(F) \cdot P(G)$.

We should observe that if E, F and G are independent, E is independent of each event made by F and G, i.e $F \cup G$ for instance.

As before independence can be extended to n distinct events E_1, $E_2 \cdots E_n$. They are called independent if each subset $E_1, \cdots E_r$, $r \leq n$ enjoys

$$P(E_1 \cap E_2 \cdots \cap E_r) = P(E_1) \cdot P(E_2) \cdots P(E_r) \tag{1.16}$$

Example 1.5.3. Consider again the tossing of two fair dice. Let E be the event "the sum of dice equals 6" and F be the event "the first die is 4". We notice they are dependent, indeed

$$P(E \cap F) = P(\{(4,2)\}) = \frac{1}{36},$$

while

$$P(E) P(F) = \frac{5}{36} \cdot \frac{1}{6} = \frac{5}{216}.$$

Example 1.5.4. Let us draw at random one card from a French deck. Let E be the event "the card is an ace" and let F be the event "the suit of the card is hearts". We have

$$P(E \cap F) = \frac{1}{52} = \frac{1}{13} \cdot \frac{1}{4} = P(E) \cdot P(F).$$

Hence, E and F are independent.

Chapter 2

Combinatorics

Consider a uniform sample space Ω. Namely for each event $E \subseteq \Omega$, $P(E) = \frac{|E|}{|\Omega|}$. Thus computing the probability of any event reduces to counting its elements. Many counting problems are simply enumerating dispositions (sorted) or combinations (not sorted) of a set I_n of n distinct element, namely $\{1, ..., n\}$.

2.1 Dispositions without repetition

A *k-disposition*, or disposition of k elements without repetition of I_n, is a sorted k-tuple $(a_1, ..., a_k)$ of distinct elements of I_n we indicate with $P(n, k)$. We have

$$P(n, k) = n(n-1)\cdots(n-(k-1)) = \frac{n!}{(n-k)!}, \tag{2.1}$$

where $0! = 1$ by definition. Moreover when $k = n$, namely a n-disposition is called *permutation*.

$$P(n, n) = n(n-1)\cdots 2 \times 1 = n!. \tag{2.2}$$

Example 2.1.1. From a urn containing n enumerated balls, we draw k balls without replacement. The outcome of the experiment is described by a k-disposition $(a_1, ..., a_k)$, where a_i is the number of the i-th ball drawn. The sample space Ω has $P(n, k)$ elements. With $k = 2$ and $n = 3$, the sample space is

$$\Omega = \{(1,2), (1,3), (2,1), (2,3), (3,1), (3,2)\}.$$

Example 2.1.2. We place k distinct objects in n cells where at most one object is allowed in a cell. The outcome of the experiment is described by a k-disposition $(a_1, ..., a_k)$, where a_i is the number of the i-th cell with the i-th object inside. The sample space has $P(n,k)$ elements. With $k = 2$ and $n = 3$, the sample space is

$$\Omega = \{(1,2), (1,3), (2,1), (2,3), (3,1), (3,2)\}.$$

2.2 Combinations without repetition

A *k-combination without repetition* of k elements of I_n is a subset of $k < n$ distinct elements $\{a_1, \ldots, a_k\}$ of I_n. With $C(n,k)$ we indicate the number of k-combinations without repetition of I_n.

$$C(n,k) = \frac{P(n,k)}{k!} = \frac{n!}{k!\,(n-k)!} := \binom{n}{k} \tag{2.3}$$

is called *binomial coefficient* and read "n choose k". For $k < 0$ or $k > n$ we set $\binom{n}{k} = 0$. It owns its name to the expansion

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k, \tag{2.4}$$

where $\binom{n}{0} = 1$ by definition.

Binomial coefficients have some useful properties:

1. $\binom{n}{k} = \binom{n}{n-k}$;
2. $\binom{n}{1} = \binom{n}{n-1} = n$.

Moreover, we have *Pascal's rule*

$$\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1},\ 0 \le k < n. \tag{2.5}$$

Example 2.2.1. From a urn containing n enumerated balls we draw k balls without replacement. The outcome of the experiment is described by a k-combination $\{a_1, ..., a_k\}$, where $a_1, \ldots a_k$ are the numbers of drawn balls. The sample space Ω has $C(n,k)$ elements. With $k = 2$ and $n = 3$, the sample space is

$$\Omega = \{\{1,2\}, \{1,3\}, \{2,3\}\}.$$

Example 2.2.2. We place k indistinguishable objects in n cells where at most one object is allowed in a cell. The outcome of the experiment is described by a k-combination $\{a_1, ..., a_k\}$, where $a_1, \ldots, a_k$ indicate the numbers of cells with an object inside. The sample space has $C(n,k)$ elements. With $k = 2$ and $n = 3$, the sample space is

$$\Omega = \{\{1,2\},\{1,3\},\{2,3\}\}.$$

2.3 Dispositions with repetition

A *k-disposition*, or disposition of k elements of I_n, is a sorted k-tuple $(a_1, ..., a_k)$ of elements of I_n eventually repeated. According to the *multiplication principle*[1] the number of k-dispositions of I_n is n^k. Recalling 2.4 where $a = b = 1$ we get

$$\sum_{k=0}^{n} \binom{n}{k} = 2^k, \tag{2.6}$$

namely the binary sequences of length k.

Example 2.3.1. From a urn containing n enumerated balls we draw k balls with replacement. The outcome of the experiment is described by a k-disposition $(a_1, ..., a_k)$, where a_i is the number of the i-th ball drawn. The sample space Ω has n^k elements. With $k = 2$ and $n = 3$, the sample space is

$$\Omega = \{(1,1)(1,2),(1,3),(2,1),(2,2),(2,3),(3,1),(3,2),(3,3)\}.$$

Example 2.3.2. We place k distinct objects in n cells where any number of objects is allowed in a cell. The outcome of the experiment is described by a k-disposition $(a_1, ..., a_k)$, where a_i is the number of the i-th cell with at least one object inside. The sample space has n^k elements. With $k = 2$ and $n = 3$, the sample space is

$$\Omega = \{(1,2),(1,2),(1,3),(2,1),(2,2),(2,3),(3,1),(3,2),(3,3)\}.$$

[1] We suppose an experiment can be split into a walk made up of m steps. Moreover r_i, namely the number of outcomes at the i-th step, is independent from the outcome of the previous step. If from different walks we gain distinct final outcomes, the experiment has $r_1 \times r_2 \times \cdots \times r_m$ different outcomes.

2.4 Combinations with repetition

A *k-combination* of k elements of I_n is a subset of $k < n$ elements $\{a_1, \ldots a_k\}$ of I_n possibly repeated. The number of k-collections of I_n is

$$C(k+n-1,k) = C(k+n-1,n-1) \tag{2.7}$$

Example 2.4.1. From a urn containing n enumerated balls we draw k balls with replacement. The outcome of the experiment is described by a k-combination $\{a_1, ..., a_k\}$, where $a_1, \ldots, a_k$ are the numbers of drawn balls. The sample space Ω has $C(k+n-1,k)$ elements. With $k = 2$ and $n = 3$, the sample space is

$$\Omega = \{\{1,1\},\{1,2\},\{1,3\},\{2,2\},\{2,3\},\{3,3\}\}.$$

Example 2.4.2. We place k indistinguishable objects in n cells where every number of objects is allowed in a cell. The outcome of the experiment is described by a k-combination $\{a_1, ..., a_k\}$, where $a_1, \ldots, a_k$ indicate the numbers of cells with at least one object inside. The sample space has $C(k+n-1,k)$ elements. With $k = 2$ and $n = 3$, the sample space is

$$\Omega = \{\{1,1\},\{1,2\},\{1,3\},\{2,2\},\{2,3\},\{3,3\}\}.$$

Chapter 3

Random variables

3.1 Random variable

Definition 3.1.1. Let $X \neq \emptyset$ be a set. A collection $\mathcal{F}$ of subsets of X is called a *σ-algebra* if it satisfies the following properties:

1. $X \in \mathcal{F}$;
2. $E \in \mathcal{F} \Rightarrow E^c \in \mathcal{F}$;
3. $E_i \in \mathcal{F},\ i = 1, 2, \ldots \Rightarrow \left(\bigcup_{i=1}^{\infty} E_i\right) \in \mathcal{F}$.

A collection that satisfies 1 and 2 above, but it is only closed under *finite* unions, is called an *algebra*.

Example 3.1.2. Let $X \neq \emptyset$. Then two (trivial) σ-algebras are $\mathcal{F} = \{\emptyset, X\}$ and $\mathcal{F} = P(X)$. For $X = \mathbb{R}$ the real numbers, $\mathcal{B}(\mathbb{R})$ denotes the *Borel sets*. This is the smallest (intersection of two σ-algebras is a σ-algebra) σ-algebra containing all open sets.

Definition 3.1.3. A triple $(\Omega, \mathcal{F}, P)$ is called *probability space* if

1. $\Omega \neq \emptyset$;
2. $\mathcal{F}$ is a σ-algebra of subsets of Ω;
3. $P : \mathcal{F} \to [0, 1]$ is such that
   (a) $P(\Omega) = 1$;
   (b) $\mathrm{P}\left(\bigcup_{i=1}^{\infty} E\right) = \sum_{i=1}^{\infty} P(E_i)$ if
      i. $E_i \in \mathcal{F}, i = 1, 2, \ldots$;

ii. $E_i \cap E_j = \emptyset,\ i \neq j$;

Remark 3.1.4. In the case when the cardinality of Ω is finite or countably infinite, the above definition coincides with the previous one when $\mathcal{F} = \mathcal{P}(\Omega)$.

Let $(\Omega, \mathcal{F}, P)$ be a probability space.

Definition 3.1.5. A *random variable* X is a function $X : \Omega \to \mathbb{R}$ such that

$$\{\omega : X(\omega) \in B\} \in \mathcal{F},\ \forall B \in \mathcal{B}(\mathbb{R}). \tag{3.1}$$

Remark 3.1.6. Condition (3.1) ensures that the probability

$$P(\{\omega : X(\omega) \in B\}) \tag{3.2}$$

has a meaning. It actually suffices to require condition (3.1) when $B = (-\infty, x]$, namely $\{\omega : X(\omega) \leq x\} \in \mathcal{F},\ \forall x \in \mathbb{R}$.

Given a random variable X, the probability measure P_X on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ given by

$$P_X(B) := P(\{\omega : X(\omega) \in B\}),\ B \in \mathcal{B}(\mathbb{R}) \tag{3.3}$$

is called *probability distribution* of X. The function F

$$F_X(x) := P(\omega : X(\omega) \leq x) \tag{3.4}$$

is called *cumulative distribution function* and represents the probability that the random variable is minus or equal x.

Moreover F enjoys some properties

1. F is not decreasing
2. $\lim_{x \to \infty} F_X(x) = 1$
3. $\lim_{x \to -\infty} F_X(x) = 0$
4. F is right-continuous and admits left limit for every $x \in \mathbb{R}$.

Definition 3.1.7. A random variable X is called *discrete* if it takes finite or countably infinite values in the set $\mathcal{X}$, namely the state space of X. Let $\mathcal{X} = \{x_1, x_2, \ldots\}$, the probability distribution p_X on $\mathcal{X}$ is defined by

$$p_X(x_i) := P(\omega : X(\omega) = x_i), \tag{3.5}$$

where

$$p_X(x_i) \geq 0 \; i = 1, 2, \ldots \quad 0 \; otherwise. \tag{3.6}$$

Since X takes at least one of x_i values

$$\sum_{i=1}^{\infty} p_X(x_i) = 1. \tag{3.7}$$

From now on let us consider random variables on a *discrete probability space* $(\Omega, \mathcal{P}(\Omega), P)$.

Definition 3.1.8. The *expected value* of X, denoted by $E(X)$, is the weighted average of the values that X takes.

$$E(X) := \sum_{x_i : p_X(x_i) > 0} x_i p_X(x_i). \tag{3.8}$$

Example 3.1.9. Let X be the outcome of tossing a die, supposing each face equiprobable. The state space is $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$and being $P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = \frac{1}{6}$ we get

$$E(X) = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = \frac{7}{2}.$$

Example 3.1.10. Let X be a constant, namely $X \equiv c$. Then $E(X) = c$.

It satisfies the following properties:

1. $|E(X)| \leq E(|X|)$;
2. $E(aX + b) = aE(X) + b, \; a, b \in \mathbb{R}$;
3. $E(y(X)) = \sum_{x_i : p_X(x_i) > 0} y(x_i) p_X(x_i)$, where y is a real valued function.

Proposition 3.1.11. Let $\mathcal{V}$ be the vector space of all random variables

1. $E(aX + bY) = aE(X) + bE(Y) \; a, b \in \mathbb{R}, \; \forall X, Y \in \mathcal{V}$;
2. $X(\omega) \geq 0, \; \forall \omega \Rightarrow E(X) \geq 0$.
   (a) In particular $X(\omega) \geq Y(\omega), \; \forall \omega \Rightarrow E(X) \geq E(Y)$;

3. $\left(E\left(|X \cdot Y|\right)\right)^2 \leq E\left(X^2\right) \cdot E\left(Y^2\right)$, namely the *Cauchy-Schwarz inequality.*

Definition 3.1.12. The *variance* of X is defined by

$$Var\left(X\right) := E\left(\left(X - E\left(X\right)\right)^2\right) \tag{3.9}$$

or

$$Var\left(X\right) = E\left(X^2\right) - \left(E\left(X\right)\right)^2 \tag{3.10}$$

and represents the dispersion of the values of the random variable about its mean. The quantity $\sigma = \sqrt{Var\left(X\right)}$ is called *standard deviation*.

In particular it is true

$$Var\left(iX + j\right) = i^2 Var\left(X\right), \forall i, j \in \mathbb{R}. \tag{3.11}$$

Example 3.1.13. Let us toss a die. $E\left(X\right) = \frac{7}{2}$ and

$$E\left(X^2\right) = 1^2\left(\frac{1}{6}\right) + 2^2\left(\frac{1}{6}\right) + 3^2\left(\frac{1}{6}\right) + 4^2\left(\frac{1}{6}\right) + 5^2\left(\frac{1}{6}\right) + 6^2\left(\frac{1}{6}\right) = \frac{91}{6}.$$

Hence $Var\left(X\right) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$ and $\sigma = \sqrt{\frac{35}{12}}$.

Theorem 3.1.14. *(Markov's Inequality)* Let X be a nonnegative random variable. Then

$$P\left(\omega : X\left(\omega\right) \geq \epsilon\right) \leq \frac{1}{\epsilon} E\left(X\right), \forall \epsilon > 0. \tag{3.12}$$

Corollary 3.1.15. (*Chebyshev's Inequality)* Let X be a random variable. Then

$$P\left(\omega : |X\left(\omega\right) - E\left(X\right)| \geq \epsilon\right) \leq \frac{1}{\epsilon^2} V\left(X\right), \forall \epsilon > 0. \tag{3.13}$$

3.2 Random variables

From now on we use the shorthand notation $P\left(X = x\right)$ to denote $P\left(\omega : X\left(\omega\right) = x\right)$.

Definition 3.2.1. Let X and Y be two random variables with state space $\mathcal{X}$ and $\mathcal{Y}$, respectively. The function $p_{XY}\left(x, y\right)$ defined on $\mathcal{X} \times \mathcal{Y}$

$$p_{XY}(x,y) := P(X=x, Y=y) \geq 0 \tag{3.14}$$

is called *joint probability distribution* of X and Y.

In particular

$$p_X(x) := P(X=x) = \sum_{y:p(x,y)>0} p(x,y) \tag{3.15}$$

and

$$p_Y(y) := P(Y=y) = \sum_{x:p(x,y)>0} p(x,y) \tag{3.16}$$

are called *marginal* distributions.

Definition 3.2.2. We define the *covariance* between X and Y

$$Cov(X,Y) := E((X - E(X))(Y - E(Y))) \tag{3.17}$$

or

$$Cov(X,Y) = E(XY) - E(X) \cdot E(Y) \tag{3.18}$$

and represents a measure of how much two variables change together.

It enjoys the following properties:

1. $Cov(X,Y) = Cov(Y,X)$;
2. $Cov(X,X) = Var(X)$ (variance is a special case of the covariance when the two variables are identical);
3. $Cov(aX,Y) = aCov(X,Y)$;
4. $Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} Cov(X_i, Y_j)$.

Remark 3.2.3. It easy to verify that

$$V(X+Y) = V(X) + V(Y) + 2Cov(X,Y). \tag{3.19}$$

The *correlation coefficient* of X and Y is defined by

$$\rho(X,Y) := \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}} \tag{3.20}$$

and

$$-1 \leq \rho(X,Y) \leq 1. \tag{3.21}$$

The correlation coefficient is a dimensionless measure of linear dependence between X and Y.

More precisely, it is 1 in the case of an increasing linear relationship, -1 in the case of a decreasing linear relationship.

In all other cases, it indicates the degree of linear dependence between the variables.

The closer the coefficient is to either -1 or 1, the stronger the correlation between the variables.

Definition 3.2.4. The expected value of X and Y is

$$E\left(f\left(x,y\right)\right)=\sum_{y}\sum_{x}f\left(x,y\right)p_{XY}\left(x,y\right), \tag{3.22}$$

where $p_{XY}\left(x,y\right)$ is the joint probability distribution.

3.3 Independence

Given two events E and F, (1.6) defines the conditional probability of E given F.

Definition 3.3.1. The *conditional probability distribution* of E *given* F is

$$p_{X|Y}\left(x|y\right)=P\left(X=x|Y=Y\right)=\frac{p\left(x,y\right)}{p_{Y}\left(y\right)},\ \forall y:p_{Y}\left(y\right)>0. \tag{3.23}$$

Hence the *conditional expectation* of X given that Y has taken the value $y\in\mathcal{Y}$ is the function $g\left(\cdot\right)$ defined on $\mathcal{Y}$ by

$$g\left(y\right)=E\left(X|Y=y\right)=\sum_{x\in\mathcal{X}}xp_{XY}\left(x|y\right),\ \forall y:p_{Y}\left(y\right)>0. \tag{3.24}$$

The conditional expectation of X given Y is the random variable $g\left(Y\right)$ which takes the value $E\left(X|Y=y\right)$ with probability $p_{Y}\left(y\right)$. If we have random variables $X,Y_{1},\ldots,Y_{n}$ taking values in $\mathcal{X},\mathcal{Y}_{1},\ldots,\mathcal{Y}_{n}$ we can define

$$g_{n}\left(y_{1},\ldots,y_{n}\right)=E\left(X|Y_{1}=y_{1},\ldots Y_{n}=y_{n}\right) \tag{3.25}$$

$$=\sum_{x\in\mathcal{X}}xP\left(X=x|Y_{1}=y_{1},\ldots Y_{n}=y_{n}\right). \tag{3.26}$$

Remark 3.3.2. Let f be any function. Then

$$E\left(X \cdot f\left(Y\right)|Y\right) = f\left(X|Y\right), \tag{3.27}$$

in particular when $X \equiv 1$, we get

$$E\left(f\left(Y\right)|Y\right) = f\left(Y\right). \tag{3.28}$$

Theorem 3.3.3. (*Iterated conditioning*) We have

$$E\left(E\left(X|Y_1, \ldots Y_n\right)\right) = E\left(X\right) \tag{3.29}$$

and when $1 \leq i_1 < \cdots < i_k \leq n$ we have

$$E\left(E\left(X|Y_1, \ldots, Y_n\right)|Y_{i_1}, \ldots Y_{i_k}\right) = E\left(X|Y_{i_1}, \ldots Y_{i_k}\right). \tag{3.30}$$

Definition 3.3.4. Two random variables are called *independent* if they satisfy

$$P\left(X = x, Y = y\right) = P\left(X = x\right) \cdot P\left(Y = y\right), \ \forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \tag{3.31}$$

they are called *dependent* otherwise.

Remark 3.3.5. Two independent random variables X and Y are *uncorrelated*, namely $\varphi\left(X, Y\right) = Cov(X, Y) = 0$, but uncorrelation does not imply independence.

Example 3.3.6. Let X and Y be two random variables such that

$$P\left(X = 0\right) = P\left(X = 1\right) = P\left(X = -1\right) = \frac{1}{3}$$

and

$$Y = \begin{cases} 0 & X \neq 0 \\ 1 & X = 0 \end{cases}.$$

We notice that $XY = 0$ which implies $E\left(XY\right) = 0$ and $E\left(X\right) = 0$. Hence by (3.18) we get

$$Cov\left(X, Y\right) = 0,$$

however X and Y are clearly dependent.

In particular (3.14) becomes

$$p_{XY}\left(x, y\right) = p_X\left(x\right) \cdot p_Y\left(y\right) \ \forall x, y. \tag{3.32}$$

Under the same assumption of independence (3.31) the conditional probability distribution (3.23) is now

$$p_{X|Y}\left(x|y\right) = p_X\left(x\right). \tag{3.33}$$

Also the expected value (3.22) becomes

$$E\left(f_1\left(X\right)f_2\left(Y\right)\right) = E\left(f_1\left(X\right)\right)\cdot E\left(f_2\left(Y\right)\right). \tag{3.34}$$

Remark 3.3.7. If X is independent of Y, the relation (3.19) is now

$$V\left(X+Y\right) = V\left(X\right)+V\left(Y\right), \tag{3.35}$$

and

$$E\left(X|Y\right) = E\left(X\right) \tag{3.36}$$

because (3.31) leads to $P\left(X=x|Y=y\right) = P\left(X=x\right)$.

Definition 3.3.8. Let $\mathcal{F}_1, \mathcal{F}_2$ be two algebras (3.1.1). They are called *independent* if all pairs $(E_1, E_2) \in \mathcal{F}_1 \times \mathcal{F}_2$ are independent. If E and F are independent, so are the algebras $\mathcal{F}_1 = \{E, E^c, \emptyset, \Omega\}$ and $\mathcal{F}_2 = \{F, F^c, \emptyset, \Omega\}$. In general the algebras $\mathcal{F}_1, \ldots \mathcal{F}_n$ are called independent if $(E_1, \ldots, E_n)$, $E_i \in \mathcal{F}_i$ are independent.

Remark 3.3.9. Pairwise independence does not imply independence.

3.4 Examples

3.4.1 Bernoulli trial

Let consider the experiment of tossing a biased coin with sample space $\Omega = \{H, T\}$. Let p and $q = 1-p$ be the probabilities of getting Head and Tail, respectively.

Definition 3.4.1. A random variable X, where

$$X\left(\omega\right) = \begin{cases} 1, & \omega = H \\ 0, & \omega = T \end{cases} \tag{3.37}$$

and

$$P\left(X=H\right) = p,\ P\left(X=T\right) = q = 1-p \tag{3.38}$$

is called *Bernoulli trial.*

Remark 3.4.2. A *Bernoulli process* consists of repeatedly performing independent but identical Bernoulli trials (n coin tosses).

3.4.2 The binomial distribution

Definition 3.4.3. The binomial distribution represents the number of successes in a sequence of n Bernoulli trials, each of which occurs with probability $p,\ 0 < p < 1$.

A discrete random variable X follows the *binomial distribution* if the probability is given by

$$p_X(k) = P(X = k) = \binom{n}{k} \cdot p^k (1-p)^{n-k}, \; k = 0, 1, \ldots n \quad (3.39)$$

and we write $X \sim B(n, p)$.

By the binomial theorem (2.4) we can prove that probabilities (3.39) sum to 1.

Proof. Indeed

$$\sum_{k=0}^{\infty} p_X(k) = \sum_{k=0}^{\infty} \binom{n}{k} \cdot p^k (1-p)^{n-k} = [p + (1-p)]^n = 1.$$

□

The expected value of X is

$$E(X) = np \quad (3.40)$$

and the variance is

$$Var(X) = np(1-p). \quad (3.41)$$

Remark 3.4.4. When $n = 1$ the binomial distribution becomes a Bernoulli trial.

3.4.3 The geometric distribution

Definition 3.4.5. The geometric distribution represents the number of Bernoulli trials needed to get one success, each of which occurs with probability p.

A random variable X follows the *geometric distribution* if the probability is given by

$$p_X(k) = P(X = k) = q^{k-1}p, \; k = 1, 2, \ldots, \quad (3.42)$$

where $q = 1 - p, \; 0 < p < 1$.

The expected value of X is

$$E(X) = \frac{1}{p} \quad (3.43)$$

and the variance is

$$Var(X) = \frac{1-p}{p^2}. \quad (3.44)$$

3.4.4 The hypergeometric distribution

Consider an urn containing N balls, where N_1 are white and N_2 are black. A group of r balls is chosen at random.

Definition 3.4.6. The hypergeometric distribution represents the probability that the group so chosen contains exactly k white balls.

A random variable X follows the *hypergeometric distribution* with parameters (N_1, N, n) if the probability is given by

$$p_X(k) = P(X = k) = \frac{\binom{N_1}{k}\binom{N_2}{n-k}}{\binom{N}{n}}, k = 0 \leq k \leq \min(N_1, n). \quad (3.45)$$

Hence $p_X(k)$ is the probability that the group so chosen contains exactly k white balls.

The expected value of X is

$$E(X) = \frac{n \cdot N_1}{N} \quad (3.46)$$

and the variance is

$$Var(X) = \frac{n \cdot N_1}{N}\left[\frac{(n-1)(N_1-1)}{N-1} + 1 - \frac{n \cdot N_1}{N}\right]. \quad (3.47)$$

Remark 3.4.7. Let X be a random variable who follows the hypergeometric distribution with parameters (N_1, N, n) and $p = \frac{N_1}{N}$. If $n = 1$ then X is a Bernoulli trial with parameter p.

Chapter 4

Markov chains

4.1 Historical background

Andrey Andreevich Markov (June 14, 1856 – July 20, 1922) was born in Ryazan as the son of the secretary of the public forest management of Ryazan, Andrey Grigorevich Markov, and his first wife Nadezhda Petrovna Markova.

In 1874 he finished the school and began his studies at the physical-mathematical faculty of St Petersburg University. He was appointed extraordinary professor in 1886 and in the same year he was elected adjunct to the Academy of Sciences.

In 1890 Markov became extraordinary member of the academy. His promotion to an ordinary professor of St Petersburg University followed in autumn 1894. In 1896, he was elected ordinary member of the academy. In 1905 he was appointed emeritus professor and got the right to retire which he immediately used.

In connection with student riots in 1908 he eventually decided to retire from the university. Markov then resumed his teaching activities and lectured probability theory and differential calculus until his death.

When Markov introduced his famous model in 1906, he was not preoccupied with applications. He just wanted to show that independence is not necessary for the law of large numbers. For example he described the alternation of consonants and vowels in Pushkin's Eugene Onegin as a two-state chain. Poincarè applied Markov theories to card shuffling. In 1907 Paul and Tatiana Ehrenfest proposed a Markov chain model to clarify thermodynamic irreversibility. Sir

Francis Galton, interested in the survival of English peerage, invented the branching process, another markov model. But the fundation of a general theory was provided during the 1930s by Andrei Kolmogorov.

Recently Markov chain theory has received an additional impetus from the advent of Monte Carlo Markov chain simulation. Generally the list of application of Markov chains is virtually infinite. Indeed today they find application to biology, genetic and population theory, social science and mobility, psychology, physics and electrical engineering.

Because of the relative simplicty of its theory and the possibility to represent extremely varied and complex behaviours the role of Markov chains may well be compared to that of ordinary differential equations.

4.2 Definitions

We want to focus on a class of discrete-time stochastic processes.

Definition 4.2.1. Let $(\Omega, \mathcal{F}, P)$ be a probability space. A discrete-time stochastic process is a sequence of random variables

$$X = \{X(n) : n \in I\} \tag{4.1}$$

defined on Ω, where I is a discrete index set.

Unless otherwise stated, from now on, we take the index set I to be a set of integers $I = \{0, 1, 2, \ldots\}$ and the state space $\mathcal{X}$ is assumed to be discrete.

Definition 4.2.2. A discrete-time, discrete state space stochastic process X is said to be a *Markov Chain* if it satisfies the property

$$\begin{aligned} P(X(n+1) = x_{n+1} \mid X(n) = x_n, \cdots, X(1) = x_1, X(0) = x_0) \\ = P(X(n+1) = x_{n+1} \mid X(n) = x_n) \\ \forall n \geq 0 \quad \forall (x_{n+1}, x_n, \ldots, x_1, x_0) \in \mathcal{X} \quad . \end{aligned} \tag{4.2}$$

Remark 4.2.3. Namely, the distribution of $X(n+1)$ depends on the past only through the present. In other words, given $X(n)$, $X(n+1)$ is conditionally independent of $X(n-1), \ldots, X(0)$.

The set $\mathcal{X}$ is either finite or denumerably infinite. Hence, we can identify x_i with i, namely the state space $\mathcal{X}$ is represented by a subset of $\mathbb{Z}_0^+$ $\{0, 1, \ldots r-1\}$ or $\mathbb{Z}$ itself.

We introduce the *transition probabilities* as

$$p_{ij}(n) := P(X(n+1) = j \mid X(n) = i). \tag{4.3}$$

They do not always depend on n, that is to say, on time and in this case the chain is said to be *time-homogeneous.* Unless the opposite is explicitly stated, we always consider Markov chains with this property.

Moreover we consider the *m-step transition probabilities*, namely the probabilities of a chain moving from a state i to state j in exactly m steps

$$p_{ij}^{(m)} := P(X(n+m) = j \mid X(n) = i). \tag{4.4}$$

When $m = 1$ they become the 1-step transition probabilities p_{ij}, we dealt with before.

Assuming $n = 0$, the m-step transition probability can be obtained as

$$\begin{aligned} p_{ij}^{(m)} &= P(X(m) = j \mid X(0) = i) = \\ &= \sum_{x_1} \cdots \sum_{x_{m-1}} P(X(m) = j, X(m-1) = i_{m-1}, \ldots, X(1) = i_1 \mid X(0) = i) \\ &= \sum_{x_1} \cdots \sum_{x_{m-1}} P(X(m) = j, X(m-1) = i_{m-1}) \ldots P(X(1) = i_1 \mid X(0) = i) \\ &= \sum_{x_1} \cdots \sum_{x_{m-1}} p_{ii_1} p_{i_1 i_2} \ldots p_{i_{m-1} j} \quad \forall m \geq 2 \end{aligned},$$

where the second equality is due to the Markovian property of the process.

It is also useful when dealing with discrete state space to identify the matrix of transition probabilities. Let the marginal distribution of $X(n)$ be

$$\pi_i(n) = P(X(n) = i). \tag{4.5}$$

By the law of total probability (1.10), we get the recursion

$$\pi_j(n+1) = \sum_i p_{ij} \pi_i(n) \tag{4.6}$$

that can also be expressed in a matrix form as

$$\pi(n+1) = P^* \pi(n), \tag{4.7}$$

where $\pi(n)^* = (\pi_0(n), \pi_1(n), \ldots)$ and the matrix operation $*$ has the significance of matrix transposition.

We define the *matrix chain*

$$P = \begin{bmatrix} p_{00} & p_{01} & p_{02} & \cdot & \cdot \\ p_{10} & p_{11} & p_{12} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \tag{4.8}$$

The (i, j) entry of P is the transition probability p_{ij}. By the three axioms of probability and the way we built it, we have

1. $p_{ij} \geq 0,\ \forall i, j$;
2. $\sum_j p_{ij} = \sum_j P\left(X\left(n+1\right) = j \mid X\left(n\right) = i\right) = 1,\ \forall i.$

That means all rows of this kind of matrix sum to one. Such matrices are called *stochastic* and have a few interesting properties:

1. At least one eigenvalue equals one.
2. The product of stochastic matrices always leads to a stochastic matrix.

Proof. To prove the latter property we use the former one; define $\mathbf{1} = (1, 1, \ldots, 1)^*$ and given two stochastic matrices P and Q we get

$$\begin{aligned} P\mathbb{1} &= \mathbb{1} \\ Q\mathbb{1} &= \mathbb{1} \\ PQ\mathbb{1} &= \mathbb{1}. \end{aligned} \tag{4.9}$$

□

Remark 4.2.4. All other eigenvalues λ of P are in the open unit disc. The *essential spectral radius* of P or SLEM, namely *second largest eigenvalue modulus*, is defined by

$$\mu\left(P\right) = \max\left\{|\lambda|; \lambda \in \sigma\left(P\right) \setminus 1\right\}, \tag{4.10}$$

it is namely the absolute value of the eigenvalue different from 1 closest to the unit circle[1].

[1] $-\log\mu$ is called the *mixing rate* of the chain. The quantity $1 - \mu\left(P\right)$ is called *spectral gap*.

Generally we have

$$p_{ij}^{(n+m)} := \sum_k P\left(X\left(n+m\right)=j \mid X\left(n\right)=k, X\left(0\right)=i\right)\cdot$$
$$\cdot P\left(X\left(n\right)=k \mid X\left(0\right)=i\right)=\sum_k p_{kj}^{(m)} p_{ik}^{(n)}$$

which is called *Chapman-Kolmogorov equation*. In simple words, it means that the probability of going from i to j in $m+n$ steps is the sum over all k of the probability of going from i to k in m steps, then from k to j in n steps.

In particular, from it we can derive the *forward equation*

$$p_{ij}^{(n+1)} := \sum_k p_{ik}^{(n)} p_{kj} \tag{4.11}$$

and the *backward equation*

$$p_{ij}^{(n+1)} := \sum_k p_{ik} p_{kj}^{(n)}. \tag{4.12}$$

Let $P^{(n)} = \left(p_{ij}^{(n)}\right)$ be the matrix of n-step transition probabilities. By (4.12) $P^{(n)} = P^n$, namely the (row times colums) product of P with itself n times (this product makes sense also when the state space is infinite).

Iterating (4.7) we get the marginal distribution of the n-th stage

$$\pi\left(n+1\right) = \left(P^*\right)^{n+1} \pi\left(0\right), \tag{4.13}$$

where $\pi\left(0\right)$ is the *initial distribution*.

Remark 4.2.5. It is now clear that this distribution $\pi\left(0\right)$ and the chain transition matrix P completely determine the Markov chain.

Definition 4.2.6. A distribution π is called *stationary* for the Markov chain X with chain transition matrix P if if satisfies

$$\pi = P^*\pi. \tag{4.14}$$

Hence (4.10) regulates how fast P^n tends to the rank one matrix with all rows equal to π as we defined in (4.14) (it may be seen that the distribution $\pi\left(n\right)$ converges to π as μ^n).

Example 4.2.7. Let the state space be $\mathcal{X} = \{0,1\}$. The chain transition matrix is necessarily of the form

$$P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix},$$

where p and q are supposed to be $0 \leq p, q \leq 1$. Its stationary distribution is $\pi(n) = (\pi_0(n), \pi_1(n))^*$ where $\pi_0(n) = \frac{q}{q+p}$ and $\pi_1(n) = \frac{p}{q+p}$. Indeed multiplying $\pi_0(n)$ by p and $\pi_1(n)$ by q we get both $\frac{pq}{q+p}$, which means the probability to change state is the same.

Definition 4.2.8. A chain matrix P is called *doubly stochastic* if also all of its colums sum to one, namely both $P\mathbb{1}$ and $P^*\mathbb{1}$ are equal to $\mathbb{1}$.

Corollary 4.2.9. A doubly stochastic $n \times n$ chain matrix has the uniform distribution $\left(\frac{1}{n}\right)\mathbb{1}$ as stationary.

Example 4.2.10. Let P be the matrix of Example 4.2.7 where $p = q$. P is clearly doubly stochastic. Then we get $\pi_0(n) = \frac{p}{p+p}$ and $\pi_1(n) = \frac{p}{q+p}$. Hence

$$\pi = \left(\frac{1}{2}, \frac{1}{2}\right)^*.$$

Consider the "past" event

$$F = \{X(0) = x_0, X(1) = x_1, \ldots, X(n-1) = x_{n-1}\}$$

and the "future" event

$$E = \{X(n+1) = x_{n+1}, X(n+2) = x_{n+2}, \ldots\}.$$

By the Markov property (4.2),

$$P(E \mid F, X(n) = x_n) = P(E \mid X(n) = x_n)$$

which implies

$$P(E \cap F \mid X(n) = x_n) = P(F \mid X(n) = x_n) \cdot P(E \mid X(n) = x_n), \tag{4.15}$$

namely, at each time n, past and future of the process are conditionally independent given the present $X(n)$. Property (4.15) in turn gives

$$P(F \mid E, X(n) = x_n) = P(F \mid X(n) = x_n),$$

which is the Markov property for the chain with time reversed.

We can introduce the *reverse-time transition probabilities* by using the Bayes' rule (1.13)

$$\begin{aligned} q_{ij}(n, \pi(0)) &= P(X(n) = j \mid X(n+1) = i) = \\ &= \frac{P(X(n) = j, X(n+1) = i)}{P(X(n+1) = i)} = \\ &= \frac{P(X(n) = j) \cdot P(X(n+1) = i \mid X(n) = j)}{P(X(n+1) = i)} = \frac{\pi_j(n)\, p_{ji}}{\pi_i(n+1)} \end{aligned} \tag{4.16}$$

stressing that they depend on both the initial distribution $\pi(0)$ and time, even when the p_{ij} don't.

Definition 4.2.11. A Markov chain is called *reversible* if there is a π such that

$$q_{ij}(n,\pi) = p_{ij},\ \forall i,j \in \mathcal{X}, \tag{4.17}$$

which is equivalent to

$$\pi_j p_{ji} = \pi_i p_{ij},\ \forall i,j \in \mathcal{X}, \tag{4.18}$$

where the latter property is called *detailed balance.*

Remark 4.2.12. If a distribution π satisfies (4.18), such a distribution π is stationary.

Proof. By summing (4.18) on both side with respect to i we obtain

$$\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji} = \pi_j. \tag{4.19}$$

□

Theorem 4.2.13. Let π be any distribution satisfying (4.18). Then π is stationary for the chain with transition probabilities $P = (p_{ij})$.

However, by the definition we just gave of reversible Markov chain and from (4.18) we get that

$$q_{ij}(n,\pi) = q_{ij}(\pi) \tag{4.20}$$

do not depend on time and

$$q_{ji} = p_{ji},\ \forall i,j \in \mathcal{X}. \tag{4.21}$$

Remark 4.2.14. If we start from $\pi(0) = \pi$, then

$$P_\pi(X(n+1) = i \mid X(n) = j) = P_\pi(X(n) = i \mid X(n+1) = j)$$

which means that the transition probabilities between two states are the same and independent of the direction of time.

4.3 Examples of Markov chains with finite state space

To each transition matrix P we can associate a *transition graph* $\mathcal{G}$. This directed graph has the states of $\mathcal{X}$ as nodes and an arc from i to j if and only if $p_{ij} > 0$. The latter probability is then displayed next to the arc.

4.3.1 Random walk with absorbing barriers

Let us consider a gambler in a casino. At any turn the player wins a chip fish with probability p and loses it with probability $q = 1 - p$. Moreover, it is supposed that the player stops gambling when he reaches $r - 1$ or 0 chip fishes.

Take $\mathcal{X} = \{0, 1, \ldots, r-1\}$, so that $X(n) = k$ is the number of chip fishes the gambler has at time n. Given a $X(n)$, it is easy to understand that the past has no influence on the future, i.e. $X(n+1), X(n+2), \ldots$, forms a Markov chain. In particular the matrix chain P_a is

$$P_a = \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ q & 0 & p & 0 & \ldots & 0 \\ 0 & q & 0 & p & \ldots & 0 \\ . & . & . & . & . & \ldots \\ . & . & . & . & . & \ldots \\ 0 & 0 & 0 & 0 & \ldots & 1 \end{bmatrix}. \tag{4.22}$$

States 0 and $r-1$ are said to be absorbing because when the system reaches them, it stays there forever. They represent the situation in which the player has 0 and $r - 1$ chip fishes respectively.

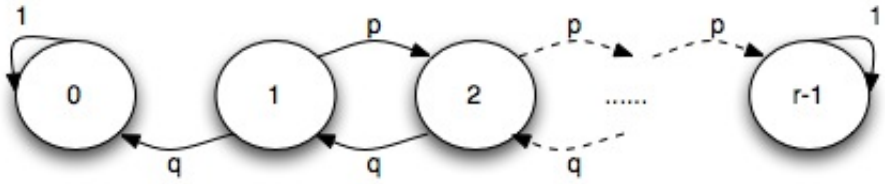

Figure 4.1: Random walk with absorbing barriers

4.3.2 Random walk with reflecting barriers

This model is similar to the previous one, indeed let the state space be $\mathcal{X} = \{0, 1, \ldots, r-1\}$. The matrix chain P_r

$$P_r = \begin{bmatrix} r_0 & p_0 & 0 & 0 & \cdots & 0 \\ q & 0 & p & \cdots & \cdots & 0 \\ 0 & q & 0 & p & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & q & 0 & p \\ 0 & 0 & 0 & \cdots & p_{r-1} & r_{r-1} \end{bmatrix} \tag{4.23}$$

is a generalization of the previous matrix chain P_a.

According to our definition of probability $r_0, p_0, r_{r-1}, p_{r-1} \geq 0$: $r_0 + p_0 = 1, r_{r-1} + p_{r-1} = 1$ and $p, q > 0 : p + q = 1$. When $p_0 = 1$ and $p_{r-1} = 1$, this is a random walk with reflecting barriers, while when $p_0 > 0$ and $p_{r-1} > 0$, the barriers are called partially reflecting.

Remark 4.3.1. When $r_0 = 1$ and $r_{r-1} = 1$, this is the random walk with absorbing barriers described in 4.3.1.

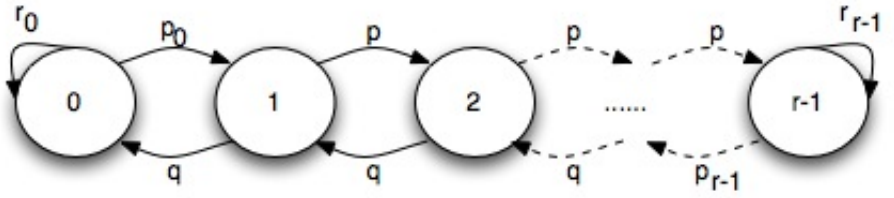

Figure 4.2: Random walk with reflecting barriers

4.3.3 Cyclical random walk

Again let $\mathcal{X} = \{0, 1, \ldots, r-1\}$. In this case each state k can reach the next state $k+1$ with probability p and the previous one, $k-1$, with probability q.

In addition we order the states cyclically so that the next state of $r-1$ is 0 and with the same meaning the previous state of 0 is $r-1$. This can be represented by the chain matrix P_c which is a circulant matrix

$$P_c = \begin{bmatrix} 0 & p & 0 & 0 & \cdots & q \\ q & 0 & p & \cdots & \cdots & 0 \\ 0 & q & 0 & p & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & q & 0 & p \\ p & 0 & 0 & \cdots & q & 0 \end{bmatrix}. \tag{4.24}$$

Notice that P_a, P_r and P_c only differ in the first and last row.

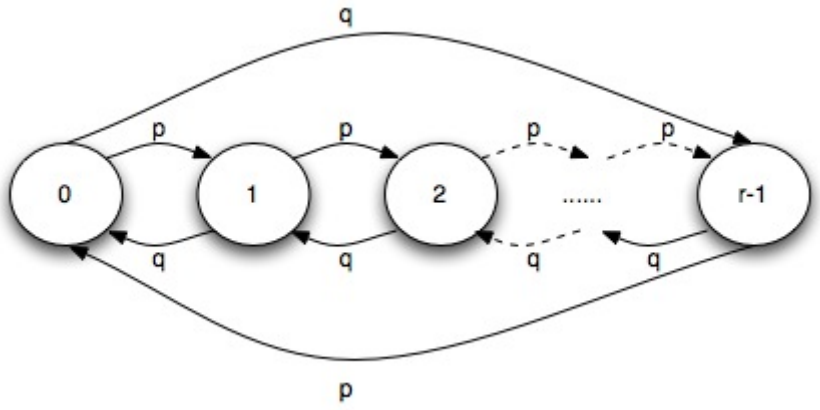

Figure 4.3: Cyclical random walk

4.3.4 The Ehrenfest model of diffusion

Given $\mathcal{X} = \{0, 1, \ldots, r-1\}$, this model is a random walk, except that in this case we set

$$\begin{aligned} p_{k,k+1} &= 1 - \frac{k}{r-1} \\ p_{k,k-1} &= \frac{k}{r-1} \\ 0 < k &< r-1 \end{aligned} \tag{4.25}$$

which leads to the chain matrix P_e so made

$$P_e = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ \frac{1}{r-1} & 0 & 1-\frac{1}{r-1} & \cdots & \cdots & 0 \\ 0 & \frac{2}{r-1} & 0 & 1-\frac{2}{r-1} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & 1-\frac{2}{r-1} & 0 & \frac{2}{r-1} \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}. \tag{4.26}$$

It represents a conceptual experiment that P. and T. Ehrenfest described in 1907.

There are two containers A e B where $r-1$ molecules are distributed at random and at each trial a molecule is picked at random and moved from its container to the other.

In this way, if we consider $X(n) = k$ as the number of molecules in A at time n, the probabilities of $X(n+1)$ to be $k-1$ or $k+1$ are respectively $\frac{(r-1-k)}{r-1}$ and $\frac{k}{r-1}$, according to whether the molecule is chosen from B or A.

As we just said, this model seems a random walk with reflecting barriers. However it has a particularity: it can be interpreted as a diffusion with central force. It is namely a random walk where the transition probabilities change with the position. From state k the chain moves to $k-1$ with higher probability if $k > \frac{r-1}{2}$, while it is more likely to move to $k+1$ if $k < \frac{r-1}{2}$. That is to say, that the system behaves as if it were subject to an attractive elastic force set in $\frac{r-1}{2}$.

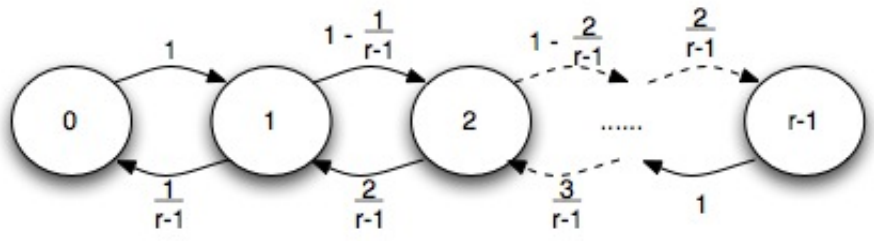

Figure 4.4: The Ehrenfest model of diffusion

4.3.5 The Bernoulli-Laplace model of diffusion

Another similar model was proposed by D. Bernoulli, who wanted to represent the flow of two incompressible liquids between two containers A and B.

In this case we deal with $2r-2$ molecules, where $r-1$ are black and $r-1$ white. Because of the incompressibility of these liquids the number $r-1$ of molecules in each containers is constant.

Let $X(n) = k$ the number of white molecules in A at time n, so that there are $r-1-k$ black molecules in A, $r-1-k$ white molecules and k black molecules in B. According to this the transition probabilities

are

$$
\begin{aligned}
p_{k,k-1} &= \left(\frac{k}{r-1}\right)^2 \\
p_{k,k+1} &= \left(\frac{r-1-k}{r-1}\right)^2 \\
p_{k,k} &= 2 \cdot \frac{k \cdot (r-1-k)}{(r-1)^2} \\
0 \leq k &\leq r-1
\end{aligned} \tag{4.27}
$$

The state space is $\mathcal{X} = \{0, 1, \ldots r-1\}$ and the matrix chain P_b has the form

$$
P_b = \begin{bmatrix}
0 & 1 & 0 & 0 & \cdots & 0 \\
\frac{1}{(r-1)^2} & 2 \cdot \frac{r-2}{(r-1)^2} & \left(\frac{r-2}{r-1}\right)^2 & 0 & \cdots & 0 \\
0 & \left(\frac{2}{r-1}\right)^2 & 2 \cdot \frac{2(r-3)}{r-1} & \left(\frac{r-3}{r-1}\right)^2 & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & \cdots & \left(\frac{r-2}{r-1}\right)^2 & 2 \cdot \frac{r-2}{(r-1)^2} & \frac{1}{(r-1)^2} \\
0 & 0 & 0 & \cdots & 1 & 0
\end{bmatrix} \tag{4.28}
$$

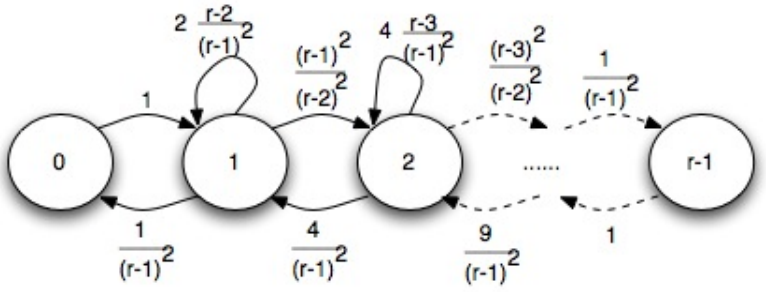

Figure 4.5: The Bernoulli-Laplace model of diffusion

4.3.6 Random placements of balls

Consider a sequence of independent trials. At any turn, we place an indistinguishable ball in one of $r-1$ containers and let $X(n) = k$

be the number of containers containing at least one ball at time n. Therefore, the transition probabilities are

$$\begin{aligned} p_{k,k} &= \frac{k}{r-1} \\ p_{k,k+1} &= \frac{r-1-k}{r-1} \\ 0 \leq k &\leq r-1 \end{aligned} \tag{4.29}$$

so the matrix chain P_p is

$$P_p = \begin{bmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & \frac{1}{r-1} & 1-\frac{1}{r-1} & 0 & \cdots & 0 \\ 0 & 0 & \frac{2}{r-1} & 1-\frac{2}{r-1} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & 1-\frac{1}{r-1} & \frac{1}{r-1} \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \end{bmatrix}. \tag{4.30}$$

State 0 represents the situation in which 0 urns contain at least one ball, so that we have a total probability to put a ball in an empty urn. State $r-1$ means that all urns have at least one ball so the system can't evolve and stays there forever.

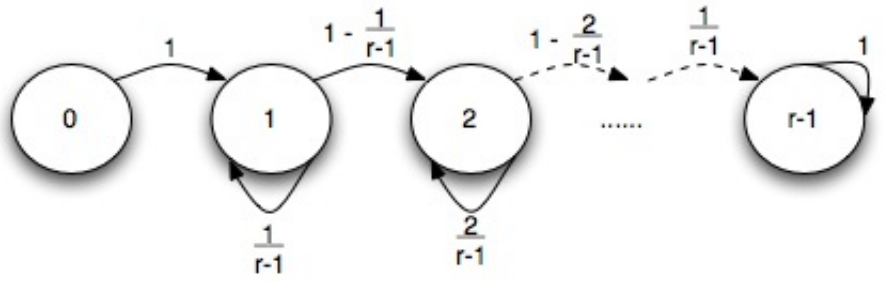

Figure 4.6: Random placements of balls

4.3.7 Wright-Fisher model

Given a fixed population of $r-1$ genes that can be of type a or A, called alleles, each gene has $2r-2$ representatives. If in the n-th generation A occurs k times and a, $2r-2-k$ times, we say the population is in state $X(n)=k$, where $\mathcal{X}=\{0,1,\ldots,2r-2\}$.

Assuming random mating, A-gene has probability $\frac{k}{2r-2}$ to be in the next generation, so that the probability of transiting by k to j with A-type genes is

$$p_{k,j} = \binom{2r-2}{j} \cdot \left(\frac{k}{2r-2}\right)^{j} \cdot \left(1 - \frac{k}{2r-2}\right)^{2r-2-j}. \qquad (4.31)$$

That is to say it is determined by the binomial distribution for $2r-2$ independent trials with success probability $\frac{k}{2r-2}$.

It is important to stress that states 0 and $2r-2$ are absorbing because all genes are of the same type, so it is not possible to exit from them.

For istance, if $r = 3$ the chain matrix P_w has the form

$$P_w = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0,3164 & 0,4219 & 0,2109 & 0,0469 & 0,0039 \\ 0,0625 & 0,2500 & 0,3750 & 0,2500 & 0,0625 \\ 0.0039 & 0,0469 & 0,2109 & 0,4219 & 0,3164 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

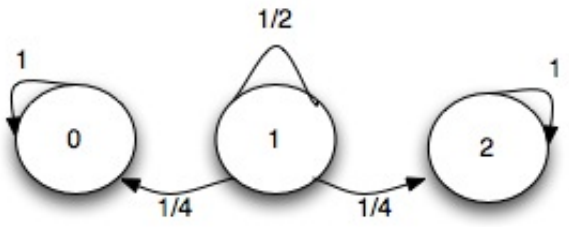

Figure 4.7: Wright-Fisher model with one gene

4.4 Examples of Markov chains with countable state space

4.4.1 Success runs

Let deal with a sequence of Bernoulli trials for a biased coin $X(n)$, namely a Bernoulli process. We also consider $X(n) = 1$ and $X(n) = 0$ a success and a failure, respectively.

We define a sequence of random variables $Y(n)$ with state space $\mathcal{Y} = \mathbb{Z}_0^+$ which represent the number of consecutive successes.

It is apparent that the only nonzero transition probabilities are

$$\begin{aligned} p_{k,0} &= q \\ p_{k,k+1} &= p \\ k &\geq 0 \end{aligned} \tag{4.32}$$

so the infinite chain matrix P_s is

$$P_s = \begin{bmatrix} q & p & 0 & 0 & 0 & \dots \\ q & 0 & p & 0 & 0 & \dots \\ q & 0 & 0 & p & 0 & \dots \\ . & . & . & . & . & \dots \end{bmatrix} \tag{4.33}$$

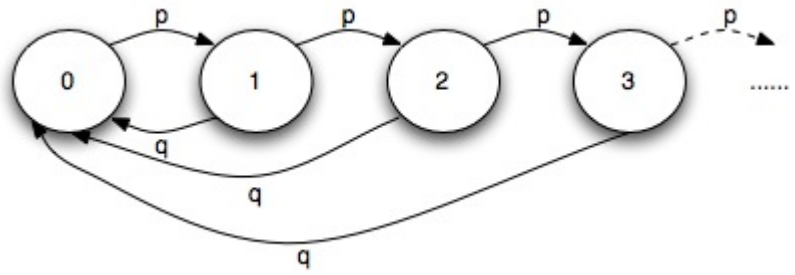

Figure 4.8: Success runs

4.4.2 Random walk on $\mathbb{Z}$

Let again X be a Bernoulli trial which takes values on $\mathcal{X} = \{1, -1\}$ instead of $\{0, 1\}$. We consider a sequence of random variables $Y(n)$ defined by

$$Y(n) = \sum_{k=0}^{n} X(k), \tag{4.34}$$

with state space $\mathcal{Y} = \mathbb{Z}$, which represents the financial status of a gambler.

The infinite chain matrix P_{rw} is

$$P_{rw} = \begin{bmatrix} \dots & . & . & . & . & . & . & \dots \\ . & 0 & q & 0 & p & 0 & 0 & \dots \\ . & 0 & 0 & q & 0 & p & 0 & \dots \\ . & 0 & 0 & 0 & q & 0 & p & \dots \\ . & . & . & . & . & . & . & \dots \end{bmatrix} \tag{4.35}$$

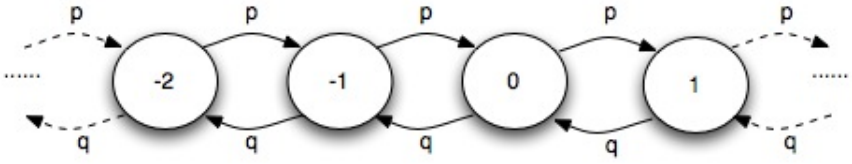

Figure 4.9: Random walk on $\mathbb{Z}$

4.4.3 Random walk on $\mathbb{Z}_0^+$

Consider a particular case on the random walk of Example 4.4.2 with state space $\mathcal{Y} = \mathbb{Z}_0^+$. The chain matrix P_{rw+} is

$$P_{rw+} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & \dots \\ q & 0 & p & 0 & 0 & \dots \\ 0 & q & 0 & p & 0 & \dots \\ . & . & . & . & . & \dots \end{bmatrix}. \tag{4.36}$$

State 0 is called reflecting because when the system reaches it, it moves to state 1 with probability $p_{01} = 1$.

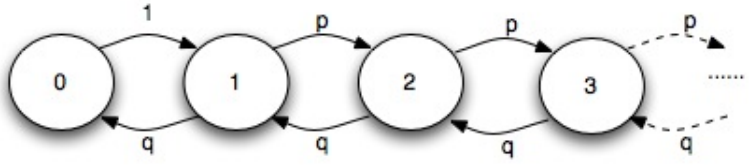

Figure 4.10: Random walk on $\mathbb{Z}_0^+$

Chapter 5

Classification of states and asymptotic analysis

5.1 Communication classes and closed sets

Let us first deal with all the *topological* properties of a Markov chain, namely those that only concern the transition graph $\mathcal{G}$.

Definition 5.1.1. We define a state j *accessible* from state i and write $i \rightsquigarrow j$, if from state i there is a strictly positive probability to eventually reach state j. It must exist namely a time $m \geq 0$ such that

$$p_{ij}^{(m)} = P\left(X\left(n+m\right) = j \mid X\left(n\right) = i\right), \tag{5.1}$$

where we set

$$p_{ij}^{(0)} = 1. \tag{5.2}$$

Remark 5.1.2. The fact that j is accessible from state i does not imply the converse.

Example 5.1.3. In Example 4.3.1 states 0 and $r-1$ are accessible from all the other states but not viceversa.

Definition 5.1.4. We say that states i and j *communicate* and write $i \leftrightsquigarrow j$ if they are accessible from each other.

We define $\mathcal{F}\left(i\right)$ and $\mathcal{B}\left(i\right)$ as the *forward and backward sets* of a state i, namely the set of states that i communicates with and the set of states that communicate with i.

Each state i communicates with itself since p_{ii}^{0} is defined to be 1. In addition if state i communicates with state j which communicates with state k, it is apparent that $i \leftrightsquigarrow k$.

This relationship of communication helps us giving a first classification of states. It is an *equivalence relation* that induces a partition of the states of $\mathcal{X}$ into equivalence classes called *communication classes.*

Example 5.1.5. In Example 4.3.1 states i, $0 < i < r-1$ communicate with all the others. There are three communication classes $E_1 = \{0\}$, $E_2 = \{1, 2, \ldots, r-2\}$, $E_3 = \{r-1\}$.

Definition 5.1.6. If the state space is only one equivalence class the Markov chain is called *irreducible.*

Example 5.1.7. Example 4.3.2 is irreducible.

Definition 5.1.8. A state i has period $d(i)$ if these properties are satisfied:

1. $p_{ii}^{(m)} > 0 \implies d(i)$ divides m;
2. $d(i)$ is the largest integer such that 1. holds.

If $p_{ii}^{(m)} = 0 \; \forall n \geq 1$, we set $d(i) = 0$, while if $d(i) = 1$ state i is called *aperiodic.*

Example 5.1.9. In Example 4.3.2, we see all states have period 2, while in Example 4.3.1 the interior states have period 2, but the absorbing ones are aperiodic.

It is not difficult to show that if $i \leftrightsquigarrow j$, states i and j have the same period.

Example 5.1.10. Example 4.3.4 yields an irreducible chain of period 2.

Definition 5.1.11. A set $C \subseteq \mathcal{X}$ of states is called *closed* if no state outside of C can be reached from any state of C. It is clear that C is closed if $p_{ij} = 0$ whenever state i belongs to C and state j doesn't. In particular it is called *absorbing* if C contains a single state i, namely $p_{ii} = 1$.

Remark 5.1.12. Each closed set C corresponds to a sub-chain in the following sense. If we delete all rows and columns corresponding to states outside C from transition matrix P we get another stochastic matrix. In particular when state i is absorbing, the matrix P is reduced to a single element.

Definition 5.1.13. Given a set C of states, the smallest closed set containing C is called *closure* of C.

Remark 5.1.14. Notice that communication classes are not necessarily closed. For instance $C_2 = \{1, 2, \ldots, r-2\}$ in Example 4.3.1 is not closed. It is indeed possible from one communication class to enter another but then it is not possible to come back. Conversely, states in a closed set need not communicate, see Example 5.2.7 below.

If (5.1) is positive $\forall i, j$, we can say that a Markov chain is irreducible, namely there exists no closed set except of the set of all states.

Example 5.1.15. Consider a chain matrix

$$P = \begin{bmatrix} 0 & 0 & 0 & * & 0 & 0 & 0 & 0 & * \\ 0 & * & * & 0 & * & 0 & 0 & 0 & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & 0 \\ * & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & * & 0 & 0 & 0 & 0 \\ 0 & * & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & * & 0 & 0 & 0 & * & * & 0 & 0 \\ 0 & 0 & * & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & * & 0 & 0 & 0 & 0 & * \end{bmatrix} \tag{5.3}$$

with state space $\mathcal{X} = \{0, 1, \ldots 8\}$, whose p_{ij} denoted by $*$ are positive elements.

We notice that:

1. in the fifth row there's only a $*$ at the fifth place: therefore $p_{44} = 1$ and state 4 is absorbing;
2. the third and the eighth row contain only one positive element each at the eighth and third place respectively, so that state 2 and 7 form a closed set;
3. states 0, 3 and 8 form another closed set because from state 0 passages are possible into state 3 and 8, and from there only to state 0, 3 and 8;
4. from state 1 direct transitions are possible to itself and to states 2, 4 and 7. According to that the closure of state 1 consists of the set $\{1, 2, 4, 7\}$;
5. the closures of states 5 and 6 consist of all nine states;

5.2 Classification of states

In addition to classifying the relationships between the pair of states, we can classify each state into other categories.

For any Markov chain X with state space $\mathcal{X} = \mathbb{N}$ we introduce

$$f_{ij}^{(n)} := P\left(X\left(n\right) = j, X\left(k\right) \neq j,\ k = 1, 2, \ldots, n-1 \mid X\left(0\right) = i\right) \tag{5.4}$$

as the probability that in a process starting from i the first entry to j occurs at the n-th step. Recalling (5.2) we put $f_{ij}^{(0)} = 0$.

We observe

$$f_{ij}^{(n)} = p_{ij}^{(n)} - \sum_{k=1}^{n-1} f_{ij}^{(k)} f_{jj}^{(n-k)} \tag{5.5}$$

and get

$$p_{ij}^{(n)} = \sum_{k=1}^{n} f_{ij}^{(k)} f_{jj}^{(n-k)}. \tag{5.6}$$

Now we can define

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)} \tag{5.7}$$

the probability that, starting from i, the system will ever reach state j, and

$$f_{ii} = \sum_{n=1}^{\infty} f_{ii}^{(n)} \tag{5.8}$$

the probability, starting from i, of ever returning to i. It is apparent that if two states communicate, $i \leftrightsquigarrow j$, $f_{ij} = f_{ji} = 1$.

Definition 5.2.1. A state i is called

1. *recurrent* if
   (a) $f_{ii} = 1$;
   (b) $\mathcal{F}\left(i\right) \subseteq \mathcal{B}\left(i\right)$;
2. *transient* if
   (a) $f_{ii} < 1$;
   (b) $\mathcal{F}\left(i\right) \nsubseteq \mathcal{B}\left(i\right)$.

Remark 5.2.2. If state i is transient, then states in $\mathcal{B}\left(i\right)$ are all transient. If state i is recurrent, on the other hand, states in $\mathcal{F}\left(i\right)$ are all recurrent. In the latter case set $\mathcal{F}\left(i\right)$ is a recurrent class, and set $\mathcal{B}\left(i\right) - \mathcal{F}\left(i\right)$, if not empty, contains only transient states.

We can also express these conditions in terms of the transition probabilities $p_{ii}^{(n)}$.

Theorem 5.2.3. State i is recurrent if and only if

$$\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty \tag{5.9}$$

and if i is recurrent and $i \leftrightsquigarrow j$, then j is recurrent.

Corollary 5.2.4. State i is transient if

$$\sum_{n=1}^{\infty} p_{ij}^{(n)} < \infty, \ \forall i. \tag{5.10}$$

The latter in turn implies

$$\lim_{n\to\infty} p_{ij}^{(n)} = 0, \ \forall i. \tag{5.11}$$

Definition 5.2.5. When a state i is recurrent it makes sense the introduction of

$$\mu_i = \sum_{n=1}^{\infty} n f_{ii}^{(n)} \leq \infty, \tag{5.12}$$

which is called the *mean recurrence time* for i.

In particular a recurrent state is called *null* if $\mu_i = \infty$, *positive* otherwise. A positive recurrent aperiodic state is called *ergodic*.

Theorem 5.2.6. Let j be recurrent aperiodic ($d(j) = 1$). Then

$$\lim_{n\to\infty} p_{ij}^{(n)} = \frac{f_{ij}}{\mu_j}. \tag{5.13}$$

In particular, if $i \leftrightsquigarrow j$, so that

$$\lim_{n\to\infty} p_{ij}^{(n)} \to (\mu_j)^{-1}, \tag{5.14}$$

where, in the case of a null recurrent state, $(\mu_j)^{-1}$ is set equal to zero. The latter result includes the important particular case

$$\lim_{n\to\infty} p_{jj}^{(n)} \to (\mu_j)^{-1}, \tag{5.15}$$

On the other hand if j has period $k = d(j) > 1$ the limit becomes

$$\lim_{n\to\infty} p_{jj}^{(nk)} = \frac{k}{\mu_i}. \tag{5.16}$$

Example 5.2.7. Now let

$$P = \begin{bmatrix} * & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & * & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & * & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & * & * & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & * & * & 0 & 0 & 0 \\ 0 & 0 & 0 & * & 0 & 0 & 0 & 0 & 0 \\ * & * & 0 & 0 & * & 0 & * & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & * & 0 & 0 \end{bmatrix}$$

be a transition matrix with $\mathcal{X} = \{0, 1, \dots 8\}$, where $*$ has the same meaning of (5.3) and transition graph

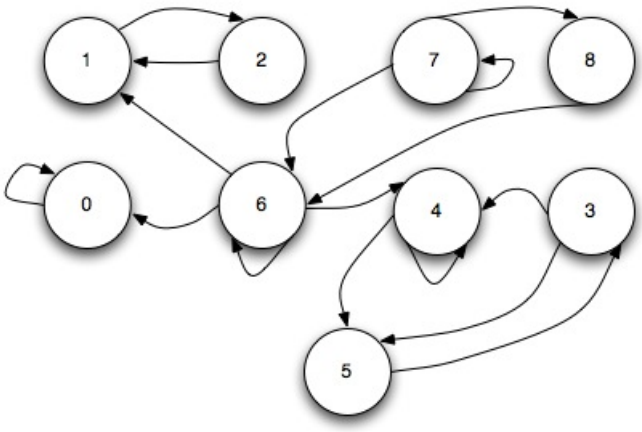

Figure 5.1: Transition graph

.

Let

First of all, we observe that there are three closed sets $E_1 = \{0\}$, $E_2 = \{1, 2\}$, $E_3 = \{3, 4, 5\}$ corresponding to the three stochastic submatrices

$$[*], \quad \begin{bmatrix} 0 & * \\ * & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & * & * \\ 0 & * & * \\ * & 0 & 0 \end{bmatrix}.$$

Then by computing

- $\mathcal{F}(i), \mathcal{B}(i)$ $i = 1, 2, 3$

- $\mathcal{B}(i) - \mathcal{F}(i)$ $i = 1, 2, 3.$

we can state:

1. state 0 is recurrent and aperiodic;
2. states 1 and 2 are recurrent and periodic of period 2 with mean recurrence time equal to 2;
3. states $3, 4, 5$ are ergodic;
4. states $6, 7, 8$ are transient.

Theorem 5.2.8. In an irreducible Markov chain, all states belong to the same class. They are either all transient, or all null recurrent, or all positive recurrent, but in any case the all have the same period.

Remark 5.2.9. In a irreducible chain $p_{jj} > 0$ for some j implies that the chain is not periodic.

Theorem 5.2.10. (Ergodic theorem for Markov chains) Consider a Markov chain X with state space $\mathcal{X}$. Then

1. There exists a stationary distribution if and only if there exists at least one positive recurrent class. In this case all stationary distributions π are such $\pi(j) = 0$ for all j transient or null recurrent;
2. There exists a unique stationary distribution if and only if there exists a unique positive recurrent class C. In this case, for $j \in C$ we have
$$\pi(j) = \frac{1}{\mu_j}. \tag{5.17}$$
Moreover, if
$$f : \mathcal{X} \to \mathbb{R} \tag{5.18}$$
satisfies
$$\sum_{i \in \chi} |f(i)| \pi_i < \infty, \tag{5.19}$$
then
$$\lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} f(X(k)) = \sum_{i \in \mathcal{X}} f(i) \pi_i; \tag{5.20}$$

3. If there are more recurrent classes C_α, let π_a be the distribution on C_a given by
$$\pi_\alpha(j) = \frac{1}{\mu_j}. \tag{5.21}$$
Then all stationary distributions are given by "mixtures" of the π_a.

4. The limit
$$\lim_{n\to\infty} (P^*)^n \pi(0) \tag{5.22}$$
exists independent from $\pi(0)$ if and only if there is a unique positive recurrent, aperiodic class. In this case the sequence $\pi(n) = P^*\pi(n-1) = (P^*)^n \pi(0)$ converges in total variation[1] to π. In detail we get
$$\lim_{n\to\infty} P^n = P^\infty, \tag{5.23}$$
where
$$P^\infty = \begin{bmatrix} \pi_1 & \pi_2 & \cdots & \pi_n \\ \pi_1 & \pi_2 & \cdots & \pi_2 \\ \cdots & \cdots & \cdots & \cdots \\ \pi_1 & \pi_2 & \cdots & \pi_n \end{bmatrix} \tag{5.24}$$
and $\pi = \{\pi_1, \pi_2, \ldots \pi_n\}$ is its stationary distribution.

Proof. Indeed
$$\begin{aligned} P^\infty P &= \left(\lim_{n\to\infty} P^n\right) P \\ &= \lim_{n\to\infty} P^{n+1} \\ &= P^\infty. \end{aligned} \tag{5.25}$$

□

[1] In probability theory, the total variation distance between two probability measures P and Q on a σ-algebra $\mathcal{F}$ is $\sup\{|P(A) - Q(A)| : A \in F\}$. Informally, this is the largest possible difference between the probabilities that the two probability distributions can assign to the same event. For a finite alphabet we can write $\delta(\mathcal{P}, \mathcal{Q}) = \frac{1}{2}\sum_x |P(x) - Q(x)|$. Sometimes the statistical distance between two probability distributions is also defined without the division by two.

Chapter 6

Finite Markov chains

6.1 Conclusions

In this chapter we examine Markov chains in the case of finite state space $\mathcal{X} = \{0, 1, \dots r-1\}$.

Proposition 6.1.1. In a finite Markov chain:

1. there are no null recurrent class;
2. A class is recurrent if and only if has no way to leave it, then all states cannot be transient;
3. there is at least one positive recurrence class.

Proof. First of all, observe that is suffices to prove the result for irreducible Markov chain. Suppose the states are all null recurrent. Then, by (5.14), $p_{ij}^{(n)} \to 0$, $\forall i, j$. By (5.11) the same happens if all states are transient, namely $p_{ij}^{(n)} \to 0$, $\forall i$. The rows of P^n, however, sum to one. Hence, its elements cannot all tend to zero.

□

Thus, after a possible re-enumeration of the states, the transition matrix of a finite Markov chain has necessarily this structure

$$P = \begin{bmatrix} P_1 & 0 & . & . & . & 0 \\ 0 & P_2 & 0 & . & . & 0 \\ . & . & . & . & . & . \\ 0 & . & . & 0 & P_m & 0 \\ A_1 & . & . & . & . & A_{m+1} \end{bmatrix}. \tag{6.1}$$

While P_k and A_{m+1} are square, $A_1, \dots A_m$ need not be. The structure of P^n is similar

$$P^n = \begin{bmatrix} P_1^n & 0 & . & . & . & 0 \\ 0 & P_2^n & 0 & . & . & 0 \\ . & . & . & . & . & . \\ 0 & . & . & 0 & P_m^n & 0 \\ * & . & . & . & . & A_{m+1}^n \end{bmatrix}. \tag{6.2}$$

Example 6.1.2. Consider Example 5.2.7. According to (6.1) we get:

1. $P_1 = \begin{bmatrix} * \end{bmatrix}$;
2. $P_2 = \begin{bmatrix} 0 & * \\ * & 0 \end{bmatrix}$;
3. $P_3 = \begin{bmatrix} 0 & * & * \\ 0 & * & * \\ * & 0 & 0 \end{bmatrix}$;
4. $A_1 = \begin{bmatrix} * & * & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$;
5. $A_2 = \begin{bmatrix} 0 & * & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$;
6. $A_4 = \begin{bmatrix} * & 0 & 0 \\ * & * & * \\ * & 0 & 0 \end{bmatrix}$.

We recall that:

1. there are three closed sets $E_1 = \{0\}$, $E_2 = \{1, 2\}$, $E_3 = \{3, 4, 5\}$
2. state 0 is recurrent and aperiodic;
3. states 1 and 2 are recurrent and periodic of period 2 with mean recurrence time equal to 2;
4. states $3, 4, 5$ are ergodic;
5. states $6, 7, 8$ are transient.

Corollary 6.1.3. As a consequence of Proposition 6.1.1 and the Ergodic theorem 5.2.10, a finite Markov chain always has at least one stationary distribution.

A complement of the Ergodic theorem 5.2.10 is the following:

Theorem 6.1.4. (Perron-Frobenius) Consider a finite Markov chain X with transition matrix P and state space $\mathcal{X} = \{0, 1, \ldots r-1\}$. Suppose there exists an integer $m \geq 1$ such that all elements of P^m are strictly positive. Such a chain is called ergodic. Then the limit

$$\lim_{m\to\infty} p_{jk}^{(m)} = \pi_k : \tag{6.3}$$

1. exists $\forall j, k$;
2. is independent from j;
3. $\pi_k > 0 \, \forall k = 0, 1, \ldots, r-1$;
4. $\pi = P^* \pi$;
5. $\sum_{k=0}^{r-1} \pi_k = 1$;

6.2 Analysis of examples

Let us now apply the results of Chapter 5 to the examples of section 4.3.

6.2.1 Random walk with absorbing barriers

Consider P_a as defined by (4.22). There are three communication classes $C_1 = \{0\}$, $C_2 \{1, 2, \ldots, r-2\}$, $C_3 = \{r-1\}$.

State 0 is the only state which forms C_1 so it is absorbing and aperiodic $\left(p_{11}^{(1)} > 0 \Rightarrow d(0) = 1\right)$. By (5.9) we get it is recurrent and by (5.12) that it is positive ($\mu_1 = 1$). Hence it is ergodic. With a similar analysis we get state $r-1$ is ergodic too ($\mu_{r-1} = 1$).

States which belong to C_2 are periodic of period 2. Indeed $p_{ii}^{(2)} > 0 \Rightarrow d(i) = 2, \forall i : 1 \leq i \leq r-2$. Moreover they are transient indeed $p_{ij}^{(n)} \to 0$, $\forall i : 0 \leq i \leq r-1$, $\forall j : 1 \leq j \leq r-2$.

Hence there are two positive recurrent classes C_1 and C_3. Let $\pi_1(1) = \frac{1}{\mu_1} = 1$ and $\pi_3(r-1) = \frac{1}{\mu_{r-1}} = 1$ the distribution on C_1 and C_3 respectively. By the Ergodic theorem 5.2.10 we get a family of stationary distributions π where $\pi_1 = \lambda, \pi_{r-1} = 1-\lambda, \forall \lambda : 0 \leq \lambda \leq 1$ and $\pi_j = 0$, $\forall j \in C_2$, namely

$$\pi^* = (\lambda, 0, \ldots, 0, 1-\lambda). \tag{6.4}$$

6.2.2 Random walk with totally reflecting barriers

Consider P_r as defined by (4.23) where $p_0 = p_{r-1} = 1$ and by consequence $r_0 = r_{r-1} = 0$. There is only one communication class $C_1 = \{0, 1, \dots r-1\}$ hence the chain is irreducible. Thus it has a unique positive recurrent class C_1.

The chain is periodic of period 2, indeed $p_{ii}^{(2)} > 0$, $\forall i : 0 \leq i \leq r-1$.

Hence by the Ergodic theorem 5.2.10 there exists a unique stationary distribution $\pi^* = \{\pi_0, \dots, \pi_{r-1}\}$.

We consider now the system of equations originating from equation (4.14). We get

$$\begin{aligned}
q \cdot \pi(1) &= \pi(0); \\
\pi(0) + q \cdot \pi(2) &= \pi(1); \\
p \cdot \pi(1) + q \cdot \pi(3) &= \pi(2); \\
&\dots \\
p \cdot \pi(r-4) + q \cdot \pi(r-2) &= \pi(r-3); \\
p \cdot \pi(r-3) + \pi(r-1) &= \pi(r-2); \\
p \cdot \pi(r-2) &= \pi(r-1).
\end{aligned} \tag{6.5}$$

Working our way from the top to the bottom we get

$$\begin{aligned}
\pi(1) &= \frac{1}{q} \cdot \pi(0); \\
\pi(2) &= \frac{p}{q^2}\pi(0); \\
&\dots \\
\pi(r-2) &= \frac{p^{r-3}}{q^{r-2}} \cdot \pi(0); \\
\pi(r-1) &= \left(\frac{p}{q}\right)^{r-2} \pi(0).
\end{aligned} \tag{6.6}$$

By the condition

$$\sum_{k=0}^{r-1} \pi(k) = 1 \tag{6.7}$$

we get

$$\pi(0) \left[1 + \frac{p^{r-2}}{q^{r-2}} + \frac{1}{q} \sum_{k=0}^{r-3} \left(\frac{p}{q}\right)^k\right] = 1. \tag{6.8}$$

The latter yields the following results:

1. if $p = q$ the random walk is *symmetric* and $\pi(0) = \pi(1) = \ldots = \pi(r-1)$ implies $\pi(0) = \frac{1}{r}$, namely π is the uniform distribution.

2. If $p \neq q$ we get

$$\pi(0) = \frac{1-2p}{2} \cdot \frac{\left(\frac{p}{q}\right)^{\frac{2-r}{2}}}{\left(q\left(\frac{p}{q}\right)^{\frac{2-r}{2}} - p\left(\frac{p}{q}\right)^{\frac{r-2}{2}}\right)}. \tag{6.9}$$

and $\pi(j)$, $j = 1, \ldots r-1$ follows (6.5).

6.2.3 Random walk with partially reflecting barriers

Consider P_r as defined by (4.23) where $p_0 = r_{r-1} = p$ and by consequence $r_0 = p_{r-1} = q$. There is only one communication class $C_1 = \{0, 1, \ldots r-1\}$ hence the chain is irreducible.

The chain is aperiodic, indeed $p_{ii}^{(1)} > 0$, $\forall i : 0 \leq i \leq r-1$. Thus it has a unique positive recurrent aperiodic class C_1.

Hence by the Ergodic theorem 5.2.10 there exists a unique stationary distribution $\pi^* = (\pi_0, \ldots, \pi_{r-1})$. Moreover every row of P^∞ tends in total variation to π^*.

We consider now the system of equations originating from equation (4.14). We get

$$\begin{aligned}
q \cdot \pi(0) + q \cdot \pi(1) &= \pi(0); \\
p \cdot \pi(0) + q \cdot \pi(2) &= \pi(1); \\
p \cdot \pi(1) + q \cdot \pi(3) &= \pi(2); \\
&\ldots \\
q \cdot \pi(r-2) + q \cdot \pi(r-1) &= \pi(r-1).
\end{aligned} \tag{6.10}$$

Working our way from the top to the bottom we get

$$\begin{aligned}
\pi(1) &= \frac{p}{q} \cdot \pi(0); \\
\pi(2) &= \left(\frac{p}{q}\right)^2 \pi(0); \\
&\ldots \\
\pi(r-1) &= \left(\frac{p}{q}\right)^{r-1} \pi(0).
\end{aligned} \tag{6.11}$$

By the condition

$$\sum_{k=0}^{r-1} \pi(k) = 1 \tag{6.12}$$

we get

$$\pi(0)\left[1 + \frac{p}{q} + \left(\frac{p}{q}\right)^2 + \ldots + \left(\frac{p}{q}\right)^{r-1}\right] = 1. \tag{6.13}$$

The latter yields the following results:

1. if $p = q$ the random walk is *symmetric* and $\pi(0) = \pi(1) = \ldots = \pi(r-1)$ implies $\pi(0) = \frac{1}{r}$, namely π is the uniform distribution.

2. If $p \neq q$ we get

$$\pi(0) = \frac{1 - \frac{p}{q}}{1 - \left(\frac{p}{q}\right)^r}$$

$$\pi(j) = \left(\frac{p}{q}\right)^j \cdot \pi(0), \ \forall j : 1 \leq j \leq r-1. \tag{6.14}$$

6.2.4 Cyclical random walk

Consider P_c as defined by (4.24). There is only one communication class $C_1 = \{0, 1, \ldots r-1\}$ hence the chain is irreducible.

By Proposition 6.1.1 and theorem 5.2.8 we get the chain is positive recurrent and periodic of period 2, indeed $p_{ii}^{(2)} > 0, \forall i : 0 \leq i \leq r-1$ thus it has a unique positive recurrent class C_1.

Hence by the Ergodic theorem 5.2.10 there exists a unique stationary distribution. Since P_c is doubly stochastic it has the uniform distribution

$$\pi^* = \left(\frac{1}{r}, \ldots, \frac{1}{r}\right) \tag{6.15}$$

as stationary.

6.2.5 The Ehrenfest model of diffusion

Consider P_e as defined by (4.26). There is only one communication class $C_1 = \mathcal{X} = \{0, 1, \ldots, r-1\}$ so the chain is irreducible.

By Proposition 6.1.1 and theorem 5.2.8 we get the chain is positive recurrent and periodic of period 2, indeed $p_{ii}^{(2)} > 0, \forall i : 0 \leq i \leq r-1$ thus it has a unique positive recurrent class C_1.

Hence by the Ergodic theorem5.2.10 there exists a unique stationary distribution $\pi^* = (\pi_0, \ldots, \pi_{r-1})$. It is the binomial (2.4) where $a = b = 1$,

$$\pi_i = \binom{r-1}{i} \cdot 2^{-(r-1)}, \ \forall r : 0 \leq i \leq r-1. \tag{6.16}$$

Remark 6.2.1. The binomial distribution is invariant with $p = \frac{1}{2}$.

This result can be interpreted as follows: whatever the initial number of molecules in the first container A, after a long time the probability of finding k molecules it is nearly the same as if the $r-1$ molecules had been distributed at random, each molecule having probability $\frac{1}{2}$ to be in the first container A.

6.2.6 The Bernoulli-Laplace model of diffusion

Consider P_b as defined by (4.28). This chain is very similar to the Ehrenfest model. There is only one communication class $C_1 = \{0, 1, \ldots, r-1\}$ so the chain is irreducible.

The chain is aperiodic, indeed $p_{ii} > 0, \forall i : 1 \leq i \leq r-2$ thus it has a unique positive recurrent aperiodic class C_1.

Hence by the Ergodic theorem 5.2.10 there exists a unique stationary distribution $\pi^* = (\pi_0, \ldots, \pi_{r-1})$. Moreover every row of P^∞ tends in total variation to π.

It is the hypergeometric distribution with parameters $(r-1, 2r-2, r-1)$.

$$\pi_i = \frac{\binom{r-1}{i}^2}{\binom{2r-2}{r-1}}, \ \forall i : 0 \leq i \leq r-1. \tag{6.17}$$

Remark 6.2.2. This means that in the state of equilibrium the distribution of liquids is the same as if the $r-1$ molecules in it had been chosen at random from a collection of $r-1$ black and $r-1$ white molecules.

6.2.7 Random placements of balls

Consider P_p as defined by (4.30). There are r communication classes $C_i = \{i-1\}, \ \forall i : 1 \leq i \leq r$.

State 0 has period 0 ($p_{00}^{(n)} = 0,\ \forall n \geq 1 \Rightarrow d(0) = 0$) while all the other states are aperiodic ($p_{ii}^{(1)} > 0,\ \forall i : 1 \leq i \leq r-1$).

State $r-1$ is absorbing ($\Rightarrow$ *recurrent*), positive ($\mu_{r-1} = 1$) and aperiodic $d(r-1) = 1$, namely ergodic. All the other states are transient ($p_{ij}^{(n)} \to 0,\ \forall i : 0 \leq i \leq r-1,\ \forall j : 0 \leq j \leq r-2$).

Hence by the Ergodic theorem 5.2.10 there exists a stationary distribution π such that $\pi(r-1) = \frac{1}{\mu_{r-1}} = 1$, $\pi(i) = 0, \forall i : 0 \leq i \leq r-2$, namely

$$\pi^* (0, \ldots, 0, 1). \tag{6.18}$$

6.2.8 Wright-Fisher model

Consider P_w with transition probabilities p_{ij} of the form (4.31). There are three communication classes $C_1 = \{0\}$, $C_2 \{1, 2, \ldots r-2\}$ and $C_3 = \{r-1\}$.

All states are aperiodic ($p_{ii}^{(1)} > 0,\ \forall i : 0 \leq i \leq r-1$).

States 0 and $r-1$ are absorbing ($\Rightarrow$ *recurrent*), positive ($\mu_0 = \mu_{r-1} = 1$) and aperiodic, namely ergodic. All the other states are transient ($p_{ij}^{(n)} \to 0,\ \forall i : 0 \leq i \leq r-1,\ \forall j : 1 \leq j \leq r-2$). Hence by the Ergodic theorem 5.2.10 we get a family of stationary distributions π where $\pi_1 = \lambda,\ \pi_{r-1} = 1-\lambda,\ \forall \lambda : 0 < \lambda < 1$ and $\pi_j = 0,\ \forall j \in C_2$, namely

$$\pi^* = (\lambda, 0, \ldots, 0, 1-\lambda). \tag{6.19}$$

Chapter 7

Analysis of examples with countable state space

Let us now apply the results of Chapter 5 to the examples of Section 4.4.

7.1 Success runs

Consider P_s as defined by (4.33). There is only one communication class $C_1 = \{0, 1, \ldots\}$ since all states can be reached through the zero state. Hence the chain is clearly irreducible.

By theorem 5.2.8 all states belong to the same class: they are either all transient or all recurrent. By Remark 5.2.9 we can also state that the chain is aperiodic ($p_{00}^1 = 1 \Rightarrow d(0) = 1$).

Hence it suffices to analyze the nature of one state to classify the chain. Let us compute $f_{00}^{(n)}$, namely the probability of reaching state 0, starting from state 0 in exactly n steps.

We get the geometric distribution

$$f_{00}^{(n)} = (1-q)^{n-1} \cdot q, \; n \geq 1 \tag{7.1}$$

where $p = 1 - q$ by definition. Hence by (5.8) state 0 is transient if $f_{00} < 1$, if $f_{00} = 1$ recurrent. In our case

$$f_{00} = \sum_{n=1}^{\infty} (1-q)^{n-1} \cdot q = 1 \tag{7.2}$$

where $0 < q < 1$.

Moreover state 0 is positive recurrent. The mean recurrence time μ_0 get by (5.12) coincides with the expected value of the geometric distribution, namely $\frac{1}{q}$.

Hence by the Ergodic theorem 5.2.10 there exists a unique stationary distribution $\pi^* = \{\pi_0, \pi_1, \ldots\}$ and every row of P^∞ tends in total variation to π^*.

Let us describe the structure of π. In general

$$\pi_k = p \cdot \pi_{k-1},\ k = 1, 2, \ldots \tag{7.3}$$

$$\pi_0 = q \cdot \pi_0 + q \cdot \pi_1 + \ldots\ . \tag{7.4}$$

From (7.3) we get

$$\pi_k = p^k \cdot \pi_0 \tag{7.5}$$

that is useful to prove the existence of the stationary distribution.

Proof. It is seen that the first k terms on the right in (7.4) add to $\pi_0 - \pi_k$. Thus by (7.3) is automatically satisfied whenever $\pi_k \to 0$, which is our case because

$$\sum_{k=1}^{\infty} p^k = \frac{p}{(1-p)} < \infty, \tag{7.6}$$

where $0 < p < 1$.

□

7.2 Random walk on $\mathbb{Z}$

Consider P_{rw} as defined by (4.35). There is only one communication class, thus the chain is irreducible. $p_{ii}^{(2)} > 0$, $\forall i \in \mathbb{Z}$ implies the chain is periodic of period 2. By theorem 5.2.8 all states belong to the same class: they are either all transitive or all recurrent.

Hence let us compute f_{ij}, namely the probability, starting from i, of reaching j. We have

$$f_{ij} = q \cdot f_{(i-1)j} + p \cdot f_{(i+1)j},\ \forall i, j \in \mathbb{Z}. \tag{7.7}$$

Since

$$q\left(\frac{q}{p}\right)^{i-2} + p\left(\frac{q}{p}\right)^{i} = \left(\frac{q}{p}\right)^{i-1}, \tag{7.8}$$

we get $\left(\frac{q}{p}\right)^{i-1}$ satisfies (7.7).

We notice that if $p > q$,

$$\begin{cases} \left(\frac{q}{p}\right)^{i-1} = 1 & i \leq 1 \\ 0 < \left(\frac{q}{p}\right)^{i-1} < 1 & otherwise \end{cases}. \tag{7.9}$$

On the other hand if $p < q$

$$\begin{cases} \left(\frac{q}{p}\right)^{i-1} = 1 & i \geq 1 \\ 0 < \left(\frac{q}{p}\right)^{i-1} < 1 & otherwise \end{cases}, \tag{7.10}$$

while if $p = q$, $\left(\frac{q}{p}\right)^{i-1} = 1$, $\forall i \in \mathbb{Z}$.

Hence the chain is recurrent if $p = q$, transient otherwise.

We consider now the system of equations originating from equation (4.14), where $p = q$. We get

$$\begin{aligned} &\dots \\ \frac{1}{2} \cdot \pi(-1) + \frac{1}{2} \cdot \pi(0) &= \pi(0); \\ \frac{1}{2} \cdot \pi(0) + \frac{1}{2} \cdot \pi(2) &= \pi(1); \\ &\dots \end{aligned} \tag{7.11}$$

In general

$$\pi(j) = \frac{1}{2} \cdot \pi(j-1) + \frac{1}{2} \cdot \pi(j+1), \ \forall j \in \mathbb{Z}. \tag{7.12}$$

All states are null recurrent, thus there is no stationary distribution.

7.3 Random walk on $\mathbb{Z}_0^+$

Consider P_{rw} as defined by (4.36). There is only one communication class, thus the chain is irreducible. $p_{00}^{(2)} > 0$, namely the chain is periodic of period 2. By theorem 5.2.8 all states belong to the same class: they are either all transitive or all recurrent.

Hence let us compute f_{i1}, namely the probability, starting from i, of reaching 1. We have

$$f_{i1} = q \cdot f_{(i-1)1} + p \cdot f_{(i+1)1}, \ \forall i > 1. \tag{7.13}$$

Since

$$q\left(\frac{q}{p}\right)^{i-2}+p\left(\frac{q}{p}\right)^{i}=\left(\frac{q}{p}\right)^{i-1}, \tag{7.14}$$

we get $\left(\frac{q}{p}\right)^{i-1}$ satisfies (7.7).

Let us now distinguish two cases:

1. if $p>q$, $f_{i1}<1$ hence all states are transient and there is no stationary distribution;
2. If $p\leq q$, $f_{i1}=1$, $\forall i\geq 1$ hence all states are recurrent.

We consider now the system of equations originating from equation (4.14). We get

$$\begin{aligned} q\cdot\pi(1)&=\pi(0);\\ \pi(0)+q\cdot\pi(2)&=\pi(1);\\ p\cdot\pi(1)+q\cdot\pi(3)&=\pi(2);\\ p\cdot\pi(2)+q\cdot\pi(4)&=\pi(3);\\ &\cdots \end{aligned} \tag{7.15}$$

Working our way from the top to the bottom we get

$$\begin{aligned} \pi(1)&=\frac{p}{q}\cdot\pi(2);\\ \pi(2)&=\frac{p}{q}\cdot\pi(3);\\ &\cdots \end{aligned} \tag{7.16}$$

Hence

$$\pi(j)=\frac{p}{q}\pi(j-1),\ j=2,3,\ldots \tag{7.17}$$

If $p=q$, we get $\pi(1)=\pi(2)=\ldots$ and there is no stationary distribution. In this case there is no stationary distribution.

If $p<q$, all states are positive recurrent and the following π is the unique stationary distribution.

By the condition

$$\sum_{k=0}^{\infty}\pi(k)=1 \tag{7.18}$$

we get

$$\pi(0)\left[q+1+\frac{p}{q}+\left(\frac{p}{q}\right)^{2}+\ldots\right]=1. \tag{7.19}$$

The latter gives

$$\pi(1) = \frac{q-p}{2q^2}. \tag{7.20}$$

In conclusion

$$\begin{aligned} \pi(0) &= \frac{q-p}{2q^2} \\ \pi(j) &= \frac{q-p}{2q^2}\left(\frac{p}{q}\right)^{j-1},\ j = 1, 2, \ldots. \end{aligned} \tag{7.21}$$

Chapter 8

Maximum entropy problems

8.1 Entropy

Consider a function of a real variable $h(x) = -x \cdot \ln x$ on $x > 0$. Notice that

$$\lim_{x \searrow 0} h(x) = 0. \tag{8.1}$$

Hence, we can extend the domain of h to all $x \geq 0$. Let p the probability distribution on the sample space $\Omega = \{\omega_1, \ldots \omega_n\}$. Then, the (Shannon) *entropy* of p is defined by

$$H(p) := \sum_{i=1}^{n} h(p(\omega_i)). \tag{8.2}$$

The base of the logarithm is not important. Entropy represents the degree of indeterminacy of p, our ignorance of the state of a system.

8.2 Convex functions

Let V be a vector space and $K \subseteq V$. The set K is *convex* if whenever $x, y \in K$, then $[x, y] \subseteq K$, where

$$[x, y] = \{z : z = \lambda x + (1 - \lambda) y,\ 0 \leq \lambda \leq 1\}. \tag{8.3}$$

Remark 8.2.1. The intersection of convex subsets of V is convex.

Let $K \subseteq V$ be convex, $f : K \to \mathbb{R}$. The function f is said to be *convex* it satisfies

$$f(\lambda x + (1 - \lambda) y) \leq \lambda f(x) + (1 - \lambda) f(y),\ \forall x, y \in K,\ \forall \lambda \in [0, 1]. \tag{8.4}$$

The function f is *strictly convex* on K if it satisfies (8.4) with strict inequality $\forall x, y : x \neq y, \lambda \in (0,1)$. The function f is called (strictly) *concave* if $-f$ is (strictly) *convex.*

Theorem 8.2.2. (Jensen 1906) Let K be a convex set of the vector space V, and let $f : K \to \mathbb{R}$. The following properties are equivalent:

1. f is convex;
2. $epif := \{(x, \alpha) \in K \times \mathbb{R} | f(x) \leq \alpha\}$ is a convex subset of $V \times \mathbb{R}$;
3. For all $n \in \mathbb{N}$, for all $(x_1, \ldots, x_n) \in K^n$ and for all $(\lambda_1, \ldots \lambda_n) \in [0,1]^n$ such that $\sum_{i=1}^{n} \lambda_i = 1$ we have the Jensen inequality

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leq \sum_{i=1}^{n} \lambda_i f(x_i) \tag{8.5}$$

Remark 8.2.3. Along the same lines as theorem 8.2.2, one can show that f is strictly convex if and only if (8.5) holds with strict inequality whenever $\sum_{i=1}^{n} \lambda_i x_i \neq x_j$, $\forall j$.

Corollary 8.2.4. Let K be a convex subset of the vector space V and let $f_\alpha : K \to (-\infty, +\infty)$, $\alpha \in I$, be a collection of convex functions. The upper *hull* of the collection is defined by

$$g(x) = \sup \{f_\alpha(x) | \alpha \in I\}, \ x \in K. \tag{8.6}$$

Then g is convex.

Proof. Observe that

$$epi\ g = \cap_\alpha epi\ f_\alpha.$$

Since each $epi\ f_\alpha$ is convex, so is their intersection.

□

Let V be a vector space, and let $A \subseteq X$. Then *convex hull* of A, written $conA$, is the intersection of all convex subsets of V containing A. The convex hull of n *affinely independent*[1] points of a Euclidean space is called an $(n-1)$-simplex.

[1]The points $x_1, \ldots x_n$ are called affinely independent if every point x in their convex hull admits a unique representation as convex combination of the points.

8.3 The simplex of probability distributions

Let $\mathcal{D}(\Omega)$ denote the family of all probability distributions on the sample space $\Omega = \{\omega_1, \ldots, \omega_n\}$. Then $\mathcal{D}(\Omega)$ is a $(n-1)$-simplex whose vertices are the singular distributions $p_i(\omega_j) = \delta_{ij}$, where δ_{ij} is the Kronecker delta. The latter distributions correspond to situations where no randomness is present.

Theorem 8.3.1. The entropy function H is strictly concave on $\mathcal{D}(\Omega)$. Moreover, it satisfies

$$0 \leq H(p) \leq \ln n. \tag{8.7}$$

In particular, $H(p) = 0$ if and only if p is a vertex of the simplex and $H(p) = n$ if and only if $p = p_u$ is the center of the simplex, where $p_u(\omega_i) \equiv \frac{1}{n}$.

8.4 Examples

8.4.1 A finite Gibbs variational principle

Consider a physical system completely described by the discrete state space $\mathcal{X} = \{1, 2 \ldots, n\}$. We can think of this mesoscopic description as originating from a microscopic description where the *phase space* $\mathbf{\Gamma}$ has undergone a "coarse graining" through subdivision into small cells.

Each of the cells represents a mesoscopic state. For each state i we consider its *energy* E_i. The function $\mathcal{H} : i \to E_i$ is called *Hamiltonian.* The thermodynamic states of the system are given by the simplex of probability distributions on $\mathcal{X}$, namely $\mathcal{S} := \mathcal{D}(\mathcal{X})$.

On $\mathcal{S}$, we define the *internal energy* as the expected value of the Energy *observable* in state p

$$U(E, p) = \sum_i E_i p_i, \tag{8.8}$$

where E denotes the n-dimensional vector with components E_i. By (8.2) let us also introduce the *Gibbs entropy*

$$S(p) = kH(p) = -k \sum_i p_i \ln p_i, \tag{8.9}$$

where k is Boltzmann's constant.

By theorem 8.3.1, S is nonnegative and strictly concave on $\mathcal{S}$. Let $\overline{E}$ be a constant satisfying

$$E_m = \min_i E_i \leq \overline{E} \leq \frac{1}{n}\sum_i E_i. \tag{8.10}$$

We can think of $\overline{E}$ as the energy of the underlying conservative microscopic system. Hence we want to study the following *Maximum Entropy* problem:

$$maximize\ \{S(p)\,;p \in \mathcal{S}\} \tag{8.11}$$

$$subject\ to\ U(E,p) = \overline{E}. \tag{8.12}$$

This is an important instance of a class of maximum entropy problems where entropy is maximized over probability distributions which give the correct expectation of certain observables in accordance with known macroscopic quantities.

In order to solve problem (8.11)-(8.12) we resort to a fundamental, although elementary, result.

Let Y be a nonempty set and let $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$ denote the extended reals. Consider the maximization of $J : Y \to \overline{\mathbb{R}}$ over the nonempty subset M of Y.

Definition 8.4.1. The map $\Lambda : Y \to \overline{\mathbb{R}}$ is called a *Lagrange functional* for the optimization problem if it is *finite* and *constant* over M.

Lemma 8.4.2. (*Lagrange Lemma*) Let $\Lambda : Y \to \overline{\mathbb{R}}$ be a Lagrange Functional and let $y_0 \in M$ maximize $\mathcal{L} = J + \Lambda$ over Y. Then y_0 maximize J over M.

Proof. For any $y \in M$, we have $J(y_0) + \Lambda(y_0) \geq J(y) + \Lambda(y) = J(y) + \Lambda(y_0)$. Hence $J(y_0) \geq J(y)$.

□

This is an incredible result. Indeed it does not require any algebraic nor topological structure on Y and the hypotheses on Λ are also minimal. Now we can attack problem (8.11)-(8.12).

Let us introduce the positive orthant $\mathbb{R}^n_+ = \{p = (p_1, p_2 \ldots, p_n)\,|p_i \geq 0,\ i = 1, 2, \ldots, n\}$. In our setting, $\mathbb{R}^n_+ = Y$, $S = Y$ and

$$M = \left\{ p = (p_1, p_2 \ldots, p_n)\,|p_i \geq 0,\ i = 1, 2, \ldots, n,\ \sum_i p_i = 1,\ U(E,p) = \overline{E} \right\}. \tag{8.13}$$

We take

$$\Lambda(p) = \lambda\left(\overline{E} - U(E,p)\right) + \mu\left(\sum_i p_i - 1\right), \tag{8.14}$$

where $\lambda, \mu \in \mathbb{R}, \lambda \geq 0$, are called *Lagrange multipliers.* Since $\Lambda(p) \equiv 0$ on M, it is a Lagrange functional for our problem. The *Lagrangian function* is given by

$$\mathcal{L}(p,\lambda,\mu) := S(p) + \lambda\left(\overline{E} - U(E,p)\right) + \mu\left(\sum_i p_i - 1\right). \tag{8.15}$$

By Lemma 8.4.2, we consider the *unconstrained* maximization of $\mathcal{L}(\cdot,\lambda,\mu)$ over $\mathbb{R}^n_+$. Observing that $\lambda\overline{E}$ and μ are constants, we can equivalently maximize over $\mathbb{R}^n_+$ the functional

$$I(p) = -k\sum_i p_i \ln p_i - \lambda\sum_i E_i p_i + \mu\sum_i p_i. \tag{8.16}$$

Observe that

$$I(p) = \sum_i f_i(p_i),\ f_i(p_i) = (-k\ln p_i - \lambda E_i + \mu)\, p_i. \tag{8.17}$$

Hence, the problem is equivalent to maximize each $f_i(p_i)$ over $\mathbb{R}_+$. Observe that $f_i(p_i)$ is strictly concave on $\mathbb{R}_+$. Hence the vanishing of f_i' is a *sufficient* condition for a maximum point. Setting $f_i'(p_i) = 0$, $\forall i$ we get

$$-k\ln p_i - k - \lambda E_i + \mu = 0. \tag{8.18}$$

It yields

$$p_i^* = \exp\left(-1 + \frac{\mu}{k} - \frac{\lambda}{k}E_i\right). \tag{8.19}$$

Since each $p_i^* \geq 0$, $p^* = (p_1^*, p_2^*, \ldots, p_n^*)$ is the maximum point of $I(p)$. In order for p^* to solve the original constrained problem, it must lie in M.

First of all we worry about condition $\sum_i p_i^* = 1$. We can choose μ such that

$$\exp\left(1 - \frac{\mu}{k}\right) = \sum_i \exp\left(-\frac{\lambda}{k}E_i\right) := Z\left(\frac{\lambda}{k}\right). \tag{8.20}$$

Z is called *partition function* in Statistical Mechanics.

Define the *absolute temperature* $T := \frac{1}{\lambda}$, $\lambda > 0$ and $T = +\infty$, $\lambda = 0$. Define also the so-called *inverse temperature* β as

$$\beta := \frac{\lambda}{k} = \frac{1}{kT}. \tag{8.21}$$

We can then rewrite (8.19)-(8.20) as

$$p_i^* = Z(\beta)^{-1} \exp(-\beta E_i), \quad Z(\beta) = \sum_i \exp(-\beta E_i), \tag{8.22}$$

which is called *Gibbs distribution* and corresponds to the equilibrium thermodynamical state at the temperature T.

Then we impose that (8.22) satisfies (8.12). Hence we get

$$\sum_i E_i p_i^*(\beta) = \sum_i E_i Z(\beta)^{-1} \exp(-\beta E_i) := G(\beta) = \overline{E}. \tag{8.23}$$

Observe that if E_i are all equal, then $E_i = \overline{E}$, $\forall i$ and (8.23) is satisfied for any β. Let us now consider the case when the E_i are not all equal.

Theorem 8.4.3. Assume that $\mathcal{H}$ is not constant. Then the function $G : \beta \to G(\beta)$ in (*8.23*) is *strictly decreasing* on $\beta \geq 0$, bijectively mapping $[0, +\infty]$ onto $(E_m, \frac{1}{N}\sum_i E_i]$. Hence there does exists a continuously differentiable inverse function $\beta = G^{-1}\left(\overline{E}\right)$.

Remark 8.4.4. The above result shows that, given the value of the internal energy

$$\overline{E} \in \left[E_m, \frac{1}{n}\sum_i E_i\right], \tag{8.24}$$

there exists one and only one value of the absolute temperature T which solves problem (8.11)-(8.12).

8.4.2 Maximum entropy problems with a given expectation

Let $\mathcal{X} = \mathbb{N}$ and let $\mathcal{D}(\mathcal{X})$ be the simplex of probability distributions on $\mathcal{X}$. Suppose now we wish to solve

$$maximize \ \{H(p); p \in \mathcal{D}(\mathcal{X})\} \tag{8.25}$$

$$subject\ to \sum_{i=1}^{\infty} i \cdot p_i = c. \tag{8.26}$$

It is namely the problem of finding the maximum entropy distribution among those with a given expectation.

Arguing as for the Gibbs distribution in 8.4.1, we get $Y = H$, $\mathbb{R}^n_+ = \{p = (p_0, p_1, \ldots) \,|p_i \geq 0,\ i = 0, 1, \ldots\}$ and

$$M = \left\{ p = (p_0, p_1,) \,|p_i \geq 0,\ i = 0, 1, \ldots \ \sum_i p_i = 1,\ \sum_i i p_i = c \right\}. \tag{8.27}$$

Hence we take

$$\Lambda(p) = \lambda \left(c - \sum_i i p_i \right) + \mu \left(\sum_i p_i - 1 \right), \tag{8.28}$$

as Lagrangian functional, where $\Lambda(p) \equiv 0$ on M and $\lambda, \mu \in \mathbb{R}, \lambda \geq 0$.

The Lagrangian function is given by

$$\mathcal{L}(p, \lambda, \mu) := H(p) + \lambda \left(c - \sum_i i p_i \right) + \mu \left(\sum_i p_i - 1 \right). \tag{8.29}$$

By Lemma 8.4.2, we consider the unconstrained maximization of $\mathcal{L}(\cdot, \lambda, \mu)$ over $\mathbb{R}^n_+$.

We again observe that λc and μ are constants. Thus we maximize over $\mathbb{R}^n_+$ each $f_i(p_i)$ of the functional

$$I(p) = \sum_i f_i(p_i),\ f_i(p_i) = (-\ln p_i - \lambda i + \mu) p_i. \tag{8.30}$$

Because $f_i(p_i)$ is strictly concave on $\mathbb{R}_+$, by setting $f_i'(p_i) = 0$ we get

$$p_i^* = \exp(-1 + \mu) \exp(-\lambda i). \tag{8.31}$$

We deal with condition $\sum_i p_i^* = 1$. We can choose μ such that

$$\sum_{i=0}^{\infty} (\exp(-\lambda))^i = \exp(1 - \mu). \tag{8.32}$$

Let $\lambda > 0$ then we deal with a convergent geometric series, namely

$$\sum_{i=0}^{\infty} (\exp(-\lambda))^i = \frac{1}{1 - \exp(-\lambda)},\ \lambda > 0. \tag{8.33}$$

On the other hand, concerning condition $\sum_{i=0}^{\infty} i p_i = c$ we get

$$\sum_{i=0}^{\infty} i \cdot \exp(-\lambda i)(1 - \exp(-\lambda)) = c. \tag{8.34}$$

We recall that

$$\sum_{i=0}^{\infty} i \cdot p^i = \frac{p}{(1-p)^2},\ |p| < 1. \tag{8.35}$$

Hence with $\lambda > 0$, (8.34) becomes

$$(1 - \exp(-\lambda)) \sum_{i=0}^{\infty} i \cdot \exp(-\lambda i) = c$$

$$\frac{\exp(-\lambda)}{1 - \exp(-\lambda)} = c, \tag{8.36}$$

that is to say

$$\lambda = \ln \frac{c+1}{c}. \tag{8.37}$$

By (8.32) we get

$$\mu = 1 - \ln(c+1). \tag{8.38}$$

This time consider problem (8.25)-(8.26) where $\mathcal{X}$ is a subset on $\mathbb{N}$, namely $\mathcal{X} = \{1, 2, \ldots 6\}$. Moreover $\mathcal{D}(\mathcal{X})$ is still the simplex of probability distributions on $\mathcal{X}$ but (8.26) becomes

$$\sum_{i=1}^{6} i \cdot p_i = \alpha, \tag{8.39}$$

where $\alpha > 0$. This is namely the case of tossing a die.

Under the same assumption made for the previous problem we get $Y = H$, $\mathbb{R}^n_+ = \{p = (p_1, p_2, \ldots, p_6) \,|p_i \geq 0,\ i = 1, 2, \ldots, 6\}$ and

$$M = \left\{ p = (p_1, p_2, \ldots p_6) \,|p_i \geq 0,\ i = 1, 2, \ldots, 6 \sum_i p_i = 1,\ \sum_{i=1}^{6} i p_i = \alpha \right\}. \tag{8.40}$$

Hence we can maximize

$$I(p) = \sum_{i=1}^{6} f_i(p_i),\ f_i(p_i) = (-\ln p_i - \lambda i + \mu) p_i. \tag{8.41}$$

which leads to (8.31) again. However we now deal with a finite summation.

Then from constraint $\sum_i p_i^* = 1$ we get the following relations:

$$\exp(1 - \mu) = \frac{\exp(-\lambda) - \exp(-7\lambda)}{1 - \exp(-\lambda)} \tag{8.42}$$

$$p_i^* = \exp(-\lambda \cdot i) \frac{1 - \exp(-\lambda)}{\exp(-\lambda) - \exp(-7\lambda)}. \tag{8.43}$$

In order to lie in M we must choose λ to satisfy constraint (8.39). Hence we get

$$\sum_{i=1}^{6} i \cdot \exp(-\lambda \cdot i) \frac{\exp(\lambda) - 1}{1 - \exp(-6\lambda)} = \alpha \tag{8.44}$$

or equivalently

$$(6 - \alpha) \exp(-7\lambda) + (\alpha - 7) \exp(-6\lambda) + \alpha \exp(-\lambda) + 1 - \alpha = 0. \tag{8.45}$$

Suppose $\alpha = 4.5$. The numerical results are

$$\{p_1^*, \ldots, p_6^*\} = \{0.0543,\ 0.0788,\ 0.1142,\ 0.1654,\ 0.2398,\ 0.3475\}, \tag{8.46}$$

where $\lambda = -0.37105$ and $\mu = -2.28330$. Hence $H_{max} = 1.61358$.

Remark 8.4.5. When $\alpha = \frac{7}{2}$, namely the dice is fair, we get $\{p_1^*, \ldots, p_6^*\} = \{\frac{1}{6},\ \frac{1}{6},\ \frac{1}{6},\ \frac{1}{6},\ \frac{1}{6},\ \frac{1}{6}\}$, $\lambda = 0$ and $\mu = -0.7918$.

Chapter 9

Schrödinger bridges

9.1 Martingales

9.1.1 Martingales and submartingales

Definition 9.1.1. Consider a discrete time stochastic process $X = \{X(n) : n \geq 0\}$ with finite or denumerable state space $\mathcal{X}$. The process $Y = \{Y(n) : n \geq 0\}$ is called a *martingale* with respect to X if

1. $E(|Y(n)|) < \infty$, $n \geq 0$;
2. $Y(n)$ is a function of $\{X(k) : 0 \leq k \leq n\}$, $\forall n \geq 0$;
3. $E(Y(n+1)|X(0), X(1), \ldots, X(n)) = Y(n)$.

It is called a *supermartingale* with respect to X when it satisfies 1., 2. and

$$E(Y(n+1)|X(0), X(1), \ldots, X(n)) \geq Y(n). \tag{9.1}$$

We can say that a martingale is *conditionally constant* and a submartingale is *conditionally increasing*. The process is called *supermartingale* when $-Y(n)$ is a martingale. Moreover we notice that the case $Y(n) = X(n)$ is also included.

Proposition 9.1.2. Let Y a martingale with respect to X. Let φ a convex function and define

$$Z(n) := \varphi(Y(n)), \ n \geq 0. \tag{9.2}$$

If

$$E(|Z(n)|) < \infty, \ \forall n \geq 0, \tag{9.3}$$

then Z is a submartingale with respect to X.

9.1.2 Space-time harmonic functions

Consider now a discrete time Markov chain X with state space $\mathcal{X}$ and chain transition matrix P. There exists a large class of X-martingales that are constructed as *instantaneous functions* of $X(n)$.

Definition 9.1.3. A function

$$h : \mathbb{N} \times \mathcal{X} \to \mathbb{R} \tag{9.4}$$

is called *space-time harmonic* if it satisfies the backward equation

$$h(n,i) = \sum_j p_{ij} h(n+1,j), \ \forall n \geq 0, \ \forall i,j \in \mathcal{X}. \tag{9.5}$$

Proposition 9.1.4. Let h be space-time harmonic for the Markov chain X. Define the stochastic process

$$Y = \{Y(n) = h(n, X(n)), \ n \geq 0\}. \tag{9.6}$$

If

$$E(|Y(n)|) < \infty, \ \forall n, \tag{9.7}$$

then Y is a martingale with respect to X.

9.2 Information divergence

Definition 9.2.1. Let p and q belong to the simplex of probability distributions on a finite or countably infinte $\mathcal{X}$. We say that the *support* of p is contained in the support of q if $q_i = 0 \Rightarrow p_i = 0$ and write $Supp(p) \subseteq Supp(q)$. The *(Information) Divergence* or *Kullback-Leibler Index* or *Relative Entropy* of q from p is defined to be

$$\mathbb{D}(p||q) = \begin{cases} \sum_i p(i) \log \frac{p(i)}{q(i)}, & Supp(p) \subseteq Supp(q), \\ +\infty & Supp(p) \nsubseteq Supp(q). \end{cases} \tag{9.8}$$

where, by definition, $0 \cdot \log 0 = 0$.

It enjoys the following properties:

1. $\mathbb{D}(p||q) \geq 0$;
2. $\mathbb{D}(p||q) = 0$ if and only if $p = q$.

9.3 Schrödinger bridges

Consider a Markov chain X with finite state space $\mathcal{X} = \{0, 1, \ldots, r-1\}$, transition probabilities $(\pi_{ij}(n))$ and marginal probabilities $\pi_i(n)$. We denote by $\mathbf{\Pi}_0^T$ the joint distribution of $\{X(0), X(1), \ldots, X(N)\}$ on $\mathcal{X}^{N+1}$. Suppose we can estimate the marginal distribution at time N, but we find a distribution p^1 which differs from $\pi(N) = (\Pi^*)^{N+1}\pi(0)$.

Let $\mathcal{D}(0, N; p^1)$ be the family of Markovian distributions P_0^N on $\mathcal{X}^{N+1}$ that have marginal p^1 at time N and have support contained in the support of Π_0^N, namely $Supp\left(P_0^N\right) \subseteq Supp\left(\Pi_0^N\right)$.

We want to find another Markovian distribution P_0^N on $\mathcal{X}^{T+1}$ which has the observed marginal p^1 at time N and is "as close as possible" to "the prior" distribution Π_0^N. We consider namely the following

Maximum Entropy Problem (MEP1):

$$minimize\ \left\{\mathbb{D}\left(P_0^N||\Pi_0^N\right);\ P_0^N \in \mathcal{D}\left(0, N; p^1\right)\right\}. \tag{9.9}$$

Remark 9.3.1. The constraint on the final marginal in Problem 9.9, may be expressed as a *linear* constraint on P_0^N as

$$\sum_{i_0}\sum_{i_1}\cdots\sum_{i_{T-1}} P_0^N(i_0, i_1, \ldots, i_N) = p^1(i_N). \tag{9.10}$$

In order to facilitate the solution of Problem 9.9, we introduce the reverse time transition probabilities $q_{ij}(n)$ (4.16) corrisponding to P_0^N and $q_{ji}^{\pi}(n)$ corrisponding to Π_0^N. Hence we get

$$P(i_0, i_1, \ldots i_N) = q_{i_1 i_0}(0) \cdot q_{i_2 i_1}(1) \cdots q_{i_N i_{N-1}}(N-1) \cdot p^1_{i_N}, \tag{9.11}$$

$$\Pi(i_0, i_1, \ldots i_N) = q^{\pi}_{i_1 i_0}(0) \cdot q^{\pi}_{i_2 i_1}(1) \cdots q^{\pi}_{i_N i_{N-1}}(N-1) \cdot \pi^1_{i_T}(N). \tag{9.12}$$

Lemma 9.3.2. Let $p_i(n)$ denote the marginals of P at time n. Then

$$\mathbb{D}\left(P_0^N||\Pi_0^N\right) = \sum_{k=1}^{N}\sum_{i_k}\mathbb{D}\left(q_{i_k i_{k-1}}(k-1)\,||q^{\pi}_{i_k i_{k-1}}(k-1)\right)p_{i_k}(k) + \mathbb{D}\left(p^1||\pi(N)\right). \tag{9.13}$$

Theorem 9.3.3. A solution to Problem 9.21 is given by the distribution $\hat{P}_0^N$ corresponding to the Markov chain with marginal distribution p^1 at time N and reverse time transition mechanism equal to that of Π_0^N, namely

$$\hat{q}_{i_k i_{k-1}}(k-1) = q^{\pi}_{i_k i_{k-1}}(k-1),\ k = 1, 2, \ldots N. \tag{9.14}$$

Proof. Since both terms in (9.13) are nonnegative, and $\mathbb{D}\left(p^1||\pi\left(N\right)\right)$ is invariant over $\mathcal{D}\left(0,N;p^1\right)$, the best we can hope for, when minimizing $\mathbb{D}\left(P_0^N||\Pi_0^N\right)$, is to make the first term equal to zero. This is the case if (9.14) holds true.

□

Let us compute the forward transition probabilities of $\hat{P}_0^N$. Let $\hat{p}_i\left(n\right)$ and $\hat{p}_{ij}\left(n\right)$ denote the marginal at time n and the forward transition probabilities of $\hat{P}_0^N$, respectively. By (4.16), we get

$$\hat{p}_i\left(n\right)\hat{p}_{ij}=\hat{p}_j\left(n+1\right)\hat{q}_{ji}\left(n\right),\ \pi_i\left(n\right)\pi_{ij}\left(n\right)=\pi_j\left(n+1\right)q_{ji}^{\pi}\left(n\right). \tag{9.15}$$

Assume now that $\pi_i\left(n\right)>0$, $\hat{p}_i\left(n\right)>0$, $\forall i,n:0\leq n\leq N$. Then (**??**) and (9.14) yield

$$\hat{p}_{ij}=\frac{\varphi\left(n+1,j\right)}{\varphi\left(n,i\right)}\pi_{ij},\ \varphi\left(t,i\right):=\frac{\hat{p}_i\left(n\right)}{\pi_i\left(n\right)}. \tag{9.16}$$

Observe, moreover that φ is space-time harmonic with respect to the transition mechanism of Π since, by (9.16),

$$\sum_j\pi_{ij}\varphi\left(n+1,j\right)=\sum_j\hat{p}_{ij}\varphi\left(n,i\right)=\varphi\left(t,i\right). \tag{9.17}$$

Hence, we can say that the optimal solution to Problem 9.9 is obtained from the *a priori* Markov chain Π_0^T through a "multiplicative functional" transformation.

Consider now the case where p^0, the marginal at time 0 is fixed.

Let p_{ij} denote the transition probabilities corrisponding of P_0^N. This time using

$$P\left(i_0,i_1,\ldots i_N\right)=p_{i_0}^0\cdot p_{i_0i_1}\left(0\right)\cdot p_{i_1i_2}\left(1\right)\cdots p_{i_{N-2}i_{N-1}}\left(N-1\right), \tag{9.18}$$

$$\Pi\left(i_0,i_1,\ldots i_N\right)=\pi_{i_0}\left(0\right)\cdot\pi_{i_0i_1}\left(0\right)\cdots\pi_{i_{N-1}i_N}\left(N-1\right), \tag{9.19}$$

similarly to Lemma 9.3.2 we get the following representation of $\mathbb{D}\left(P_0^N||\Pi_0^N\right)$:

$$\mathbb{D}\left(P_0^N||\Pi_0^N\right)=\mathbb{D}\left(p^0||\pi\left(N\right)\right)+\sum_{k=0}^{N-1}\sum_{i_k}\mathbb{D}\left(p_{i_ki_{k+1}}\left(k\right)||\pi_{i_ki_{k+1}}\left(k\right)\right)p_{i_k}\left(k\right). \tag{9.20}$$

Consider now the

Maximum Entropy Problem (MEP2):

$$maximize\ \left\{\mathbb{D}\left(P_0^N||\Pi_0^N\right);\ P_0^N \in \mathcal{D}\left(0,N;p^0\right)\right\}. \tag{9.21}$$

We get the following result.

Theorem 9.3.4. A solution to Problem 9.21 is given by the distribution $\hat{P}_0^N$ corresponding to the Markov chain with marginal distribution p^0 at time 0 and forward transition mechanism equal to that of Π_0^N, namely

$$\hat{p}_{ij}(n) = \pi_{ij}(n),\ n = 0,1,\dots N. \tag{9.22}$$

Let us compute the reverse-time transition probabilities of $\hat{P}_0^N$. Let $\hat{p}_i(n)$ and $\hat{q}_{ji}(n)$ denote the marginal at time n and the reverse-time transition probabilities of $\hat{P}_0^N$, respectively. By (4.16), we get

$$\hat{p}_j(n+1)\,\hat{q}_{ji}(n) = \hat{p}_i(n)\,\pi_{ij}(n),\ \pi_i(n)\,\pi_{ij} = \pi_j(n+1)\,q_{ji}^{\pi}(n). \tag{9.23}$$

Assume now that $\pi_i(n) > 0$, $\hat{p}_i(n) > 0$, $\forall i,n : 0 \leq n \leq N$. Then (**??**) and (9.22) yield

$$\hat{q}_{ji} = \frac{\theta(n,i)}{\theta(n+1,j)} q_{ji}^{\pi}(n),\ \theta(n,i) := \frac{\hat{p}_i(n)}{\pi_i(n)}. \tag{9.24}$$

Observe, moreover that θ is space-time harmonic with respect to the reverse-time transition mechanism of Π since, by (9.24),

$$\sum_i q_{ji}^{\pi}(n)\,\theta(n,i) = \sum_i \frac{\pi_i(n)}{\pi_j(n+1)}\pi_{ij}(n)\,\frac{\hat{p}_i(n)}{\pi_i(n)} = \theta(n+1,j). \tag{9.25}$$

Again the optimal solution to Problem 9.21 is obtained from the *a priori* Markov chain Π_0^T through a "multiplicative functional" transformation.

We now attack the initial-final marginal problem. Let $\mathcal{D}(0,N;p^0,p^1)$ denote the family of Markovian distribution P_0^N on $\mathcal{X}^{N+1}$ that have marginals p^0 at time 0 and p^1 at time N, respectively, and have support contained in the support of Π_0^N, namely $Supp\left(P_0^N\right) \subseteq Supp\left(\Pi_0^N\right)$.

We consider the following

Maximum Entropy Problem (MEP3):

$$minimize\ \left\{\mathbb{D}\left(P_0^N||\Pi_0^N\right);P_0^N \in \mathcal{D}\left(0,N;p^0,p^1\right)\right\}. \tag{9.26}$$

Consider (9.20). Since $\mathbb{D}\left(p^0||\pi\left(0\right)\right)$ is invariant over $\mathcal{D}\left(0,N;p^0,p^1\right)$, $\mathbb{D}\left(P_0^N||\Pi_0^N\right)$ is now

$$\mathbb{D}\left(P_0^N||\Pi_0^N\right)=\sum_{k=0}^{N-1}\sum_{i_k}\mathbb{D}\left(p_{i_ki_{k+1}}\left(k\right)||\pi_{i_ki_{k+1}}\left(k\right)\right)p_{i_k}\left(k\right). \tag{9.27}$$

Moreover consider a *space-time harmonic function* φ for the *reference* stochastic evolution Π_0^N, namely

$$\varphi\left(n,i\right)=\sum_j\pi_{ij}\varphi\left(n+1,j\right),\ 0\leq n\leq\left(N-1\right) \tag{9.28}$$

Assuming that $\varphi\left(n,i\right)>0,\ \forall n:0\leq n\leq N,\ \forall i\in\mathcal{X}$. Problem (9.26) is equivalent to

$$minimize\ \left\{J\left(P_0^N\right);P_0^N\in\mathcal{D}\left(0,N;p^0,p^1\right)\right\}, \tag{9.29}$$

where

$$\begin{aligned}J\left(P_0^N\right)=&\sum_{k=0}^{N-1}\sum_{i_k}\mathbb{D}\left(p_{i_ki_{k+1}}\left(k\right)||\pi_{i_ki_{k+1}}\left(k\right)\right)p_{i_k}\left(k\right)+\\&+\sum_{i_0}\log\varphi\left(0,i_0\right)p_{i_0}^0-\sum_{i_N}\log\varphi\left(N,i_N\right)p_{i_N}^1,\end{aligned} \tag{9.30}$$

since the last two terms are invariant over $\mathcal{D}\left(0,N;p^0,p^1\right)$.

It is possible to dimostrate that $J\left(P_0^N\right)$ may be rewritten as

$$J\left(P_0^N\right)=\sum_{k=0}^{N-1}\sum_{i_k}\mathbb{D}\left(p_{i_ki_{k+1}}\left(k\right)||\pi_{i_ki_{k+1}}\left(k\right)\frac{\varphi\left(k+1,i_{k+1}\right)}{\varphi\left(k,i\right)}\right)p_{i_k}\left(k\right). \tag{9.31}$$

Define

$$\hat{p}_{ij}\left(k\right)=\pi_{ij}\left(k\right)\frac{\varphi\left(k+1,j\right)}{\varphi\left(k,i\right)}. \tag{9.32}$$

Notice that $\hat{p}_{ij}\geq 0$ and, by (9.28),

$$\sum_j\hat{p}_{ij}\left(k\right)=\sum_j\pi_{ij}\left(k\right)\frac{\varphi\left(k+1,j\right)}{\varphi\left(k,i\right)}=\frac{\varphi\left(k,i\right)}{\varphi\left(k,i\right)}=1. \tag{9.33}$$

Consider the probabilities $\hat{p}\left(n\right)$ defined by the ricursion

$$\hat{p_j}\left(n+1\right)=\sum_i\hat{p_{ij}}\left(n\right)\hat{p}_i\left(n\right),\ \hat{p}_i\left(0\right)=p_i^0. \tag{9.34}$$

Define

$$\hat{\varphi}(n,i) := \frac{\hat{p_i}(n)}{\varphi(n,i)}, \tag{9.35}$$

by (9.32) and (9.34) we get

$$\hat{\varphi}(n+1,j) = \frac{\hat{p_j}(n+1)}{\varphi(n+1,j)} = \sum_i \pi_{ij}(n)\,\hat{\varphi}(n,i), \tag{9.36}$$

namely $\hat{\varphi}(n,i)$ is *space-time co-harmonic*. Thus we establish the following theorem:

Theorem 9.3.5. Suppose there exists a pair of nonnegative functions $(\varphi, \hat{\varphi})$ defined on $[0,N] \times \mathcal{X}$ and satisfying the system

$$\varphi(n,i) = \sum_j \pi_{ij}(n)\,\varphi(n+1,j),\ 0 \le n \le (N-1), \tag{9.37}$$

$$\hat{\varphi}(n+1,j) = \sum_i \pi_{ij}(n)\,\hat{\varphi}(n,i),\ 0 \le n \le (N-1), \tag{9.38}$$

as well as the boundary conditions

$$\varphi(0,i)\cdot\hat{\varphi}(0,i) := p_i^0,\ \varphi(N,i)\cdot\hat{\varphi}(N,i) := p_i^N,\ \forall i \in \mathcal{X}. \tag{9.39}$$

Suppose moreover that $\varphi(n,i) > 0$, $\forall n : 0 \le n \le N$, $\forall i \in \mathcal{X}$. Then, the Markov distribution $\hat{P}$ in $\mathcal{D}(0,N;p^0,p^1)$ having transition probabilities

$$\hat{p}_{ij} = \pi_{ij}\frac{\varphi(n+1,j)}{\varphi(n,i)} \tag{9.40}$$

solves Problem 9.26. From (9.40) we get

$$\hat{P}(n) = \begin{bmatrix} \pi_{00}\frac{\varphi(n+1,0)}{\varphi(n,0)} & \pi_{01}\frac{\varphi(n+1,1)}{\varphi(n,0)} & \cdots & \pi_{0(r-1)}\frac{\varphi(n+1,r-1)}{\varphi(n,0)} \\ \pi_{10}\frac{\varphi(n+1,0)}{\varphi(n,1)} & \pi_{11}\frac{\varphi(n+1,1)}{\varphi(n,1)} & \cdots & \pi_{1(r-1)}\frac{\varphi(n+1,r-1)}{\varphi(n,1)} \\ \cdot & \cdot & \cdots & \cdot \\ \pi_{(r-1)0}\frac{\varphi(n+1,0)}{\varphi(n,r-1)} & \cdots & \cdots & \pi_{(r-1)(r-1)}\frac{\varphi(n+1,r-1)}{\varphi(n,r-1)} \end{bmatrix}. \tag{9.41}$$

Notice that if $(\varphi, \hat{\varphi})$ satisfy (9.37)-(9.38)-(9.39), so does the pair $\left(c\varphi, \frac{1}{c}\varphi\right)$ for all $c > 0$. Hence, uniqueness for the Schrödinger system is always intended up to such multiplications. As in the diffusion case, the problem is now reduced to establish, under suitable assumption, existence and uniqueness for the Schrödinger system (9.37)-(9.38)-(9.39).

Corollary 9.3.6. Given an initial distribution $\pi(0)$ and $\Pi^n = \left(\Pi_{ij}^{(n)}\right)$, namely the prior matrix of n-step transition probabilities, in view of the cancellations that occur, $\hat{p}(N)$ is given by

$$\hat{p}(N) = \prod_{n=0}^{N-1} \hat{P}(n)\,\pi(0) = diag\left(\varphi(N,0), \varphi(N,1), \ldots, \varphi(N,r-1)\right)\left(\Pi^*\right)^N \cdot \\ \cdot diag\left(\varphi(0,0), \varphi(0,1), \ldots, \varphi(0,r-1)\right)^{-1}\pi(0). \qquad (9.42)$$

Example 9.3.7. Consider the following chain matrix

$$\Pi = \begin{bmatrix} 0 & 0.3893 & 0 & 0 & 0.6107 \\ 0.6107 & 0 & 0.3893 & 0 & 0 \\ 0 & 0.6107 & 0 & 0.3893 & 0 \\ 0 & 0 & 0.6107 & 0 & 0.3893 \\ 0.3893 & 0 & 0 & 0.6107 & 0 \end{bmatrix},$$

and the initial distribution

$$\pi(0) = \begin{bmatrix} 0.2834 \\ 0.1592 \\ 0.1422 \\ 0.2147 \\ 0.2005 \end{bmatrix}.$$

The chain matrix has the form of (4.24) and represents a cyclical random walk. Notice that Π is doubly stochastic hence the Markov chain has

$$\overline{\pi} = \begin{bmatrix} 0.2000 \\ 0.2000 \\ 0.2000 \\ 0.2000 \\ 0.2000 \end{bmatrix}$$

as unique stationary distribution.

Since the SLEM is $\mu(\Pi) = 0.8194$, that is very close to one, the chain converges to $\overline{\pi}$ slowly. Indeed, starting from $\pi(0)$, the system effectively reaches $\overline{\pi}$ only after $T = 34$ steps, namely

$$\left(\Pi^*\right)^T \pi(0) = \overline{\pi}.$$

Notice moreover that with $N = 4$ we get $\Pi^N > 0$. Hence, we solve the Schrödinger system (9.37)-(9.38)-(9.39) with $p^0 = \pi(0)$, $p^N = \overline{\pi}$ and $N = 4$.

We get

$$\varphi(0) = \begin{bmatrix} 0.3977 \\ 0.4334 \\ 0.4225 \\ 0.4043 \\ 0.4445 \end{bmatrix}$$

and

$$\varphi(4) = \begin{bmatrix} 0.3931 \\ 0.4126 \\ 0.4544 \\ 0.3683 \\ 0.4740 \end{bmatrix}.$$

Then we compute $\prod_{n=0}^{3} \hat{P}(n)$ according to (9.42), verifying that

$$\hat{P}(0) \cdot \hat{P}(1) \cdot \hat{P}(2) \cdot \hat{P}(3) \cdot \pi(0) = \begin{bmatrix} 0.2000 \\ 0.2000 \\ 0.2000 \\ 0.2000 \\ 0.2000 \end{bmatrix},$$

the invariant distribution which considerably differs from

$$(\Pi^*)^4 \pi(0) = \begin{bmatrix} 0.2090 \\ 0.2017 \\ 0.1872 \\ 0.2207 \\ 0.1813 \end{bmatrix}.$$

Remark 9.3.8. The solution of problems 9.9 and 9.21 may be viewed as a special cases of the solution of problem 9.26 where the role of the space-time co-harmonic function $\hat{\rho}(n,i)$ is played by the prior probabilities $\pi_i(n)$.

Indeed, let us compute the reverse-time transition probabilities $\hat{q}_{ji}(n)$ of $\hat{P}_0^N$. By (4.16) and using $\hat{p}_i(n) = \varphi(n,i) \cdot \hat{\varphi}_i(n)$, we get

$$\hat{q}_{ji}(n) = q_{ji}^{\pi}(n) \frac{\xi(n,i)}{\xi(n+1,j)},\ \xi(n,i) := \frac{\hat{\varphi}(n,i)}{\pi_i(n)}. \tag{9.43}$$

Moreover

$$\sum_i q_{ji}^{\pi}(n)\, \xi(n,i) = \sum_i \frac{\pi_i(n)}{\pi_j(n+1)} \pi_{ij}(n) \frac{\hat{\varphi}_i(n)}{\pi_i(n)} = \xi(n+1,j), \tag{9.44}$$

namely, ξ is space-time harmonic with respect to the reverse-time transition mechanism of Π_0^N.

We then get the suggestive formula

$$\hat{p}_i(n) = \varphi(n,i) \cdot \xi(n,i) \cdot \pi_i(n), \tag{9.45}$$

showing that the "a posterior" marginals are obtained from the "a priori" marginals through multiplication by a space-time harmonic function for the forward transition mechanism and by a space-time harmonic function for the reverse time transition mechanism.

In detail, in Problems 9.9 and 9.21, we have $\xi(n,i) \equiv 1$ and $\xi(n,i) \equiv 1$, respectively. Namely the solutions to (9.9) and (9.21) appear as particular cases of the solution to Problem 9.26, where only the forward or only the backward transition mechanism undegoes a multiplicative functional transformation induced by a space-time harmonic function, whereas in (9.26) *both* transition mechanism are subject to such a transfomation.

Finally, notice that if we exchange p^0 and p^1, the solution will simply be to be time reversal of $\hat{P}_0^N$.

9.4 Existence and uniqueness for the Schrödinger system

As argued in [1, , pp. 13-14] we can prove the existence and uniqueness of a Schrödinger system. In particular we mark the following theorem.

Theorem 9.4.1. Let $X = \{X(0), X(1), \ldots\}$ be a Markov chain with state space $\mathcal{X}$ chain matrix $P = (p_{ij})$ and let $p(k,x,m,y) := P(X(m) = y | P(k) = x)$. The *Schrödinger system* (9.37)-(9.38)-(9.39) has a unique solution with $\varphi(n,x) > 0$, $\forall 0 \leq n \leq N$, $\forall x \in \mathcal{X}$ assuming that:

1. p^1 is a distribution on $\mathcal{X}$ with $p_x^1 > 0$, $\forall x \in \mathcal{X}$;
2. $p(0,x,N,y) > 0$, $\forall x, y \in \mathcal{X}$.

In many important applications, the prior transition probabilities do not depend on time. We get the following result for finite, irreducible and aperiodic Markov chains.

Corollary 9.4.2. Let $X = \{X(0), X(1), \ldots\}$ be a Markov chain with finite state space $\mathcal{X}$ and transition matrix $P = (p_{ij})$. The *Schrödinger system* (9.37)-(9.38)-(9.39) has a unique solution with $\varphi(n, x) > 0$, $\forall 0 \leq n \leq N$, $\forall x \in \mathcal{X}$ assuming that:

1. p^1 is a distribution on $\mathcal{X}$ with $p_x^1 > 0$, $\forall x \in \mathcal{X}$;
2. The matrix P^N has all positive elements.

Chapter 10

Mixing Markov chain

10.1 The average consensus problem

Animal aggregation, such as school of fish, flocks of birds, groups of bees, etc. are believed to use local coordination rules that result in complex intelligent behaviour at the group level. These can be associated with problems of multi-agent systems *consensus*, *flocking* and *synchronization* such as mobile robots coordination, estimation with distributed sensors and load balancing in computer networking.

However, the position of the agents and/or technological and cost limitations make the exchanging of information very limited. We suppose to have N agents whose state at time k is represented by a vector $x(k) \in \mathbb{R}^{\mathbb{N}}$. In such a way their dynamics is simply given by

$$x(k+1) = x(k) + u(k). \tag{10.1}$$

where $u(k) \in \mathbb{R}^{\mathbb{N}}$ represents the control inputs.

The goal is to design a feedback control law $u = Kx$ such that, for any initial condition $x_0 \in \mathbb{R}^{\mathbb{N}}$, the closed loop system $x(k+1) = (I+K)x(k)$ satisfies

$$\lim_{k\to\infty} x(k) = \alpha(x_0)\mathbb{1}, \tag{10.2}$$

where $\mathbb{1} = (1, 1, \ldots, 1)^*$.

If (10.2) is satisfied we reach the consensus because all the agents are in the same position.

Definition 10.1.1. Let $\mathcal{G}_K = (V, E)$, $|V| = N$ the *communication graph* associated to K, whose arcs k_{ij} differs from 0 if and only if agent j has to communicate x_j to agent i to compute its feedback

control. Moreover it is always assumed that each agent has access to its own state, namely $k_{ii} \neq 0$. Let $P = I + K$.

If we want that (10.2) holds, we need to impose $P\mathbf{1} = \mathbf{1}$ and if we add the requirement that the elements $p_{ij} \geq 0$, P becomes stochastic.

Definition 10.1.2. Given a directed graph $\mathcal{G} = (V, E)$, $|V| = N$ that represents the communication network, we say K is *compatible* with $\mathcal{G}$ if $\mathcal{G}_K \subseteq \mathcal{G}$.

The consensus problem is said to be solvable on $\mathcal{G}$ if there exists a feedback matrix K compatible with $\mathcal{G}$ and (10.2) holds. If $\mathcal{G}_K$ is strongly connected, the chain is irreducible.

It is often required that consensus is attained as the average of the initial states

$$\bar{x}(0) = \frac{1}{N}\sum_i x_i(0). \tag{10.3}$$

This is the case if $K\mathbb{1} = \mathbb{1}$ and K is symmetric, namely P is doubly stochastic.

Since of the speed of convergence is essential in these applications it is important to choose K so that the SLEM (4.10) is as small as possible. In particular consider a strongly connected graph $\mathcal{G} = (V, E)$, $|V| = N$ such that

$$\begin{cases} (j,j) \in \mathcal{G} & \forall j \\ (i,j) \in \mathcal{G} \Leftrightarrow (j,i) \in \mathcal{G} & \forall i,j \end{cases}. \tag{10.4}$$

We should find a stochastic matrix $P = (p_{ij})$ such that $(i,j) \notin \mathcal{G} \Rightarrow p_{ij} = 0$ that has minimum SLEM. This is always the case if the stochastic matrix is symmetric

10.2 Fastest mixing Markov chain problem

Consider a Markov chain $X = \{X(0), X(1), \ldots\}$ with finite state space $\mathcal{X} = \{0, 1, \ldots, r-1\}$. Let $\mathcal{G} = (V, E)$ be its transition graph with nodes set $V = \mathcal{X}$ and edges set $E \subseteq V \times V$, with $(i,j) \in E \Leftrightarrow (j,i) \in E$. We also assume that each node has a self-loop.

Suppose we deal with Markov chains whose chain matrix P is doubly stochastic. By Definition 4.2.8 we get

$$\pi = \left(\frac{1}{r}\right)\mathbb{1}$$

as a stationary distribution for the chain. In addition, if the chain is aperiodic and irreducible, by the Perron-Frobenius theorem 6.1.4 we get:

1. π is the unique stationary distribution;
2. the chain converges in total variation to π, as n increases, independently of the initial distribution.

We also know that the smaller the SLEM μ (4.10), the faster the Markov chain converges to its stationary distribution. The quantity $-\log\mu$ is called the *mixing rate* of the chain and $1-\mu(P)$ is called *spectral gap.*

It is possible to consider the following problem:

Fastest mixing Markov chain $(FMMC)$ problem:

$$minimize\ \mu(P) \tag{10.5}$$

$$subject\ to\ P \geq 0,\ P\mathbb{1} = \mathbb{1},\ P = P^*,\ (i,j) \notin E \Rightarrow p_{ij} = 0 \tag{10.6}$$

Here P is the optimization variable and the graph $\mathcal{G}$ is the problem data, where (10.6) summarizes all the above properties of P.

Remark 10.2.1. Since μ is continuous and the set of possible transition matrices is compact, there is at least one optimal transition matrix $\overline{P}$ such that $\mu\left(\overline{P}\right) = \overline{\mu}$ the optimal SLEM.

As shown in [2, pp. 669-670], there are several methods to obtain transition probabilities that give fast mixing.

We propose the one which applies the *Metropolis-Hastings algorithm* (see B.2.1.12).

Definition 10.2.2. Let d_i be the *degree* of node i, not counting the self-loop. It is namely the number of neighbor nodes of node i not counting i itself.

This algorithm modifies the transition probabilities p_{ij} in the following way

$$P_{ij}^{mh} = \begin{cases} \frac{1}{\max(d_i,d_j)} & if\ (i,j) \in E,\ i \neq j \\ \sum\limits_{(i,k)\in E} \max\left(0, \frac{1}{d_i} - \frac{1}{d_k}\right) & if\ i = j \\ 0 & if\ (i,j) \notin E \end{cases}, \tag{10.7}$$

which we call the *Metropolis-Hastings chain matrix*.

Namely the transition probability between two distinct, connected edges is the reciprocal of the larger degree, and the self loop probabilities are chosen to ensure P^{mh} to be stochastic.

Remark 10.2.3. It is interesting to notice that in P^{mh} the transition probability of an edge only depends on local information, i.e. the degrees of its adjacent nodes.

Determining or bounding the SLEM of Markov chains is very important in Markov chain Monte Carlo simulation (see A.3). The efficiency of such simulation depends on how fast the constructed Markov chain converges to its stationary distribution.

As numerically shown in [2], the fastest mixing Markov chain on a given graph can be computed exactly by a polynomial time optimization algorithm. In practice, this is feasible at least for graphs with a modest number of edges, such as 1000.

Example 10.2.4. Let $\mathcal{G}_1$ be the graph of figure 10.1. Its optimal transition matrix is

$$\overline{P} = \begin{bmatrix} \frac{6}{11} & \frac{5}{11} & 0 & 0 \\ \frac{5}{11} & 0 & \frac{3}{11} & \frac{3}{11} \\ 0 & \frac{3}{11} & \frac{4}{11} & \frac{4}{11} \\ 0 & \frac{3}{11} & \frac{4}{11} & \frac{4}{11} \end{bmatrix}$$

and the SLEM of this matrix is $\overline{\mu} = \frac{7}{11}$. Notice that it is really close to $\mu^{mh} = \frac{2}{3}$, namely the slem of the matrix

$$P^{mh} = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \\ 0 & \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \end{bmatrix},$$

obtained by Metropolis-Hastings algorithm.

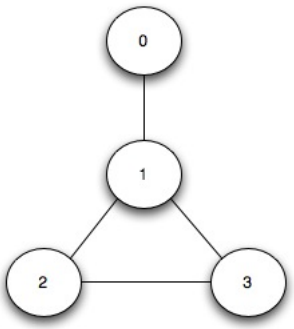

Figure 10.1: Fast mixing

10.3 A class of slow mixing Markov chain: Random geometric graph

First of all let us introduce and discuss a particular type of graph called *random geometric graph.*

In graph theory, a random geometric graph is an undirected graph drawn on a bounded region, i.e. the unit torus $[0,1) \times [0,1)$.

It is generated by

1. Placing n nodes at random uniformly and independently on the chosen region;
2. Connecting two nodes, u, v if and only if the distance between them is at most a threshold r, namely $d(u,v) \leq r$.

Remark 10.3.1. Such graphs can be represented by the *adjacency matrix*

$$E_{ij} = \begin{cases} 1 & d(i,j) \leq r \\ 0 & d(i,j) > r \end{cases}, \tag{10.8}$$

where $d(i,j) = d(j,i)$ by definition and we assume the absence of self-loops.

However the way of constructing a random geometric graph does not ensure to generate a connected graph. Thus a Markov chain associated to such a graph may not be irreducible, which is an important condition we require.

So we must find an easy way to control if an undirected graph is connected.

Definition 10.3.2. Given a graph $\mathcal{G} = (V,E)$, $|V| = N$, the *Laplacian matrix* $L = (l_{ij})$ is defined by

$$l_{ij} := \begin{cases} d_i & i = j \\ -1 & i \neq j \text{ and } v_i \text{ is adiacent to } v_j \\ 0 & otherwise \end{cases} \tag{10.9}$$

for every $i,j \in \{0,1,\dots n-1\}$, where d_i is the degree of node i, see Definition 10.2.2.

Moreover L is always *positive-semidefinite,* namely its eigenvalues are such that $\lambda_i \geq 0$, $0 \leq i \leq n-1$, and by

$$L\mathbb{1} = 0\mathbb{1}, \tag{10.10}$$

we get $\lambda_0 = 0$.

Remark 10.3.3. The number of times $\lambda_i = 0$ is the number of connected components in the graph.

Corollary 10.3.4. The *algebraic connectivity* λ_1 of a graph $\mathcal{G}$ is the second-smallest eigenvalue of the Laplacian matrix of $\mathcal{G}$. This eigenvalue is greater than zero if and only if $\mathcal{G}$ is connected.

Example 10.3.5. Consider the following graph.

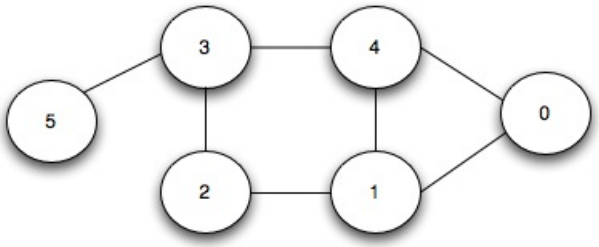

Figure 10.2: Undirected graph

We get the adjacency matrix

$$E = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}, \tag{10.11}$$

which is obviously symmetric, and the Laplacian matrix

$$L = \begin{bmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{bmatrix}.$$

The former has the following eigenvalues $\{0,\ 0.7216,\ 1.6826,\ 3,\ 3.7046,\ 4.8912\}$.

The algebraic connectivity is $\lambda_1 = 0,7216 > 0$. Hence by Corollary 10.3.4 the graph is connected as figure 10.2 has previously shown.

This is an important result which gives us the possibility of dealing only with connected graphs.

Remark 10.3.6. The distance r may be choosen at random, but in order to avoid the problem of generating a huge number of graphs it is generally a function of n, namely $r \propto \frac{1}{\sqrt{n}}$.

Now suppose to assign random probabilities to the edges of a random geometric graph such that it becomes a transition graph. The adjacency matrix E changes into a stochastic chain matrix P. Notice that even if E is symmetric, P is not necessarily doubly stochastic.

Remark 10.3.7. Stochastic matrices generated on a random geometric graph have the SLEM really close to one.

This fact is very important because it makes us deal with irreducible finite Markov chains whose convergence to their stationary distribution is very slow.

Example 10.3.8. Consider Example 10.3.5. A stochastic chain matrix related to (10.11) is

$$P = \begin{bmatrix} 0 & 0.4740 & 0 & 0 & 0.5260 & 0 \\ 0.0769 & 0 & 0.5444 & 0 & 0.3787 & 0 \\ 0 & 0.2632 & 0 & 0.7368 & 0 & 0 \\ 0 & 0 & 0.2218 & 0 & 0.3871 & 0.3911 \\ 0.0762 & 0.4667 & 0 & 0.4571 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix},$$

whose SLEM is $\mu(P) = 0.9747$.

Chapter 11

Simulation for Schrödinger bridges

11.1 Simulation for Schrödinger bridges

Hereafter we concentrate on the simulations for Schrödinger bridges.

Using MATLAB (see Appendix B), we simulated some Markov chain models described so far.

Remark 11.1.1. Our interest is on irreducilble finite Markov chains whose chain matrix P is such that $P^N > 0$ for an integer $N > 1$. In this way, by the Perron-Frobenius Theorem 6.1.4, we know that there exists a unique stationary distribution π, which can be found on the rows on P^∞.

First we focus on some examples that we described in Section 4.3, in particular:

1. random walks with partially reflecting barriers (see B.2.2.5);
2. cyclical random walks (see B.2.2.4);
3. the Bernoulli-Laplace model of diffusion (see B.2.2.3).

Then we choose to represent a particular case of success runs (see B.2.2.1) described in Section 4.4.

Because of the constraints in Remark11.1.1, the chain matrix rep-

resenting a success runs (4.33) is now

$$P_s = \begin{bmatrix} q & p & 0 & 0 & 0 & \dots & 0 \\ q & 0 & p & 0 & 0 & \dots & 0 \\ q & 0 & 0 & p & 0 & \dots & 0 \\ . & . & . & . & . & \dots & 0 \\ 1 & 0 & \dots & \dots & \dots & \dots & 0 \end{bmatrix},$$

which is evidently irreducible and finite.

After that "classic" models, we want to create Markov chains whose properties satisfy our requirements and try to represent *complex networks.*

These networks are graphs with a huge number of nodes. An important class of complex networks is *small-world networks*. They are graphs in which most nodes are not neighbors of one another, but most nodes can be reached from every other by a small number of steps.

Posterior to the implementation of the random geometric graphs described above (see B.2.2.6), we choose to represent a Markov chain with a completely random stochastic chain matrix (see B.2.2.7).

Remark 11.1.2. This way of constructing stochastic matrices does not ensure a matrix to be irreducibile and very sparse. Hence we need to test that before starting a simulation (see B.2.2.2).

Given m irreducible random stochastic matrices $P_0,\ P_1,\ \dots, P_{m-1}$, whose transition graphs are $\mathcal{G}_0, \mathcal{G}_2, \dots, \mathcal{G}_{m-1}$, let us connect at least one state of $\mathcal{G}_i$ with another one of $\mathcal{G}_j$, $j = i\, mod\ (m-1)$. We get a new connected transition graph $\overline{\mathcal{G}}$ and thus a new stochastic matrix $\overline{P}$ of the following form

$$\overline{P} = \begin{bmatrix} P_0 & 1_0^* & 0 & \dots & 0 \\ 0 & P_1 & 1_1^* & 0 & . \\ . & 0 & \ddots & \ddots & 0 \\ . & . & . & \ddots & 1_{m-2}^* \\ 1_{m-1}^* & 0 & \dots & 0 & P_{m-1} \end{bmatrix}, \tag{11.1}$$

where $1_i^*, \forall i : 0 \leq i \leq m-1$ is a square matrix of the same dimension of P_i with at least one element different from zero (see B.2.2.8).

Remark 11.1.3. This important method that permits us to construct matrices of huge dimensions which are surely irreducible and very sparse.

At this point it is interesting to explain our simulation algorithm (see B.2.3). Hence using the implemented functions of B.2.1 we can generalize the following:

Algorithm 11.1 Simulation

Generate a stochastic matrix Π;

Find the SLEM S;

Compute the number of steps N such that $\Pi^N > 0$;

Generate a random marginal distribution $\pi(0)$;

Compute the stationary distribution π the system reaches in $T \neq N$ steps starting at $\pi(0)$;

Notice that $(\Pi^*)^N \pi(0) \neq \pi$ and it is not a stationary distribution;

Solve the Schrödinger bridge with parameters π, $\pi(0)$, N, getting the stochastic matrices $P^*(n)$, $n = 0, \dots N-1$;

Verify that $\left(\prod_0^{N-1} P^*(n)\right) \pi(0) = \pi$.

Example 11.1.4. Consider the following three random stochastic matrices

$$P_0 = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix},$$

$$P_1 = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 1 & 0 \end{bmatrix},$$

$$P_2 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

and their transition graphs.

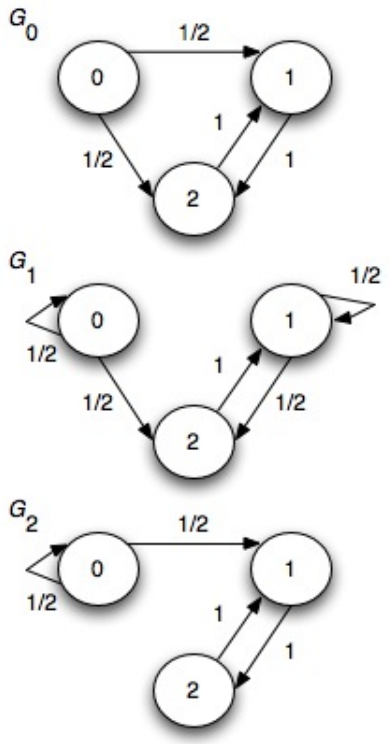

Figure 11.1: Transition graphs $\mathcal{G}_0$, $\mathcal{G}_1$, $\mathcal{G}_2$

By connecting them each other we get

$$\overline{P} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 1 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix},$$

and the transition graph $\overline{\mathcal{G}}$.

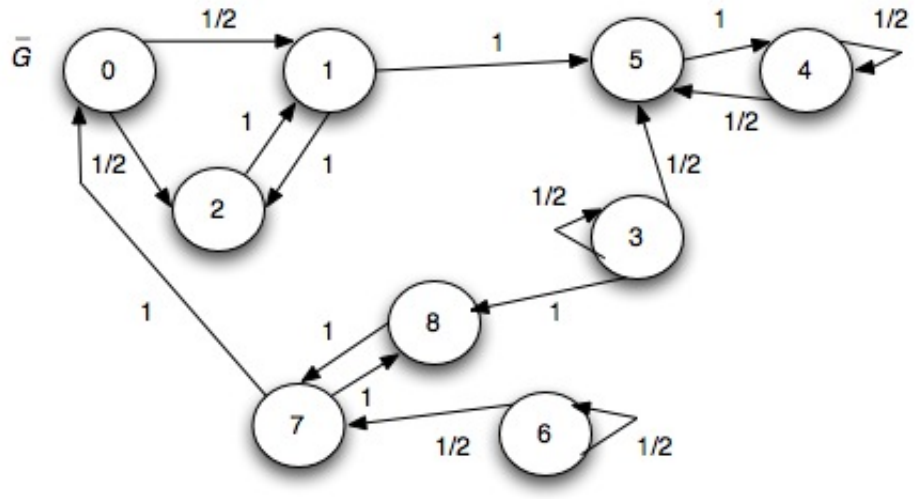

Figure 11.2: Transition graph $\overline{\mathcal{G}}$

Notice that $\overline{\mathcal{G}}$ is connected and $\overline{P}$ is irreducible. The SLEM is $\mu = 0.9462$ and the matrix $\overline{P}^7 > 0$.

Example 11.1.5. Following Algorithm 11.1, we generate a random geometric graph. By weighting its edges at random so that it becomes a transition graph, we get the following chain matrix

$$\Pi = \begin{bmatrix} 0.2479 & 0.2080 & 0 & 0.2821 & 0 & 0 & 0 & 0 & 0 & 0.2621 \\ 0.2705 & 0.3816 & 0 & 0 & 0 & 0 & 0.2512 & 0 & 0 & 0.0966 \\ 0 & 0 & 0.4615 & 0 & 0.4154 & 0 & 0 & 0.1231 & 0 & 0 \\ 0.2568 & 0 & 0 & 0.3716 & 0 & 0 & 0 & 0 & 0.3716 & 0 \\ 0 & 0 & 0.2000 & 0 & 0.1459 & 0 & 0.5135 & 0.1405 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.6533 & 0 & 0 & 03467 & 0 \\ 0 & 0.1881 & 0 & 0 & 0.4356 & 0 & 0.3564 & 0 & 0 & 0.0198 \\ 0 & 0 & 0.6167 & 0 & 0.583 & 0 & 0 & 0.3250 & 0 & 0 \\ 0 & 0 & 0 & 0.6711 & 0 & 0.2148 & 0 & 0 & 0.1141 & 0 \\ 0.1519 & 0.6329 & 0 & 0 & 0 & 0 & 0.1835 & 0 & 0 & 0.0316 \end{bmatrix}.$$

Then, we find its SLEM $\mu(\Pi) = 0.9356$, which is very close to neo circle and compute $T = 7$, namely the number of steps which guarantees $\Pi^T > 0$.

Now we generate at random the initial distribution

$$\pi(0) = \begin{bmatrix} 0.1199 \\ 0.0414 \\ 0.1383 \\ 0.1325 \\ 0.0656 \\ 0.1421 \\ 0.0576 \\ 0.2033 \\ 0.0916 \\ 0.0077 \end{bmatrix}.$$

Starting from $\pi(0)$, only after $N = 55 \neq T$ steps the system effectively reaches its unique stationary distribution

$$\overline{\pi} = \begin{bmatrix} 0.0902 \\ 0.1324 \\ 0.1274 \\ 0.0998 \\ 0.1655 \\ 0.0345 \\ 0.1956 \\ 0.0576 \\ 0.0554 \\ 0.0416 \end{bmatrix}$$

which considerably differs from the marginal distribution

$$(\Pi^*)^7 \pi(0) = \begin{bmatrix} 0.0924 \\ 0.1122 \\ 0.1168 \\ 0.1363 \\ 0.1425 \\ 0.0623 \\ 0.1625 \\ 0.0521 \\ 0.0848 \\ 0.0382 \end{bmatrix}.$$

Hence, we set up and solve the Schrödinger system (9.37)-(9.38)-(9.39) with $p^0 = \pi(0)$, $p^N = \overline{\pi}$ and $N = 7$.

We get

$$\varphi(0) = \begin{bmatrix} 0.6142 \\ 0.6988 \\ 0.8469 \\ 0.4928 \\ 0.8238 \\ 0.3897 \\ 0.7917 \\ 0.8578 \\ 0.4483 \\ 0.7026 \end{bmatrix}$$

and

$$\varphi(7) = \begin{bmatrix} 0.5142 \\ 0.7244 \\ 0.8887 \\ 0.3454 \\ 0.9018 \\ 0.2447 \\ 0.8744 \\ 0.8961 \\ 0.2993 \\ 0.6122 \end{bmatrix}.$$

Then, we compute $\prod_{n=0}^{6} \hat{P}(n)$ according to (9.42), verifying that

$$\hat{P}(0) \cdots \hat{P}(6)\, \pi(0) = \begin{bmatrix} 0.0902 \\ 0.1324 \\ 0.1274 \\ 0.0998 \\ 0.1655 \\ 0.0345 \\ 0.1956 \\ 0.0576 \\ 0.0554 \\ 0.0416 \end{bmatrix} = \overline{\pi}.$$

11.2 Time analysis

We test all the implemented algorithms (see B.2.1) with larger matrix dimension.

In particular, we focus on the random geometric graph algorithm

(see B.2.2.6). It is interesting to make a time analysis of this algorithm, since it features a large SLEM (see Remark 4.2.4).

In this way, according to Algorithm 11.1, we can get an idea of how much time it basically takes to:

1. create a random but connected geometric graph;
2. associate a stochastic matrix to this graph;
3. generate an initial and a final distribution for the Schrödinger system;
4. solve the Schrödinger system, with special attention to step 4.

The graph in Figure 11.3 shows the time needed to compute all four steps,

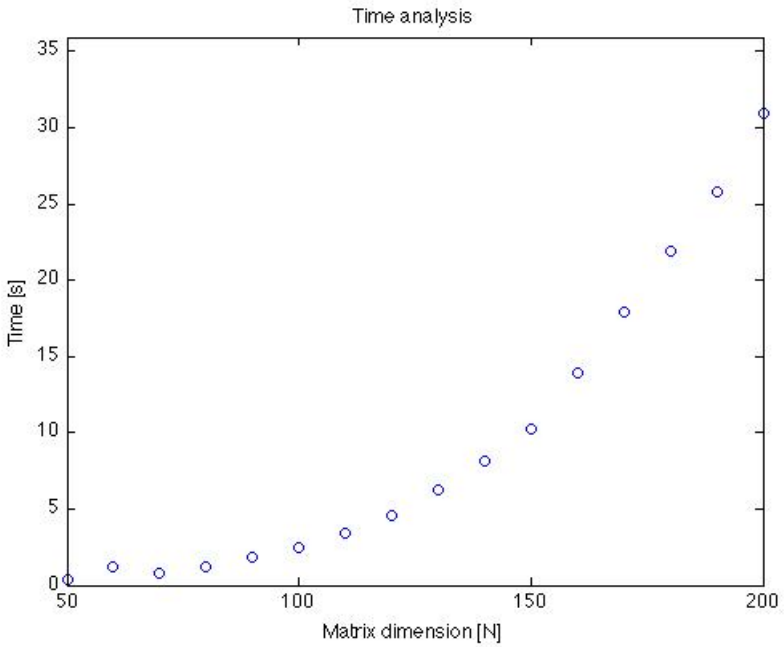

Figure 11.3: Time analysis

where we have limited our simulations to matrices with dimension between 50×50 and 200×200. Notice that computing time obviously increases with matrix dimensions.

Figure 11.3 shows that, on the average, Algorithm11.1 takes about 30 seconds with matrices of dimensions 200×200.

Let $N \times N$ be the matrix dimensions of Π and T the number of steps such that $\Pi^T > 0$. We recall that solving Schrödinger bridges means solving a $[0, T] \times N$ systems.

The complexity of solving Schrödinger bridges is the principal reason why the following figure,

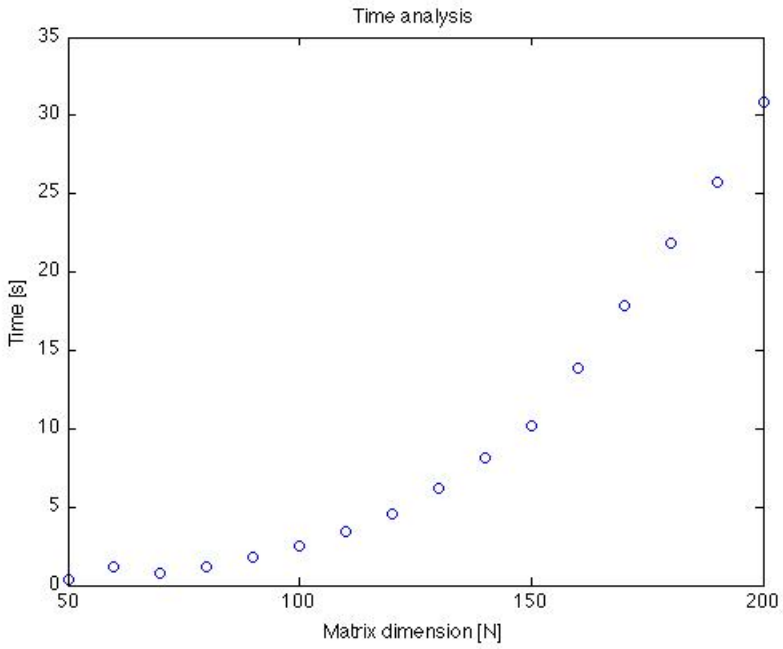

Figure 11.4: Step four time analysis

that is the time analysis of the only step 4 of Algorithm 11.1, does not differ so much from Figure 11.3 just described.

Remark 11.2.1. Our simulations are made on a PC with:

1. MATLAB version 7.8.0 (R2009a);
2. LINUX FEDORA 11;
3. AMD Athlon(tm) 64 X2 DUAL CORE PROCESSOR 5200+;
4. DDR2 RAM 1.9 GB.

11.3 Conclusions

It is interesting to notice that Schrödinger bridges are a great tool to drive in finite time Markov chains to a desired distribution.

Moreover, differently from the Hastings-Metropolis algorithm, they do not change the graph topological properties. They represent therefore some kind of "cheap" control action.

Schrödinger bridges, however, depend on the initial distribution $\pi(0)$, a fact that apparently makes them insuitable to solve the average consensus problem.

Acknowledgments

The author wishes to thank Dr. F. Ramponi for explaining some important features of MATLAB in order to reduce simulation times.

He also wishes to thank D. Varagnolo and S. Del Favero for introducing him the theory of random geometric graphs in order to create very sparse, irreducible Markov chains with a high SLEM.

Bibliography

[1] M. Pavon and F. Ticozzi. Discrete-time classical and quantum Markovian evolutions: Maximum entropy problems on path space, April 2009, submitted for publication, preprint arXiv:math-ph/0811.0933v2.

[2] S. Boyd, P. Diaconis and L. Xiao, Fastest mixing Markov chain on a graph.

[3] M. Pavon, Lectures on finite probabilistic models, Department of Pure and Applied Mathematics, University of Padua, Italy, 2009.

Appendix A

Some modern application of Markov chains

A.1 Identifying genes in genomic DNA

Markov chains play a central role also in recent attempts to find genes through statistical modeling. For example in *prokaryotes* (bacteria and archaea), genes appear as stretches in an enormously long sequence of four symbols $\mathcal{A} := \{A, C, G, T\}$ which represent the initials of the basis contained in nucleotides: *Adenine*, *Cytosine*, *Guanine* and *Thymine*. This model can be viewed as a k-step Markov chains, whose values belong to $\mathcal{A}^k$. In general $k = 5$, in which case the state space χ has cardinality 4^5.

A.2 The Google Page Rank algorithm

During the period 1995-1998 Larry Page and Sergey Brin developed at Stanford University an algorithm design to weight the pages of the World Wide Web which has since been known as *Google's PageRank*. Let N be the number of current pages. Ideally, the algorithm should mimic the actual navigation of web users. Define a $N \times N$ matrix P as

$$p = (p_{ij})_{i,j=1}^{N}, \quad p_{ij=p}\frac{g_i j}{\sum_j g_{ij}} + (1-p)\,\frac{1}{N}, \quad 0 < p < 1, \qquad \text{(A.1)}$$

where $p = 0.85$ in the PageRank algorithm. P is stochastic and the chain is clearly irreducible. Hence there is a unique stationary distribution π with $\pi_j > 0$, $\forall j$. Moreover P is nearly sparse.

The stationary distribution, which is also calculated as asymptotic distribution of the chain once a month, provides the ranking.

A.3 Markov chain Monte Carlo

Suppose we like to compute the area of a planar figure contained in the square $[0,1] \times [0,1]$, We know that points in $[0,1]$ are in one to one correspondence with infinite binary sequences. Hence, if we toss two coins a large number of times we get a minuscule square in $[0,1] \times [0,1]$ that we can consider as a point which may or not be contained in our figure. If we can repeat the experiment many times, we can approximate the area by the fraction of points that fall in the figure. This is the essence of integration by simulation which becomes extremely effective in higher dimension, namely $\mathbb{R}^p$, $p \geq 2$. *Monte Carlo methods* are computational algorithms that rely on repeated random sampling. Monte Carlo simulation treats deterministic problems by first finding a probabilistic analog. On the other hand previous methods of simulation and statistical sampling generally did the opposite, namely they used simulation to test a previously understood deterministic problem. Monte Carlo algorithms are the most prominent class of *Randomized algorithms.* They make random choises during their execution. It is possible that they may give different answers for the same input at difference runs and the result may be incorrect.

The Markov chain Monte Carlo (MCMC) methods seeks to construct a finite irreducible, aperiodic Markov chain whose unique stationary distribution is the distribution π we like to sample from. Applications of MCMC include simulating noisy images, textures, protein structures, approximate counting for polymer models, Bayesian statistics and scientific computing.

A.4 Distribution of epithelial cells

A *simple epithelium* is a tissue composed of a single layer of cells. It has been observed that cells are organized in monolayer epithelia as an array of irregular polygonal forms, with a majority of hexagonal shapes. The pattern in proliferating epithelia is, however, much more irregular. The fact that the *average number* of cells sides exponentially six had been known for some time to be a consequence of the following fact: each cell division produces two new nodes and three new edges. The model is provided by a simple Markov chain,

where a cell's state is its number of sides. The transition matrix of the chain is

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & \cdots \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 & \cdots \\ 0 & \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} & 0 & \cdots \\ 0 & \frac{1}{16} & \frac{4}{16} & \frac{6}{16} & \frac{4}{16} & \frac{1}{16} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix},$$

where $i, j = 4, 5, 6, \ldots$ and the state space is $\mathcal{X} = \{4, 5, 6, \ldots\}$. The state 4 is transient, all the other states are recurrent. The stationary distribution π of cell shape has $\pi_4 = 0$, $\pi_5 \sim 0.289$, $\pi_6 \sim 0.464$, $\pi_7 \sim 0.208$, $\pi_8 \sim 0.0359$, $\pi_9 \sim 0.0028$ and other states having equilibrium probabilities less than 10^{-4}.

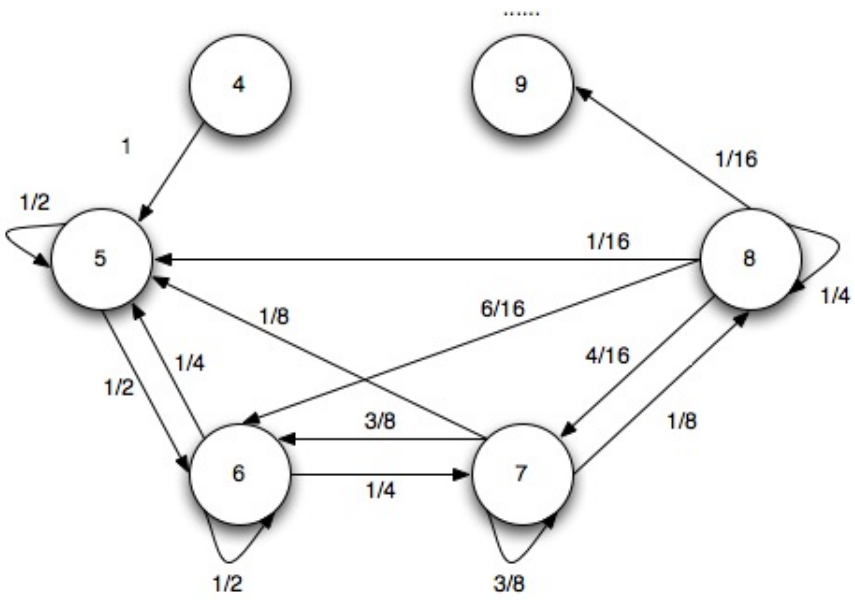

Figure A.1: Distribution of epithelial cells

Appendix B

Matlab

Let us list all the used Matlab functions.

B.1 Package functions

B.1.1 size

Algorithm B.1 size

```
SIZE Size of array.
   D = SIZE(X), for M-by-N matrix X, returns the two-element
   row vector. D = [M,N] containing the number of rows and
   columns in the matrix.
```

B.1.2 disp

Algorithm B.2 disp

```
DISP Display array.
   DISP(X) displays the array, without printing the array
   name. If X is a string, the text is displayed.
```

B.1.3 fsolve

Algorithm B.3 fsolve

```
1 FSOLVE solves systems of nonlinear equations of several
2     variables.
3     FSOLVE attempts to solve equations of the form:
4     F(X) = 0 where F and X may be vectors or matrices.
5     X = FSOLVE(FUN,X0) starts at the matrix X0 and tries to
6     solve the equations in FUN. FUN accepts input X and returns
7     a vector (matrix) of equation values F evaluated at X.
```

B.1.4 rand

Algorithm B.4 rand

```
1 RAND Uniformly distributed pseudorandom numbers.
2     R = RAND(N) returns an N-by-N matrix containing
3     pseudorandom values drawn from the standard uniform
4     distribution on the open interval(0,1). RAND returns a
5     scalar. RAND(SIZE(A)) returns an array thesame size as A.
```

B.1.5 randi

Algorithm B.5 randi

```
1 RANDI Pseudorandom integers from a uniform discrete
2     distribution.
3     R = RANDI(IMAX,N) returns an N-by-N matrix containing
4     pseudorandom integer values drawn from the discrete
5     uniform distribution on 1:IMAX. RANDI(IMAX) returns a scalar.
6     RANDI(IMAX,SIZE(A)) returns an array the same size as A.
7     R = RANDI([IMIN,IMAX],...) returns an array containing
8     integer values drawn from the discrete uniform
9     distribution on IMIN:IMAX.
```

B.1.6 diag

Algorithm B.6 diag

```
DIAG Diagonal matrices and diagonals of a matrix.
    DIAG(V,K) when V is a vector with N components is a square
    matrix of order N+ABS(K) with the elements of V on the
    K-th diagonal. K = 0 is the main diagonal, K > 0 is above
    the main diagonal and K < 0 is below the main diagonal.
    DIAG(V) is the same as DIAG(V,0) and puts V on the main
    diagonal. DIAG(X,K) when X is a matrix is a column vector
    formed from the elements of the K-th diagonal of X. DIAG(X)
    is the main diagonal of X. DIAG(DIAG(X)) is a diagonal matrix.
```

B.1.7 sum

Algorithm B.7 sum

```
SUM Sum of elements.
    S = SUM(X) is the sum of the elements of the vector X.
    If X is a matrix, S is a row vector with the sum over
    each column. S = SUM(X,DIM) sums along the dimension DIM.
```

B.1.8 zeros

Algorithm B.8 zeros

```
ZEROS Zeros array.
    ZEROS(N) is an N-by-N matrix of zeros. ZEROS(SIZE(A)) is
    the same size as A and all zeros. ZEROS with no arguments
    is the scalar 0
```

B.1.9 ones

Algorithm B.9 ones

```
ONES Ones array.
    ONES(N) is an N-by-N matrix of ones. ONES(SIZE(A)) is
    the same size as A and all ones. ONES with no arguments
    is the scalar 1.
```

B.1.10 eye

Algorithm B.10 eye

```
EYE Identity matrix.
   EYE(N) is the N-by-N identity matrix. EYE(SIZE(A)) is the
   same size as A. EYE with no arguments is the scalar 1.
```

B.1.11 abs

Algorithm B.11 abs

```
ABS Absolute value.
  ABS(X) is the absolute value of the elements of X. When X
   is complex, ABS(X) is the complex modulus(magnitude) of
   the elements of X.
```

B.1.12 eig

Algorithm B.12 eig

```
EIG Eigenvalues and eigenvectors.
   E = EIG(X) is a vector containing the eigenvalues of a
   square matrix X. [V,D] = EIG(X) produces a diagonal
   matrix D of eigenvalues and a full matrix V whose columns
   are the corresponding eigenvectors so that X*V = V*D.
```

B.1.13 sort

Algorithm B.13 sort

```
SORT Sort in ascending or descending order.
   For vectors, SORT(X) sorts the elements of X in
   ascending order. For matrices, SORT(X) sorts each
   column of X in ascending order. For N-D arrays, SORT(X)
   sorts the along the first non-singleton dimension of X.
   Y = SORT(X,DIM,MODE) has two optional parameters.
      DIM selects a dimension along which to sort.
      MODE selects the direction of the sort
                'ascend' results in ascending order
          'descend' results in descending order
   The result is in Y which has the same shape and type as X.
```

B.1.14 rank

Algorithm B.14 rank

```
RANK Matrix rank.
   RANK(A) provides an estimate of the number of linearly
   independent rows or columns of a matrix A. RANK(A,tol) is
   the number of singular values of A that are larger than tol.
   RANK(A) uses the default tol = max(size(A)) * eps(norm(A)).
```

B.1.15 all

Algorithm B.15 all

```
ALL True if all elements of a vector are nonzero.
   For vectors, ALL(V) returns logical 1 (TRUE) if none of
   the elements of the vector are zero.
Otherwise it returns
   logical 0 (FALSE). For matrices, ALL(X) operates on the
   columns of X, returning a row vector of logical 1's and
   0's. ALL(X,DIM) works down the dimension DIM.
```

B.1.16 find

Algorithm B.16 find

```
FIND Find indices of nonzero elements.
   I = FIND(X) returns the linear indices corresponding to
   the nonzero entries of the array X. X may be a logical
   expression. I = FIND(X,K) returns at most the first K
   indices corresponding to the nonzero entries of the array
   X. K must be a positive integer, but can be of any numeric
   type.
```

B.2 Implemented functions

B.2.1 System

B.2.1.1 isfull

Algorithm B.17 isfull

```
1  function full = isfull(A)
2  % ISFULL controls if A is a strickly full matrix/vector.
3  %   ISFULL(A) returns 1 if A is strickly full, 0
4  %   otherwise.
5  full=1;
6  for r=1:size(A,1),
7          for c=1:size(A,2)
8                  if A(r,c)==0
9                  full=0;
10                 break;
11                 end
12         end
13         if full==0
14         break;
15         end
16 end
17 end
```

B.2.1.2 pit

Algorithm B.18 pit

```
1  function pitres=pit(P,pi0,T)
2  %PIT calculates the marginal pi at time T.
3  %   PIT(P,PI0,T) calculates the marginal pi at time T
4  %   using P as stochastic matrix and PI0 as initial marginal
5  %   distribution.
6  pitres=((P')^T)*pi0;
```

B.2.1.3 pi0

Algorithm B.19 pi0

```
function pi0 = pi0(dim)
%PI0 creates a random stochastic vector.
%   PIO(DIM) creates a DIMX1 vector whose column sums
%   to 1.
pi0=zeros(dim,1);
for a=1:dim,
    pi0(a,1)=randi(1000,1,1);
end
plus=sum(pi0);
pi0=pi0/plus;
end
```

B.2.1.4 periodt

Algorithm B.20 periodt

```
function T = periodt(A)
%PERIODT finds the period T (= the number of steps) in
%which a square matrix becomes full.
%  PERIODT(A) returns a value T between 0 and 7 if
%  matrix A is square, displays 'not square matrix'
%  otherwise. If T=0 it means that matrix A becomes
%  strictly full in 8 steps at least.
maxT=7;
if size(A,1)~=size(A,2)
        disp('not square matrix');
        return;
end
A1 = A;
T=0;
for t = 1:maxT
        if isfull(A1)==0
        A1 = A1*A;
    else
    T = t;
    break;
    end
end
end
```

B.2.1.5 slem

Algorithm B.21 slem

```
function s = slem(P)
%SLEM finds the slem(second largest eigenvalue magnitude)
%of a given matrix.
%   SLEM(P) returns the slem of the given matrix P which
%   must be square because of the eigenvalues calculus.
%   It returns 0 and displays 'not square matrix' if P
%   isn't square, while it displays 'scalar matrix' and
%   returns 1 if P is scalar.
if size(P,1)~=size(P,2)
   disp('not square matrix');
   s=0;
elseif
   size(P,1)==1 && size(P,2)==1
   disp('scalar matrix');
   s=P(1);
else
   E=eig(P);
   A=abs(E);
   Asorted=sort(A,'descend');
   s=Asorted(2);
end
```

B.2.1.6 schsys

Algorithm B.22 schsys

```
function F = schsys(P,pi0,pibar,phi,T)
%SCHSYS builds a Schrodinger system.
%   SCHSYS(P,PI0,PIBAR,PHI,T) builds a Schrodinger
%   system with P as stochastic matrix,PI0 and PIBAR as
%   the inital and final marginal distributions,T as the
%   period in which P becomes strickly full and PHI as
%   the vector made up of the harmonic function at time T
%   (phi(T)) and the coharmonic function at time 0
%   (phihat(0)).
pi=[pi0;pibar];
c=size(P,2);
r=2*c;
Pb=zeros(r,c);
Pb(1:c,1:c)=P^T;
Pb(c+1:r,1:c)=(P')^T;
F=[((Pb(1:c,1:c)*phi(1:c)).*phi(c+1:r))-pi(1:c);
    ((Pb(c+1:r,1:c)*phi(c+1:r)).*phi(1:c))-pi(c+1:r)];
```

B.2.1.7 pibar

Algorithm B.23 pibar

```
1  function [p,T] = pibar(P,pi0)
2  %PIBAR calculates the number of steps to reach a
3  %stationary distribution
4  %  PIBAR(P,PI0) calculates the number of steps in which
5  %  matrix P reaches its stationary distribution starting
6  %  at PI0. It also returns the reached distribution.
7  Pt=P';
8  flag=1;
9  p=pi0;
10 T=0;
11 while flag==1
12      pn=Pt*p;
13      err=abs(pn-p);
14      if all(err<(1/10000))==1
15          flag=0;
16          p=pn;
17      else
18          T=T+1;
19          p=pn;
20      end
21 end
22 end
```

B.2.1.8 pibart

Algorithm B.24 pibart

```
1  function pibart=pibart(P,phi,pi0)
2  % PIBART calculates the final marginal distribution.
3  %   PIBART(P,PHI,PI0) calculates the final marginal
4  %   distribution given a stochastic matrix P, an initial
5  %   marginal distribution PI0 and a matrix PHI that is
6  %   made up of the harmonic and coharmonic functions as
7  %   the solution of the Schrodinger system
8  %   'solveschsys(P,PI0)'.
9  dim=size(P,1);
10 T=periodt(P);
11 Pt=P';
12 pibart=diag(phi(1:dim,1))*((Pt)^T)*(inv(diag(phi(1:dim,2))))*pi0;
13 end
```

B.2.1.9 lpibar

Algorithm B.25 lpibar

```
function [ps,T] = lpibar(n,pi0)
%LPIBAR calculates the stationary distribution
%of a Laplace model of diffusion.
%   LPIBAR(N,PI0) calculate the stationary
%   distribution of a Laplace model of diffusion
%   with 2*N molecules starting at PI0. If N=0 it
%   displays 'Too few molecules'.
if n<=1
    disp('Too few molecules');
end
ps=zeros(n+1,1);
for r=0:n
    ps(r+1)=(factorial(n)^4)/(factorial(r)^2*
            *factorial(2*n)*(factorial(n-r)^2));
end
P=laplace(n);
Pt=P';
dim=size(P,1);
flag=0;
T=0;
p=pi0;
if sum(abs(ps-p)<(1/10000))==dim
    return;
end
while flag==0
    pn=Pt*p;
    if dim==sum(abs(pn-ps)<(1/10000))
        flag=1;
        p=pn;
    else
        T=T+1;
        p=pn;
    end
end
end
```

B.2.1.10 schsyssolve

Algorithm B.26 schsyssolve

```
function [schres,schres2,schres0] = schsyssolve(P,pi0,pibar)
%SOLVESCHSYS solves a Schrodinger system
%   SOLVESCHSYS(P,PI0) solves a Schrodinger system,
%   where P must be a stochastic matrix, PI0 an
%   initial marginal distribution and PIBAR its
%   stationary distribution. It returns as output
%   parameter the harmonic and coharmonic function
T=periodt(P);
dim=size(P,1);
schres=zeros(2*dim,2);
guess=sqrt([pibar;pibar]);
schres(:,1)=fsolve(@(phi) schsys(P,pi0,pibar,phi,T),guess);
schres(:,2)=[(P^T)*schres(1:dim,1);
((P')^T)*schres(dim+1:2*dim,1)];
schres2=schres(1:dim,1);
schres0=schres(1:dim,2);
end
```

B.2.1.11 binomial

Algorithm B.27 binomial

```
function bc = binomial(n,m)
%BINOMIAL computes the binomial coefficient n choose m.
%   BINOMIAL(N,M) calculates the binomial coefficient n
%   choose m if m<=n, otherwise displays 'm must be > n'.
if m>n
disp('m must be > n');
end
if 2*m<=n
    m1=m;
else
    m1=n-m;
end
if n <= 100
    size=abs(n-m1)+1;
    ps=pascal(size);
    bc=ps(size,m1+1);
else
    bc=exp(gammaln(n+1)-gammaln(n-m+1)-gammaln(m+1));
end
```

B.2.1.12 metropolis

Algorithm B.28 metropolis

```
function P = metropolis(E)
%METROPOLIS applies Metropolis-Hastings algorithm
%   METROPOLIS(E) returns a stochastic chain matrix P
%   get by applying Metropolis-Hastings algorithm to
%   the adjacency matrix E.
N=length(E);
P=zeros(N,N);
deg=sum(E,2);
for r=1:N
    for c=1:N
        if E(r,c)==1 && r~=c
            P(r,c)=1/(max(deg(r),deg(c)));
        end
    end
plus=sum(P(r,:),2);
P(r,r)=1-plus;
end
end
```

B.2.2 Models

B.2.2.1 Success runs

Algorithm B.29 Success runs

```
1  function R = successruns(dim)
2  %SUCCESS RUNS creates a stochastic matrix which
3  %represents a success runs model.
4  %   SUCCESSRUNS(DIM) returns a DIMxDIM stochastic matrix
5  %   which represents a success runs.
6      R=zeros(dim,dim);
7      R(1:dim,1)=randi(10000,dim,1);
8      upperdia=randi(10000,dim-1,1);
9      R=R+diag(upperdia,1);
10     plus=sum(R,2);
11 for r=1:dim
12     R(r,1:dim)=R(r,1:dim)/plus(r);
13 end
14 end
```

B.2.2.2 Rsparse

Algorithm B.30 Rsparse

```
1  function P = rsparse(dim)
2  %RSPARSE creates an irreducible stochastic matrix.
3  %   RSPARSE(N) continues creating a new random NxN
4  %   stochastic matrix by using function randomsparse until
5  %   it is not irreducible.
6  T=0;
7  while T==0
8          P=randomsparse(dim);
9          flag=0;
10         while flag==0
11                 if rank(P)==dim
12                     flag=1;
13                 else
14                     P=randomsparse(dim);
15                 end
16         end
17         T=periodt(P);
18 end
19 end
```

B.2.2.3 Laplace

Algorithm B.31 Laplace

```
function L = laplace(n)
%LAPLACE creates a stochastic matrix which represents
%a Bernoulli-Laplace model of diffusion.
%  LAPLACE(N) returns a (N+1)x(N+1) stochastic matrix
%  which represents a Bernoulli-Laplace model of
%  diffusion with 2*N molecules (N white and N black).
maindiag=zeros(n+1,1);
upperdiag=zeros(n,1);
for r=0:n
    maindiag(r+1)=((2*r)*(n-r))/(n^2);
    end
for r=0:n-1
    upperdiag(r+1)=((n-r)^2)/(n^2);
end
lowerdiag=flipud(upperdiag);
L=diag(maindiag)+diag(upperdiag,1)+diag(lowerdiag,-1);
end
```

B.2.2.4 Cyclical random walk

Algorithm B.32 Cyclical random walk

```
function c = cyclicalrw(dim)
%CYCLICALRW creates a stochastic matrix which
%represents a cyclical random walk.
%   CYCLICAL(DIM) returns a DIMXDIM stochastic
%   matrix which represents a cyclical random
%   walk where probabilities p and q chosen at
%   random.
p=randi(10000,1,1);
q=randi(10000,1,1);
maindiag=zeros(dim,1);
upperdiag=p*ones(dim-1,1);
lowerdiag=q*ones(dim-1,1);
c=diag(maindiag)+diag(upperdiag,1)+diag(lowerdiag,-1);
c(1,dim)=q; c(dim,1)=p;
for row=1:dim
  plus=sum(c(row,1:dim));
  c(row,1:dim)=c(row,1:dim)/plus;
  end
end
```

B.2.2.5 Random walk

Algorithm B.33 Random walk

```
1  function r = randomwalk(dim)
2  %RANDOMWALK creates a stochastic matrix which represents
3  %a random walk.
4  %   RANDOMWALK(DIM) returns a DIMXDIM stochastic
5  %   matrix which represents a random walk where
6  %   probabilities p and q are chosen at random.
7  maindiag=zeros(dim,1);
8  upperdiag=zeros(dim-1,1);
9  lowerdiag=zeros(dim-1,1);
10 maindiag(1)=randi(10000,1,1);
11 maindiag(dim)=randi(10000,1,1);
12 lowerdiag(dim-1)=randi(10000,1,1);
13 upperdiag(1)=randi(10000,1,1);
14 p=randi(10000,1,1);
15 q=randi(10000,1,1);
16 for row=1:dim-1
17     lowerdiag(row)=q;
18 end
19 for row=2:dim-1
20     upperdiag(row)=p;
21 end
22 r=diag(maindiag)+diag(upperdiag,1)+diag(lowerdiag,-1);
23 for row=1:dim
24     plus=sum(r(row,1:dim));
25     r(row,1:dim)=r(row,1:dim)/plus;
26 end
27 end
```

B.2.2.6 Random geometric graph

Algorithm B.34 Random geometric graph

```
function P = randgeomgraph(N,r)
%RANDGEOMGRAPH creates a random geometric graph.
%   RANDGEOMGRAPH(N,R) creates a random geometric
%   graph by placing N nodes at random uniformly
%   and independently on [0,1)x[0,1) and connecting
%   all node whose distance is at most R.
P=zeros(N,N);
   err=(1/1000000)*ones(N,1);
while err(2)<(1/100000)
        x=rand(N,1);
        y=rand(N,1);
             E=(x*ones(1,N)-ones(N,1)*(x')).^2;
             E=sqrt(E+((y*ones(1,N)-ones(N,1)*(y')).^2));
             E=E<(r*ones(N,N));
        deg=sum(E,2);
             L=diag(deg)-E;
        eigval=sort(eig(L),'ascend');
        err=abs(eigval);
end
for r=1:N
    for c=1:N
         if E(r,c)==1 && r~=c
            P(r,c)=1/(max(deg(r),deg(c)));
         end
    end
plus=sum(P(r,:),2);
P(r,r)=1-plus;
end
end
```

B.2.2.7 Randomsparse

Algorithm B.35 Randomsparse

```
1  function S = randomsparse(dim)
2  %RANDOMSPARSE creates a random stochastic matrix.
3  %   RANDOMSPARSE(N) returns a NxN stochastic matrix
4  %   not strictly full.
5  S=zeros(dim,dim);
6  for r=1:dim
7      int=randi(round(dim/2),1);
8      for a=1:int
9          flag=0;
10         ind=randi(dim,1);
11         if int==1
12                 while ind==r
13                   ind=randi(dim,1,1);
14             end
15         elseif ind==r && flag==0
16                 flag=1;
17         elseif ind==r && flag==1
18                 while ind==r
19                       ind=randi(dim,1,1);
20                 end
21         end
22         S(r,ind)=randi(10000,1,1);
23     end
24 end
25 for r=1:dim
26     plus=sum(S(r,1:dim));
27     S(r,1:dim)=S(r,1:dim)/plus;
28 end
29 end
```

B.2.2.8 Blocksparse

Algorithm B.36 Blocksparse part I

```
function M = blockrsparse(details)
%BLOCKSPARSE creates a random stochastic matrix made up
%of smaller matricies.
%   BLOCKSPARSE(DETAILS) returns a random stochastic
%   diagonal blocks matrix, composed by N smaller
%   matricies, where N is the number of elements of
%   vector DETAILS, whose i-th element is the dimension
%   of the i-th matrix. Then the i-th matrix of the
%   diagonal is linked with the i+1-th matrix while the
%   last matrix is linked with the first one. It returns
%   0 and displays 'Too few blocks' if DETAILS=0.
%   If details has at most N-1 zero elements, DETAILS
%   becomes DETAILS without any zero element. If DETAILS
%   isn't a column or a row vector it returns 0 and
%   displays 'Vector expected'.
rows=size(details,1);
columns=size(details,2);
if rows>1 && columns>1
    disp('Vector expected');
   M=0;
else
    if columns>1
        details=details';
    end
    zero=zeros(rows,1);
    if details==zero
        disp('Too few blocks');
       M=0;
    else
      if all(details)==0
         disp('details');
         disp(' ');
         disp(details);
         disp('becomes');
         disp(' ');
         inotzeros=find(details>0);
         dim=size(inotzeros,1);
         temp=zeros(dim,1);
         for r=1:dim
             temp(r)=details(inotzeros(r));
         end
         details=temp;
         disp(details);
      end
```

Algorithm B.37 Blocksparse part II

```
1         T=0;
2         while T==0
3                 numblocks=size(details,1);
4                 M=rsparse(details(1));
5                 for r=2:numblocks
6                     M=blkdiag(M,randomsparse(details(r)));
7                 end
8                 dim=size(M,1);
9                 row=1;
10                for s=1:(numblocks-1)
11                    M(row,row+details(s))=rand(1);
12                    plus=sum(M(row,1:dim));
13                    M(row,1:dim)=M(row,1:dim)/plus;
14                    row=row+details(s);
15                end
16                M(dim,1)=rand(1);
17                plus=sum(M(dim,1:dim));
18                M(dim,1:dim)=M(dim,1:dim)/plus;
19                T=periodt(M);
20        end
21    end
22 end
23 end
```

B.2.3 Simulations

B.2.3.1 Wsimul

Algorithm B.38 Wsimul

```
function sim = wsimul(dim)
if dim>8
    sim='Too many steps';
else
    P=randomwalk(dim);
    S=slem(P);
    T=periodt(P);
    pi0res=pi0(dim);
    pitres=pit(P,pi0res,T);
    [pibarres,Ts]=pibar(P,pi0res);
    [phi,phi2,phi0]=schsyssolve(P,pi0res,pibarres);
    pibartres=pibart(P,phi,pi0res);
    disp('A ''Random walk'' stochastic matrix P is')
    disp(' '),disp(P)
    disp('The SLEM is'),disp(' '),disp(S)
    disp('Period T is'),disp(' '),disp(T)
    disp('A random initial marginal distribution pi0 is')
    disp(' '),disp(pi0res)
    disp('Its stationary distribution starting at pi0 is')
    disp(' '),disp(pibarres)
    disp('which reaches in Ts steps'),disp(' '),disp(Ts)
    disp('The distribution at time T pi(T) is'),disp(' ')
    disp(pitres)
    disp('The harmonic function at time 0 phi(0) is')
    disp(' '),disp(phi0)
    disp('The harmonic function at time T phi(T)'),disp(' ')
    disp(phi2)
    disp('The distribution at time T pibar(T) is'),disp(' ')
    disp(pibartres)
    disp(' ')
    sim='simulation complete';
end
end
```

B.2.3.2 Csimul

Algorithm B.39 Csimul

```
1  function sim = csimul(dim)
2  if dim>15
3     sim='Too many steps';
4  else
5     P=cyclicalrw(dim);
6     S=slem(P);
7     T=periodt(P);
8     pi0res=pi0(dim);
9     pitres=pit(P,pi0res,T);
10    [pibarres,Ts]=pibar(P,pi0res);
11    [phi,phi2,phi0]=schsyssolve(P,pi0res,pibarres);
12    pibartres=pibart(P,phi,pi0res);
13    disp('A ''cyclical random walk'' stochastic matrix P is')
14    disp(' '),disp(P)
15    disp('The SLEM is'),disp(' '),disp(S)
16    disp('Period T in which P becomes full is'),disp(' ')
17    disp(T)
18    disp('A random initial marginal distribution pi0 is')
19    disp(' '),disp(pi0res)
20    disp('Its stationary distribution starting at pi0 is')
21    disp(' '),disp(pibarres)
22    disp('which reaches in Ts steps'),disp(' '),disp(Ts)
23    disp('The distribution at time T pi(T) is'),disp(' ')
24    disp(pitres)
25    disp('The harmonic function at time 0 phi(0) is')
26    disp(' '),disp(phi0)
27    disp('The harmonic function at time T phi(T)')
28    disp(' '),disp(phi2)
29    disp('The distribution at time T pibar(T) is')
30    disp(' '),disp(pibartres)
31    disp(' ')
32    sim='Simulation complete';
33 end
34 end
```

B.2.3.3 Ssimul

Algorithm B.40 Ssimul

```
function sim = ssimul(dim)
if dim>7
    sim='Too many steps';
else
    P=successruns(dim);
    S=slem(P);
    T=dim;
    pi0res=pi0(dim);
    pitres=pit(P,pi0res,T);
    [pibarres,Ts]=pibar(P,pi0res);
    [phi,phi2,phi0]=schsyssolve(P,pi0res,pibarres);
    pibartres=pibart(P,phi,pi0res);
    disp('A random ''Success runs'' stochastic matrix P is')
    disp(' '),disp(P)
    disp('The SLEM is'),disp(' '),disp(S)
    disp('Period T in which P becomes full is'),disp(' ')
    disp(T)
    disp('A random initial marginal distribution pi0 is')
    disp(' '),disp(pi0res)
    disp('Its stationary distribution starting at pi0 is')
    disp(' '),disp(pibarres)
    disp('which reaches in Ts steps'),disp(' '),disp(Ts)
    disp('The distribution at time T pi(T) is'),disp(' ')
    disp(pitres)
    disp('The harmonic function at time 0 phi(0) is')
    disp(' '),disp(phi0)
    disp('The harmonic function at time T phi(T)')
    disp(' '),disp(phi2)
    disp('The distribution at time T pibar(T) is'),disp(' ')
    disp(pibartres)
    disp(' ')
    sim='Simulation complete';
end
end
```

B.2.3.4 Rsimul

Algorithm B.41 Rsimul

```
1  function sim = rsimul(dim)
2  if dim<2
3     sim='Too many steps';
4  else
5     P=rsparse(dim);
6     S=slem(P);
7     T=periodt(P);
8     pi0res=pi0(dim);
9     [pibarres,Ts]=pibar(P,pi0res);
10    pitres=pit(P,pi0res,T);
11    [phi,phi2,phi0]=schsyssolve(P,pi0res,pibarres);
12    pibartres=pibart(P,phi,pi0res);
13    disp('A random stochastic matrix P is'),disp(' ')
14    disp(P) disp('The SLEM is'),disp(' '),disp(S)
15    disp('Period T in which P becomes full is')
16    disp(' '),disp(T)
17    disp('A random initial marginal distribution pi0 is')
18    disp(' '),disp(pi0res)
19    disp('Its stationary distribution  starting at pi0 is')
20    disp(' '),disp(pibarres)
21    disp('which reaches in Ts steps'),disp(' '),disp(Ts)
22    disp('The distribution at time T pi(T) is'),disp(' ')
23    disp(pitres)
24    disp('The harmonic function at time 0 phi(0) is')
25    disp(' '),disp(phi0)
26    disp('The harmonic function at time T phi(T)')
27    disp(' '),disp(phi2)
28    disp('The distribution at time T pibar(T) is'),disp(' ')
29    disp(pibartres)
30    disp(' ')
31    sim='Simulation complete';
32 end
33 end
```

B.2.3.5 Lsimul

Algorithm B.42 Lsimul

```
function sim = lsimul(n)
if n<2
    sim='Too few molecules';
else
    P=laplace(n);
    if P==0
    return;
    end
    dim=n+1;
    S=slem(P);
    T=periodt(P);
    pi0res=pi0(dim);
    pitres=pit(P,pi0res,T);
    [pibarres,Ts]=lpibar(n,pi0res);
    [phi,phi2,phi0]=schsyssolve(P,pi0res,pibarres);
    pibartres=pibart(P,phi,pi0res);
    disp('A ''Laplace'' matrix P with'),disp(' '),disp(n)
    disp(' '),disp('white balls and'),disp(' '),disp(n)
    disp(' '),disp('black balls'),disp(' '),disp('is')
    disp(P)
    disp('The SLEM is'),disp(' '),disp(S)
    disp('Period T in which P becomes full is')
    disp(' '),disp(T)
    disp('A random initial marginal distribution pi0 is')
    disp(' '),disp(pi0res)
    disp('Its stationary distribution starting at pi0 is')
    disp(' '),disp(pibarres)
    disp('which reaches in Ts steps starting at pi0')
    disp(' '),disp(Ts)
    disp('The distribution at time T pi(T) is'),disp(' ')
    disp(pitres)
    disp('The harmonic function at time 0 phi(0) is')
    disp(' '),disp(phi0)
    disp('The harmonic function at time T phi(T)')
    disp(' '),disp(phi2)
    disp('The distribution at time T pibar(T) is'),disp(' ')
    disp(pibartres)
    disp(' ')
    sim='Simulation complete';
end
end
```

B.2.3.6 Rggsimul

Algorithm B.43 Rggsimul

```
1  function sim = rggsimul(n,r)
2  if n<2
3     sim='Too few points';
4  else
5     P=randgeomgraphr(n,r);
6     S=slem(P);
7     T=periodt(P);
8     dim=size(P,1);
9     pi0res=pi0(dim);
10    pitres=pit(P,pi0res,T);
11    [pibarres,Ts]=pibar(P,pi0res);
12    [phi,phi2,phi0]=schsyssolve(P,pi0res,pibarres);
13    pibartres=pibart(P,phi,pi0res);
14    disp('A random stochastic matrix P is'),disp(' ')
15    disp(P)
16    disp('The SLEM is'),disp(' '),disp(S)
17    disp('Period T in which P becomes full is')
18    disp(' '),disp(T)
19    disp('A random initial marginal distribution pi0 is')
20    disp(' '),disp(pi0res)
21    disp('Its stationary distribution starting at pi0 is')
22    disp(' '),disp(pibarres)
23    disp('which reaches in Ts steps'),disp(' '),disp(Ts)
24    disp('The distribution at time T pi(T) is')
25    disp(' '),disp(pitres)
26    disp('The harmonic function at time 0 phi(0) is')
27    disp(' '),disp(phi0)
28    disp('The harmonic function at time T phi(T)')
29    disp(' '),disp(phi2)
30    disp('The distribution at time T pibar(T) is')
31    disp(' '),disp(pibartres)
32    disp(' ')
33    sim='Simulation complete';
34 end
35 end
```

B.2.3.7 Bsimul

Algorithm B.44 Bsimul

```
function sim = bsimul(a)
if lenght(a)<1
    sim='Too few blocks';
else
    P=blockrsparse(a);
    S=slem(P);
    T=periodt(P);
    dim=size(P,1);
    pi0res=pi0(dim);
    pitres=pit(P,pi0res,T);
    [pibarres,Ts]=pibar(P,pi0res);
    [phi,phi2,phi0]=schsyssolve(P,pi0res,pibarres);
    pibartres=pibart(P,phi,pi0res);
    disp('A random ''blocks diagonal'' stochastic matrix P is')
    disp(' '),disp(P)
    disp('The SLEM is'),disp(' '),disp(S)
    disp('Period T in which P becomes full is'),disp(' ')
    disp(T)
    disp('A random initial marginal distribution pi0 is')
    disp(' '),disp(pi0res)
    disp('Its stationary distribution starting at pi0 is')
    disp(' '),disp(pibarres)
    disp('which reaches in Ts steps'),disp(' '),disp(Ts)
    disp('The distribution at time T pi(T) is'),disp(' ')
    disp(pitres)
    disp('The harmonic function at time 0 phi(0) is')
    disp(' '),disp(phi0)
    disp('The harmonic function at time T phi(T)')
    disp(' '),disp(phi2)
    disp('The distribution at time T pibar(T) is'),disp(' ')
    disp(pibartres)
    disp(' ')
    sim='Simulation complete';
end
end
```

Appendix C

Matrix properties

Let us remind something of matrix theory.

C.1 Definitions

A set of numbers put by rows and columns is called matrix. Considering A a matrix with r rows and c columns, we denote it as a $r \times c$ matrix and it is represented by such a form

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1c} \\ a_{21} & a_{22} & \cdots & a_{2c} \\ \cdots & \cdots & \cdots & \cdots \\ a_{r1} & a_{r2} & \cdots & \cdots a_{rc} \end{bmatrix}, \tag{C.1}$$

where a_{ij} are its elements. If $r = c$ A is called square matrix, otherwise rectangular. On the other hand if $r = 1$ or $c = 1$ matrix A, sometimes indicated as $A = [a_{ij}]$, is respectively reduced to a row vector or a column vector. A sub matrix of A is matrix A without any rows or columns.

C.2 Operations

If A and B are matrix with the same dimension, their summation can be defined as

$$C = A + B, \tag{C.2}$$

where

$$c_{ij} = a_{ij} + b_{ij}. \tag{C.3}$$

This operation is commutative and associative, so that

$$A + B = B + A \tag{C.4}$$

$$(A + B) + C = A + (B + C) \tag{C.5}$$

Two matrix can also be multiplied if they are compatible, that is, given a matrix $A = m' \times n'$ and $B = n" \times m"$ they are compatible if $m' = m"$ or $n' = n"$. In this last case product matrix

$$C = AB \tag{C.6}$$

has dimension $m' \times m"$, where

$$c_{ij} = \sum_{k=1}^{m'} a_{ik} b_{kj}. \tag{C.7}$$

Multiplication of matrices satisfies associative property but in general not the commutative one.

$$A\,(BC) = (AB)\,C \tag{C.8}$$

$$AB = BA \tag{C.9}$$

In addition a $r \times c$ matrix can be transposed by switching its rows and columns. We write A^* as the transposed of A,

$$A^* = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{r1} \\ a_{12} & a_{22} & \cdots & a_{r2} \\ \cdots & \cdots & \cdots & \cdots \\ a_{1c} & a_{2c} & \cdots & a_{rc} \end{bmatrix} \tag{C.10}$$

with $c \times r$. A matrix is called symmetric if

$$A^* = A \tag{C.11}$$

and satisfies these properties:

$$(AB)^* = B^* A^*; \tag{C.12}$$

$$(ABC)^* = C^* B^* A^*; \tag{C.13}$$

$$(A + B)^* = A^* + B^*. \tag{C.14}$$

C.3 Determinant and inverse matrix

We define as determinant of a $n \times n$ square matrix

$$det\,(A) = \sum_{k=1}^{n} a_{ij}\gamma_{ij}$$
$$i = 1, 2, \ldots n, \tag{C.15}$$

where γ_{ij} is called co-factor and

$$\gamma_{ij} = (-1)^{i+j}\, det\,(M_{ij})\,, \tag{C.16}$$

where $det\,(M_{ij})$ is called minor and coincides with matrix A except its i-th row and j-th column. We notice that M_{ij} is always a $(n-1)\times(n-1)$ and minors differ from co-factors at most by sign. The adjoint of a matrix A is the transposition of its co-factors matrix:

$$adj\,(A) = [\gamma_{ij}]^{*}\,. \tag{C.17}$$

It can be shown that

$$A\,adj\,(A) = (det\,(A))\,I \tag{C.18}$$

where I is the identity matrix, that is an element a_{ij} is 1 if $i = j$, namely if belongs to the diagonal, 0 otherwise. If $det\,(A) \neq 0$ we say the inverse of a matrix A, called singular, is

$$A^{-1} = \frac{adj\,(A)}{det\,(A)} \tag{C.19}$$

with the property that

$$AA^{-1} = A^{-1}A = I. \tag{C.20}$$

In addiction the inverse matrix of a product is the product of the singular inverse matrices:

$$(ABC)^{-1} = C^{-1}B^{-1}A^{-1}. \tag{C.21}$$

C.4 Rank

The number of rows or columns linearly independent is called rank of a matrix. If A has rank r all the $(r+1)\times(r+1)$ submatrices of A are singular, while at least one submatrix $r \times r$ isn't. It is true that

$$row\,rank\,(A) = column\,rank\,(A) \tag{C.22}$$

C.5 Eigenvalues and eigenvectors

Every scalar λ and every vector not null vector v which satisfies

$$Av = \lambda v \tag{C.23}$$

are respectively called eigenvalue and right eigenvector associated with A. By $(B.23)$ we gain

$$(\lambda I - A)\, v = 0 \tag{C.24}$$

and considering that v is not null we have

$$det\,(\lambda I - A) = 0. \tag{C.25}$$

Let v be the eigenvector associated with λ, hence αv is an eigenvector. Eigenvectors are generally normalized to have unity length, that is $\parallel v \parallel^2 = v^* v = 1$. If w^* is a row vector not null so that

$$w^* A = \lambda w^*, \tag{C.26}$$

w is called left eigenvector of A or with the same meaning w is a right eigenvector of A^*

$$A^* w = \lambda w. \tag{C.27}$$