The Islamic University of Gaza
Deanery of Higher Studies
Faculty of Science
Department of Mathematics

Markov Chain Monte Carlo Method and Perfect Simulation

Presented By
Arwah Noaman Karam

Supervisor
Professor Mohamed I. Riffi

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF MATHEMATICS
2007

To my parents,,,

Contents

1 Markov Chains **3**

1.1 Definitions and Basic Properties 3
1.2 Higher Transition Probabilities 6
1.3 Irreducible and Aperiodic Markov Chains 11
1.4 Important Properties of Markov Chains 13
1.5 Stationary Distributions 16
1.6 Detailed Balance and Time Reversal 21

2 Markov Chain Monte Carlo Algorithms **25**

2.1 Metropolis Algorithm 26
2.2 Metropolis-Hastings Algorithm 28
2.3 The Gibbs Sampler 31

3 Coupling **36**

3.1 Convergence in Variation and Coupling 36
3.2 Coalescence 38
3.3 Forward Coupling 45

4 Perfect Simulation **49**

4.1 Introduction 49
4.2 Coupling From The Past Algorithm 50
4.3 Monotonicity and Anti-Monotonicity 60
4.4 The Ising Model 66
4.5 Dominated Coupling From The Past 70

List of Figures

1.1 Transition graphs for the Markov chain in Examples (1.1.1), (1.2.1). 11

3.1 Forward simulation with t=4 of Example (3.3.1). 46
3.2 Forward simulation for Example (3.3.2) with t=8. 48

4.1 All possible transitions for Example (4.2.1). 52
4.2 A run of the Propp-Wilson algorithm "CFTP" for the Markov chain of Example (4.2.1). Transitions that are carried out in the running of the algorithm are indicated with solid lines; others are dashed. 53
4.3 Simulation from the past for Example (3.3.1) with t=1. 54
4.4 CFTP for Example (3.3.1) with t=4. 54
4.5 CFTP for Example (3.3.2). The paths started in state 0 and in state 3 are shown as solid lines. The dotted lines are the paths started from intermediate states. 55
4.6 CFTP for the Markov chain in Example (4.3.2). The dotted line show the maximal and minimal path. 65

Acknowledgements

First of all, praise is to God who gives me the chance to complete my Master Degree. I am grateful to my supervisor, Professor Mohammed I. Riffi, for a substantial amount of guidance, advice and support in producing this thesis. I am also grateful to all the members of the Mathematics Department in the Islamic University of Gaza, in particular Professor Eissa D. Habil, and Dr. Samir K. Safi for checking and valuable suggestions.

My sincere gratitude to my family, specially my mother for her love and great deal of support and patience and my brother Dr. Emad, for his effort to provide me with many references which helped me. A special thanks to my little brother Mohammed for excellent cooperation. He has been helpful. Finally, I am also thankful to all my friends for their love and encouragement.

Abstract

Markov Chain Monte Carlo method is used to sample from complicated multivariate distribution with normalizing constants that may not be computable and from which direct sampling is not feasible. Recent years have seen the development of a new, exciting generation of Markov Chain Monte Carlo method: perfect simulation algorithms.
In this thesis, we give a review of the new perfect simulation algorithms using Markov chains, focussed on the method called Coupling From The Past, since it allows not only an approximate but perfect (exact) simulation of the stationary distribution of finite state space Markov chain.

Introduction

The method of simulating a Markov chain whose stationary distribution is the distribution we want to sample from is known in general as Markov chain Monte Carlo or in short **(MCMC)** algorithms such as the Metropolis-Hastings algorithm and the Gibbs sampler are widely used in statistics, chemistry and physics for exploring complicated probability distributions, also they have become extremely popular for Bayesian inference problems ([7], [9], [20], [23]) and for problems in other areas, such as spatial statistics, statistical physics, and computer science ([5], [18]).

The traditional way to proceed is to run the Markov chain for a long time (called the burn-in time) in the hope that by the end of this period the Markov chain will be sufficiently close to stationarity that we may assume that we are now sampling from the required distribution. The problem in **MCMC** is that it is rarely possible to know when the Markov chain which is used for simulation has reached equilibrium. This problem was solved by the introduction of Coupling From The Past (**CFTP**) which was introduced by Propp and Wilson [18] (see also [7] and [9]) and has been studied and used by a number of authors ([12], [16], [17]). By searching backwards in time until paths from all starting states have coalesced, this algorithm uses the Markov transition kernel $\mathbf{P}$ to sample exactly from the stationary distribution π.

The main purpose of this thesis is to give a detailed introduction to the general idea of **MCMC** methods and perfect simulation, and concentrate on one particular method called **CFTP** and its extensions.

This thesis is organized in four chapters. In the first chapter we present the theory of Markov chains, basic definitions, important properties of Markov chains and give the conditions for stationary distribution to exist.

In the second chapter, we talk about **MCMC**, and their algorithms which

attempt to simulate direct draws from some complex distributions of interest, explain the need to **MCMC**.

The third chapter describes coupling for Markov chains which are a basic tool in perfect simulation from the desired distribution.

Finally, in the fourth chapter, we discuss perfect simulation focusing on CFTP as developed in Propp and Wilson [18], and moving on to very useful extensions of the method, Dominated Coupling From The Past.

Notation: Throughout this thesis we will use the symbols:

- π for the stationary distribution of some Markov chain.
- **MCMC** for Markov chain Monte Carlo.
- **CFTP** for Coupling From The Past.
- **DCFTP** for Dominated Coupling From The Past.
- $\mathcal{L}(X)$ for the probability distribution of a random variable X.

Chapter 1

Markov Chains

To understand the properties of a probability measure, which may have a complicated state space, or may be hard to deal with by explicit calculation, we construct a suitable Markov chain whose long run stationary distribution is the target probability measure.

In this chapter, we will introduce the concepts of Markov chains and common algorithms for their simulation, that are necessary to understand the rest of the chapters. So we will be dealing with Markov chains which have a finite state space, and so the definitions we give here will apply to this discrete case, for continuous case, we change summation to integration.

1.1 Definitions and Basic Properties

Suppose we have some experiments $X_1, X_2, X_3, \ldots$ whose results (outcomes) fulfil the following properties:

(1) Any outcome belong to a set of outcomes $\{x_1, x_2, \ldots, x_m\}$, which we will call the sample space or the state space for this system. For example: if the outcome of the experiment numbered t is x_i, then we say the system is in state x_i at step t if $X_t = x_i$.

(2) The outcome of any experiment is dependent only upon the immediate previous outcome.

For every couple of states (x, y), we can find the probability $P(x, y) = P(x \to y)$ which is the probability of moving from one state x to another state y at a given time t.

More precisely, $P(x, y) = \mathbb{P}(X_{t+1} = y \mid X_t = x)$. Such these stochastic experiments are called finite **Markov Chains**.

Definition 1.1.1. *[10] A Markov chain is a sequence of random variables $\{X_0, X_1, X_2, \dots\} : \Omega \longrightarrow E$ defined on a probability space and mapping to a finite state space E with the property (**Markov property**) the conditional probability distribution of the next future state X_{t+1} given the present and past states is a function of the present state X_t alone, i.e.,*

$$\mathbb{P}(X_{t+1} = y \mid X_0 = x_0, X_1 = x_1, \dots, X_t = x)$$

$$= \mathbb{P}(X_{t+1} = y \mid X_t = x) = P(x, y)$$

for all steps t, all states $x_0, x_1, \dots, x, y$.

It is called a Homogeneous Markov Chain (HMC) if the right hand side of this equation is independent of t. The range of the variables, i.e., the set of their values, is called the state space.

Definition 1.1.2. *[10] An $|E| \times |E|$ matrix $\mathbf{P}$ with elements $P(x, y) : x, y \in E$ is **a transition matrix** for a Markov chain with finite state space E if*

$$P(x, y) = \mathbb{P}(X_{t+1} = y \mid X_t = x)$$

for all $x, y \in E$.

*The elements of the transition matrix $\mathbf{P}$ are called transition probabilities (**transition kernel**). The transition probability $P(x, y)$ is the conditional probability of being in state y " tomorrow" given that we are in state x "today". Every transition matrix is a stochastic square matrix with zero or positive elements less than or equal to one such that the summation of elements in each row is unity, i.e.,*

$$P(x, y) \geq 0, \quad x, y \in E, \qquad \text{and} \qquad \sum_{y \in E} P(x, y) = 1, \quad x \in E. \tag{1.1.1}$$

A stochastic matrix $\mathbf{P}$ is *regular* if some matrix power $\mathbf{P}^k$ contains only strictly positive entries.

Remark 1.1.1. *If the transition probabilities are fixed for all t, then the chain is considered homogeneous; meaning they don't change over time, so the probability of going from state x at time $t+1$ to state y at time $t+k+1$ is the same as the probability of going from state x at time t to state y at time $t+k$.*

Example 1.1.1. ***A very simple weather model.***
The probabilities of weather conditions, given the weather on the preceding day, can be represented by a transition matrix $\mathbf{P}=\begin{pmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{pmatrix}$.

The matrix $\mathbf{P}$ represents the weather model in which a sunny day is 90% likely to be followed by another sunny day, and a rainy day is 50% likely to be followed by another rainy day. The columns can be labelled "sunny" and "rainy" respectively, and the rows can be labelled in the same order.

$P(x, y)$ is the probability that, if a given day is of type x, it will be followed by a day of type y.

Notice that the rows of $\mathbf{P}$ sum to 1, this is because $\mathbf{P}$ is a stochastic matrix.

Example 1.1.2. *At an intersection, a working traffic light will be out of order the next day with probability* 0.07*, and an out-of order light will be working the next day with probability* 0.88*. Let $X_t = 1$ if on day t the light will work; $X_t = 0$ if on day t it will not work. Then the transition probability matrix is given by* $\mathbf{P} = \begin{pmatrix} 0.12 & 0.88 \\ 0.07 & 0.93 \end{pmatrix}$.

Definition 1.1.3. *In mathematics and statistics, a probability vector or a stochastic vector is a vector with non-negative entries that add up to one. Here are examples of probability vectors*

$$x_0 = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix},\ x_1 = \begin{pmatrix} 0.65 \\ 0.35 \end{pmatrix}.$$

1.2 Higher Transition Probabilities

Definition 1.2.1. *The "**one-step**" transition probability is defined as* $\mathbb{P}(X_{t+1} = y \mid X_t = x) = P(x, y)$, *which is known as the "**short term**" behavior of a Markov chain. Suppose now that the probability that the system is in state* x_i *at an arbitrary time (step)* t *is* $p_i^t = \mathbb{P}(X_t = x_i)$ *which is known as the "**long term**" behavior of the Markov chain, and if this probability is a probability vector like* P^t, *then the initial probability distribution at time zero is* $p_i^0 = \mathbb{P}(X_0 = x_i)$ *represented as a row vector of the state space probabilities at step zero is given by*

$$\begin{aligned} P^0 &= (p_1^0, p_2^0, \ldots, p_m^0) \\ &= (\mathbb{P}(X_0 = x_1), \mathbb{P}(X_0 = x_2), \ldots, \mathbb{P}(X_0 = x_m)), \end{aligned}$$

and $\sum_{i=1}^{m} p_i^0 = 1$.

Often all elements of P^0 *are zero except for a single element of one, corresponding to the process starting in that particular state.*
Similarly, the probability distribution at the first time (step) is

$$P^1 = (p_1^1, p_2^1, \ldots, p_m^1).$$

And the probability distribution of the Markov chain at time t *is*

$$\begin{aligned} P^t &= (p_1^t, p_2^t, \ldots, p_m^t) \\ &= (\mathbb{P}(X_t = x_1), \mathbb{P}(X_t = x_2), \ldots, \mathbb{P}(X_t = x_m)). \end{aligned}$$

Suppose we know the transition matrix and the distribution of the initial state $X_0 = x_i$, then we can compute $P^n(x_i, x_j) = \mathbb{P}(X_n = x_j | X_0 = x_i)$, the probability for moving from state x_i to state x_j at n-steps exactly.

First: if the chain starts in state $x_i \neq x_j$ and then moves to state x_j in

one step, then,

$$\begin{aligned} p_j^1 &= \mathbb{P}(X_1 = x_j) \\ &= \sum_{i=1}^{m} \mathbb{P}(X_0 = x_i, X_1 = x_j) \\ &= \sum_{i=1}^{m} \mathbb{P}(X_0 = x_i).\mathbb{P}(X_1 = x_j \mid X_0 = x_i) \\ &= \sum_{i=1}^{m} p_i^0 P(x_i, x_j). \end{aligned}$$

Second: if the chain starts in state $x_i \neq x_j$ and then moves to state x_j after two additional transitions, we can compute this probability from the one step as follows:

$$\begin{aligned} P^2(x_i, x_j) &= \mathbb{P}(X_2 = x_j | X_0 = x_i) \\ &= \sum_{k=1}^{m} \mathbb{P}(X_2 = x_j, X_1 = x_k | X_0 = x_i) \\ &= \sum_{k=1}^{m} \mathbb{P}(X_2 = x_j \mid X_1 = x_k, X_0 = x_i).\mathbb{P}(X_1 = x_k | X_0 = x_i) \quad \text{by Bayes'formula} \\ &= \sum_{k=1}^{m} \mathbb{P}(X_2 = x_j \mid X_1 = x_k).\mathbb{P}(X_1 = x_k | X_0 = x_i) \quad \text{by Markov property} \\ &= \sum_{k=1}^{m} P(x_i, x_k) P(x_k, x_j). \end{aligned} \tag{1.2.1}$$

In general, we define the n-step transition probabilities $P^n(x, y)$ by

$$P^n(x, y) = \mathbb{P}(X_{t+n} = y \mid X_n = x),$$

which is just the x, y^{th} element of the matrix $\mathbf{P}^n$, where $\mathbf{P}^n$ is the n^{th} power of the matrix $\mathbf{P}$.

Let us show that the transition probabilities $P^n(x, y)$ satisfy Chapman-Kolmogorov equation.

Theorem 1.2.1. *[19]* ***Chapman-Kolmogorov Equation****.*

For all $x, y \in E = \{x_1, x_2, \ldots, x_m\}$*, we have that*

$$P^{n+1}(x,y) = \sum_{z=1}^{m} P^n(x,z)P^1(z,y)$$

Proof.

$$\begin{aligned}
P^{n+1}(x,y) &= \mathbb{P}(X_{n+1} = y | X_0 = x) \\
&= \sum_{z=1}^{m} \mathbb{P}(X_{n+1} = y, X_n = z | X_0 = x) \\
&= \sum_{z=1}^{m} \frac{\mathbb{P}(X_{n+1} = y, X_n = z, X_0 = x)}{\mathbb{P}(X_0 = x)} \\
&= \sum_{z=1}^{m} \frac{\mathbb{P}(X_{n+1} = y | X_n = z, X_0 = x).\mathbb{P}(X_n = z, X_0 = x)}{\mathbb{P}(X_0 = x)} \\
&= \sum_{z=1}^{m} \mathbb{P}(X_{n+1} = y \mid X_n = z).\mathbb{P}(X_n = z | X_0 = x) \quad \text{by Markov property} \\
&= \sum_{z=1}^{m} P^n(x,z)P^1(z,y).
\end{aligned} \tag{1.2.2}$$

□

If $\mathbf{P}^n = P^n(x,y)$ denotes the matrix of the $n-$ step transition probabilities $P^n(x,y)$, then

$$\mathbf{P}^n = \mathbf{P}^{n-1}.\mathbf{P}^1$$

Example 1.2.1. *Let* $\boldsymbol{P}$ *be the four by four transition matrix*

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

If the Markov chain is in state 3 *at time* t*, then to find the probability that it will stay in state* 3 *at time* $t+2$*, we calculate* $P^2(3,3)$ *from Chapman-Kolmogorov equation*

$$\begin{aligned} P^2(3,3) &= \sum_{z=1}^{4} P^1(3,z)P^1(z,3) \\ &= P^1(3,1)P^1(1,3) + P^1(3,2)P^1(2,3) + P^1(3,3)P^1(3,3) + P^1(3,4)P^1(4,3) \\ &= (0)(0) + (\frac{1}{2})(\frac{1}{2}) + (0)(0) + (\frac{1}{2})(\frac{1}{2}) \\ &= \frac{1}{2}. \end{aligned}$$

And if the Markov chain is in state 4 *at time* t*, then the probability that it will go to state* 1 *at time* $t+2$ *is*

$$\begin{aligned} P^2(4,1) &= \sum_{z=1}^{4} P^1(4,z)P^1(z,1) \\ &= P^1(4,1)P^1(1,1) + P^1(4,2)P^1(2,1) + P^1(4,3)P^1(3,1) + P^1(4,4)P^1(4,1) \\ &= (\frac{1}{2})(0) + (0)(\frac{1}{2}) + (\frac{1}{2})(0) + (0)(\frac{1}{2}) \\ &= 0. \end{aligned}$$

The answer may also be calculated as the appropriate elements of the product of the transition matrix with itself, that is $\mathbf{P}^2$, which is given as

$$\mathbf{P}^2 = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}.$$

The probability $P^2(3,3)$ is the vector product of the third row of the first matrix with the third column of the second matrix, that is,

$$P^2(3,3) = \mathbb{P}(X_{t+2} = 3 | X_t = 3) = \frac{1}{2}.$$

Similarly,

$$P^2(4,1) = \mathbb{P}(X_{t+2} = 1 | X_t = 4) = 0.$$

The Chapman-Kolmogorov equation in matrix form becomes

$$P^{n+1} = P^n \mathbf{P}.$$

Using the matrix form, we find that

$$P^n = P^{n-1}\mathbf{P} = (P^{n-2}\mathbf{P})\mathbf{P} = P^{n-2}\mathbf{P}^2 = \ldots = P^0\mathbf{P}^n.$$

Theorem 1.2.2. *[10] Let* $\mathbf{P}$ *be the transition matrix of a Markov chain with initial distribution* P^0*, we have for any* t *that the distribution* P^t *at time* t *satisfies*

$$P^t = P^0 \boldsymbol{P}^t.$$

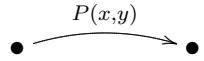
Example 1.2.2. ***Predicting the weather***.
Refer to Example (1.1.1). The weather on day 0 *is known to be sunny. This is represented by a vector* $P^0 = \begin{pmatrix} 1 & 0 \end{pmatrix}$ *in which the "sunny" entry is* 100%*, and the "rainy" entry is* 0%*.*
The weather on day 1 *can be predicted by:*

$$P^1 = P^0 \boldsymbol{P} = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{pmatrix} = \begin{pmatrix} 0.9 & 0.1 \end{pmatrix}.$$

Thus, there is an 90% *chance that day* 1 *will also be sunny.*
The weather on day 2 *can be predicted in the same way:*

$$P^2 = P^1 \boldsymbol{P} = P^0 \boldsymbol{P}^2 = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{pmatrix}^2 = \begin{pmatrix} 0.86 & 0.4 \end{pmatrix}.$$

As we saw before, if the state space is finite, the transition probability distribution can be represented as a matrix, called the transition matrix, with the x, y^{th} element equal to $P(x, y) = \mathbb{P}(X_{t+1} = y \mid X_t = x)$. But the transition probabilities can be described by a (directed) graph whose vertices are the states, and an arrow from state x to state y with the number $P(x, y)$ over it indicates that it is possible to pass from point x to point y with probability $P(x, y)$.
When $P(x, y) = 0$, the corresponding arrow is omitted.



This is explained by showing the transition graph of the transition matrix of Examples (1.1.1), (1.2.1).

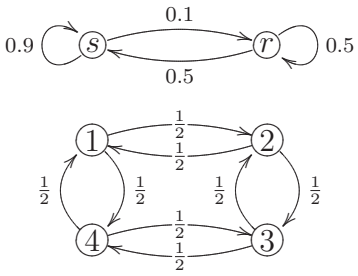


Figure 1.1: Transition graphs for the Markov chain in Examples (1.1.1), (1.2.1).

1.3 Irreducible and Aperiodic Markov Chains

We will discuss two such conditions on Markov chains: irreducibility and aperiodicity. These conditions are of central importance in Markov theory, and they play a key role in the study of stationary distributions.

Suppose we have a Markov chain with state space E, and transition matrix $\mathbf{P}$. Two states, x and y, communicate writing $x \leftrightarrow y$, if there exist finite m, n such that $P^m(x, y) > 0$ and $P^n(y, x) > 0$.

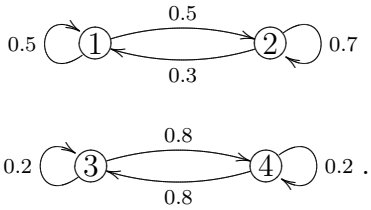
Definition 1.3.1. *([10], [19]) A Markov chain with state space E and a transition matrix $\mathbf{P}$ is connected or irreducible if for all states $x, y \in E$ there exists a time $t \leq |E|$ such that $P^t(x, y) > 0$.*

In an irreducible Markov chain there is a positive probability of going from every state to any other state in a finite number of steps, that is every state y is eventually reachable from any start state x [$\mathbb{P}(X_t = y \mid X_0 = x) > 0$, for some $t \geq 0$]; i.e., all states communicate, as one can always go from any state to any other state although it may take more than one step.
An easy way to verify that a Markov chain is irreducible, is to look at its transition graph, and check that from each state there is a sequence of arrows leading to any other state, for example the graph of Example(1.1.1) is irreducible.
Let us also have a look at an example which is not irreducible.

Example 1.3.1. *[10]**A reducible Markov Chain**.*

Consider a Markov chain with state space $\{1, 2, 3, 4\}$ and transition matrix

$$\mathbf{P}=\begin{pmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.3 & 0.7 & 0 & 0 \\ 0 & 0 & 0.2 & 0.8 \\ 0 & 0 & 0.8 & 0.2 \end{pmatrix}$$ *which has a transition graph as follows:*



We see that if the chain starts in state 1 *or* 2*, then it is restricted to state* 1 *or* 2 *forever. Similarly, if it starts in state* 3 *or* 4*, then it can never leave the subset* $\{3, 4\}$ *of the state space. Hence, the chain is reducible.*

Definition 1.3.2. *([10], [19]) A state* x *is said to have period* d *if* $P^t(x, x) = 0$ *whenever* t *is not divisible by* d*. The period* $d(x)$ *of a state* $x \in E$ *is defined as*

$$d(x) := \gcd\{t \geq 1 : P^t(x, x) > 0, \forall x\}.$$

That is

$$\gcd\{t : \mathbb{P}(X_t = x \mid X_0 = x) > 0\},$$

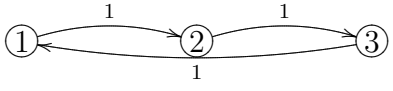
i.e., a return to state x *after* t *transitions has positive probability only if* t *is a multiple of* d*. If* $d(x) = 1$*, then the state* x *is aperiodic, that is a Markov chain is said to be* ***aperiodic*** *if all its states have period* 1*, otherwise the chain is said to be periodic.*

Example 1.3.2. *[10]*

The Markov chain in Example (1.1.1) is aperiodic since $P^t(x, x) > 0$ *for all* t *and all states* x*.*

Example 1.3.3. *This Markov chain on the state space* $\{1, 2, 3\}$ *with transition matrix*

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



is periodic. It has period 3 *and it alternates at three distributions.*

The transition graph of the Markov chain in Example (1.2.1) has period $d = 2$, since a particle can return to each state after $2, 4, 6, \dots$ steps.

Lemma 1.3.1. *If the states x and y communicate, and if x has period d, then so has state y.*

This means that the state x is aperiodic if $x \leftrightarrow y$ and $P(y, y) > 0$.

One reason for the usefulness of aperiodicity is the following theorem.

Theorem 1.3.1. *[10] Suppose that we have an aperiodic Markov chain $\{X_0, X_1, \dots\}$ with state space E and transition matrix $\mathbf{P}$. Then there exists an $T < \infty$ such that*

$$P^t(x, x) > 0$$

for all $x \in E$ and all $t \geq T$.

By combining aperiodicity and irreducibility, we get the following important corollary, which will be used to prove the so-called Markov chain convergence theorem, see Theorem 3.2.1.

Corollary 1.3.1. *[10] Now suppose our Markov chain is aperiodic and irreducible. Then there exist an $T < \infty$ such that*

$$P^t(x, y) > 0$$

for all $x, y \in E$ and all $t \geq T$.

1.4 Important Properties of Markov Chains

In the following section we will summarize some of the most common properties of Markov chains that are used in the context of **MCMC**. We always refer to a Markov chain $\{X_0, X_1, X_2, \dots\}$ with transition matrix $\mathbf{P}$ on a finite state space E.

Given an initial state $X_0 = x$ of a Markov chain with transition matrix $\mathbf{P}$, the probability of first return to state x at time t is given by

$$f^t(x, x) = \mathbb{P}(X_t = x, \text{ and for } \ 1 \leq s \leq t-1, X_s \neq x | X_0 = x)$$

and for $x \neq y$,

$$f^t(x,y) = \mathbb{P}(X_t = y, \text{ and for } 1 \leq s \leq t-1, X_s \neq y | X_0 = x)$$

is the probability of first arrival at state y at time t, (i.e., first transition into state y occurs at time t).

For each $x \in E$, let

$$f(x,x) = \sum_{t=1}^{\infty} f^t(x,x)$$

be the probability that leaves state x sooner or later, return to that state, (i.e., there is a transition into state x at some time $t > 0$).
In other words, $f(x,x) = \mathbb{P}(T < \infty)$, where $T = \inf\{t \geq 1 : X_t = x\}$ is the number of steps for reaching state x for the first time known as the hitting time with $T = \infty$ if no such time exists.
The expected number of time steps to reach state y starting at x is

$$h(x,y) = \sum_{t=1}^{\infty} t f^t(x,y).$$

Definition 1.4.1. *[19] A state, x, is **a recurrent state** if $f(x,x) = 1$. It is called **transient** if $f(x,x) < 1$, and therefore $h(x,x) = \infty$. Thus, x is a transient state if given that we start in state x, there is a non-zero probability that we will never return back to x. **Positive recurrent** means that the expected return time is finite for every state.*

Lemma 1.4.1. *[19]*

(a) *The state x is recurrent if and only if* $\sum_{t=1}^{\infty} P^t(x,x) = \infty$.

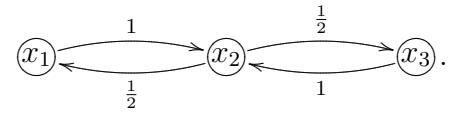
(b) *If state y is transient, then* $\sum_{t=1}^{\infty} P^t(x,y) < \infty$.

Theorem 1.4.1. *[3]If x is recurrent and $P(x,y) > 0$, then y is recurrent and $P(y,x) = 1$.*

A simple example of Markov chain with two recurrent classes [a class will be called recurrent, if all (one) of the states are (is) recurrent] can be represented by transition matrix of Example (1.3.1), since states $1 \leftrightarrow 2$ but neither state 1 nor state 2 communicates with state 3 or 4.

Example 1.4.1. *Here is a Markov chain on state space* $E = \{x_1, x_2, x_3\}$, *with transition matrix*

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \end{pmatrix}$$



The transition graph of this Markov chain has three recurrent states, and one recurrent class.

Note: Sometimes the terms indecomposable, a cyclic, and persistent are used as synonyms for "irreducible", "aperiodic", and "recurrent" respectively.

Example 1.4.2. *Consider the Markov chain with state space* $\{0, 1, 2, 3, 4\}$ *and transition matrix*

$$\mathbf{P} = \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 & \frac{2}{3} \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & \frac{3}{5} & \frac{2}{5} \end{pmatrix}.$$

Since $P(4,4) > 0$ *and states* 3 *and* 4 *communicate, it follows from Lemma (1.3.1) that these states are aperiodic. State* 0 *has period* 3. *From* 0 *the chain moves to either* 1 *or* 4. *From* 4 *it never moves back to* 0 *while from* 1 *the chain moves to either* 2 *or* 3. *From* 3 *it never moves back to* 0. *From* 2 *it moves to either* 0 *or* 4. *This shows that* 0 *is transient and that* $P^n(0,0) > 0$ *only if* n *is a multiple of* 3. *As a result, the transient state* 0 *has period* 3. *Since* $0 \leftrightarrow 1 \leftrightarrow 2$. *By Lemma (1.3.1) the states* 1 *and* 2 *also have period* 3. $P(4,3) > 0$ *and* $P(3,4) = 1$ *so* 4 *must be recurrent, but* $P(1,3) > 0$ *and* $P(3,1) = 0$ *so* 1 *must be transient, or we would contradict Theorem (1.4.1).*

In the context of **MCMC** a question of particular interest is the question of the long-term behavior of a Markov chain. Given certain conditions, does the distribution of the chain converge to a well defined and unique limit?. The concept of irreducibility and aperiodicity will provide an answer.

1.5 Stationary Distributions

We consider one of the central issues in Markov theory: asymptotic for the long-term behavior of Markov chains.

Example 1.5.1. *Suppose the state space are (Rain, Sunny, Cloudy) and weather follows a Markov chain, that is, the probability of tomorrow's weather simply depends on today's weather, and not any other previous days.*

Suppose the probability transition given today is rainy are

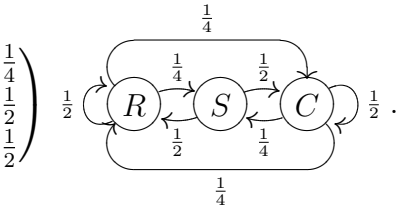
$$\mathbb{P}(Rain\ tomorrow \mid Rain\ today) = \frac{1}{2}$$

$$\mathbb{P}(Sunny\ tomorrow \mid Rain\ today) = \frac{1}{4}$$

$$\mathbb{P}(Cloudly\ tomorrow \mid Rain\ today) = \frac{1}{4}$$

The first row of the transition probability matrix thus becomes $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$. *Suppose the rest of the transition matrix is given by*

$$\boldsymbol{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix}$$



This Markov chain is irreducible, as all states communicate with each other. Suppose today is cloudy. What is the expected weather two days from now ? seven days?

Here $P^0 = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}$. Hence

$$\begin{aligned} P^2 = P^0\mathbf{P}^2 &= (001) \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{7}{16} & \frac{3}{16} & \frac{3}{8} \\ \frac{3}{8} & \frac{1}{4} & \frac{3}{8} \\ \frac{3}{8} & \frac{3}{16} & \frac{7}{16} \end{pmatrix} \\ &= \begin{pmatrix} \frac{3}{8} & \frac{3}{16} & \frac{7}{16} \end{pmatrix} \\ &= \begin{pmatrix} 0.375 & 0.1875 & 0.4375 \end{pmatrix}, \end{aligned}$$

and

$$\begin{aligned} P^7 = P^0\mathbf{P}^7 &= \begin{pmatrix} \frac{819}{2048} & \frac{819}{4096} & \frac{1639}{4096} \end{pmatrix} \\ &= \begin{pmatrix} 0.3999 & 0.1999 & 0.4001 \end{pmatrix} \\ &= \begin{pmatrix} 0.4 & 0.2 & 0.4 \end{pmatrix}. \end{aligned}$$

Conversely, suppose today is rainy, so that $P^0 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$.

The expected weather becomes

$$\begin{aligned} P^2 &= \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{7}{16} & \frac{3}{16} & \frac{3}{8} \\ \frac{3}{8} & \frac{1}{4} & \frac{3}{8} \\ \frac{3}{8} & \frac{3}{16} & \frac{7}{16} \end{pmatrix} \\ &= \begin{pmatrix} \frac{7}{16} & \frac{3}{16} & \frac{3}{8} \end{pmatrix} \\ &= \begin{pmatrix} 0.4375 & 0.1875 & 0.375 \end{pmatrix}, \end{aligned}$$

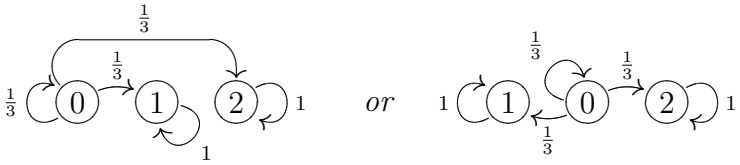
and $P^7 = \begin{pmatrix} 0.4 & 0.2 & 0.4 \end{pmatrix}$.

After a sufficient amount of time, the expected weather is independent of the starting value. In other words, the chain has reached a stationary distribution, where the probability values are independent of the actual starting value. The previous example illustrates, a Markov chain may reach a stationary distribution π, where the vector of probabilities of being in any particular given state is independent of the initial condition.

Example 1.5.2. *Consider the Markov chain on state space* $E = \{0, 1, 2\}$ *with the transition matrix*

$$\mathbf{P}=\begin{pmatrix}\frac{1}{3} & \frac{1}{3} & \frac{1}{3}\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}. \text{ Then } \mathbf{P}^t=\begin{pmatrix}\frac{1}{3^t} & \frac{1-3^{-t}}{2} & \frac{1-3^{-t}}{2}\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix} \xrightarrow{t\to\infty} \begin{pmatrix}0 & \frac{1}{2} & \frac{1}{2}\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix},$$

which is called the limiting probabilities which are not independent of the initial state, because the states 1 *and* 2 *do not communicate with each other, that is, the Markov chain is reducible as we see in its transition graph.*



Definition 1.5.1. *[10] Let* $\{X_0, X_1, \dots\}$ *be a Markov chain with state space* E *and transition matrix* $\boldsymbol{P}$*. A row vector* $\pi(x), x \in E$ *is said to be a stationary distribution for the Markov chain, if it satisfies*

1. $\pi(x) \geq 0$ *for all* x *and* $\sum_{x\in E} \pi(x) = 1$.

2. $\pi\boldsymbol{P} = \pi$, *that is* $\sum_{x\in E} \pi(x)P(x,y) = \pi(y)$ *for all* y.

Property (1) means that π should describe a probability distribution on E, and property (2) implies that if the initial distribution P^0 equal π, then the distribution P^1 of the chain at time 1 satisfies $P^1 = P^0\mathbf{P} = \pi\mathbf{P} = \pi$, and by iterating we see that $P^t = \pi$ for every t.

Example 1.5.3. ***Steady state of the weather.***

Consider the Markov chain in Example (1.1.1) which is irreducible, aperiodic and regular. In this example, predictions for the weather on more distant days are increasingly inaccurate and tend towards a steady state vector. This vector represents the probabilities of sunny and rainy weather on all days, and is independent of the initial weather.

The stationary distribution π defined as:

$$\pi = \lim_{t\longrightarrow\infty} \mathbf{P}^t$$

but only converges if $\mathbf{P}$ is a regular transition matrix (that is, there is at least one P^t with all non-zero entries). For the weather example:

$$\begin{aligned}
\mathbf{P} &= \begin{pmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{pmatrix}. \\
\text{And since } \pi\mathbf{P} &= \pi \qquad [\pi \text{ is unchanged by} \mathbf{P}]. \\
\text{Then } \pi[\mathbf{P} - I] &= \mathbf{0} \\
&= \pi\left[\begin{pmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right] \\
&= \pi \begin{pmatrix} -0.1 & 0.1 \\ 0.5 & -0.5 \end{pmatrix}.
\end{aligned}$$

Therefore

$$\begin{pmatrix} \pi(1) & \pi(2) \end{pmatrix} \begin{pmatrix} -0.1 & 0.1 \\ 0.5 & -0.5 \end{pmatrix} = \begin{pmatrix} 0 & 0 \end{pmatrix}.$$

So

$-0.1\pi(1) + 0.5\pi(2) = 0$ and since they are a probability vector we know that

$\pi(1) + \pi(2) = 1$. Solving this pair of simultaneous equations gives the steady state distribution:

$$\begin{pmatrix} \pi(1) & \pi(2) \end{pmatrix} = \begin{pmatrix} 0.833 & 0.167 \end{pmatrix}.$$

In conclusion, in the long term, 0.83 of days are sunny.

Example 1.5.4. *Markov chains are used in the study of probabilities connected to genetic models. Genes come in pairs; and for any trait governed by a pair of genes, an individual may have genes of the gene type GG (dominant), Gg (hybrid), or gg (recessive). Each offspring inherits one gene from one parent, at random and independently.*

1. Suppose that an individual of unknown genetic makeup is matched with a hybrid. Set up a transition matrix to describe the possible states of a resulting offspring and their probabilities.

2. What will happen to the genetic makeup of the offspring after many generations of matching with a hybrid?

Solution: If the unknown is dominant (GG) and is matched with a hybrid (Gg), the offspring has a probability of $\frac{1}{2}$ of being dominant and a probability of $\frac{1}{2}$ of being hybrid. If two hybrids are matched, the offspring may be dominant, hybrid or recessive with probabilities $\frac{1}{4}, \frac{1}{2}$, and $\frac{1}{4}$, respectively. If the unknown is recessive (gg), the offspring of it and a hybrid has probability $\frac{1}{2}$ of being recessive and a probability of $\frac{1}{2}$ of being hybrid. So, a transition matrix from the unknown parent to an offspring is given by

$$\mathbf{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix},$$

where the states are $E = \{d, h, r\}$. Since the matrix $\mathbf{P}^2$ has all positive entries, and hence a stationary distribution $\pi = (\pi(1), \pi(2), \pi(3))$ exists. From the matrix equation

$$\pi = \pi\mathbf{P}$$

along with $\pi(1) + \pi(2) + \pi(3) = 1$ we obtain

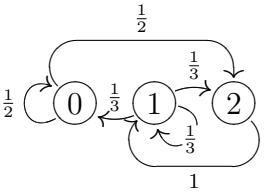
$$\pi = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4}).$$

No matter what the genetic makeup of the unknown parent happens to be, the ratio to dominant to hybrid to recessive offspring among its descendants, after many generations of mating with hybrids, should be $1 : 2 : 1$.

Definition 1.5.2. *A Markov chain is ergodic if it is both irreducible and aperiodic.*

Example 1.5.5. *The Markov chain on state space $E = \{0, 1, 2\}$ with transition matrix*

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



is an ergodic since it is irreducible and aperiodic.

Definition 1.5.3. *A Markov chain* $\{X_0, X_1, X_2, \dots\}$ *is called ergodic if the limit*

$$\pi(y) = \lim_{t \longrightarrow \infty} P^t(x, y)$$

(1) *exists for all* $y \in E$

(2) *is positive and does not depend on* x

(3) $\pi(x), x \in E$ *is a probability distribution on* E.

Ergodic Markov chain are useful for convergence theorem:

Theorem 1.5.1. *[10]**Existence of stationary distribution.***

For any irreducible and aperiodic Markov chain, there exists at least one stationary distribution.

1.6 Detailed Balance and Time Reversal

We are interested in constructing Markov chains for which the distribution we wish to sample from, given by π, is invariant (stationary), so we need to find transition matrices $\mathbf{P}$ that satisfies $\pi = \pi\mathbf{P}$. That is, if $P(x, y)$ denotes the probability of a transition from $x \in E$ to $y \in E$ under $\mathbf{P}$, we require that $\sum_x \pi(x)P(x, y) = \pi(y)$ for all y. We introduced a special class of Markov chains known as the reversible ones.

Definition 1.6.1. *A stationary Markov chain* $\{X_0, X_1, X_2, \dots\} : \Omega \longrightarrow E$ *with transition matrix* $\mathbf{P}$ *and stationary distribution* π *is called reversible if its finite-dimensional distributions do not depend on the orientation of the time axis, i.e., if* $\mathbb{P}(X_0 = x_0, X_1 = x_1, \dots, X_{t-1} = x_{t-1}, X_t = x_t) = \mathbb{P}(X_t = x_0, X_{t-1} = x_1, \dots, X_1 = x_{t-1}, X_0 = x_t)$ *for arbitrary* $t \geq 0$ *and* $x_0, \dots, x_t \in E$.

The reversibility of Markov chains is a particularly useful property for the construction of dynamic simulation algorithms, see Section (4.2).

Definition 1.6.2. *[10] A probability distribution π on the state space E is reversible for the Markov chain $\{X_0, X_1, \dots\}$ with transition matrix $\boldsymbol{P}$ if for all $x, y \in E$ we have*

$$\pi(x)P(x,y) = \pi(y)P(y,x),$$

which is also known as detailed balance equation.

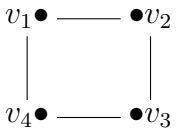
Definition 1.6.3. *A Markov chain with state space $E = \mathbb{Z}$ is called a random walk if, for some number $p \in (0,1)$ we have*

$$P(x, x+1) = p \quad \textit{and} \quad P(x, x-1) = 1 - p \quad \textit{for} \;\; x \in E.$$

We can call it a random walk because we may think of it as being a model for an individual walking on a straight line who at each point of time either takes one step to the right with probability p or one step to the left with probability $1 - p$.

Example 1.6.1. *[10] Consider a "**random walk**" on the vertices of a graph $G = (V, \mathcal{E})$(a square) consists of a vertex set V together with an edge set $\mathcal{E}$. Each edge connects two of the vertices; an edge connecting the vertices v_i and v_j is denoted (v_i, v_j). No two edges are allowed to connect the same pair of vertices. Two vertices are said to be neighbours if they share an edge.*

This random walk is a Markov chain with state space $V = \{v_1, v_2, v_3, v_4\}$ and here $\mathcal{E} = \{(v_1, v_2), (v_2, v_3), (v_3, v_4), (v_4, v_1)\}$.



If we denote the number of neighbours of a vertex v_i by d_i, then the elements of the transition matrix are given by

$$P(i,j) = \begin{cases} \frac{1}{d_i} = \frac{1}{2} & \textit{if } v_i \textit{ and } v_j \textit{ are neighbours} \\ 0 & \textit{otherwise,} \end{cases} \tag{1.6.1}$$

which is a reversible Markov chain, with reversible distribution π given by

$$\pi = \left(\frac{d_1}{d}, \ldots, \frac{d_4}{d}\right) = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right),$$

where $d = \sum_{i=1}^{4} d_i$ since it satisfies detailed balance equation. To see that, we calculate

$$\pi(i)P(i,j) = \begin{cases} \frac{d_i}{d}\frac{1}{d_i} = \frac{1}{d} = \frac{1}{8} = \frac{d_j}{d}\frac{1}{d_j} = \pi(j)P(j,i) & \text{if } v_i \text{ and } v_j \text{ are neighbours} \\ 0 = \pi(j)P(j,i) & \text{otherwise.} \end{cases}$$

It is possible for a distribution to be stationary without detailed balance holding.

Example 1.6.2. ***A nonreversible Markov chain.***

Let $\{X_1, \ldots, X_t\}$ be the Markov chain in Example (1.3.3), then the uniform distribution $\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ on the state space $\{1, 2, 3\}$ is stationary (invariant) but detailed balance does not hold, to see this: let $x = 1$, $y = 2$, we get

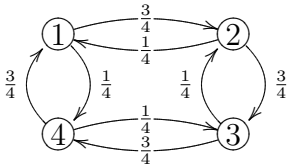
$$\pi(1)P(1,2) = \frac{1}{3}.1 = \frac{1}{3} > 0 = \frac{1}{3}.0 = \pi(2)P(2,1),$$

so that $\pi(1)P(1,2) \neq \pi(2)P(2,1)$ and reversibility fails.

Also, if we let $X_0 = 1$, then $X_t = 1$ whenever $t = 0, 3, 6, 9, ...$(a multiple of 3), thus $\mathbb{P}(X_t = 1) = 0$ or 1, so $\mathbb{P}(X_t = 1) \not\longrightarrow \pi(3)$, and there is again no convergence to π.

Example 1.6.3. *The following example is not reversible. Let $E = \{1, 2, 3, 4\}$ and*

$$\mathbf{P} = \begin{pmatrix} 0 & \frac{3}{4} & 0 & \frac{1}{4} \\ \frac{1}{4} & 0 & \frac{3}{4} & 0 \\ 0 & \frac{1}{4} & 0 & \frac{3}{4} \\ \frac{3}{4} & 0 & \frac{1}{4} & 0 \end{pmatrix}$$



.

The transition matrix $\boldsymbol{P}$ is irreducible, but not aperiodic, and stationary distribution (which is uniquely determined by Definition (1.5.1)) is given by

$\pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. *However,*

$$\pi(1)P(1,2) = \frac{1}{4}.\frac{3}{4} = \frac{3}{16} > \frac{1}{16} = \frac{1}{4}.\frac{1}{4} = \pi(2)P(2,1).$$

This cyclic random walk is not reversible as clockwise steps are much more likely than counterclockwise movements.

The next simple theorem will help us in constructing **MCMC** algorithms that (approximately) sample from a given distribution π.

Theorem 1.6.1. *[10]**Detailed Balance Test.***
If the probability distribution π is reversible for a Markov chain, then it is also a stationary distribution for the chain.

Proof. Since π is reversible, we have $\forall\ x, y$,

$$\pi(x)P(x,y) = \pi(y)P(y,x).$$

For fixed $x \in E$, we sum this equation with respect to $y \in E$ to get

$$\sum_y \pi(x)P(x,y) = \sum_y \pi(y)P(y,x).$$

But the left hand side equals $\pi(x)\sum_y P(x,y) = \pi(x)$, since rows sum to one. Thus, $\forall\ x, y$,

$$\pi(x) = \sum_y \pi(y)P(y,x),$$

and this implies $\pi = \pi\mathbf{P}$, which makes π a stationary distribution. □

If a Markov chain satisfies the detailed balance condition, then it is time reversible, i.e., one could not tell whether a sequence of samples is being acquired forward or backward in time. That is, at stationarity, the probability of transition being from state x to state y is the same as the probability of it being from state y to state x.

Chapter 2

Markov Chain Monte Carlo Algorithms

It is often difficult to sample directly from the distribution π; the reason is that they are defined as $\pi(x) = \frac{f(x)}{z}$, where $f(x)$ is an easily computable density function, and z is unknown normalizing constant that is often very difficult to compute. In such circumstance we can use algorithms (Markov Chain Monte Carlo **MCMC** simulation) that help us with sampling from complicated and high dimensional distributions, so the method of simulating a Markov chain whose stationary distribution is the distribution we want to sample from is known in general as Markov Chain Monte Carlo.

MCMC method is a class of algorithms for sampling from probability distributions based on constructing a Markov chain $\{X_0, X_1, X_2, \dots\}$ having state space E and has the desired (target) distribution as its stationary distribution π. When used in **MCMC** algorithms, Markov chains are usually constructed from a Markov transition kernel $\mathbf{P}$, a conditonal probability density on E such that $X_{t+1}|X_t \sim P(X_t, X_{t+1})$. As we saw in Section 1.2, $\forall\ x \in E, P(x, y)$ represents the probability of moving from x to y, and $P^t(x, y)$ is the probability that the Markov chain with initial state x is in state y after t iterations. The transition kernel must satisfy $P^t(x, y) \longrightarrow \pi$ as $t \longrightarrow \infty$, for all initial states x.

Now suppose we want to draw a sample from a density distribution π. The main idea behind **MCMC** is to start from an arbitrary state in E, run a

Markov chain, whose stationary distribution equals to π, for a long time, say t iterations, and note what state the chain is in after these t iterations, and then take the result"output" as an approximate sample from π. The ergodicity means that, by taking t large enough, we can ensure that the distribution of the output state is arbitrarily close to the desired distribution π. So if the Markov chain has reached equilibrium, the output is distributed according to the stationary distribution, as desired. **MCMC** techniques provide a powerful tool for statistical analysis for data.

In this chapter, we construct Markov chains with the required stationary distribution by presenting two common techniques which have ignited **MCMC**: The Metropolis-Hastings algorithms and the Gibbs sampler.

2.1 Metropolis Algorithm

The classic paper of Metropolis et al. [15], was the first to use Markov chain sampling, in the form now known as the Metropolis algorithm. This algorithm has since been applied extensively to problems in statistical physics. The Metropolis algorithm, was born in 1953, uses a symmetric candidate generating distribution (a transition matrix $Q = q(x, y)$) on E for which $q(x, y) = q(y, x)$. Suppose our goal is to draw samples from some distribution $\pi(x)$ where $\pi(x) = \frac{f(x)}{z}$, such that the normalizing constant z may not be known, and very difficult to compute.

Define a Markov chain with the following process:

(1) Start with any initial value x_0 satisfying $f(x_0) > 0$.

(2) Using current value x_t, sample a candidate point x_* from a symmetric candidate distribution $q(x, y)$ which is the probability of returning a value of y given a previous value of x.

(3) Calculate the ratio r of the density at the candidate x_* and current point x_t (this is easy since the normalization constant z cancels)

$$r = \frac{\pi(x_*)}{\pi(x_t)} = \frac{\frac{f(x_*)}{z}}{\frac{f(x_t)}{z}}.$$

(4) If the above calculation gives $(r > 1)$, accept the candidate point (set $x_{t+1} = x_*$), and return to step 2.

In detail, this can be done by generating a random number, U, from the uniform distribution on $[0, 1]$, and then sitting the next state as follows:

$$x_{t+1} = \begin{cases} x_* & \text{if } U < \alpha(x_t, x_*) \\ x_t & \text{otherwise.} \end{cases} \tag{2.1.1}$$

This generates a Markov chain $\{x_0, x_1, \ldots, x_k, \ldots\}$ as the transition probabilities from x_t to x_{t+1} depends only on x_t and not on $\{x_0, \ldots, x_{t-1}\}$.

Following a sufficient burn-in period (of k steps), the chain approaches its stationary distribution and samples from $\{x_{k+1}, \ldots, x_{k+m}\}$ are samples from $\pi(x)$.

Example 2.1.1. *Consider the scalar inverse* χ^2 *distribution,*

$$\pi(x) = C.x^{\frac{-n}{2}} \exp\left(\frac{-a}{2x}\right) \tag{2.1.2}$$

and suppose we wish to simulate draws from the distribution with $(n = 5)$ *degrees of freedom, and scaling factor* $a = 4$ *using the Metropolis algorithm. Suppose that a uniform distribution on* $[0, 100]$ *is the candidate distribution. Take* $x_0 = 1$ *as starting (initial) value, and suppose the uniform returns a candidate value of* $x_* = 39.82$*. Here*

$$\alpha = \min(r, 1) = \min\left(\frac{f(x_*)}{f(x_t)}, 1\right) = \min\left(\frac{(39.82)^{-2.5}\exp(\frac{-2}{39.82})}{(1)^{-2.5}\exp(\frac{-2}{1})}, 1\right) = 0.0007$$

$\alpha = 0.0007 < 1$*, implies* x_* *is accepted with probability* 0.0007*.*

Thus, we randomly draw U *from a uniform* $[0, 1]$ *and accept* x_* *if* $U \leq \alpha$*. In this case, the candidate is rejected, and we draw another candidate value from the candidate distribution (which turns out to be* 71.36*), then*

$$\alpha = \min\left(\frac{(71.36)^{-2.5}\exp(\frac{-2}{71.36})}{(1)^{-2.5}\exp(\frac{-2}{1})}, 1\right) = 0.0002 < 1,$$

implies 71.36 *is accepted with probability* 0.0002*, and continue as above for (say)* 500 *values of* x*.*

2.2 Metropolis-Hastings Algorithm

If we want to sample from some distribution π that has a sample space of high dimension, then we want a Markov chain whose unique stationary distribution is π, and run this chain long enough and then take the result as an approximate sample from π. A very general way to construct such a Markov chain is the Metropolis-Hastings algorithm. Since the symmetry requirement of the Metropolis proposal distribution can be hard to satisfy, Hastings in (1970)[11] extended the Metropolis algorithm to a more general candidate transition which converges to $\pi(x)$ by using an arbitrary transition probability function $q(x, y) = \mathbb{P}(x \longrightarrow y)$, (i.e., a Markov transition kernel).

The algorithm proposes a new point (state) on the Markov chain which is either accepted or rejected.

- If the state is accepted, the Markov chain moves to the new state.
- If the state is rejected, the Markov chain remains in the same state.

By choosing the acceptance probability correctly, we create a Markov chain which has π as a stationary distribution. We begin with a state space E and a probability distribution π on E. Then we choose a candidate distribution $q(x, y) : x, y \in E$ with $q(x, y) \geq 0$ and $\sum_y q(x, y) = 1$ for each $x \in E$.

The algorithm works as follows:

(1) Sample a candidate point " state" x_* from the candidate distribution $q(x, y)$ which is not necessarily symmetric.

(2) Compute the ratio $r = \dfrac{\pi(x_*)q(x_*, x_t)}{\pi(x_t)q(x_t, x_*)}$.

(3) With probability $\alpha(x_t, x_*) = \min(r, 1)$, transition to x_* " accept the candidate" . Otherwise, stay in the same state x_t " reject the candidate", where

$$\alpha(x_t, x_*) = \begin{cases} \min(r, 1) & \text{if } \pi(x_t)q(x_t, x_*) > 0 \\ 1 & \text{if } \pi(x_t)q(x_t, x_*) = 0. \end{cases} \tag{2.2.1}$$

(4) Discard initial " burn in" values.

(5) Remaining $x's$ are $\sim$ independent and identically distribution (i.i.d.) $\pi(x)$.

The initial value is arbitrarily selected, which means Metropolis-Hastings typically has " burn in" period at the start of a simulation.

Now for reversible chains it is impossible to distinguish between forward and backward running of the chain. So, to construct the transition probabilities $P(x, y)$, we set

$$P(x, y) = q(x, y)\alpha(x, y), \tag{2.2.2}$$

where $q(x, y)$'s are the transition probabilities for another Markov chain fulfilling

$$q(x, y) > 0 \Longrightarrow q(y, x) > 0 \quad \forall \quad x, y \in E. \tag{2.2.3}$$

Finally, to see why the Metropolis-Hastings algorithm work (generates a Markov chain whose equilibrium density is that candidate density $f(x)$), it is sufficient to show that the implied transition kernel $q(x, y)$ of any Metropolis-Hastings algorithm satisfies the detailed balance equation, i.e., we want to show that the Markov chain resulting from the Metropolis-Hastings algorithm is reversible with respect to π.

Theorem 2.2.1. *[4] The Metropolis-Hastings algorithm produces a Markov chain* $\{X_0, X_1, \dots\}$ *which is reversible with respect to* π.

Proof. We must show that $\pi(x)P(x, y) = \pi(y)P(y, x)$.

Obviously this holds if $x = y$, so we will consider $x \neq y$, then

$$\begin{aligned} \pi(x)P(x, y) &= \pi(x)q(x, y)\alpha(x, y) \\ &= \pi(x)q(x, y)\min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\} \\ &= \min\{\pi(x)q(x, y), \pi(y)q(y, x)\}. \end{aligned} \tag{2.2.4}$$

Similarly, we find that

$$\pi(y)P(y, x) = \min\{\pi(x)q(x, y), \pi(y)q(y, x)\}.$$

This implies that

$$\pi(x)P(x,y) = \pi(y)P(y,x).$$

Therefore, the chain is reversible. □

It follows that the Metropolis-Hastings algorithm converges to the target distribution π which is the stationary distribution by detailed balance.

Example 2.2.1. *If x has a χ^2 distribution with n degrees of freedom, shorthand notation for this $x \sim \chi^2_n$, then the pdf of x is given by*

$$f(x) = \begin{cases} \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} x^{(\frac{n}{2})-1} e^{\frac{-x}{2}} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2.2.5}$$

Now suppose we wish to use a χ^2 distribution as our candidate density, by simply drawing from a χ^2 distribution independent of the current position.

Thus, $q(x,y) = f(y) = \dfrac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} y^{(\frac{n}{2})-1} e^{\frac{-y}{2}}$ which implies that $q(x,y)$ is not symmetric, since $q(x,y) = f(y) \neq f(x) = q(y,x)$. Hence, we must use the Metropolis-Hastings sampling, with acceptance probability

$$\alpha(x,y) = \min\left(\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right) = \min\left(\frac{\pi(y)x^{(\frac{n}{2})-1}e^{\frac{-x}{2}}}{\pi(x)y^{(\frac{n}{2})-1}e^{\frac{-y}{2}}}, 1\right),$$

move to y, otherwise stay at x.

Using the target distribution as in (2.1.2),

$$\pi(x) = C.x^{\frac{-5}{2}} \exp\left(\frac{-2}{x}\right),$$

the rejection probability becomes

$$\alpha(x,y) = \begin{cases} \min\left[\frac{y^{\frac{-5}{2}} \exp(\frac{-2}{y}) x^{(\frac{n}{2})-1} e^{\frac{-x}{2}}}{x^{\frac{-5}{2}} \exp(\frac{-2}{x}) y^{(\frac{n}{2})-1} e^{\frac{-y}{2}}}, 1\right] & \text{if } \pi(x)q(x,y) > 0 \\ 1 & \text{if } \pi(x)q(x,y) = 0. \end{cases} \tag{2.2.6}$$

2.3 The Gibbs Sampler

Gibbs sampling is a statistical method that allows us to generate random variables from highly complicated distributions without calculating their density functions. Gibbs sampling was introduced by [15]. Later, the approach was modified and improved by [11]. A good introduction into Gibbs sampling can be found in [2]. One of the first applications of Gibbs sampling is shown in [8], where it is used for the digital restoration of images.

The Gibbs sampler introduced by Geman and Geman [8], also known as the **Heat Bath** Algorithm for simulating a Markov chain $\{X_0, X_1, \dots\}$ which is converging to the target distribution $\pi(x)$, by successively sampling from the full conditional component distribution $\pi(x_i \mid x_{-i}), i = 1, \dots, n$, where x_{-i} denotes the components of x other than x_i. The idea behind the Gibbs sampler is that one only considers univariate conditional distributions-the distribution when all of the random variables but one are assigned fixed values. Such conditional distributions are far easier to simulate than complex joint distributions and usually have simple forms often being normals, or inverse χ^2. The Gibbs sampler is a method for generating samples from the joint distribution of two or more variables and it is a special case of the Metropolis-Hastings algorithm where the proposals q (the candidate distributions) are the full conditionals and the acceptance probability is always 1.

To show this let the candidate distribution be

$$q(x, y) = \begin{cases} \pi(y_i \mid x_{-i}) & \text{if } y_{-i} = x_{-i} \\ 0 & \text{otherwise.} \end{cases} \tag{2.3.1}$$

The importance ratio is

$$\begin{aligned} r &= \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} \\ &= \frac{\pi(y)\pi(x_i \mid y_{-i})}{\pi(x)\pi(y_i \mid x_{-i})} && \text{Definition of the candidate distribution} \\ &= \frac{\pi(y)\pi(x_i \mid x_{-i})}{\pi(x)\pi(y_i \mid y_{-i})} \\ &= \frac{\pi(y)\pi(x_i, x_{-i})\pi(y_{-i})}{\pi(x)\pi(y_i, y_{-i})\pi(x_{-i})} && \text{Definition of conditional probability} \\ &= \frac{\pi(y_{-i})}{\pi(x_{-i})} = 1. && \text{We did not change other variables.} \end{aligned} \tag{2.3.2}$$

So the acceptance probability is $\alpha(x,y) = \min\left(1, \dfrac{\pi(y)\pi(x_i|x_{-i})}{\pi(x)\pi(y_i|y_{-i})}\right) = 1.$

This means that the Gibbs sampler is the best, since the candidate is always accepted.

Let $\pi(X)$ be the joint distribution of $X = (X_1, \ldots, X_n)$, and let $\pi(X_i \mid X_{-i}) = \pi(X_i \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ be the pdf for X_i given all other components. These distributions are called the full conditionals. To use the Gibbs sampler, the marginal distribution of all the variables must be known. Now for simplicity consider the case when $n = 2$.

The Gibbs sampler generates a Markov chain $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_k, Y_k)\}$ converging to $\pi(X, Y)$, by successively sampling

$$\begin{array}{lll} X_1 & \text{from} & \pi(X \mid Y_0) \\ Y_1 & \text{from} & \pi(Y \mid X_1) \\ X_2 & \text{from} & \pi(X \mid Y_1) \\ & \vdots & \\ X_k & \text{from} & \pi(X \mid Y_{k-1}) \\ Y_k & \text{from} & \pi(Y \mid X_k). \end{array}$$

To get started, prespecify an initial value for Y_0.

Example 2.3.1. *[2] Suppose the joint distribution of $x = 0, 1, ..., n$ which is discrete, and $0 \leq y \leq 1$ which is continuous is given by*

$$f(x,y) = \frac{n!}{(n-x)!x!}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{x+\alpha-1}(1-y)^{n-x+\beta-1},$$

where the joint density is complex.

The Gibbs sampler, having f as its stationary distribution proceeds by successively sampling from the conditional distributions $X|y \sim b(n,y)$, and the conditional density $y|x \sim Beta(x+\alpha, n-x+\beta)$. To see this, first recall that a binomial random variable z has a density function

$$f(z|p,n) = \frac{n!}{(n-z)!z!}p^z(1-p)^{n-z} \quad \text{for} \quad z = 0, 1, 2, \ldots, n,$$

where $0 < p < 1$ is the success parameter and n is the number of trails, and we denote $z \sim b(n,p)$.
Likewise recall the density for $z \sim Beta(a,b)$, a beta distribution with shape parameters a and b is given by

$$f(z|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}z^{a-1}(1-z)^{b-1} \quad \text{for} \quad 0 \leq z \leq 1.$$

The marginal distribution of x is given by

$$\begin{aligned} f(x) = \mathbb{P}(X = x) = \int_0^1 f(x,y)dy &= \int_0^1 \frac{n!}{(n-x)!x!}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{x+\alpha-1}(1-y)^{n-x+\beta-1}dy \\ &= \frac{n!}{(n-x)!x!}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 y^{x+\alpha-1}(1-y)^{n-x+\beta-1}dy \\ &= \frac{n!}{(n-x)!x!}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha+\beta+n)}. \end{aligned}$$

Which is called the beta-binomial (n, α, β).

And the marginal distribution of y is given by

$$\begin{aligned} f(y) = \mathbb{P}(Y = y) = \sum_{x=0}^{n} f(x,y) &= \sum_{x=0}^{n} \frac{n!}{(n-x)!x!} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{x+\alpha-1}(1-y)^{n-x+\beta-1} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \sum_{x=0}^{n} \frac{n!}{(n-x)!x!} y^{x}(1-y)^{n-x} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{x}(1-y)^{n-x}.1 \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{x}(1-y)^{n-x}. \end{aligned}$$

Hence, $y \sim Beta(\alpha, \beta)$.

With these probability distributions, the conditional distribution of x (y is a fixed constant) is

$$\begin{aligned} f(X|y) &= \frac{f(x,y)}{f(y)} \\ &= \frac{\frac{n!}{(n-x)!x!} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}}{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}} \\ &= \frac{n!}{(n-x)!x!} y^{x}(1-y)^{n-x}. \end{aligned}$$

Thus, $X|y \sim b(n, y)$. While

$$\begin{aligned} f(y|x) &= \frac{f(x,y)}{f(x)} \\ &= \frac{\frac{n!}{(n-x)!x!} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}}{\frac{n!}{(n-x)!x!} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha+\beta+n)}} \\ &= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}. \end{aligned}$$

Therefore, $y|x \sim Beta(x+\alpha, n-x+\beta)$.

To illustrate the Gibbs sampler for the above, suppose $n = 10$ and $\alpha = 1, \beta = 2$.

The algorithm of the sampler is as follows:

(1) Start with some initial value $y_0 = Beta(\alpha, \beta) = Beta(1, 2) = \frac{\Gamma(1)\Gamma(2)}{\Gamma(3)} = \frac{0!1!}{2!} = \frac{1}{2}$ for y and obtain X_0 by generating a random variable from the conditional distribution $X|y = y_0 = \frac{1}{2} \sim b(n, y_0) = b(10, \frac{1}{2})$, giving $X_0 = 5$ in our simulation.

(2) Use $x_0 = 5$ to generate a new value of y_1, drawing from the conditional distribution $y|x = 5 \sim Beta(x_0+\alpha, n-x_0+\beta) = Beta(5+1, 10-5+2) = Beta(6, 7)$ based on the value $x_0 = 5$, giving $y_1 = 0.33$.

(3) Repeat step (1) to obtain $X_1 = 3$ which is a realisation of $b(n, y_1) = b(10, 0.33)$ random variable.

(4) Repeat step (2) to obtain $y_2 = 0.56$ from $Beta(x_1 + \alpha, n - x_1 + \beta) = Beta(3 + 1, 10 - 3 + 2) = Beta(4, 9)$.

(5) Similarly $X_2 = 7$ is obtained from $b(n, y_2) = b(10, 0.56)$ random variable.

Thus, after repeating this process three times "iterations", we generate a Gibbs sequence "a Markov chain" of length $k = 3$, where a subset of points $(5, 0.5), (3, 0.33), (7, 0.56)$ are taken as our simulated draws from the full joint distribution.

Remark 2.3.1. *Note that, the initial values in the chain (the Gibbs sequence) are dependent on the y_0 value chosen to start the chain. This dependence decays as the sequence length increases and so we typically start after a sufficient number of "burn-in" iterations have occurred to remove any effects of the starting conditions (initial sampling values).*

The Gibbs sequence converges to a stationary (equilibrium) distribution that is independent of the starting values, and by construction this stationary distribution is the target distribution we are trying to simulate [23].

Remark 2.3.2. *As we have seen, we use the Gibbs sampler and the Metropolis-Hastings algorithms to simulate a Markov chain $X_1, \ldots, X_k$ which converges in distribution to $\pi(X)$, [i.e., as k increases, the distribution of X_k gets closer and closer to $\pi(X)$]. Simulation of a Markov chain requires a starting value X_0 such that if the chain is converging to $\pi(X)$, then the dependence between say X_j and X_0 diminishes as j increases. After a suitable "burn in" period of N iterations, $X_N, \ldots, X_k$ behaves like a dependent sample from $\pi(X)$.*

Chapter 3

Coupling

A fundamental problem of **MCMC** algorithms is the determination of the number of iterations required, so that the result will be approximately a sample from the distribution of interest.
Recently, a new variant of **MCMC** method, which is called perfect simulation algorithms, have been developed. These are algorithms which automatically ensure that the Markov chain is only sampled after equilibrium has been reached. The basis of perfect simulation are couplings. The coupling method, deals with comparison of probability measures on a measurable space. It is an important tool in probability theory and its applications, is primarily used in estimates of total variation distances. The method also works well in establishing inequalities and has been highly successful in the study of Markov chain.

In this chapter, we will discus how to couple paths of a Markov chain which are started in different initial states.

3.1 Convergence in Variation and Coupling

Let μ and ν be two probability distributions on a finite state space E. That is, $\mu(x)=$ probability that x occurs. A distribution ω on $E \times E$ is a coupling if:

For all $x \in E$, $\sum_{y \in E} \omega(x, y) = \mu(x)$, and

For all $y \in E$, $\sum_{x \in E} \omega(x, y) = \nu(y)$.

In other words, ω is a joint distribution whose marginal distributions are the appropriate distributions.

Definition 3.1.1. *[14] A coupling of the probability measures P and P' on a measurable space $(E, \mathcal{E})$ is a probability measure $\hat{P}$ on $(E^2, \mathcal{E}^2)$ such that*

$$P = \hat{P}\pi^{-1} \quad \textit{and} \quad P' = \hat{P}\pi'^{-1},$$

where $\pi(x, x') = x, \pi'(x, x') = x'$ for $(x, x') \in E^2$.

Thus P and P' are the marginal distributions of $\hat{P}$. The marginal distributions in this chapter are the distributions of Markov chains with different initial states. Coupling provides a method to bound the variation distance between a pair of distributions.

Definition 3.1.2. *The distance between two probability distributions μ and ν is variation distance, defined as*

$$\|\mu - \nu\| = \sup_A |\mu(A) - \nu(A)|.$$

Definition 3.1.3. *[10]Let μ and ν be two probability distributions on the finite state space E. The total variation distance between μ and ν is defined as*

$$d_{TV}(\mu, \nu) = \frac{1}{2} \sum_{x \in E} |\ \mu(x) - \nu(x)| = \sup_{A \subseteq E} |\ \mu(A) - \nu(A)|. \tag{3.1.1}$$

The bias or error in the distribution of the Markov chain after running the chain for t steps can be measured by the total variation distance between the distribution of the state at that time P^t and the stationary distribution π. This means for Markov chains with initial state x, transition matrix $\mathbf{P}$ and stationary distribution π, we are interested in how close the distribution of states at time t is to π, i.e.,

$$d_{TV}(P^t, \pi) = \|P^t - \pi\| = \frac{1}{2} \sum_{x \in E} |\ P^t(x) - \pi(x)| = \sup_{A \subseteq E} |P^t(A) - \pi(A)|,$$

as we will see it in Theorem (3.2.1).

Theorem 3.1.1. *[14]**Coupling Inequality.***
Suppose we have two random variables X and Y, defined jointly on state space E. Let $\mathcal{L}(X)$ and $\mathcal{L}(Y)$ be their respective probability distributions, then

$$\|\mathcal{L}(X) - \mathcal{L}(Y)\| \leq \mathbb{P}(X \neq Y). \tag{3.1.2}$$

Proof.

$$\begin{aligned}
\|\mathcal{L}(X) - \mathcal{L}(Y)\| &= \sup_A |\ \mathbb{P}(X \in A) - \mathbb{P}(Y \in A)| \\
&= \sup_A |\ \mathbb{P}(X \in A, X = Y) + \mathbb{P}(X \in A, X \neq Y) \\
&\quad - \mathbb{P}(Y \in A, Y = X) - \mathbb{P}(Y \in A, Y \neq X)| \\
&= \sup_A |\ \mathbb{P}(X \in A, X \neq Y) - \mathbb{P}(Y \in A, Y \neq X)| \\
&\leq \mathbb{P}(X \neq Y).
\end{aligned}$$

□

Since we have used that $\mathbb{P}(X \in A, X = Y) = \mathbb{P}(Y \in A, Y = X)$ and the difference between two nonnegative quantities each less than or equal $\mathbb{P}(X \neq Y)$ must itself be less than or equal $\mathbb{P}(X \neq Y)$.

3.2 Coalescence

The first step in obtaining a perfect sample is to make X_t independent of the starting value by coupled parallel chains. Suppose there are m states in E, and we start a Markov chain in each state at time $t = 0$. These are *parallel chains*, which can be coupled through the following [10]:

1. Construct valid initiation function $\psi : [0, 1] \longrightarrow E$, to generate the starting value X_0, where $X_0 = \psi(U_0)$, and $\mathbb{P}(X_0 = x) = \mathbb{P}(\psi(U_0) =$

$x) = P^0(x)$.

$$\psi(u) = \begin{cases} x_1 & \text{for } u \in [0, P^0(x_1)) \\ x_2 & \text{for } u \in [P^0(x_1), P^0(x_1) + P^0(x_2)) \\ \vdots & \\ x_i & \text{for } u \in \left[\sum_{j=1}^{i-1} P^0(x_j), \sum_{j=1}^{i} P^0(x_j)\right) \\ \vdots & \\ x_m & \text{for } u \in \left[\sum_{j=1}^{m-1} P^0(x_j), 1\right]. \end{cases} \tag{3.2.1}$$

2. Construct updating function $f : E \times [0,1] \longrightarrow E$, which takes an input state $x \in E$ and a number between 0 and 1, to produce another state $x' \in E$ as output, such that

$$f(x_i, u) = \begin{cases} x_1 & \text{for } u \in [0, \mathbb{P}(i,1)) \\ x_2 & \text{for } u \in [\mathbb{P}(i,1), \mathbb{P}(i,1) + \mathbb{P}(i,2)) \\ \vdots & \\ x_j & \text{for } u \in \left[\sum_{l=1}^{j-1} \mathbb{P}(i,l), \sum_{l=1}^{j-1} \mathbb{P}(i,l)\right) \\ \vdots & \\ x_m & \text{for } u \in \left[\sum_{l=1}^{m-1} \mathbb{P}(i,l), 1\right]. \end{cases} \tag{3.2.2}$$

Then set

$$X_0 = \psi(U_0)$$
$$X_1 = f(X_0, U_1)$$
$$X_2 = f(X_1, U_2),$$

and so on.

Definition 3.2.1. *The transition of a Markov chain* $\{X_0, X_1, \dots\}$ *with transition matrix* $\boldsymbol{P}$ *can be described by a deterministic update function* f *by* $X_{t+1} = f(X_t, U_{t+1})$ *if*

$$\mathbb{P}(f(x, U) = y) = \mathbb{P}(X_{t+1} = y \mid X_t = x) = P(x, y) \tag{3.2.3}$$

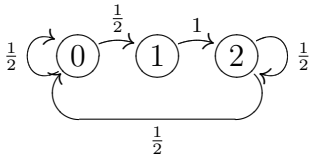
for all $x, y \in E$ *and for a uniform random number* U*.*

Definition 3.2.2. *A random map* $f : E \to E$ *is consistent with a Markov chain with transition matrix* $\boldsymbol{P}$ *if* $\mathbb{P}(f(x, U) = y) = P(x, y)$*, for all* x, y*.*

The transition rule is a random map which specifies for each state of E the transition of the chain from time t to time $t + 1$.

Example 3.2.1. *Consider the Markov chain* $\{X_0, X_1, \dots\}$ *with state space* $E = \{0, 1, 2\}$ *and transition matrix*

$$\mathbf{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}$$



Suppose we start the Markov chain on $x_1 = 0$*, so that* $P^0 = (1, 0, 0)$*. We can simulate this Markov chain using initiation function*

$$\psi(u) = x_1 = 0, \ \textit{for all} \ \ u,$$

and update function given by

$$f(x_1, u) = f(0, u) = \begin{cases} x_1 = 0 & \textit{if } u \in [0, \frac{1}{2}) \\ x_2 = 1 & \textit{if } u \in [\frac{1}{2}, 1], \end{cases}$$

$$f(x_2, u) = f(1, u) = x_3 = 2, \quad \textit{if} \ \ u \in [0, 1], \ \textit{and}$$

$$f(x_3, u) = f(2, u) = \begin{cases} x_1 = 0 & \textit{if} \quad u \in [0, \frac{1}{2}) \\ x_3 = 2 & \textit{if} \quad u \in [\frac{1}{2}, 1]. \end{cases} \tag{3.2.4}$$

A simulation is straightforward: Given one initial start value, then draw a random number from a uniform distribution, and according to the transition probabilities for the current state, decide the next state of the chain and then repeat the same process, as we see in Section (3.3).

Definition 3.2.3. *Two Markov chains* $\{X_0, X_1, \dots\}$, $\{X_0', X_1', \dots\}$ *taking their values in the same state space* E *are said to couple if there exists an almost surely finite random time* T *such that*

$$X_t = X_t^{'} \quad for \quad t \geq T.$$

The random time T *is called coupling time of the two chains, and obtain from Theorem (3.1.1) that*

$$\| \mathbb{P}(X_t = x) - \mathbb{P}(X_t^{'} = x)\| \leq \mathbb{P}(T > t),$$

since $\{X_t \neq X_t'\} \subset \{T > t\}$.

Definition 3.2.4. *[13]**Coalescence***
A family of random processes $X, Y, Z, \dots$ *are said to coalesce if there is some random time* T *(the coalescence time) at which they are all equal:* $X(T) = Y(T) = Z(T) = \dots$ *Sometimes called grand coupling, since two processes* X, Y *are said to couple if* $X(T) = Y(T)$ *for some random time* T.

Definition 3.2.5. *We say that two chains have coalesced if they at one step reach the same state. From that step and onward the two chains will follow the same sample path, since they used the same updating function.*

Definition 3.2.6. ***Convergence in Variation***
Let $\{X_0, X_1, \dots\}$ *be a Markov chain with state space* E. *If for some probability distribution* π *on* E, *the distribution* μ *of* X_t *converges in variation to* π, *that is*

$$\lim_{t\to\infty} |\mathbb{P}(X_t = x) - \pi(x)| = \lim_{t\to\infty} d_{TV}(\mu, \pi) = 0,$$

then $\{X_0, X_1, \dots\}$ *is said to converge in variation to* π.

We are now ready to state the main result about convergence to stationarity.

Theorem 3.2.1. *[10]* ***Markov chain convergence theorem.***
Consider an irreducible aperiodic Markov chain $\{X_0, X_1, \dots\}$. *If we denote the chains distribution after the* t^{th} *transition by* P^t, *we have for any initial distribution* P^0 *and a stationary distribution* π

$$P^t \xrightarrow{TV} \pi \tag{3.2.5}$$

In words: If we run the Markov chain for a long time, its distribution will be very close to the stationary distribution π. This is often referred to as the Markov chain approaching equilibrium as $t \longrightarrow \infty$.

Proof. Suppose we have an irreducible aperiodic Markov chain $\{X_0, X_1, \dots\}$. We use the coupling idea to prove (3.2.5) by introducing a parallel Markov chain $\{X'_0, X'_1, \dots\}$, independent of $\{X_0, X_1, \dots\}$, having the stationary distribution π. That is, the two chains obtained by the simulation method.

$$\begin{aligned} X_0 &= \psi_{P^0}(U_0) & X'_0 &= \psi_\pi(U'_0) \\ X_1 &= f(X_0, U_1) & X'_1 &= f(X'_0, U'_1) \\ X_2 &= f(X_1, U_2) & X'_2 &= f(X'_1, U'_1) \\ &\vdots \end{aligned}$$

where ψ_{P^0}, ψ_π are a valid initiation function for P^0 and the distribution π, respectively, f is a valid update function for $\mathbf{P}$, and $(U_0, U_1, \dots)$ independent of $(U'_0, U'_1, \dots)$ are an i.i.d. sequences of uniform $[0, 1]$ random variables. Since π is a stationary distribution, we have that X'_t has distribution π for any t. We want to show that, if the hitting time "first meeting time"

$$T = \min\{k; X_k = X'_k\} \tag{3.2.6}$$

is finite with probability 1, then the two chains will meet, with the convention that $T = \infty$ if the chains never meet.

Since the Markov chain $\{X_0, X_1, \dots\}$ is irreducible and aperiodic, we can find, using Corollary (1.3.1), an $M < \infty$ such that

$$P^M(x, y) = \mathbb{P}(X_M = y | X_0 = x) > 0 \qquad for \quad all \quad x, y \in E.$$

Set

$$\alpha = \min\{P^M(x, y) > 0 \quad for\ all \quad x, y \in E\}.$$

Then we get,

$$\begin{aligned}
\mathbb{P}(T \leq M) &\geq \mathbb{P}(X_M = X'_M) \\
&\geq \mathbb{P}(X_M = x_1, X'_M = x_1) \\
&= \mathbb{P}(X_M = x_1)\mathbb{P}(X'_M = x_1) \\
&= \left(\sum_{i=1}^{m} \mathbb{P}(X_0 = x_i, X_M = x_1)\right)\left(\sum_{i=1}^{m} \mathbb{P}(X'_0 = x_i, X'_M = x_1)\right) \\
&= \left(\sum_{i=1}^{m} \mathbb{P}(X_0 = x_i)\mathbb{P}(X_M = x_1 \mid X_0 = x_i)\right)\left(\sum_{i=1}^{m} \mathbb{P}(X'_0 = x_i)\mathbb{P}(X'_M = x_1 | X'_0 = x_i)\right) \\
&\geq \left(\alpha \sum_{i=1}^{m} \mathbb{P}(X_0 = x_i)\right)\left(\alpha \sum_{i=1}^{m} \mathbb{P}(X'_0 = x_i)\right) = \alpha^2.
\end{aligned}$$

So that

$$\mathbb{P}(T > M) \leq 1 - \alpha^2.$$

Similarly, given every thing that has happened up to time M, we have conditional probability at least α^2 of having $X_{2M} = X'_{2M} = x_1$. This implies

$$\begin{aligned}
\mathbb{P}(T > 2M) &= \mathbb{P}(T > M)\mathbb{P}(T > 2M \mid T > M) \\
&\leq (1 - \alpha^2)\mathbb{P}(T > 2M \mid T > M) \\
&\leq (1 - \alpha^2)\mathbb{P}(X_{2M} \neq X'_{2M} \mid T > M) \\
&= (1 - \alpha^2)(1 - \mathbb{P}(X_{2M} = X'_{2M} \mid T > M)) \\
&\leq (1 - \alpha^2)^2.
\end{aligned}$$

By iterating this argument, we get for any l that

$$\mathbb{P}(T > lM) \leq (1 - \alpha^2)^l,$$

which tend to 0 as $l \to \infty$. Hence,

$$\lim_{t \to \infty} \mathbb{P}(T > t) = 0. \tag{3.2.7}$$

Therefore, the two chains will meet with probability 1.

Now, Define $\{X''_0, X''_1, \dots\}$ to make a coupling of X and X' at T which is called the coupling time, by

$$X''_t = \begin{cases} X_t & \text{if} \quad t < T \\ X'_t & \text{if} \quad t \geq T, \end{cases} \tag{3.2.8}$$

where T as in (3.2.6). This implies that for all $x \in E$

$$\begin{aligned}
|\ \mathbb{P}(X_t = x) - \pi(x)| &= |\ \mathbb{P}(X_t'' = x) - \mathbb{P}(X_t^{'} = x)| \\
&= |\ \mathbb{P}(X_t'' = x, T \leq t) + \mathbb{P}(X_t'' = x, T > t) \\
&\quad - \mathbb{P}(X_t^{'} = x, T \leq t) - \mathbb{P}(X_t^{'} = x, T > t)| \\
&= |\ \mathbb{P}(X_t = x, T > t) - \mathbb{P}(X_t' = x, T > t)| \quad \text{due to (3.2.8)} \\
&\leq |\ \mathbb{P}(X_t = x, T > t) - \mathbb{P}(X_t = x, X_t' = x, T > t)| \\
&= \mathbb{P}(X_t = x, X_t' \neq x) \\
&\leq \mathbb{P}(X_t \neq X_t') \\
&\leq \mathbb{P}(T > t),
\end{aligned}$$

which tend to 0 as $t \to \infty$. Hence,

$$\lim_{t\to\infty} |P^t(x) - \pi(x)| = \lim_{t\to\infty} |\ \mathbb{P}(X_t = x) - \pi(x)| = 0. \tag{3.2.9}$$

This implies that

$$\lim_{t\to\infty} d_{TV}(P^t, \pi) = \lim_{t\to\infty} \left(\frac{1}{2} \sum_{i=1}^{m} |P^t(x) - \pi(x)| \right) = 0 \qquad \textit{due to (3.2.9).}$$

Hence, $P^t \xrightarrow{TV} \pi$.

□

Theorem 3.2.2. *[10]* ***Uniqueness of the stationary distribution.*** *Any irreducible and aperiodic Markov chain has exactly one stationary distribution.*

Proof. By Theorem (1.5.1), an irreducible and aperiodic Markov chain has at least one stationary distribution. So assume the chain $\{X_0, X_1, \dots\}$ has more than one stationary distribution, say π and π'. Then the chain distribution after t transitions is $P^t = \pi'$. On the other hand we have got

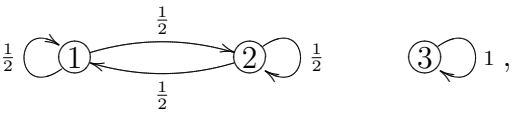
$$P^t \xrightarrow{TV} \pi.$$

Since $P^t = \pi'$, meaning that $\lim d_{TV}(\pi, \pi') = 0$ and this does not depend on t at all, this implies $d_{TV}(\pi, \pi') = 0$. We conclude $\pi = \pi'$. □

Remark 3.2.1. *Even if a Markov chain has stationary distribution π, it may still fail to converge to stationarity.*

Example 3.2.2. *Take a Markov chain on a state space $E = \{1, 2, 3\}$, and transition matrix*

$$\mathbf{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}$$



with $\pi(1) = \pi(2) = \pi(3) = \frac{1}{3}$. Then π is a stationary distribution by Definition (1.5.1). However, let $X_0 = 1$, then $X_t \in \{1, 2\}$ for all t. This implies that $\mathbb{P}(X_t = 3) = 0$ for all t, so $\mathbb{P}(X_t = 3) \not\longrightarrow \pi(3)$; and the distribution of X_t does not converge to π.

Remark 3.2.2. *The Markov chain in Example (3.2.2) is reducible, since the chain can never get from state 1 or 2 to state 3, in any number of steps, so the stationary distribution is not unique, and the distribution of X_t converges to a different stationary distribution defined by $\pi = (\frac{1}{2}, \frac{1}{2}, 0)$.*

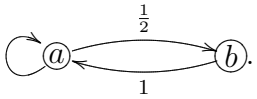
3.3 Forward Coupling

Coupling is the soul of probability, and is often used to determine convergence properties of Markov chains. Let X_t and Y_t be two copies of a Markov chain on a discrete finite state space E. The idea of coupling to the future or *forward coupling* is to start a path of the chain from each possible state $X_0 = x_0$ and $Y_0 = y_0$. The two chains then evolve via an update function defined in (3.2.1) and then wait until they all produce the same results. These results don't depend on the starting value of the chain, so the coupling method is to give the paths started from different initial states a chance to agree. We say two chains are coupled if they use the same sequence of random numbers for the transitions, in other wards if their paths coincide after a random time T; the coupling time, then the chains will merge or coalesce, that is $X_t = Y_t,\ \forall t \geq T$, so the effect of the initial state has worn off and equilibrium has been reached. However, when the chains coalesce, they need

not have reached the stationary distribution, i.e., the state distribution of the Markov chain at the coupling time is in general not equal to the stationary distribution π, and this will not yield a correct algorithm, as the following example show.

Example 3.3.1. *[12]Let $\{X_0, X_1, \dots\}$ be a Markov chain on state space $E = \{a, b\}$, with transition matrix*

$$\mathbf{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{pmatrix}$$



By solving the equation system of Definition(1.5.1), the stationary distribution is

$$\pi = (\pi(1), \pi(2)) = (\frac{2}{3}, \frac{1}{3}). \tag{3.3.1}$$

The transition rule of the chain is given by

$$f(x_1, U_{t+1}) = f(a, U_{t+1}) = \begin{cases} x_1 = a & \text{if } U_{t+1} \leq \frac{1}{2} \\ x_2 = b & \text{if } U_{t+1} > \frac{1}{2} \end{cases} \tag{3.3.2}$$

and $f(x_2, U_{t+1}) = f(b, U_{t+1}) = x_1 = a$, for uniform random numbers $U_{t+1} \in [0, 1]$.

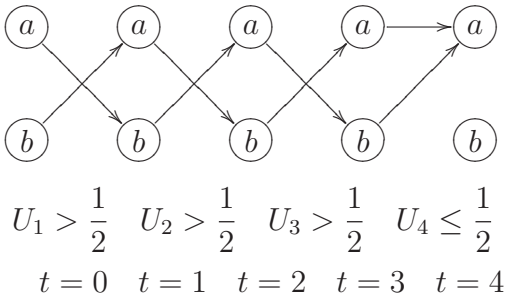


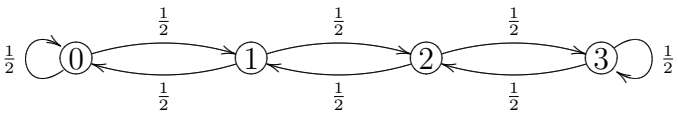
Figure 3.1: Forward simulation with t=4 of Example (3.3.1).

Starting one chain in each of the two states, state a and state b at time 0 and output the state of the chains once they had coalesced, then the chains always coalesce in state a for the first time at some random time T. If

$X_{T-1}^{(1)} \neq X_{T-1}^{(2)}$, we necessarily obtain $X_{T-1}^{(1)} = b$ or $X_{T-1}^{(2)} = b$ and therefore $X_T^{(1)} = X_T^{(2)} = a$, because the chain moves deterministically to state a from state b.
Figure 3.1 shows the forward simulation with stopping time $T = 4$ in state a ($X_T = a$), i.e., in distribution $(1, 0)$. This does not agree with the stationary distribution in (3.3.1), and hence $X_T \not\sim \pi$, so the distribution at the time of coalescence is not correct, that is the algorithm is incorrect.

Example 3.3.2. *[22] Consider a Markov chain X with transition matrix*

$$\mathbf{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$



and state space $E = \{0, 1, 2, 3\}$. This Markov chain has the uniform distribution as a stationary distribution.
We would like to start a path of the chain from different initial states, so to produce paths of the chain we can use a fair coin such that when the coin comes up heads (H) all paths go a step upwards or stay in state 3 if in state 3. Alternatively, if the coin comes up tails (T) then all paths move a step downwards or stay in state 0 if in state 0. Figure (3.2) illustrates the procedure. Each of the resulting paths behaves like a path of the random walk started in the corresponding initial state. If paths meet then they merge, we say they coalesce. So, if we continue evolving the paths they all merge eventually and we reach complete coalescence. At this stage the current state of the chain is the same regardless in which state it was started. We may describe the coupling using a transition rule given by

$$f(x, C) = \begin{cases} \min(x + 1, 3) & \text{if } C = H \\ \max(x - 1, 0) & \text{if } C = T, \end{cases} \qquad x \in \{0, 1, 2, 3\}, \tag{3.3.3}$$

where C describes whether the coin comes up H or T. We achieve the coupling of paths by applying the same realisation of the transition rule f to all paths.

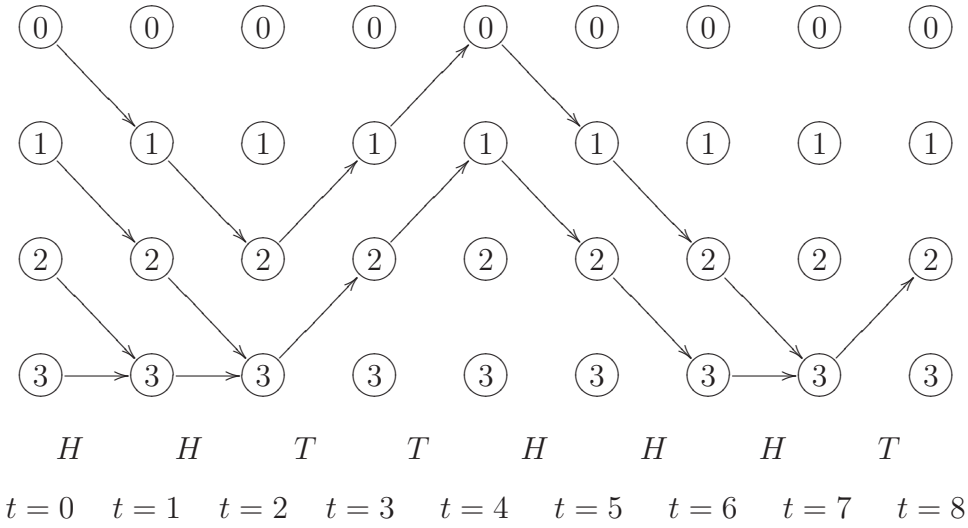


Figure 3.2: Forward simulation for Example (3.3.2) with t=8.

Recall from Equation (3.2.3) that $\mathbb{P}(f(x, C) = y) = P(x, y)$, for all $x, y \in E = \{0, 1, 2, 3\}$ where $P(x, y)$ are the transition probabilities of the target chain X. Furthermore, at time t we update all states using the same realisation of C. This is motivated by the fact that in this case

$$\forall\ x, y \in \{0, 1, 2, 3\}, \quad \mathbb{P}(f(x, C) = f(y, C)) \geq \sum_{z=0}^{3} P(x, z)P(y, z).$$

This means that at each step of our coupling the probability of two paths merging when using the transition rule f is greater or equal than the probability of two paths merging in an independent coupling. An independent coupling is achieved if we use an independent coin for each path.

At the time of complete coalescence, the chain is necessarily either in state 0 or 3, which is not a sample of the stationary distribution $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. To solve this problem, there is a very effective modification enables us to sample X in equilibrium. This modification is called *backward coupling* or Coupling From The Past and is due to Propp and Wilson (1996).

Chapter 4

Perfect Simulation

4.1 Introduction

In 1906, Andry Markov produced the first result about a class of stochastic processes. In such a process, the previous states are irrelevant for predicting the subsequent states, given knowledge of the current state. This property is now known as the Markov property. A generalization to countably infinite state spaces was given by Kolmogorov(1936). Markovian systems are related to Brownian motion and the ergodic hypothesis, two topics in physics which were important in the early years of the twentieth century. Markov chain methods were then used as the building stone for generating sequences of random numbers to accurately reflect very complicated desired probability distribution-a process called Markov Chain Monte Carlo (**MCMC**). **MCMC** method is a class of algorithms for sampling from probability distributions, based on constructing a Markov chain that has the desired distribution as its stationary distribution.

In the last ten years, the extensive use of **MCMC** methods to sample from complicated multivariate probability distribution attracts much attention among statisticians ([6],[20],[23]).

In (1992) Asmussen et al.[1] gave the first algorithm for sampling from the stationary distribution π of a Markov chain without knowledge of the mixing time of the Markov chain "time that guarantees that the distribution

will not be much far from π whenever the chain starts". They give a general procedure, which given m and a Markov chain on m states, simulates the Markov chain for a while, stops after a finite time, and then outputs a random state distributed exactly according to π. Recent years have seen the development of a new, exciting generation of **MCMC**: perfect or *exact* simulation algorithms. In contrast to conventional **MCMC**, perfect simulation produces samples which are guaranteed to have exact equilibrium distribution, i.e., the feature of sampling after equilibrium has been reached is that it produced samples which are guaranteed to have the target (desired) distribution and solve the problem of choosing an adequate burn in period.

Given a Markov chain which is irreducible and aperiodic that starts from any probability distribution P, and run the chain for t steps, obtain the probability distribution P^t which we want to be close to the desired distribution π. As we mentioned in the introduction one of the long standing problems in **MCMC** is that it is rarely possible to know when the Markov chain which is used for simulation has reached equilibrium. This problem was solved in $(1995, 1996)$ by the introduction of Coupling From The Past which was introduced by Propp and Wilson [18].

In this chapter, we discuss perfect simulation, beginning with ordinary CFTP and moving on to the Dominated Coupling From The Past.

4.2 Coupling From The Past Algorithm

Coupling From The Past is an MCMC method devised by Propp and Wilson [18], also called ***backward coupling*** to produce a perfect sample from the equilibrium distribution of a Markov chain through coupling multiple Markov chains with the same equilibrium distribution but different initial states. Consider a finite state space E with m states and an ergodic Markov chain (irreducible and aperiodic) with stationary distribution π that can be described by a deterministic update function f as in definition (3.2.1).

The idea of CFTP is to couple the Markov chain starting in all possible states of E at some time $-T < 0$ in the past, and to run them until time

0. If all copies have coalesced by time 0, then the value at time 0 does not depend on the starting value, and the output must be a sample from the stationary distribution. We use the same sequence of random numbers for all these chains. Since E is finite and the chains are ergodic with the same unique stationary distribution π, all the chains will coalesce almost surely and be stationary by time 0. CFTP is an algorithm for finding $-T$ and X_0, and goes as follows:

(1) Start chains at time $t = -1$ from every state of E. Generate U_0.

(2) Simulate each chain to time $t = 0$ by applying the transition rule $X_0 = f(x_{-1}, U_0)$. If the chains have coalesced at time $t = 0$, then $-T = -1$ and the common value X_0 is a draw from π.

(3) Otherwise, move back to time $t = -2$, generate U_{-1}, and simulate each chain using the transition rule $X_{-1} = f(x_{-2}, U_{-1})$ and $X_0 = f(x_{-1}, U_0)$. If the chains have coalesced at time $t = 0$, then $-T = -2$ and the common value X_0 is a draw from π.

(4) Otherwise, move back to time $t = -3$ and continue until coalescence occurs.

Example 4.2.1. *We consider a simple example with* $x \in E = \{0, 1, 2\}$ *and* $X_{t+1} = f(x_t, U_{t+1})$. *Draw* $U_{t+1} \sim$ *Uniform* $[0, 1]$ *and the transitions will be illustrated by Figure (4.1).*

Begin at time $t = -1$ and draw U_0. Suppose $U_0 \in [0, 0.3)$. As in the top part of Figure (4.2) it turns out that

$$\begin{cases} f(0, U_0) = 0 \\ f(1, U_0) = 1 \\ f(2, U_0) = 1. \end{cases}$$

Hence the state at time 0 can take two different values (0 or 1) depending on the state at time -1, so we go to time $t = -2$ and draw U_{-1}. Suppose

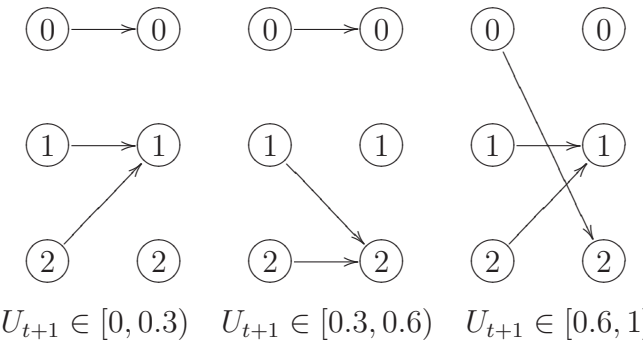


Figure 4.1: All possible transitions for Example (4.2.1).

$U_{-1} \in [0.3, 0.6)$. We then get

$$\begin{cases} f(f(0, U_{-1}), U_0) = f(0, U_0) = 0 \\ f(f(1, U_{-1}), U_0) = f(2, U_0) = 1 \\ f(f(2, U_{-1}), U_0) = f(2, U_0) = 1. \end{cases}$$

which again produces two different values at time 0. That is the chains have not coalesced, so we go to time $t = -3$ and draw $U_{-2} \in [0.6, 1]$.

This yields

$$\begin{cases} f(f(f(0, U_{-2}), U_{-1})), U_0) = f(f(2, U_{-1}), U_0) = f(2, U_0) = 1 \\ f(f(f(1, U_{-2}), U_{-1})), U_0) = f(f(1, U_{-1}), U_0) = f(2, U_0) = 1 \\ f(f(f(2, U_{-2}), U_{-1})), U_0) = f(f(1, U_{-1}), U_0) = f(2, U_0) = 1. \end{cases}$$

This time, we get to state 1 at time 0, regardless of the starting value at time -3. So all chains have coalesced into $X_0 = 1$. The algorithm therefore stops with output equal to 1. Note: Even though the chains have coalesced at $t = -1$, we do not accept $X_{-1} = 2$ as an output.

Remark 4.2.1. *When going back to time $t = -2$, we use the same U_0 that was already drawn, starting chains from every state; drawing U_{-1}; we use it to update all chains to time $t = -1$; we use the same U_0 to update all chains to time $t = 0$; then, we check for coalescence; and we either accept $T = -2$ and X_0 if the chains have coalesced or we back to time -3 if they have not. The algorithm continues backing through time until coalescence occurs.*

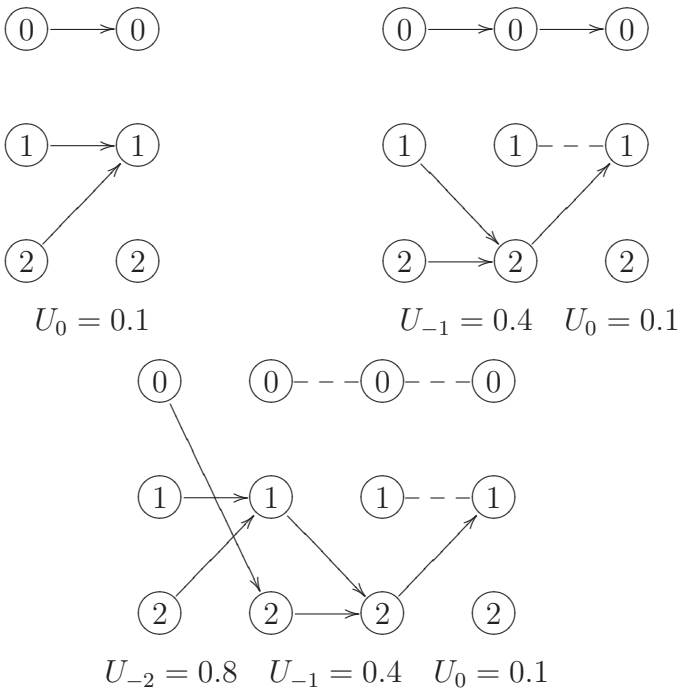


Figure 4.2: A run of the Propp-Wilson algorithm "CFTP" for the Markov chain of Example (4.2.1). Transitions that are carried out in the running of the algorithm are indicated with solid lines; others are dashed.

Note: In this algorithm T is successively equal to $-1, -2, \ldots$ Propp and Wilson (1996) argue that $T = -1, -2, -4, -8, \ldots$ is near optimal.

Example 4.2.2. ***CFTP*** *for Example (3.3.1). In both examples the same sequence of random numbers is used. In Figure (4.3) we begin our backward simulation start running the chain from time* -1 *to time* 0*, coalesce might occur with* $t = 1$ *in state* a*. While in Figure (4.4), the left diagram illustrates CFTP where the chains coalesce in state* a *but the state at time* 0 *is state* b*, since if* $U_{-2} \leq \frac{1}{2}$ *then* $X_0 = a$ *as we see in the right diagram.*

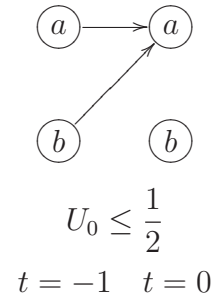


Figure 4.3: Simulation from the past for Example (3.3.1) with t=1.

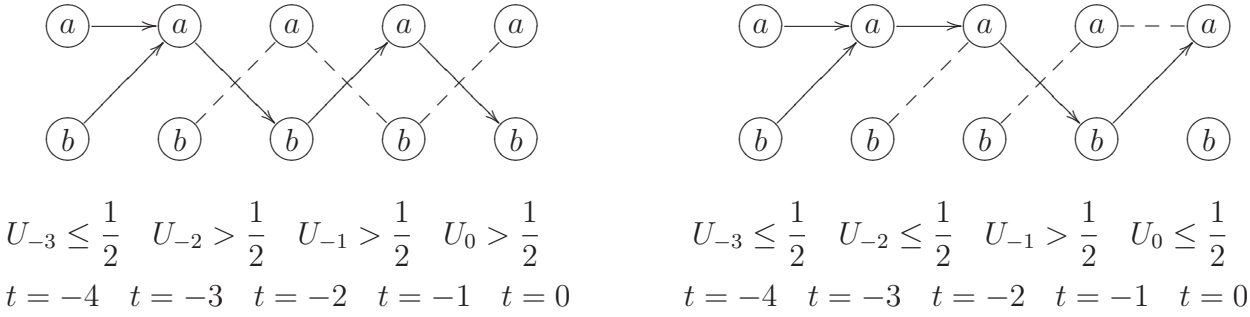


Figure 4.4: CFTP for Example (3.3.1) with t=4.

Example 4.2.3. ***CFTP*** *for Example (3.3.2). The algorithm as follows :*

1. *At time $-t, t \in \mathbb{N}$, we independently flip a fair coin C_{-t}.*
2. *Then, we start a path from all initial states $\{0, 1, 2, 3\}$ and evolve them from time $-t$ till time 0 according to the coin flips $C_{-t}, C_{-t+1}, \ldots, C_{-1}$.*
3. *If all paths coalesce at time 0, then we return their common state as a sample from π.*
4. *If the paths don't coalesce, we go a step backwards in time and repeat the above steps.*

Figure (4.5) illustrates the CFTP algorithm for Example (3.3.2).

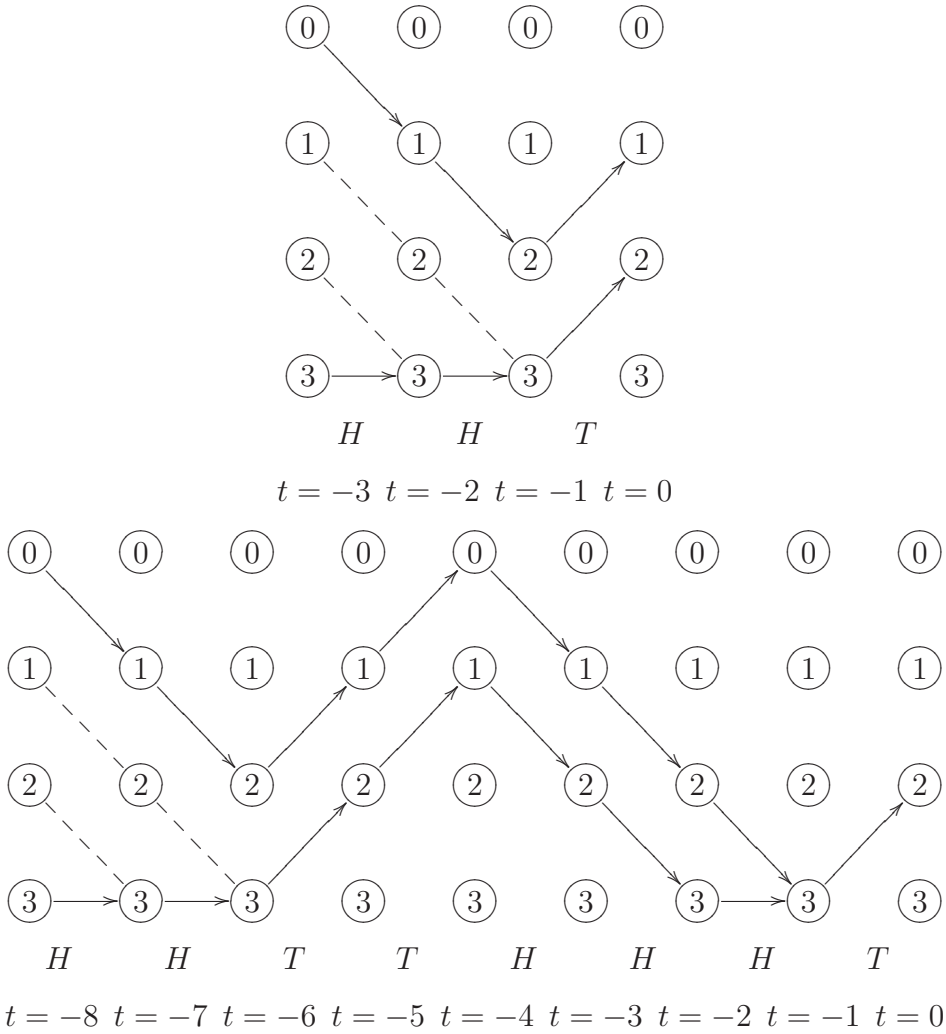


Figure 4.5: CFTP for Example (3.3.2). The paths started in state 0 and in state 3 are shown as solid lines. The dotted lines are the paths started from intermediate states.

Recall from Example (3.3.2) that we produced coupled sample paths of the chain X started from every initial state $x \in E$ by sampling transition rules $f_t, t \in \mathbb{N}$ by setting

$$X_t(x) = f(X_{t-1}(x), C_t) = (f_t \circ f_{t-1} \circ \ldots \circ f_1)(x) = F_0^t,$$

which becomes a constant function as $t \to \infty$ and f is defined as in (3.3.3). In Figure (3.1) $F_0^t(x) = a$ always, and in Figure (3.2) $F_0^t(x) = 2$.

Propp and Wilson reverse the order of composition and have the equivalent procedure

$$F_{-t}^0 = f_{-1} \circ \ldots \circ f_{-t}, \qquad t \in \mathbb{N}.$$

Let $F_{-t}^s = X_s^{-t}(x) = f_{-s-1} \circ \ldots \circ f_{-t}(x)$, then $\{X_s^{-t}(x), -t \leq s \leq 0\}$ behaves like a path of X started at time $-t$ in state x. Thus $F_{-t}^0(x)$ is the state at time 0 of a path of X started at time $-t$ in state x. The transition rules $f_{-t}, t \in \mathbb{N}$ describe transitions according to the transition kernel $\mathbf{P}$ of X. As $\pi\mathbf{P} = \pi$ the transition rules preserve the equilibrium distribution and so it follows that $F_{-T}^0(x)$ is also a sample from π, where $T = \min\{t : F_{-t}^0$ is a constant function$\}$.

In general: The idea of Coupling From The Past is simple.

Let $U_t, t \in \mathbb{Z}$ be independent, each chosen uniformly at random from $[0, 1]$ associated with time t. Consider the simulation of a Markov chain on E using independent identically distributed transition map $f_t : E \longrightarrow E$ such that $f_t(x) = f(x, U_t)$, where f is a deterministic function defined in (3.2.1). For a chain (X_t), the function f_t will define the evolution at time t. That is, we construct a Markov chain $X_{-t}, \ldots, X_0$ by constructing a set of update functions $f_s(x, u)$, such that:

$$X_s = f_s(X_{s-1}, U_s), \qquad s = 0, -1, \ldots, -t$$

where U_s is uniform at random variable on $[0, 1]$.

For any realisation of the set of independent random variables $U_0, U_{-1}, \ldots, U_{-t}$ these update functions specify the transition from a state X_{s-1} at time s to a state X_s at time $s + 1$. If $P(x, y)$ are the transition probabilities of the Markov chain to move from x to y then the common distribution of the f_t should be such that $\mathbb{P}(f_t(x, u) = y) = P(x, y)$.

Let $t_1 < t_2$ be integers, and

$$F_{t_1}^{t_2}(x) = (f_{t_2-1} \circ f_{t_2-2} \circ \ldots \circ f_{t_1})(x) = f_{t_2-1}(f_{t_2-2}(\ldots f_{t_1}(x)\ldots)).$$

Thus, $F_{t_1}^{t_2}$ defines the evolution from times t_1 to t_2, so F_0^t is the standard forwards simulation for t steps starting at time 0, and F_{-t}^0 is the $t-$step

evolution of the Markov chain from time $-t$ to time 0 (from the past). The compositions F^0_{-t} can be updated via the rule $F^0_{-t} = F^0_{-t+1} \circ f_{-t}$. More to the point is the observation that if the map F^0_{-t} ever becomes a constant map, with $F^0_{-t}(x) = F^0_{-t}(y)$ for all x, y.

From (3.2.3) it follows that for every $x, y \in E$

$$\mathbb{P}(F^{t_2}_{t_1}(x, U) = y) = P^{t_2 - t_1}(x, y).$$

As $F^0_{-t} = X^{-t}_0 = f_{-1} \circ \ldots \circ f_{-t}(x)$, then these update functions determine a value of $F^0_{-t} = X^{-t}_0$ for all values of X^{-t}_s. If t is such that, for some x, regardless of the value of X^{-t}_s, $F^0_{-t} = X^{-t}_0 = x$, then x is a draw from the stationary distribution of the Markov chain. Therefore, if F^0_{-t} maps the state space to a single value, then for any $t' > t$, $F^0_{-t'}$ will also map the state space to this same value. So with probability 1, all but finitely many of the random variables $F^0_{-1}, F^0_{-2}, \ldots$ will take the same value.

Definition 4.2.1. *Suppose that a Markov chain is defined via a random of functions $f_t, t \in \mathbb{Z}$ consistent with a transition matrix $\boldsymbol{P}$ as in Definition (3.2.2). We refer to such a chain as a complete coupling chain. If $F^0_{-t}(x) = F^0_{-t}(y)$, we say that the two Markov chains X and Y have coupled. If F^0_{-t} is a constant function, then we say that the chain has completely coupled at time 0, and we denote the first time this occurs by T.*

Theorem 4.2.1. *[18] Suppose an ergodic Markov chain is constructed so that there exists a time $L < \infty$ such that*

$$\mathbb{P}(F^L_0(x, U) = y) = P^L(x, y) > 0,$$

then,

(a) *with probability 1 the CFTP algorithm returns a value, and*

(b) *this value is a realisation of a random variable distributed according to the stationary distribution of the Markov chain.*

Proof. **(a)** Since the chain is ergodic, there is an L such that for all states x and y, there is a positive chance of going from x to y in L steps. let F^t_{t-L} be the event that the chain started from all states in E at time $t-l$

have all coalesced by time t. Hence, $F_{t-L}^{t}, \quad \forall t$ has a positive chance of being constant. Since each of the maps $F_{-kL}^{-(k-1)L}$, $k \in \{1, 2, \dots\}$ has some positive probability of being constant, and since these map are independent [because coalescence in (-kL, -(k-1)L)) only depends on $U_{-kL}, U_{-kL+1}, \dots, U_{-(k-1)L}$ which are independent of all of the other U_s and does not depend on the initial states] and by assumption of the theorem, it will happen with probability 1 that at least one of them is constant. This implies that the algorithm terminates with probability 1.

(b) let $T = \min\{t : F_{-t}^{0} \text{ is a constant function}\}$. Define

$$T' = \min\{t : F_{-t}^{1} \text{ is a constant}\}.$$

Couple the Markov chain for F_{-t}^{0} and F_{-t}^{1}, using the same $f_{t'}$ at some time t' (hence $T' \leq T$). Let $Z_{-\infty}^{1} = F_{-T'}^{1}(X)$ be the constant value of the function $F_{-T'}^{1}$, and π_0, π_1 be the distribution of $Z_{-\infty}^{0}$ and $Z_{-\infty}^{1}$, respectively. Since $Z_{-\infty}^{0}, Z_{-\infty}^{1}$ have the same probability distribution (because they are the value of the constant function obtained by CFTP up to some fixed time 0 or 1 respectively), we have $\pi_0 = \pi_1$. Since $Z_{-\infty}^{1} = f(Z_{-\infty}^{0})$ is obtained from $Z_{-\infty}^{0}$ by running the Markov chain one step, the common distribution $\pi_0 = \pi_1$ is a fixed point of f, so it must be the unique stationary distribution of the Markov chain.

□

A similar theorem can be found in [21],[24] with a slightly different proof.

Definition 4.2.2. *A random variable T is called a stopping time with respect to U_{-t} , $t = 0, 1, \dots$ if for any t, the event $T \leq t$ is determined by $U_0, U_{-1}, \dots, U_{-t}$.*

The following theorem is an important tool in determining the mixing time of a Markov chain.

Theorem 4.2.2. *Suppose that we have a completely coupling chain, X_0 has some distribution over E and Y_0 has a stationary distribution π. Then*

$$d_{TV}(F_0^t(X_0), \pi) = \|F_0^t(X_0) - \pi\| \leq P(X_t \neq Y_t) = \mathbb{P}(T > t).$$

Proof. By assumption, since Y_0 has a stationary distribution π, then $\mathbb{P}(Y_t \in A) = \pi(A)$ for all $A \subset E$, so

$$\begin{aligned}
\|F_0^t(X_0) - \pi\| &= \sup_{A\subset E} |\ \mathbb{P}(X_t \in A) - \pi(A)| \\
&= \sup_{A\subset E} |\ \mathbb{P}(X_t \in A, T > t) + \mathbb{P}(X_t \in A, T \leq t) - \mathbb{P}(Y_t \in A)| \\
&= \sup_{A\subset E} |\ \mathbb{P}(X_t \in A, T > t) + \mathbb{P}(Y_t \in A, T \leq t) \\
&\quad - (\mathbb{P}(Y_t \in A, T > t) + \mathbb{P}(Y_t \in A, T \leq t))| \\
&= \sup_{A\subset E} |\mathbb{P}(X_t \in A, T > t) - \mathbb{P}(Y_t \in A, T > t)| \\
&\leq \sup_{A\subset E} [\sup\{\mathbb{P}(X_t \in A, T > t), \mathbb{P}(Y_t \in A, T > t)\}] \\
&\leq \sup_{A\subset E} \mathbb{P}(T > t) \\
&= \mathbb{P}(T > t).
\end{aligned}$$

□

Theorem 4.2.3. *Assume that*

1) *the distribution of* $U_t\ , t \in \mathbb{Z}$ *is a stationary in time,*

2) *there exists a state* $x \in E$ *so that for any event* $A \subseteq E$, $\mathbb{P}(F_0^t(x) \in A) \to \pi(A)$ *as* $t \to \infty$, *and*

3) $T \geq 0$ *is an almost surely finite stopping time with respect to* $U_{-t}\ , t = 0, 1, \ldots$ *such that* $F_{-t}^0(x) = F_{-T}^0(x)$ *whenever* $T \leq t$. *Then* $F_{-T}^0(x) \sim \pi$

Proof.

$$\begin{aligned}
\mathbb{P}(F_{-T}^0(x) \in A) &= \mathbb{P}(\lim_{t\to\infty} F_{-t}^0(x) \in A) \\
&= \lim_{t\to\infty} \mathbb{P}(F_{-t}^0(x) \in A) \quad \text{by monotone convergence theorem} \\
&= \lim_{t\to\infty} \mathbb{P}(F_0^t(x) \in A) \quad \text{due to (1)} \\
&= \pi(A) \quad \text{due to (2).}
\end{aligned}$$

□

4.3 Monotonicity and Anti-Monotonicity

As we saw, the method of simulating a Markov chain whose stationary distribution is the distribution we want to sample from is known in general as **MCMC**. The traditional way to proceed is to run the Markov chain for a long time (called the burn in time), so that by the end of this period the Markov chain will be sufficiently close to stationarity that we may assume that we are now sampling from the required distribution. The question is "how long is long enough?". One partial solution, (**CFTP**) was developed by Propp and Wilson [18]. Other methods of sampling from the equilibrium distribution of a Markov chain have since been developed, and the general class of such algorithms is now commonly referred to as "Perfect Simulation", a name first coined by Kendall [12].

In Section (3.3), we showed how to couple paths of a Markov chain from different initial states such that after a random time T, the paths have merged into one. Although the state of any of these paths at this time doesn't depend on its starting value, it is still biased. This is because T is not a fixed time but a random time. The solution to the biased sampling problem is to run the chains from the past to the present, that is, we assume that at time $-\infty$ one chain is started in each of the states of E and run until 0. Also, we use the same sequence of random numbers for all these chains. To see why this gives a correct sampling scheme, Propp and Wilson assume that the state space is finite with the following conditions:

Condition 1: There exist a partial ordering $\preceq$ on the state space E : that is, if $X_t \preceq Y_t$, then

$$X_{t+1} = f(X_t, U_{t+1}) \preceq f(Y_t, U_{t+1}) = Y_{t+1}, \quad \forall \ t.$$

Condition 2: $\exists\ x^{min}$ and x^{max} such that $x^{min} \preceq x \preceq x^{max}$ for any state x.

Monotonicity of **P** is guaranteed when one can couple transitions using a monotone transition rule, and some Markov chains exhibits the property of monotonicity, which depends on the existence of a partial ordering of the state space E.

Definition 4.3.1. *A partial order on E is a binary relation $\preceq$ on E such that, for all $x, y, z \in E$,*

(i) $x \preceq x$,

(ii) $x \preceq y$ *and* $y \preceq x$ *imply* $x = y$,

(iii) $x \preceq y$ *and* $y \preceq z$ *imply* $x \preceq z$.

The set $(E, \preceq)$ *is said to be a partially ordered set (abbreviated: poset).*

Definition 4.3.2. *Suppose the functions* $f_t, t \in \mathbb{Z}$ *determine a completely coupling Markov chain. The Markov chain is said to be preserve the partial order* $\preceq$ *if for all* $x, y \in E$ *and all times* t, $f_t(x) \preceq f_t(y)$ *whenever* $x \preceq y$.

Now by induction, if the Markov chain preserves the partial order, that is $X_0 \preceq Y_0$, then $F^0_{-t}(X_0) \preceq F^0_{-t}(Y_0)$.

Definition 4.3.3. *A monotone transition rule for a transition matrix* $\mathbf{P}$ *on a partially ordered state space* $(E, \preceq)$ *is a measurable function* $f : E \times U \longrightarrow E$, *together with a random number* U *taking values in an arbitrary probability space* U, *such that:*

(i) $f(x, u) \preceq f(y, u)\ \forall\ u \in U$ *whenever* $x \preceq y$,

(ii) $\mathbb{P}(f(x, u) = y) = \mathbb{P}(X_{t+1} = y \mid X_t = x) = P(x, y)$, *for all* $x, y \in E$,

(iii) $X_{t+1} = f(X_t, U_{t+1})$ *which present a Markov chain,*

where the function f *is the updating function, and the* U_{t+1} *are i.i.d. sequence distributed uniformly on* $[0, 1]$ *defined in Section (3.2).*

If all realisations of the random map are monotone functions, then we call it a monotone transition rule, and when it exists, we can simultaneously generate transitions from various states in such a way as to maintain relations for each realisation. In other words, if we consider two paths generated by the updating rule, using the same random sequence, and started in the states $x \preceq y$, respectively, then with respect to $\preceq$, the path started in x remains beneath the path started in y.

Definition 4.3.4. *When a monotone transition rule and* x^{min}, x^{max} *exist for a given transition matrix* $\mathbf{P}$, *we shall say that monotone case obtains.*

When the transition rule is monotone, we need only start the Markov chain in the extreme states x^{min}, x^{max}. When these two chains have coalesced, so have all the others. That is, at time 0, if $F^0_{-t}(x^{min}) = F^0_{-t}(x^{max})$, where $x^{min} \preceq x \preceq x^{max}$ for all $x \in E$, then since for all times t, $F^0_{-t}(x^{min}) \preceq F^0_{-t}(x) \preceq F^0_{-t}(x^{max})$, we have that $F^0_{-t}(x^{min}) = F^0_{-t}(x) = F^0_{-t}(x^{max})$ for all $x \in E$. The algorithm works as follows:

1. Run two copies of the chain from time $t = -1, -2, -4, -8, \ldots$ up to time $t = 0$, one in X^{max} (the top chain) and the other in X^{min} (the bottom chain) respectively using update function as in Definition (3.2.1) with random numbers U_t which is the same for all chains.

2. If at time 0 all chains "coalesce" end in the same state z, then use z as a sample. Otherwise, repeat but start at time $-2, -4, -8, \ldots$, re-using randomness whenever possible.

Remark 4.3.1. *We use the same sequence of random numbers in going from time t to $t+1$ for all paths, else the samples that we generate will be biased.*

Example 4.3.1. *Consider a Markov chain on state space $E = \{0, 1, 2\}$, with the following transition matrix* $\mathbf{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$.

By solving the equation system of definition(1.5.1), we find the stationary distribution $\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Monotonicity is obtained if we have the partial ordering $0 \preceq 1 \preceq 2$, and we may run the Propp-Wilson algorithm with the two chains X and Y with initial states $X_0 = 0, Y_0 = 2$. To ensure the monotone case obtains the chains evolves according a transition rule using a uniform variable U in the following way

$$X_{t+1} = f(X_t, U_{t+1}) = \begin{cases} \min(X_t + 1, 2) & \text{if } U_{t+1} > \frac{1}{2} \\ \max(X_t - 1, 0) & \text{if } U_{t+1} \leq \frac{1}{2}. \end{cases} \tag{4.3.1}$$

With 2 iterations, let I be a categorical random variable taking value

$$\begin{cases} 1 & \text{if } X, Y \text{ have coalesced within 2 iterations} \\ -1 & \text{if no coalescence have occurred.} \end{cases} \tag{4.3.2}$$

Let z denote the output of the simulation conditionally given completion.

By investigating the transition matrix $\mathbf{P}$, we find that X and Y can not coalesce in just one step, and with 2 steps we have the four following possible outcomes:

If $U_1 \leq \frac{1}{2}$, $U_2 \leq \frac{1}{2}$, then $I = 1$ and $X_1 = \max(-1, 0) = 0 \Rightarrow X_2 = \max(-1, 0) = 0$, $Y_1 = \max(1, 0) = 1 \Rightarrow Y_2 = \max(0, 0) = 0$, therefore $z = 0$.

If $U_1 \leq \frac{1}{2}$, $U_2 > \frac{1}{2}$, then $I = -1$, and $X_1 = \max(-1, 0) = 0 \Rightarrow X_2 = \max(-1, 0) = 0$, $Y_1 = \min(3, 2) = 2 \Rightarrow Y_2 = \min(3, 2) = 2$, thus z is not defined.

If $U_1 > \frac{1}{2}$, $U_2 \leq \frac{1}{2}$, then $I = -1$, and $X_1 = \min(1, 2) = 1 \Rightarrow X_2 = \min(2, 2) = 2$, $Y_1 = \max(1, 0) = 1 \Rightarrow Y_2 = \max(0, 0) = 0$, implies z is not defined.

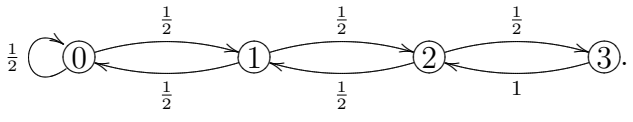
If $U_1 > \frac{1}{2}$, $U_2 > \frac{1}{2}$, then $I = 1$, and $X_1 = \min(1, 2) = 1 \Rightarrow X_2 = \min(2, 2) = 2$, $Y_1 = \min(3, 2) = 2 \Rightarrow Y_2 = \min(3, 2) = 2$, therefore $z = 2$.

Thus, we see that this algorithm will yield output z governed by the distribution $\pi' = (\frac{1}{2}, 0, \frac{1}{2})$ whereas the desired distribution was $\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. That is, the samples we get will be biased in favor of the extreme states 0 and 2.

Example 4.3.2. *[22] Suppose we have three balls which are distributed over two urns. With probability $\frac{1}{2}$ we pick a ball from the left urn and put it into the right urn. Alternatively, we take a ball from the right urn and put it into the left urn. If we find a chosen urn empty we do nothing. Assume that we choose the left urn and it is empty, then we take a ball from the right urn and put into the left one. If we find the right urn empty we do nothing.*

The number of balls in the left urn is a Markov chain with state space $E = \{0, 1, 2, 3\}$ and transition matrix

$$\mathbf{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 \end{pmatrix}$$



Again we would like to start a path of the chain from different initial states, so to simulate the chain we can use a fair coin such that when the coin comes up heads (H) *paths remain in state* 0 *if in state* 0*, and move step up if in state* 1 *or* 2 *and move step down if in state* 3*. Alternatively, if the coin comes up tails* (T) *paths move a step up from state* 0 *and move a step down if in state* 1*,* 2 *or* 3*.*

Similar to Example (3.3.2) we may produce a coupling of paths by using the same realisation of a coin flip when updating the paths. As before, we may describe the coupling using a transition rule given by

$$f(x,C) = \begin{cases} x+1 & \text{if} \quad C = H \quad \text{and} \quad x \in \{1,2\} \\ x-1 & \text{if} \quad C = T \quad \text{and} \quad x \in \{1,2\} \\ 0 & \text{if} \quad C = H \quad \text{and} \quad x = 0 \\ 1 & \text{if} \quad C = T \quad \text{and} \quad x = 0 \\ 2 & \text{if} \quad x = 3, \end{cases} \tag{4.3.3}$$

where C is the realisation of the coin flip. Suppose we impose the following partial order on E: $2 \preceq 0 \preceq 1 \preceq 3$. Then for fixed C and $x \preceq y$ we have $f(x,C) \succeq f(y,C) \quad \forall \;\; x,y \in E$.

To see that, take $C = H$, then $0 = f(0,H) \succeq f(1,H) = 2$, and if $C = T$, then $1 = f(0,T) \succeq f(1,T) = 0$. This transition rule is called anti-monotone, which allows us to monitor complete coalescence by evolving two paths only, we denote them by X^{min} and X^{max}, that is

$$X_{-t}^{min} = 2 \;\; \textit{and} \;\; X_{-t}^{max} = 3.$$

We then evolve the two paths as follows

$$X_{-t+1}^{min} = f_{-t}(X_{-t}^{max}, C_{-t+1}), \quad X_{-t+1}^{max} = f_{-t}(X_{-t}^{min}, C_{-t+1}),$$

where C_t is the t^{th} coin toss. At time $-t$ we start a path in the minimal state 2 and a path in the maximal state 3. Hence the two paths evolve as a two-component Markov chain in which the update of one component is made according to the current state of the other component. If coalescence of the

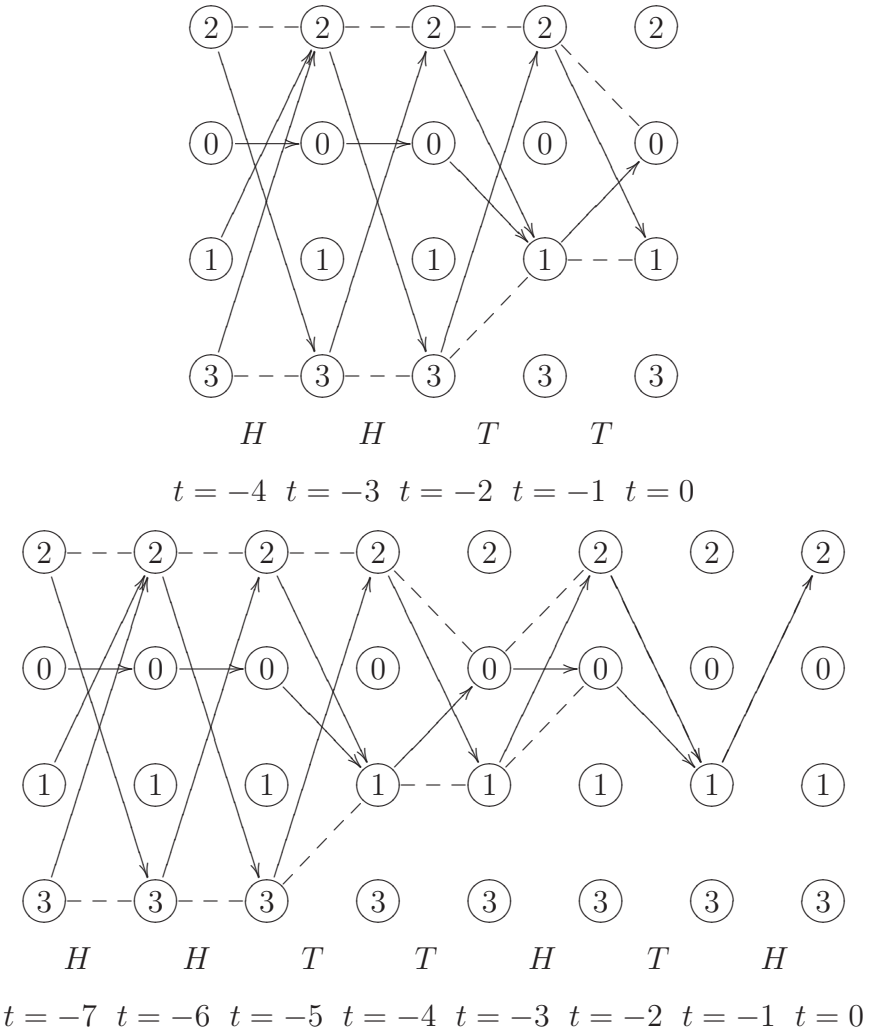


Figure 4.6: CFTP for the Markov chain in Example (4.3.2). The dotted line show the maximal and minimal path.

minimal and maximal paths occurs by time 0, then their common state at time 0 has the equilibrium distribution.

In the top of Figure (4.6), the state at time 0 can take two different values (0 or 1), so the chains have not coalesced, we go far in the past and reuse

the coin toss realisation.

$$X_{-6}^{min} = f_{-7}(3, H) = 2, \qquad X_{-6}^{max} = f_{-7}(2, H) = 3$$

$$X_{-5}^{min} = f_{-6}(3, H) = 2, \qquad X_{-5}^{max} = f_{-6}(2, H) = 3$$

$$X_{-4}^{min} = f_{-5}(3, T) = 2, \qquad X_{-4}^{max} = f_{-5}(2, T) = 1$$

$$X_{-3}^{min} = f_{-4}(1, T) = 0, \qquad X_{-3}^{max} = f_{-4}(2, T) = 1$$

$$X_{-2}^{min} = f_{-3}(1, H) = 2, \qquad X_{-2}^{max} = f_{-3}(0, H) = 0$$

$$X_{-1}^{min} = f_{-2}(0, T) = 1, \qquad X_{-1}^{max} = f_{-2}(2, T) = 1$$

$$X_{0}^{min} = f_{-1}(1, H) = 2, \qquad X_{0}^{max} = f_{-1}(1, H) = 2,$$

i.e., $(2, 3), (2, 3), (2, 3), (2, 1), (0, 1), (2, 0)$.

Complete coalescence occurs at time -1, however we continue till time 0 and sample state 2.

For many Markov chains on large state spaces, the determination of complete coalescence is not practical if we have to monitor the paths from all initial states, and **CFTP** algorithm will not be feasible. Efficient simulation is possible only if we can reduce the number of chains to run, which can be done by monotonicity. One example for such a chain is the Gibbs sampler for the Ising model.

4.4 The Ising Model

The Ising Model was introduced in statistical physics in the study of ferromagnetic substances, and was the first model for perfect simulation ([5],[18]). This model is difficult to sample from, and **MCMC** algorithms must be used for this. The statistical foundation of the Ising model is the theory of Markov random fields.

Definition 4.4.1. ***Random Field***

Let I be a vertex set, and let $X = \{x_i, i \in I\}$ be a random vector with state space E defined on I. If we have $\mathbb{P}(X = x) = \pi(x) > 0$ for all $x \in E$, then we say we have a random field.

Definition 4.4.2. *A collection $\delta = \{\delta(i); i \in I\}$ of subsets of I is called a neighbourhood system if*

(i) *$i \notin \delta(i)$.*

(ii) *$i \in \delta(j)$ if and only if $j \in \delta(i)$.*

The sites $i \in \delta(j)$ are called the neighbours of j, and we write $i \sim j$ if i and j are neighbours.

Definition 4.4.3. *Consider a finite set I of vertices, regarded as sites. A configuration $X = (x_i; i \in I)$ assigns to each site i a spin x_i, either $+1$ "upward" or -1 "downward".*

We denote $(x_i = \sigma \mid x_{-i})$ for the configuration whose spin at site i is σ and whose spins at other sites agree with these for x.

Let π be a probability distribution on E and let $x \in E$ with $\pi(x) > 0$.

If the site i is to be updated, the spin at this site is therefore set to

$$\begin{cases} +1 & \text{with probability } \mathbb{P}(x_i = +1 \mid x_{-i}) = \dfrac{\pi(x_i = +1, x_{-i})}{\pi(x_i = +1, x_{-i}) + \pi(x_i = -1, x_{-i})} \\ -1 & \text{with the complementary probability } \mathbb{P}(x_i = -1 \mid x_{-i}). \end{cases} \tag{4.4.1}$$

These transition probabilities do not depend on x_i.

We order the set of configuration by putting $x \preceq y$ if and only if

$$x_j \preceq y_j \quad \text{for all} \quad j \in I.$$

Definition 4.4.4. *[24] We say that π is attractive if the conditional probability of the event $\pi(x_i = 1 \mid x_{-i})$ is an increasing function of the values of x_{-i} (in the partial order).*

Example 4.4.1. *([22],[24])**Ising model**:*

The Ising model has energy function

$$E(x) = -J \sum_{i \sim j} x_i x_j - \sum_i \alpha_i x_i,$$

where

x_i*= spin at site* i *of* I*,*

$i \sim j$ *=nearest neighbours in* I*,*

J*= the interaction strength between* i *and* j*, such that: when* $J > 0$*, the state of lowest energy is when all spins are aligned. This is called a ferromagnetic, and* $J < 0$ *is an anti-ferromagnetic,*

α_i*= external field measured at site* i*.*

A certain spin configuration is chosen according to the probability distribution π *with*

$$\pi(x) = \frac{1}{z} exp(-E(x)),$$

where z *is the normalizing constant known as a "partition function". Sampling can be done effectively via Gibbs sampling with no need of calculating the partition function* z*, to produce a Markov chain whose distribution converges to the Ising model. Recall that the Gibbs sampling method updates one site at a time, and that the conditional distribution is used as proposal.*

Let x_{-i} *denotes the configuration* x *on* I *excluding the site* i*, then the conditional probability of the Ising model having an upwards spin at site* i *given the current spin configuration on all other sites is*

$$\begin{aligned} \mathbb{P}(x_i = +1 \mid x_{-i}) &= \pi(x_i = 1 \mid x_{-i}) \\ &= \frac{\pi(x_i = +1, x_{-i})}{\pi(x_i = +1, x_{-i}) + \pi(x_i = -1, x_{-i})}. \end{aligned} \tag{4.4.2}$$

Let $(x_i = +1, x_{-i})$ *be the configuration which we obtain by setting* $x = +1$ *and leaving the spins of all other sites in* x *unchanged, then from Equation*

(4.4.2), we have

$$\begin{aligned}
\mathbb{P}(x_i = +1 \mid x_{-i}) &= \pi(x_i = 1 \mid x_{-i}) \\
&= \frac{exp(J\sum_{i\sim j} x_i x_j + \sum_i \alpha_i x_i)}{exp(J\sum_{i\sim j} x_i x_j + \sum_i \alpha_i x_i) + exp(-J\sum_{i\sim j} x_i x_j - \sum_i \alpha_i x_i)} \\
&= \left[\frac{exp(J\sum_{i\sim j} x_i x_j + \sum_i \alpha_i x_i) + exp(-J\sum_{i\sim j} x_i x_j - \sum_i \alpha_i x_i)}{exp(J\sum_{i\sim j} x_i x_j + \sum_i \alpha_i x_i)}\right]^{-1} \\
&= [1 + exp(-2J\sum_{j\sim i} x_j - 2\alpha_i)]^{-1} \\
&= [1 + exp(-2(\alpha_i + J\sum_{j\sim i} x_j))]^{-1},
\end{aligned} \tag{4.4.3}$$

which does not depend on the partition function z.

Remark 4.4.1. *Since Equation (4.4.3) is increasing in x_{-i}, the Gibbs sampler is monotone for the Ising model with an external field.*

Proof. We can couple paths of the Gibbs Sampler started from different initial states by reusing the sampled random variates U_{t+1} and N_{t+1}, such that U_{t+1} is uniformly distributed on $[0, 1]$ and N_{t+1} is drawn uniformly on $(1, \ldots, n)$. At time t we update the same site N_{t+1} in each path using the same realisation of U_{t+1} for all paths.

We may defined the Gibbs sampler through the update function f given by

$$f(x, U, N) = \begin{cases} (x_N = +1, x_{-N}) & \text{if } U \leq \mathbb{P}(x_N = +1 \mid x_{-N}) \\ (x_N = -1, x_{-N}) & \text{otherwise.} \end{cases} \tag{4.4.4}$$

To show that f is monotone, we need to show that $f(x, U, N) \preceq f(y, U, N)$ for fixed U, N, when $x \preceq y$.

Now if the system is attractive then, $\pi(x_N = 1 \mid x_{-N}) \preceq \pi(y_N = 1 \mid y_{-N})$ for $x \preceq y$, hence $\mathbb{P}(x_N = +1 \mid x_{-N}) \leq \mathbb{P}(y_N = +1 \mid y_{-N})$. Thus, for fixed U, N, we have

1. Whenever f assigns an upwards spin to x_N then, it also assigns an upwards spin to y_N, as $U \leq \mathbb{P}(x_N = +1 \mid x_{-N}) \leq \mathbb{P}(y_N = +1 \mid y_{-N})$.

2. Analogously, whenever f assigns a downward spin to y_N then, it assigns a downward spin to x_N.

Thus, if $x \preceq y$ then, the updated configurations maintain their partial ordering, that is, $f(x, U, N) \preceq f(y, U, N)$ for fixed U, N. Therefore our transition rule f is monotone with respect to $\preceq$. □

The basic idea of the Propp-Wilson method is not to run the Markov chain forward in time, but rather to run the Markov chain from time $-t$ in the past up to time $t = 0$. **CFTP** can be used in the Ising model by running two coupled realisations X and Y of the Markov chains (two spin configurations), starting in the maximal and minimal states at some time $-t$ in the past, with all spin up or all spin down, using the updating function f defined in (4.4.4), such that $X_{t+1} = f(X_t, U_{t+1}, N_{t+1})$ and $Y_{t+1} = f(Y_t, U_{t+1}, N_{t+1})$. [The same random numbers are used for each spin configurations]. If $X_0 = Y_0$, then this value is our sample. That is, if the two chains have coupled at time 0, the state at time 0 is an exact sample from the Markov chain's stationary distribution.

Since f is monotone, we have that $F_{t_1}^{t_2}(E)$ is constant $\iff F_{t_1}^{t_2}(x^{min}) = F_{t_1}^{t_2}(x^{max})$, where $x^{min}, x^{max} \in E$. The coupling time of f is the smallest T such that $F_0^T(x^{min}) = F_0^T(x^{max})$. The coupling from the past time of f is the smallest M such that $F_{-M}^0(x^{min}) = F_{-M}^0(x^{max})$. We have that

$$\mathbb{P}(T > t) = \mathbb{P}(F_0^t(x^{min}) \neq F_0^t(x^{max})) = \mathbb{P}(F_{-t}^0(x^{min}) \neq F_{-t}^0(x^{max})) = \mathbb{P}(M > t).$$

Hence M has the same distribution as T.

4.5 Dominated Coupling From The Past

It follows from the very existence of **CFTP** constructions that all the chains discussed so far have been uniformly ergodic.

Definition 4.5.1. *A Markov chain having stationary distribution π is uniformly ergodic if*

$$d_{TV}(P^t(x), \pi) = \|P^t(x) - \pi\| \leq M\rho^t, \qquad t = 1, 2, \ldots$$

for some $\rho < 1$ and $M < \infty$.

If the Markov chain is not uniformly ergodic, what should we use? Suppose $T \geq 0$ is the smallest random time such that coupled paths of the target chain started in all initial states at time $-T$ have coalesced by time 0. As pointed in [6], uniform ergodicity is a sufficient and necessary property for applying the Propp and Wilson algorithm, the failure of uniform ergodicity implies that coalescence of the Markov chain can not happen in finite time, so the original **CFTP** algorithm of finite state space E of [18] needs modification. This modification is an extension of the original **CFTP** algorithm, called Dominated coupling from the past (**DCFTP**) or *horizontal* CFTP, which is an example of perfect simulation for sampling from the equilibrium distribution of a Markov chain.

In [6], *vertical* CFTP is a coupling of realisations started at a fixed time for all possible initial states, while *horizontal* CFTP is a coupling of realisations started at the minimal state at all sufficiently early initial times.

The idea of **DCFTP** is as follows: generate target chains coupled to dominating process for which equilibrium is known. We consider a discrete time monotic chain defined on $[0, \infty)$.

Definition 4.5.2. *[13]**DCFTP***
Consider X, an ergodic Markov chain on $[0, \infty)$. Suppose it can be coupled as follows: for each $x \geq 0, -t < 0$ we can construct $X^{-t}(x)$ to be X begun at x at time $-t$, such that if $s \geq -t, s \geq -u$, then

$$X_s^{-t}(x) \leq X_s^{-u}(y) \qquad \text{implies} \qquad X_{s+1}^{-t}(x) \leq X_{s+1}^{-u}(y).$$

Furthermore, suppose we can build a dominating process Y on $[0, \infty)$, which is stationary, defined for all time, and coupled to the $X^{-t}(x)$ by

$$X_s^{-t}(x) \leq Y_s \qquad \text{implies} \qquad X_{s+1}^{-t}(x) \leq Y_{s+1},$$

whenever $s \geq -t$. Then the following algorithm delivers a perfect sample from the equilibrium distribution of X, so long as it terminates almost surely:

(1) Draw Y_0 from its equilibrium distribution;

(2) Simulate Y backwards in time to time $-T$;

(3) Set $y = Y_{-T}$, and simulate the upper process $X^{-T}(y)$ and the lower process $X^{-T}(0)$ forwards in time to time 0 (The upper and lower must be coupled to each other, to Y, and at later stages of the algorithm they must be coupled to other simulations of X at the same process time);

(4) If $X_0^{-T}(y) = X_0^{-T}(0)$, then return their common value as a perfect draw from the equilibrium distribution of X. Otherwise, extend the previous simulation of Y back to time $-2T$, update $-T$ to $-2T$, and repeat from step (3).

Theorem 4.5.1. *[13],[22] If coalescence is almost sure then dominated CFTP samples from equilibrium.*

Proof. Suppose the target process X is monotone, and that it and the dominating process Y are nonnegative.

Let $X^{upper,-t}, X^{lower,-t} = X^{-t}$ be versions of the target chain started at time $-t$ at $Y(-t), 0$ respectively. Let

$$T = \inf\{t \geq 0 : X^{upper,-t}(0) = X^{lower,-t}(0)\},$$

so, $-T$ (the coalescence time) is the latest time such that $X^{upper,-T}(0) = X^{lower,-T}(0) = X^{-T}(0)$. So,

$$X_0^{-t} = X_0^{-T} \quad \text{whenever} - t \leq -T;$$
$$\mathcal{L}(X_0^{-t}) = \mathcal{L}(X_t^0), \text{ where } X_s^{-t} = F_{-t}^s(0) \qquad \text{for } -t \leq s.$$

If X converges to an equilibrium π in total variation d_{TV}, then

$$d_{TV}(\mathcal{L}(X_0^{-T}), \pi) = \lim_t d_{TV}(\mathcal{L}(X_0^{-t}), \pi) = \lim_t d_{TV}(\mathcal{L}(X_t^0), \pi) = 0.$$

This proves the theorem. □

Conclusion

Random sampling has found numerous applications in computer science, statistics, and physics. The most widely applicable method of random sampling is to use a Markov chain whose stationary distribution is the probability distribution π from which we wish to sample. After the Markov chain has been run for long enough, its state is distributed exactly according to π. The principal problem with this approach is that it is often difficult to determine how long to run the Markov chain. This thesis provides the motivation for perfect simulation. It begins with the concept of a Markov chain on finite state space, and attempts to explain and summaries **MCMC** algorithms. Then gives a review of the basic ideas in the new methods of perfect simulation, the concepts of Coupling From The Past, which is a class of algorithms for generating perfectly random samples from a Markov chain in the discrete case, produces samples which are guaranteed to have the exact equilibrium distribution. We have seen that **CFTP** of uniformly ergodic Markov chain is based on couplings of paths of the target Markov chain started in different initial states such that after an almost surely finite time the paths coincide (they merge into one or coalesce), and does not depend on its starting value. Monotonicity and anti-monotonicity with respect to a partial order of the chosen coupling help in checking for complete coalescence by reducing the number of paths to run into lower and upper path, such that if they coalesce at time 0, their common state is an exact sample from the stationary distribution. If the Markov chain is not uniformly ergodic, we have to use **DCFTP**.

Bibliography

[1] Asmussen, S., Glynn, P.W. and Thorisson, H. *Stationary detection in the initial transient problem*, ACM Transactions on Modeling and Computer Simulation, **2**(2)(1992), 130-157.

[2] Casella, G. and George, E. I. *Explaining the Gibbs sampler* , Am. Stat. **46**(1992), 167-174.

[3] Durrett, R. *Probability: Theory and Examples*, Wadsworth Publishing Company, Second Edition, 1996.

[4] Evans, Michael J. and Rosenthal, J. S. *probability and statistics*, The science of Uncertainty, W. H. Freeman and Company, NewYork, 2003.

[5] Fill, J.A. *An interruptible algorithm for perfect sampling via Markov chains*, Annals of Applied Probability, **8**(1998), 131-162.

[6] Foss, S.G. and Tweede, R.L. *Perfect simulation and backward coupling*, Stochastic models **14**(1998), 187-203.

[7] Gelfand, A.E. and Smith, A.F.M. *Sampling based approaches to calculating marginal densities*, Journal of the American Statistical Association, **85**(1990), 398-409.

[8] Geman, S. and Geman, D. *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images*, I.E.E.E. Trans. On Patten Analysis Machine Intelligence, **6**(1984), 721-741.

[9] Gilks, W.R., Richardson, S. and Spiegelhater, D.J. *Markov Chain Monte Carlo In Practice*, Chapman and Hall, London, 1996.

[10] Haggstrom, O. *Finite Markov Chains and Algorithmic Applications*, Cambridge University Press, 2000.

[11] Hastings, W.K. *Monte Carlo sampling using Markov chains and their Applications* Biometrika, **57**(1970), 97-109.

[12] Kendall, W.S. and Thonnes, E. *Perfect Simulation in Stochastic Geometry*, 1998.

[13] Kendall, W.S., Liang, F. and Wang, J-S. *Markov Chain Monte Carlo Innovations and Applications*, World Scientific Publishing Co. Pte. Ltd., 2005.

[14] Lindvall, T. *Lectures on the Coupling Method*, Dover Publications Inc., Mineola, NY, 2002.

[15] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. *Equations Of State Calculations by Fast Computing Machine*, Journel of chemical physics, Vd., **21**(1953), 1087-1092.

[16] Mϕller, J. *A Note On Perfect Simulation of Conditionally Specified Models*, 1997.

[17] Murdoch, D.J. and Rosenthal, J.S. *An extension of Fill's exact sampling algorithm to non-monotone chains*, 1998.

[18] Propp, J.G. and Wilson, D.B. *Exact Sampling With Coupled Markov Chains and applications to statistical mechanics*, Random Structures Algorithms, **9**(1996), 223-252.

[19] Shiryayev, A.N. *Probability*, Springer-Verlag , NewYork, 1984.

[20] Smith, A.F.M. and Robert, G.G. *Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods*, Journel of the Royal Statistical Society, Series B(1993), 553-23.

[21] Thonnes, E. *A primer on Perfect Sampling*, International Statistical Review, **69**(1)(2000), 27-48.

[22] Thonnes, E. *A primer on Perfect Simulation*, Technical report, Department of Mathematical Statistics, Chalmers University of Technology, 1999.

[23] Tierney, L. *Markov Chains for exploring posterior distributions*, (with discussion). Annals of statistics, **22**(1994), 1701-1762.

[24] Xeni, k. Dimakos. *A Guide To Exact Simulation*, Statistical Physics and Spatial Statistic, ed. Klaus R. Mecke and D. stoyan, Springer Lecture Notes in physics, **554**(2001), 349-378.