Coupling: Cutoffs, CFTP and Tameness

Stephen Connor

A Thesis presented for the degree of
Doctor of Philosophy

Department of Statistics
University of Warwick
England

Submitted: June 2007

CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

My heartfelt thanks goes out to everyone who has helped and encouraged me throughout my PhD. Especially warm thanks go to my supervisor, Wilfrid Kendall, for his boundless patience, enthusiasm and sense of humour. I don't believe that anyone could have been a more supportive supervisor.

I acknowledge the support of the EPSRC, who funded this research. I would also like to express my gratitude to Neil O'Connell (Cork), and Gersende Fort and Randal Douc (Paris) for their friendly hospitality and interest in my work.

Thanks to everyone in the Department of Statistics at Warwick, from those who contributed to my work to those who simply made the place more fun to be around. I won't name you all here, but you know who you are. Thanks in particular to Paula, Julia and Sue for refreshing conversation and general kindness.

Thanks to Chris and Katherine for uncountable chats over coffee, cake and crosswords (as well as the odd alcoholic beverage): I couldn't have asked for better friends with whom to celebrate and commiserate. Also to Julie, who has somehow suffered my company for nearly eight years at Warwick, and whose friendship continues to mean a great deal to me.

I'd like to say a particular thank you to my family who, despite not understanding what I've been doing this past few years, have unconditionally supported me throughout.

Finally, my greatest thanks of all go to my wonderful wife, Bert, for her love and inspiration, and without whom this would have been a far greater struggle.

DECLARATION

I declare that this thesis is my own work, except where explicitly stated, and has not been submitted elsewhere. The material at the start of Chapter 4 is taken from the encyclopedia article of the author (Connor 2007), and Chapter 5 is based upon the article of Connor and Kendall (2007).

ABSTRACT

The principal theme underlying this work is that of coupling. Coupling is a general technique with applications in many areas of probability, as well as being an active area of research in its own right. In this thesis a number of problems involving coupling are investigated: some new results, as well as an indication of exciting possibilities for future research, are given in each case.

Our journey into the world of coupling begins with the topic of the cutoff phenomenon for random walks on groups. Chapter 2 investigates the behaviour of a coupling for a general random walk on the hypercube, proving the existence under a simple condition of a new type of threshold behaviour called a *coupling-cutoff*. Chapter 3 is concerned with the theory of *maximal* couplings of Markov chains. This concept is generalised to maximal coalescent couplings, and an explicit description of an optimal co-adapted coupling for the symmetric random walk on $\mathbb{Z}_2^n$ is presented. The difference between optimal co-adapted and maximal couplings is also investigated for Brownian motion and the Ornstein-Uhlenbeck process.

Coupling is at the heart of the simulation technique known as perfect simulation, and this subject forms the focus of the second half of the thesis. Some consideration is given to the efficiency of Coupling from the Past (CFTP) algorithms, but the principal novel contribution to this area is an investigation into the existence of a dominated CFTP algorithm for subgeometrically ergodic Markov chains. This question turns out to be significantly harder than that for geometrically ergodic chains: we introduce a class of positive recurrent chains, named *tame chains*, for which a perfect simulation algorithm is shown to exist.

LIST OF SYMBOLS AND ACRONYMS

Everything starts somewhere, although many physicists disagree.

Hogfather, by Terry Pratchett

1. AN INTRODUCTION TO COUPLING

In this introductory chapter we present some background material on the theory and applications of coupling. Precise definitions shall be given, but a more intuitive view of coupling methods will often be taken, as it is this approach which usually serves to be the most enlightening in applications. In preparation for the work in Chapters 2 and 3, we introduce the concept of random walks on groups and briefly review some methods for analysing their convergence rate. The relatively simple but important example of a random walk on the hypercube will also be introduced: this process is the main subject of analysis in the first half of this thesis.

1.1 Coupling: art meets science

The term coupling, in the field of probability, refers to the practice of constructing two (or more) probability measures on a single measurable space in order to compare them. Usually this is performed in order to deduce properties of the individual measures, or to investigate similarities between the two. The two principal applications of coupling that will be exploited in this thesis are: bounding the rate of convergence to equilibrium of ergodic Markov chains; and proving stochastic domination statements.

Formally speaking, a coupling may be defined as follows (Lindvall 2002).

Definition 1.1. Let P and P' be two probability measures on a measurable space $(E, \mathcal{E})$. A *coupling of P and P'* is a probability measure $\hat{P}$ on $(E^2, \mathcal{E}^2)$ such that the marginals of $\hat{P}$ are P and P'. That is, if we define the natural projections π and π' by $\pi(x, x') = x$ and $\pi'(x, x') = x'$ for $(x, x') \in E^2$, then

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

Rather than simply working with probability measures, throughout this thesis we shall primarily be concerned with coupling random processes, such as Markov

chains. To that end, the following definition of the coupling of two random elements is more appropriate.

Definition 1.2. Let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, \mathcal{E})$ and $X' : (\Omega', \mathcal{F}', \mathbb{P}') \to (E, \mathcal{E})$ be $\mathcal{F}/\mathcal{E}$- and $\mathcal{F}'/\mathcal{E}$-measurable functions respectively. A *coupling of X and X'* is a measurable function $(\hat{X}, \hat{X}') : (\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}}) \to (E^2, \mathcal{E}^2)$ such that

$$\hat{X} \stackrel{\mathcal{D}}{=} X \quad \text{and} \quad \hat{X}' \stackrel{\mathcal{D}}{=} X' .$$

Thus $\hat{\mathbb{P}}(\hat{X}, \hat{X})^{-1}$ is simply a coupling of the probability measures $\mathbb{P}X^{-1}$ and $\mathbb{P}'X'^{-1}$.

When there is no danger of confusion, we shall normally say simply that (X, X') is a coupling, or that X and X' are *coupled*. (This is a slight abuse of notation, since in such cases we will strictly be working with a coupled version $(\hat{X}, \hat{X}')$ of (X, X'), but we shall be explicit about this if necessary.) Similarly, we shall also simply write $(\Omega, \mathcal{F}, \mathbb{P})$ for the joint space on which (X, X') is defined. It will be assumed throughout this thesis (unless explicitly stated to the contrary) that all state spaces are Polish (and so regular versions of conditional probability measures may be assumed to exist.)

Note that the existence of a coupling for any two such random elements is trivial: simply take X and X' to be independent of one another. This idea forms the basis of the first application of coupling (Doeblin 1938), but is commonly not of great use: the true power of the coupling method lies in the number of possibilities for the joint distribution of $(\hat{X}, \hat{X}')$: careful construction of this distribution can often be extremely informative about the marginals. Although coupling is a well-defined mathematical technique, and rigorous descriptions can be given using measure-theoretic language as above, the choice of an *informative* coupling is something of an art form as we shall see, and is often based on intuition.

Throughout this thesis, much use will be made of the total variation metric as a means of measuring the distance between two probability measures.

Definition 1.3. Let μ and ν be probability measures defined on a space E. Define the *total variation distance* between μ and ν by

$$\|\mu - \nu\| = \sup_{A \subset E} |\mu(A) - \nu(A)| = \frac{1}{2} \sup_{|h| \leq 1} \left| \int h \, d(\mu - \nu) \right| .$$

This metric assumes values in $[0, 1]$. If E is countable, then

$$\|\mu - \nu\| = \frac{1}{2} \sum_{x \in E} |\mu(x) - \nu(x)| \ .$$

There are a number of useful identities satisfied by this metric, the most convenient of which (for the work in later chapters) is the following:

Lemma 1.4. *Let $\mu \wedge \nu$ be the greatest common component of μ and ν, and let λ be a measure that dominates μ and ν. Write*

$$f = \frac{d\mu}{d\lambda}, \qquad g = \frac{d\nu}{d\lambda}.$$

Then the total variation metric satisfies the following equality:

$$\|\mu - \nu\| = 1 - \int (f \wedge g) \, d\lambda \,. \tag{1.1}$$

Proof. Define the set $C_\mu \subseteq E$ by $C_\mu = \{x \ : \ f(x) \geq g(x)\}$. Then

$$\begin{aligned} 2 \, \|\mu - \nu\| = \sup_{|h| \leq 1} \left| \int h \, d(\mu - \nu) \right| \quad &= \int_{C_\mu} d(\mu - \nu) - \int_{E \setminus C_\mu} d(\mu - \nu) \\ = \int |f - g| \ d\lambda \quad &= \int (f - f \wedge g) \, d\lambda + \int (g - f \wedge g) \, d\lambda \\ = 2 \left(1 - \int (f \wedge g) \, d\lambda \right) . \end{aligned}$$

□

Two other notions of distance between measures that are commonly used are those of ℓ^p and separation distance.

Definition 1.5. Let μ and ν be probability measures defined on a space E. Let λ, f and g be as in the statement of Lemma 1.4. For $1 \leq p \leq \infty$ the *ℓ^p-distance* between μ and ν is given by

$$\|\mu - \nu\|_p = \begin{cases} \left(\int |f - g|^p \ d\lambda \right)^{1/p} & \text{if } 1 \leq p < \infty \\ \operatorname{ess\,sup} |f - g| & \text{if } p = \infty. \end{cases} \tag{1.2}$$

Note that

$$\|\mu - \nu\|_1 = 2 \, \|\mu - \nu\|$$

by definition, and that Jensen's inequality yields:

$$\|\mu - \nu\|_p \leq \|\mu - \nu\|_q \,, \quad \text{for all } 1 \leq p \leq q \leq \infty.$$

Definition 1.6. Let μ and ν be probability measures defined on a countable space E. The *separation distance* between μ and ν is defined by

$$d_S(\mu, \nu) = \max_{x \in E} \left(1 - \frac{\mu(x)}{\nu(x)}\right).$$

Equivalently, $d_S(\mu, \nu)$ is the smallest $s \geq 0$ such that $\mu = (1-s)\nu + s\rho$ for some probability measure ρ. This distance takes values in $[0, 1]$ but is not a metric. It is simple to show (see for example Diaconis (1988)) that

$$\|\mu - \nu\| \leq d_S(\mu, \nu).$$

Now suppose that the random variables X and X' are coupled. The following inequality provides a useful upper bound on the total variation distance between the laws of X and X' on $(E, \mathcal{E})$.

Lemma 1.7. *Let the coupled random variables X and X' have laws μ and μ' respectively on $(E, \mathcal{E})$. Then*

$$\left\|\mu - \mu'\right\| \leq \mathbb{P}\left(X \neq X'\right). \tag{1.3}$$

Proof. For any set $A \in \mathcal{E}$,

$$\begin{aligned}
\mathbb{P}(X \in A) - \mathbb{P}\left(X' \in A\right) &= \mathbb{P}\left(X \in A,\ X = X'\right) + \mathbb{P}\left(X \in A,\ X \neq X'\right) \\
&\quad - \mathbb{P}\left(X' \in A,\ X = X'\right) - \mathbb{P}\left(X' \in A,\ X \neq X'\right) \\
&= \mathbb{P}\left(X \in A,\ X \neq X'\right) - \mathbb{P}\left(X' \in A,\ X \neq X'\right) \\
&\leq \mathbb{P}\left(X \neq X'\right).
\end{aligned}$$

The result follows by Definition 1.3. □

In Chapter 3 the existence of a coupling which shows that inequality (1.3) is sharp will be demonstrated.

Taking this idea a little further, suppose now that $X = \{X_n\}_0^\infty$ and $X' = \{X'_n\}_0^\infty$, where X_n and X'_n are elements in $(E, \mathcal{E})$. Let $(\hat{X}, \hat{X}')$ be a coupling of (X, X'), and suppose that T is a random time such that

$$\hat{X}_n = \hat{X}'_n \qquad \text{for } n \geq T.$$

Direct application of Lemma 1.7 (using the fact that $\{\hat{X}_n \neq \hat{X}'_n\} \subseteq \{T > n\}$) then yields:

Lemma 1.8 (The coupling inequality)**.** *Let $(\hat{X}, \hat{X}')$ be a coupling of X and X' as above. Then*

$$\left\|\mathbb{P}\left(X_n \in \cdot\right) - \mathbb{P}\left(X'_n \in \cdot\right)\right\| \leq \mathbb{P}\left(T > n\right) . \tag{1.4}$$

Such a random time T is called a *coupling time*. A 'good' coupling is usually one that has a 'small' coupling time. The coupling is called *successful* if

$$\mathbb{P}\left(T < \infty\right) = 1 .$$

Now, if X is a Markov chain on a countable space E, with transition kernel P, then X is said to be *weakly ergodic* if

$$\lim_{n\to\infty} \sum_{k\in E} |\delta_x P^n(k) - \delta_y P^n(k)| = 0 \quad \text{for all } x, y \in E .$$

(Here δ_x is the Dirac point mass at x.) That is, X is weakly ergodic if and only if the total variation distance at time n between the law of the chain started at x and that of the chain started at y converges to zero as $n \to \infty$, for all $x, y \in E$. Thus the coupling inequality, despite its simplicity, immediately leads to a very important application of coupling theory: if there exists a successful coupling of these two chains for any pair of starting states $x, y \in E$, then X is weakly ergodic. The existence of the reverse implication to this statement forms the basis of the work in Chapter 3.

Furthermore, if X and X' are two versions of an ergodic chain with stationary distribution π, such that $X_0 = x$ and $X'_0 \sim \pi$, then inequality (1.4) provides a bound on the rate at which X approaches equilibrium:

$$\|\mathbb{P}\left(X_n \in \cdot\right) - \pi(\cdot)\| \leq \mathbb{P}\left(T > n\right) .$$

Remark 1.9. An analogous result to Lemma 1.8 holds if X and X' are continuous-time processes in D_E, the space of càdlàg functions defined on $[0, \infty)$ with values in a Polish space E, endowed with the Skorokhod topology.

An important class of couplings is that of *co-adapted* couplings:

Definition 1.10. Let $(\hat{X}, \hat{X}')$ be a coupling of the two random processes X and X'. This coupling is called *co-adapted* if the processes $\hat{X}$ and $\hat{X}'$ have a common past expressed by a fixed filtration of σ-algebras.

Note that if X and X' are Markov processes then this condition implies that $\hat{X}$ and $\hat{X}'$ viewed separately are both Markov processes, but not that the joint process $(\hat{X}, \hat{X}')$ is Markovian. More informally, a co-adapted coupling is one for which the evolution of either process is not allowed to depend upon the future of the other. Co-adapted couplings tend to be easier to work with, but non-co-adapted couplings certainly have their uses: this topic will be further discussed in Chapter 3.

Finally, a more general notion in the theory of coupling is that of *distributional* coupling (called *weak* coupling in Lindvall (2002)). Let $X = \{X_n\}$ be a discrete time stochastic process on $(E, \mathcal{E})$. For non-negative integer-valued random variables T, let $\theta_T X$ be the shift operator defined by

$$\theta_T X = \begin{cases} \{X_{T+n}\}_{n=0}^{\infty} & \text{on } \{T < \infty\} \\ (x, x, \dots) & \text{on } \{T = \infty\}, \end{cases}$$

where x is a fixed element of E. Writing $\stackrel{\mathcal{D}}{=}$ to denote identity in distribution, we have the following definition:

Definition 1.11. Let X and X' be discrete time stochastic processes on $(E, \mathcal{E})$. We say that $(\hat{X}, \hat{X}')$ is a *distributional* coupling of X and X', with coupling times T and T', if

(a) $\hat{X} \stackrel{\mathcal{D}}{=} X$ and $\hat{X}' \stackrel{\mathcal{D}}{=} X'$;

(b) $(\theta_T \hat{X}, T) \stackrel{\mathcal{D}}{=} (\theta_{T'} \hat{X}', T')$.

An analogous definition holds for processes $X, X' \in D_E$.

Thus distributional coupling serves to weaken the requirement that $\hat{X}$ and $\hat{X}'$ eventually coincide, and asks instead that $\hat{X}$ behaves probabilistically from time T as $\hat{X}'$ does from time T'. Further discussion of distributional coupling can be found in Thorisson (2000): we shall briefly meet this coupling again in Chapter 3, where the notion of stitching together distributions at randomised stopping times is considered.

1.2 Random walks on finite groups

We have seen in the above that the coupling inequality may be used to bound the total variation distance between two Markov chains. In this section we consider a special class of chains: random walks on finite groups. This class contains many interesting processes, a large proportion of which are very well suited to the coupling method.

Given a finite group G, a random walk X may be defined on G by first defining a probability measure P on a generating subset $H \subseteq G$. Repeated independent draws from the distribution P yield random elements $h_1, h_2, \ldots \in H$. $X = \{X_k\}_{k \geq 0}$, started at some element $x \in G$, is then defined as follows:

$$X_0 = x; \qquad X_k = h_k X_{k-1} \,.$$

The transition kernel for this walk, denoted $P(\cdot, \cdot)$, satisfies

$$P(x, y) = P(yx^{-1}) \,,$$

and the distribution of X_k is given by

$$\mathbb{P}_x \left(X_k = g \right) = P^k(gx^{-1}) = \sum_{h \in H} P^{k-1}(hx^{-1}) P(gh^{-1}) = \sum_{h \in H} P^{k-1}(x, h) P(h, g) \,.$$

Since H generates G it follows that, if X is aperiodic, $P^k(\cdot)$ converges to the uniform distribution on G, $\mathit{Uniform}(G)$, as $k \to \infty$.

On many occasions it will be more convenient to work with a continuous-time version of the random walk, $X = \{X_t\}$. The kernel $P(\cdot, \cdot)$ has associated to it a continuous-time semigroup $\mathcal{P}^t = e^{-t(I-P)}$: this has kernel

$$\mathcal{P}^t(x, y) = e^{-t} \sum_{k=0}^{\infty} \frac{t^k P^k(x, y)}{k!} \,.$$

X may be realised by holding the walk in its present state for an $\mathit{Exp}(1)$ amount of time, and then making a transition according to the discrete-time kernel $P(\cdot, \cdot)$. For this reason, we will say that the kernel $P(\cdot, \cdot)$ (or the measure P) generates both the discrete and continuous time walks.

We now introduce a very important example of a Markov chain, which shall be the subject of much analysis in the first three chapters of this thesis.

Example 1.12 (Simple symmetric random walk on a hypercube). Let $\mathbb{Z}_2^n$ be the group of binary n-tuples under coordinate-wise addition modulo 2: this can be viewed as the vertices of a cube in n dimensions. $\mathbb{Z}_2^n$ is one of the simplest groups on which to study random walks, due to its high level of symmetry.

A random walk X on $\mathbb{Z}_2^n$ may be described as follows. For $x \in \mathbb{Z}_2^n$, write $x = (x(1), \ldots, x(n)) \in \{0,1\}^n$, and define elements $\{e_i\}_0^n$ by

$$e_0 = (0, 0, \ldots, 0); \qquad e_i(k) = \mathbf{1}_{[i=k]}, \quad i = 1, \ldots, n\,.$$

Consider then the probability measure P_n given by

$$P_n(e_i) = \frac{1}{n+1}, \quad i = 0, \ldots, n\,.$$

For $x, y \in \mathbb{Z}_2^n$ let

$$|x - y| = \sum_{i=1}^{n} |x(i) - y(i)|$$

be the Hamming distance between x and y. It follows that the transition kernel corresponding to P_n is given by

$$P_n(x,y) = \begin{cases} (n+1)^{-1} & \text{if } |x-y| \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

This kernel then describes an aperiodic, irreducible nearest-neighbour symmetric random walk on $\mathbb{Z}_2^n$: its state diagram is the Cayley graph of $\mathbb{Z}_2^n$ with generating set $H = \{e_i\}_0^n$, and its unique equilibrium distribution is $U_n = \mathit{Uniform}(\mathbb{Z}_2^n)$.

The continuous-time version of this walk can be realised as follows: flip coordinate i ($1 \le i \le n$) to its opposite value (zero or one) whenever there is an incident on Λ_i, where $\{\Lambda_i\}_1^n$ are a set of mutually independent Poisson processes, each of rate $1/n$.

Variations on this walk, produced by allowing $P_n(e_i)$ to depend on i, will be considered in Chapter 2.

This random walk was analysed in detail during the 1980s. The following examples describe two co-adapted coupling schemes for this chain. The first of these is based on a method (which can be applied to both the discrete and continuous-time chains) due to Aldous (1983).

Example 1.13 (Aldous coupling). At time $k \ge 0$, let

$$U_k = \left\{1 \le i \le n \;:\; X_k(i) \ne X'_k(i)\right\}$$

be the set of coordinates on which X_k and X'_k disagree, where X and X' are the two random walks to be coupled. At step $k+1$, draw $0 \leq i \leq n$ uniformly at random and set $X_{k+1} = X_k + e_i$. Then define X'_{k+1} as follows:

- if $|X_k - X'_k| > 1$:
  - if $i \notin U_k$ set $X'_{k+1} = X'_k + e_i$;
  - otherwise, choose a coordinate $j \in U_k \setminus \{i\}$ uniformly at random and set $X'_{k+1} = X'_k + e_j$;
- if $|X_k - X'_k| = 1$, with $U_k = \{j\}$:
  - if $X_k(i) = X'_k(i)$ set $X'_{k+1} = X'_k + e_i$;
  - if $i = 0$ set $X'_{k+1} = X'_k + e_j$;
  - if $i = j$ set $X'_{k+1} = X'_k + e_0$;
- finally, if $|X_k - X'_k| = 0$, set $X'_{k+1} = X'_k + e_i$.

This update scheme marginally updates X' according to the transition kernel P_n, and so this is a valid coupling. It may be described more simply as follows: if X_k and X'_k agree on coordinate i then this match is preserved at time $k+1$; if not, then another unmatched coordinate of X'_k (if it exists) is moved so as to decrease the number of unmatched coordinates by two; when (if) a single unmatched coordinate j remains, this coordinate is coupled as soon as $i \in \{0, j\}$.

The second coupling we shall consider here is defined for the continuous-time walks X and X'.

Example 1.14 (Partial-independence coupling). This coupling treats each coordinate of the n-tuple X separately, and is very simple to describe. For $1 \leq i \leq n$, let Λ_i and Λ'_i be independent Poisson processes of rate $1/n$. Coordinates $X(i)$ and $X'(i)$ are made to flip independently (at incident times of Λ_i and Λ'_i respectively) until the first time that they agree, after which they remain equal forever (with transitions driven by Λ_i, say). Due to the memoryless property of the Exponential distribution this defines a valid coupling. In order to distinguish this coupling from the trivial independence coupling (where X and X' are completely independent until they

agree on all coordinates), we shall refer to this scheme as the *partial-independence* coupling.

Both of these coupling schemes will prove useful in the next two chapters of this work. In the final section of this introductory chapter, we look at how the Aldous coupling may be used to bound the time taken for the walk to approach equilibrium. There are many other ways of analysing the convergence rate of random walks on groups, including eigenvalue analysis and group representation theory (Diaconis 1988): this latter technique can sometimes provide very tight bounds on the time taken to reach stationarity, especially in examples where the measure P driving the walk is constant on conjugacy classes of G. One final method of analysing convergence rates, which is more probabilistic in nature, is that of strong uniform times: this technique will also (briefly) be reviewed below.

1.2.1 Time to reach equilibrium

Definitions 1.3, 1.5 and 1.6 provide a number of ways of measuring the distance between two probability measures, and hence between the distribution of a random walk at any given time and its stationary distribution. In his paper of 1983, Aldous defined the following parameter, which measures the time until the random walk is within a fixed distance of equilibrium (with respect to the total variation metric).

Definition 1.15. Let X be a random walk on a group G, driven by a probability measure P, with equilibrium distribution $U = \mathit{Uniform}(G)$. We define

$$\tau^{mix}(\varepsilon) = \tau^{mix}(G, P, \varepsilon) = \inf \{t > 0 : \|\mathbb{P}(X_t \in \cdot) - U(\cdot)\| \leq \varepsilon\} , \tag{1.5}$$

and set

$$\tau^{mix} = \tau^{mix}(1/e) . \tag{1.6}$$

τ^{mix} will be referred to as the *mixing time* of X.

The choice of $1/e$ in this definition is fairly arbitrary, and is made primarily for algebraic convenience: with this choice it can be shown (see for example Aldous and Fill (2002)) that

$$\|\mathbb{P}(X_t \in \cdot) - U(\cdot)\| \leq \exp\left(-\left\lfloor t/\tau^{mix} \right\rfloor\right) .$$

We now use the coupling of Example 1.13 (applied to the continuous-time chain) to bound the mixing time for the symmetric random walk on $\mathbb{Z}_2^n$, following the calculation in Aldous (1983). Let

$$N_t = \left|X_t - X_t'\right|$$

count the number of coordinates on which the coupled chains X and X' disagree at time $t \geq 0$. The coupling scheme ensures that N is a decreasing process, and that the coupling time T is given by the time taken for N to be absorbed at zero. Now, the transition rates for N are given by

$$Q(k, k-2) = \frac{k}{n}, \quad \text{for } 2 \leq k \leq n; \qquad Q(1,0) = \frac{2}{n}\,.$$

Let $\{T_k\}$ be independent $Exp(k/n)$ random variables. Then the coupling time T is bounded above by

$$T^* = T_m + \cdots + T_5 + T_3 + T_1; \quad m = \begin{cases} n & n \text{ odd} \\ n-1 & n \text{ even.} \end{cases}$$

Since the $\{T_k\}$ are independent, it follows that

$$\begin{aligned} \mathbb{E}\left[T^*\right] &= n\left(\frac{1}{m} + \cdots + \frac{1}{5} + \frac{1}{3} + 1\right) \sim \frac{1}{2} n \log n \\ \mathrm{Var}\left(T^*\right) &= n^2\left(\frac{1}{m^2} + \cdots + \frac{1}{5^2} + \frac{1}{3^2} + 1\right) \sim Cn^2\,. \end{aligned}$$

Finally, for any $\alpha > 1/2$, the coupling inequality and Chebyshev's inequality between them yield

$$\begin{aligned} \left\| P_n^{\alpha n \log n} - U_n \right\| &\leq \mathbb{P}\left(T^* > \alpha n \log n\right) \\ &\leq \frac{\mathrm{Var}\left(T^*\right)}{\left((\alpha - 1/2) n \log n\right)^2} \xrightarrow[n\to\infty]{} 0\,. \end{aligned} \tag{1.7}$$

Therefore τ^{mix} (and indeed $\tau^{mix}(\varepsilon)$ for all $0 < \varepsilon < 1$) is asymptotically bounded above by $(n/2)\log n$.

An alternative method of bounding the mixing time is to use a strong uniform time. This is a particular type of randomised stopping time (first defined in Pitman and Speed (1973)).

Definition 1.16. A random variable T defined on $(\Omega, \mathcal{F})$ with values in the extended time set $\mathbb{N} \cup \{\infty\}$ is said to be a *randomised stopping time* (RST) of the Markov chain X, if for each $n \in \mathbb{N}$ the event $\{T > n\}$ is conditionally independent of the future $(X_{n+1}, X_{n+2}, \dots)$ given the past $(X_0, \dots, X_n)$.

Note that if for all n the event $\{T > n\}$ is completely determined by $(X_0, \dots, X_n)$ then T is simply a stopping time of X. A RST differs from a stopping time in general since the event $\{T > n\}$ is allowed to depend upon auxiliary randomness. As a simple example of a RST, let X be a real-valued Markov chain and suppose that V is a real-valued random variable independent of X. Then

$$T = \inf \{n \,:\, X_n \geq V\}$$

is a RST of X.

Now suppose that X is an ergodic random walk on the group G, with stationary distribution *Uniform*(G).

Definition 1.17. A *strong uniform time* (SUT) for X is a randomised stopping time T such that

$$\mathbb{P}\left(X_k = x \,|\, T = k\right) = \frac{1}{|G|}\,,$$

for all $x \in G$ and $k \geq 0$.

That is, X_T has distribution *Uniform*(G) and is independent of T. (This concept can of course be generalised to Markov chains on a more general state space: this leads to the definition of a strong *stationary* time.) Strong uniform times are useful for bounding the mixing time of a random walk due to the following result (Aldous and Diaconis 1987).

Proposition 1.18. *If T is a SUT for X then*

$$d_S\left(\mathcal{L}(X_n), \mathit{Uniform}(G)\right) \leq \mathbb{P}\left(T > n\right), \quad n \geq 0 \tag{1.8}$$

(where $\mathcal{L}(X)$ is the law of X). Conversely, there exists a SUT T such that inequality (1.8) *holds with equality.*

Inequality (1.8) shows that SUTs are to separation distance what coupling times are to total variation distance (recall inequality (1.4)). SUTs are connected to coupling via the following proposition.

Proposition 1.19 (Aldous and Diaconis (1987))**.** *Let T be a SUT for the Markov chain X with starting state x. Then there exists a coupling of X and the stationary chain Y with coupling time T.*

Proof. Given T and $\{X_n\}$, we construct Y as follows. On each non-null event $\{T = m\}$, define $Y_n^{(m)} = X_n$, for $n \geq m$. Conditional on $\{T = m\}$, the future process $\left\{Y_n^{(m)} : n \geq m\right\}$ is distributed as the stationary Markov chain (since T is a SUT), and so can be extended backwards to $\left\{Y_n^{(m)} : n \geq 0\right\}$ as the stationary chain. Finally, define $Y_n = Y_n^{(m)}$ on $\{T = m\}$, for each m. It follows that $\{Y_n\}$ is a stationary Markov chain, with $Y_n = X_n$ for all $n \geq T$. □

Note that this coupling is not co-adapted, since the chain $\{Y_n : n \geq 0\}$ depends upon the values of T and X_T.

SUTs also have a link to the distributional coupling of Definition 1.11 (due to Thorisson, but reported in Aldous and Diaconis (1987)). This follows from the observation that if T is a SUT of X, and Y is a stationary version of the chain, then

$$(T, X_T, X_{T+1}, \dots) \stackrel{\mathcal{D}}{=} (T, Y_T, Y_{T+1}, \dots) .$$

Thus X and Y form a distributional coupling, with coupling time T.

Finally, Aldous and Diaconis (1987) construct a simple SUT T for the symmetric random walk on $\mathbb{Z}_2^n$, and show that

$$\mathbb{P}\left(T \geq (n+1)(\log n + c)\right) \leq \frac{1}{(c-1)^2}, \quad c > 1 .$$

Thus the time taken to make the separation distance small is asymptotically bounded above by $n \log n$ for this walk. Note that this is twice the bound on τ^{mix} obtained above using coupling.

1.3 Summary

This introductory chapter has summarised some of the general coupling theory that underlies the work in this thesis. Thorisson (2000) and Lindvall (2002) contain a wealth of information on the coupling method, while a more comprehensive study of distances between probability measures can be found in Gibbs and Su (2002). Readers interested in the fascinating topic of random walks on groups (which will

really only be touched upon in this thesis) are advised to consult any of the following references: Aldous (1983), Diaconis (1988), Aldous and Fill (2002), Saloff-Coste (2004).

The layout of the rest of this thesis is as follows. Chapter 2 investigates the partial-independence coupling of Example 1.14 for a generalised version of the random walk on $\mathbb{Z}_2^n$, proving the existence under a simple condition of a new type of threshold behaviour called a *coupling-cutoff*. Chapter 3 is concerned with the theory of *maximal* couplings of Markov chains. This idea is generalised to maximal coalescent couplings, and an explicit description of an optimal co-adapted coupling for the symmetric random walk on $\mathbb{Z}_2^n$ is presented.

The tone of the thesis changes at this point, as Chapters 4 and 5 deal with the subject of perfect simulation. The first of these chapters provides an introduction to Coupling from the Past and associated algorithms: this review is based upon the encyclopedia article of the author (Connor 2007). It concludes with a summary of the paper by Kendall (2004), which proves the existence of a perfect simulation algorithm for geometrically ergodic Markov chains. Chapter 5 is based upon the article of Connor and Kendall (2007), which extends this result to a class of positive recurrent chains.

All generalizations are dangerous, even this one.

Alexandre Dumas

2. THE COUPLING-CUTOFF PHENOMENON

A natural question in the study of random walks on groups is the following: how many steps of the walk are *necessary* for the chain to be within a small distance of its stationary distribution? Classical theory says that asymptotically equilibrium is approached exponentially, with the rate governed by the second-largest eigenvalue (in modulus) of the transition matrix P (see, for example, Rosenthal (1995a)). However, such an answer only gives a bound (which is often very conservative) on how many steps are sufficient, and does not provide any quantitative description of how the distance from stationarity changes over the relatively short term.

Initial interest in the above problem was sparked by the question: how many shuffles of a pack of cards is necessary for the deck to be "well-shuffled"? Of course, an answer to such a question depends on the definition of a 'shuffle' and on the precise meaning of 'well-shuffled'. Card shuffling questions can be re-phrased in terms of random walks on the symmetric group S_n (the group of permutations of n elements) with the uniform distribution U_n as the unique equilibrium distribution. 'Well-shuffled' can then be defined in terms of the distance to stationarity (usually measured using one of the distances introduced in Section 1.1).

The first random walk for which a conclusive result was obtained was that of the transposition shuffle. This walk evolves in discrete time by choosing two positions $1 \leq i, j \leq n$ uniformly at random (with replacement) and swapping the cards in positions i and j. (Equivalently, the cards labelled i and j may be swapped.) In the notation of Section 1.2, the random walk X is generated by the following distribution P_n:

$$P_n(id) = \frac{1}{n}\,, \quad P_n(\sigma) = \frac{2}{n^2} \quad \text{for all transpositions } \sigma \in S_n.$$

This random walk was studied by Diaconis and Shahshahani (1981). They proved the remarkable result that $(n/2)\log n$ steps are both necessary and sufficient for X to be close to uniform when n is large.

Theorem 2.1 (Diaconis and Shahshahani 1981)**.** *Let $\tau_n = (n/2)\log n$. For $c > 0$,*

$$\left\| P_n^{\tau_n + cn} - U_n \right\| \leq a e^{-2c}$$

for a universal constant a. Conversely, for $c < 0$, as n tends to infinity,

$$\left\| P_n^{\tau_n + cn} - U_n \right\| \geq \frac{1}{e} - e^{-e^{-2c}} + o(1)\,.$$

(Here, and throughout this thesis, we write $f_n = o(g_n)$ if $f_n/g_n \to 0$ as $n \to \infty$, and $f_n = O(g_n)$ if f_n/g_n is eventually bounded.) Such behaviour is called a *cutoff phenomenon*: the total variation distance from stationarity stays close to its maximum value of one for a time, before dropping rapidly towards zero. This type of result really concerns a sequence of random walks: for the transposition shuffle this sequence is $\{S_n, P_n\}$, and Theorem 2.1 states that convergence takes place asymptotically in a window of length $O(n)$ centred at time $(n/2)\log n$.

More formally, the following definitions are commonly used in the setting of continuous-time random walks (see, for example, Diaconis and Saloff-Coste (2006), Chen (2006)).

Definition 2.2. For $n \geq 1$, let P_n be a probability measure on a finite group G_n, such that the continuous-time random walk $X^{(n)}$ generated by P_n (recall Section 1.2) has stationary distribution π_n. We say that the sequence $\{G_n, P_n, \pi_n\}_1^\infty$ exhibits:

(1) a *τ_n-pre-cutoff* if there exist $0 < a < b$ and a sequence of positive numbers $\{\tau_n\}_1^\infty$ such that

$$\liminf_{n\to\infty} \|\mathcal{P}_n^{a\tau_n} - \pi_n\| > 0\,, \quad \text{and} \quad \lim_{n\to\infty} \left\|\mathcal{P}_n^{b\tau_n} - \pi_n\right\| = 0\,;$$

(2) a *τ_n-cutoff* if there exists a sequence of positive numbers $\{\tau_n\}_1^\infty$ such that

$$\begin{aligned} \forall c \in (0,1), \quad & \lim_{n\to\infty} \|\mathcal{P}_n^{c\tau_n} - \pi_n\| = 1 \\ \text{and } \forall c > 1, \quad & \lim_{n\to\infty} \|\mathcal{P}_n^{c\tau_n} - \pi_n\| = 0\,; \end{aligned}$$

(3) a *(τ_n, b_n)-cutoff* if $\tau_n, b_n > 0$ satisfy $b_n = o(\tau_n)$ and

$$\begin{aligned} d_-(c) = \liminf_{n\to\infty} \left\|\mathcal{P}_n^{\tau_n + cb_n} - \pi_n\right\| \quad & \text{satisfies } \lim_{c\to-\infty} d_-(c) = 1\,, \\ d_+(c) = \limsup_{n\to\infty} \left\|\mathcal{P}_n^{\tau_n + cb_n} - \pi_n\right\| \quad & \text{satisfies } \lim_{c\to\infty} d_+(c) = 0\,. \end{aligned}$$

The three parts of this definition are given in increasing order of precision: (3) implies (2) implies (1). The weakest statement (that a pre-cutoff exists) simply means that the convergence to equilibrium takes place over a period of time bounded by $(a\tau_n, b\tau_n]$. The sequence $\{\tau_n\}$ will be called the cutoff time in cases (2) and (3), with the exact type of cutoff being made explicit when necessary. We will often informally say simply that the sequence of random walks $\left\{X^{(n)}\right\}$ has a cutoff at time τ_n. In case (3) the term b_n will be referred to as the *window* of the cutoff. Note that two sequences $\{\tau_n\}$ and $\{\tau'_n\}$ are both cutoff times for a sequence $\{G_n, P_n, \pi_n\}$ if and only if $\tau_n \sim \tau'_n$ as n tends to infinity (Chen 2006).

Recall from Definition 1.15 that $\tau^{mix}(G, P) = \tau^{mix}(G, P, 1/e)$ is the time taken for the random walk on G with transitions driven by P to be within $1/e$ of its stationary distribution (with respect to total variation). Thus if $\{G_n, P_n, \pi_n\}$ is a sequence of walks which exhibits a τ_n-cutoff, it follows that for any $\varepsilon \in (0, 1)$,

$$\tau^{mix}(G_n, P_n, \varepsilon) \sim \tau^{mix}(G_n, P_n) \sim \tau_n \quad \text{as } n \to \infty.$$

In what follows we shall write τ_n^{mix} for $\tau^{mix}(G_n, P_n, 1/e)$. Thus for a sequence presenting a cutoff, it is always possible to take τ_n to be the mixing time τ_n^{mix}: in this case therefore, there is strong motivation for saying that the mixing time is 'the time taken to reach equilibrium'.

An analogous set of definitions exist of course for discrete-time random walks. If a set of continuous chains $\left\{X^{(n)}\right\}$ exhibits a τ_n-cutoff with $\tau_n \to \infty$ then the discrete-time family of walks will display a cutoff at the same instant (Chen 2006). This will be the case for all the walks we are interested in below, most of which will have a uniform stationary distribution, $\mathit{Uniform}(G_n)$: this will usually be written simply as U_n.

There is also no need to restrict attention to random walks on groups when studying cutoffs, nor to use total variation to measure distance from stationarity. Work in this latter setting has been carried out very recently by Chen (2006), who uses eigen-analysis to prove the existence of ℓ^p-cutoffs for a large number of sequences of Markov chains. In this thesis we are principally concerned with total variation distance (due to its link with coupling via the coupling inequality, Lemma 1.8). It should be remarked though, that mixing times and cutoff phenomena are of course

specific to the distance by which convergence is measured. More information on the cutoff phenomenon can be found in any of the following definitive references, from which much of this introductory material is taken: Diaconis (1988), Diaconis (1996), Saloff-Coste (1997), Saloff-Coste (2004).

The most famous example of a cutoff comes from analysis of the riffle shuffle: this is the shuffle commonly used by card players, whereby the deck is cut into two piles which are then merged into one whilst maintaining the relative order of the cards in each pile. A mathematical model for this shuffle was proposed by Gilbert and Shannon (Gilbert 1955), and later, independently, by Reeds (1981). Following earlier work by Aldous (1983), Bayer and Diaconis (1992) proved that for the Gilbert-Shannon-Reeds riffle shuffle, with $\tau_n = (3/2)\log_2 n$,

$$\left\| P_n^{\tau_n + c} - U_n \right\| = 1 - 2\Phi\left(\frac{-2^{-c}}{4\sqrt{3}} \right) + O\left(n^{-1/2} \right).$$

Thus this random walk exhibits a $(\tau_n, 1)$-cutoff. This result made national newspaper headlines when it was first published: it shows that it takes about 7 riffle shuffles to adequately randomise a standard deck of 52 playing cards.

The great interest in the cutoff phenomenon arises from the fact that such behaviour has been shown to occur for a number of random walks on groups (see Figure 2.1). It is certainly not true, however, that all random walks on groups exhibit this behaviour: the walk on $\mathbb{Z}/n\mathbb{Z}$ driven by the uniform measure on $\{-1, 0, 1\}$ does not present a cutoff (see Diaconis (1988)). Despite this interest, there is as yet a limited amount of theory for predicting exactly when a cutoff will exist (although a few heuristic arguments have been put forward - see the discussions in Diaconis (1996) and Ycart (1999)). One recent conjecture, due to Yuval Peres, is that a

G_n	Walk	τ_n	Reference
S_n	Top-to-random	$n \log n$	Aldous and Diaconis (1986)
S_n	Random transpositions	$\frac{1}{2} n \log n$	Diaconis and Shahshahani (1981)
S_n	Riffle shuffle	$\frac{3}{2} \log_2 n$	Bayer and Diaconis (1992)
$\mathbb{Z}_2^n$	Symmetric random walk	$\frac{1}{4} n \log n$	Aldous (1983)

Fig. 2.1: Table of some known total variation cutoff phenomena for random walks on groups.

necessary and sufficient condition for a set of Markov chains to exhibit a cutoff is that

$$\tau_n^d(1/4)\lambda_n \to \infty \quad \text{as } n \to \infty,$$

where λ_n is the spectral gap of P_n (i.e. 1 minus the second largest eigenvalue), and τ_n^d is the mixing time measured by the distance d. (If d is the total variation metric then $\tau_n^d(\varepsilon) = \tau_n^{mix}(\varepsilon)$.) This conjecture is proved when d is the ℓ^p-distance ($1 < p \leq \infty$) by Chen (2006), and in the case of birth-death chains when d is separation distance by Diaconis and Saloff-Coste (2006). There do exist counterexamples to this conjecture however, due to Aldous (Chen 2006) and Peres (P. Diaconis, personal communication).

Of particular interest to us is the fact that the simple random walk on $\mathbb{Z}_2^n$, introduced in Example 1.12, is known to exhibit a $((n/4)\log n, n)$-cutoff. This was first shown in Aldous (1983), with the following refined version being proved by Diaconis and Shahshahani (1987).

Theorem 2.3 (Diaconis and Shahshahani (1987))**.** *For the simple random walk on $\mathbb{Z}_2^n$, with $\tau_n = \frac{n+1}{4}\log n$, the following statements hold:*

$$\left\|P_n^{\tau_n+cn} - U_n\right\|^2 \leq \frac{1}{2}\left(e^{e^{-4c}} - 1\right);$$

and as $n \to \infty$, for any $\varepsilon > 0$ there exists $C < 0$ such that for $c < C$,

$$\left\|P_n^{\tau_n+cn} - U_n\right\| \geq 1 - \varepsilon\,.$$

The upper bound of this theorem is proved using group representation theory. The lower bound is obtained by a probabilistic argument. The upper bound of this cutoff has also been proved using coupling and strong uniform times (Matthews 1987). Chen (2006) shows that this random walk exhibits a $((n/4)\log n, n)$-ℓ^p-cutoff for all $1 \leq p \leq 2$, and also a $((n/2)\log n, n)$-ℓ^∞-cutoff.

The result of Theorem 2.3 was made even more precise in the paper of Diaconis et al. (1990). They analysed the 'shape' of the cutoff: that is, how $\|P_n^{\tau_n+cn} - U_n\|$ behaves as a function of c. They proved the following:

Theorem 2.4 (Diaconis et al. (1990))**.** *For the simple random walk on $\mathbb{Z}_2^n$, let $\tau_n = (n/4)\log n$. Then for fixed $c \in (-\infty, \infty)$, as $n \to \infty$,*

$$\left\|P_n^{\tau_n+cn} - U_n\right\| \sim \mathrm{Erf}\left(\frac{e^{-2c}}{\sqrt{8}}\right), \tag{2.1}$$

where

$$\mathrm{Erf}\,(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt$$

denotes the error function.

This theorem provides a far greater level of detail about the convergence to stationarity around the cutoff time than is available for most other chains. Such a detailed analysis is possible thanks to the facts that $\mathbb{Z}_2^n$ is abelian and benefits from a highly symmetric structure.

2.1 Coupling-cutoffs for random walks on $\mathbb{Z}_2^n$

The discussion at the start of this chapter concerns the behaviour of the distance from stationarity of a sequence of random walks. In the results stated above, this distance is measured using the total variation metric, for which the simple random walk on $\mathbb{Z}_2^n$ exhibits a $((n/4)\log n, n)$-cutoff. The question that we now ask is: when does a coupling time for two such walks exhibit similar cutoff behaviour? This depends upon the coupling of course! For the continuous-time symmetric random walk on the hypercube, one coupling strategy encountered in Chapter 1 is particularly accessible to analysis: this is the partial-independence coupling of Example 1.14. In this chapter we investigate the behaviour of this coupling when n is large. The analysis is then extended to random walks where each coordinate may move at a different rate.

To define what is meant by 'similar cutoff behaviour', let us introduce a little notation. Let $X^{(n)}$ and $Y^{(n)}$ be continuous-time random walks on a group G_n, each generated by the probability measure P_n. Let T_n be a coupling time for $X^{(n)}$ and $Y^{(n)}$. For $t \geq 0$, define

$$F_n(t) = \mathbb{P}\,(T_n \leq t) \tag{2.2}$$

to be the distribution function of T_n. It is then possible to define the following three types of behaviour:

Definition 2.5. For $n \geq 1$, let T_n and F_n be defined as above. We say that the sequence $\{G_n, P_n, T_n\}_1^\infty$ (or simply the sequence $\left\{X^{(n)}\right\}$, when it is clear what coupling strategy is being used) exhibits:

(1) a τ_n-*coupling-pre-cutoff* if there exist $0 < a < b$ and a sequence of positive numbers $\{\tau_n\}_1^\infty$ such that

$$\limsup_{n\to\infty} F_n(a\tau_n) < 1\,, \quad \text{and} \quad \lim_{n\to\infty} F_n(b\tau_n) = 1\,;$$

(2) a τ_n-*coupling-cutoff* if there exists a sequence of positive numbers $\{\tau_n\}_1^\infty$ such that

$$\begin{aligned} \forall c \in (0,1), \quad & \lim_{n\to\infty} F_n(c\tau_n) = 0 \\ \text{and } \forall c > 1, \quad & \lim_{n\to\infty} F_n(c\tau_n) = 1\,; \end{aligned}$$

(3) a (τ_n, b_n)-*coupling-cutoff* if $\tau_n, b_n > 0$ satisfy $b_n = o(\tau_n)$ and

$$\begin{aligned} F_+(c) = \limsup_{n\to\infty} F_n(\tau_n + cb_n) \quad &\text{satisfies } \lim_{c\to-\infty} F_+(c) = 0\,, \\ F_-(c) = \liminf_{n\to\infty} F_n(\tau_n + cb_n) \quad &\text{satisfies } \lim_{c\to\infty} F_-(c) = 1\,. \end{aligned}$$

As with total variation cutoffs (Definition 2.2), a coupling-cutoff for discrete-time walks may be defined in the obvious way. Note that in this definition no restrictions are placed upon the sequence $\{G_n, P_n, T_n\}$: this is in keeping with the general definition of the cutoff phenomenon. Although this sequence *will* have a natural structure in all of the examples considered below, no such structure is demanded in general: the coupling-cutoff phenomenon is a new area of research and we do not wish to restrict ourselves to working with a particular set of sequences at this early stage.

Recall from Example 1.12 that the continuous-time walk on $\mathbb{Z}_2^n$ evolves as follows: coordinate i ($1 \le i \le n$) flips to its opposite value (zero or one) whenever there is an incident on Λ_i, where $\{\Lambda_i\}_1^n$ are a set of mutually independent Poisson processes, each of rate $1/n$. In what follows these rates will be allowed to differ between processes Λ_i and Λ_j: in general the rate of Λ_i will be denoted by λ_i. Unless otherwise stated, it will not be assumed that the rates $\{\lambda_i\}$ are normalised.

The partial-independence coupling for two such walks allows unmatched coordinates to evolve independently until the time that they first agree, whereafter they move synchronously. Let $X^{(n)}$ and $Y^{(n)}$ be two random walks coupled in this way, with $X_0^{(n)}$ equal to some fixed state and $Y_0^{(n)} \sim U_n$. If $X_0^{(n)}$ and $Y_0^{(n)}$ do not agree on the i^{th} coordinate (which happens with probability $1/2$), then the time taken for

agreement on this coordinate is equal to the time of the first incident on a Poisson process of rate $2\lambda_i$. This yields

$$\mathbb{P}\left(X_t^{(n)}(i) = Y_t^{(n)}(i)\right) = \frac{1}{2} + \frac{1}{2}\left(1 - e^{-2t\lambda_i}\right) = 1 - \frac{1}{2}\,e^{-2t\lambda_i}\,.$$

Therefore, if T_n is the partial-independence coupling time,

$$F_n(t) = \mathbb{P}\left(T_n \leq t\right) = \prod_{i=1}^{n} \mathbb{P}\left(X_t^{(n)}(i) = Y_t^{(n)}(i)\right) = \prod_{i=1}^{n}\left(1 - \frac{1}{2}\,e^{-2t\lambda_i}\right). \tag{2.3}$$

(Note that this is independent of the choice of $X_0^{(n)}$.)

For the rest of this chapter, T_n will always be the partial-independence coupling time. Since the coupling strategy is consistent throughout, we will simply say that the sequence $\left\{X^{(n)}\right\}$ does/does not exhibit a coupling-cutoff.

2.1.1 A simple example: the symmetric random walk

We begin this investigation into coupling-cutoffs with two results concerning the *symmetric* random walk, when $\lambda_i = 1/n$ for all i.

Proposition 2.6. *Suppose that the rates $\{\lambda_i\}$ are normalised, so that $\sum \lambda_i = 1$. Then $F_n(t)$ is maximised for all $t \geq 0$ when $\lambda_i = 1/n$ for all $i = 1, \ldots, n$.*

Proof. Recall the classical inequality relating geometric and arithmetic means (Hardy et al. 1952): for any set of real non-negative numbers $a_1, \ldots, a_n$,

$$\left(\prod_{i=1}^{n} a_i\right)^{1/n} \leq \frac{1}{n}\sum_{i=1}^{n} a_i, \tag{2.4}$$

with equality if and only if all the a_i's are equal. Therefore,

$$\begin{aligned} F_n(t) = \prod_{i=1}^{n}\left(1 - \frac{1}{2}\,e^{-2t\lambda_i}\right) \quad &\leq \left(\frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{1}{2}\,e^{-2t\lambda_i}\right)\right)^n \\ = \left(1 - \frac{1}{2n}\sum_{i=1}^{n} e^{-2t\lambda_i}\right)^n &\leq \left(1 - \frac{1}{2}\left(\prod_{i=1}^{n} e^{-2t\lambda_i}\right)^{\frac{1}{n}}\right)^n \\ = \left(1 - \frac{1}{2}\,e^{-\frac{2t}{n}}\right)^n\,, & \end{aligned}$$

where both inequalities follow by application of inequality (2.4). □

The following result shows that the partial-independence coupling time for the symmetric random walk is $(n/2)\log n$: Proposition 2.6 shows that this is therefore a lower bound on the expected coupling time for a random walk driven by any set of normalised rates $\{\lambda_i\}$.

Proposition 2.7. *The random walk with $\lambda_i = 1/n$ for all $i = 1, \ldots, n$ exhibits a $((n/2)\log n, n)$-coupling-cutoff.*

Proof. Define $\tau_n = (n/2)\log n$ and $b_n = n$. From equation (2.3) it follows easily that

$$\begin{aligned} F_n(\tau_n + cb_n) &= \left(1 - \frac{1}{2}\exp\left(-\frac{2}{n}(\tau_n + cb_n)\right)\right)^n \\ &= \left(1 - \frac{1}{2}\frac{e^{-2c}}{n}\right)^n \\ &\xrightarrow[n\to\infty]{} \exp\left(-\frac{1}{2}e^{-2c}\right). \end{aligned}$$

Thus

$$F_+(c) = F_-(c) = \exp\left(-\frac{1}{2}e^{-2c}\right)$$

for this random walk, and Definition 2.5(3) is satisfied. □

This simple example demonstrates very clearly what is meant by a coupling-cutoff. Figure 2.2 shows plots of $F_n(c\tau_n)$ (with $\tau_n = (n/2)\log n$) as a function of c for a range of n. It is evident that the transition from zero to one, when the time axis is scaled by τ_n in this way, takes place over a shorter interval as n increases: the sequence of functions $\{F_n(c\tau_n)\}$ converges to a step function (with unit jump at $c = 1$) as $n \to \infty$.

Note the difference between the shape of the coupling-cutoff around τ_n and that of the true cutoff (around $\tau_n/2$). Theorem 2.4 showed that the change in total

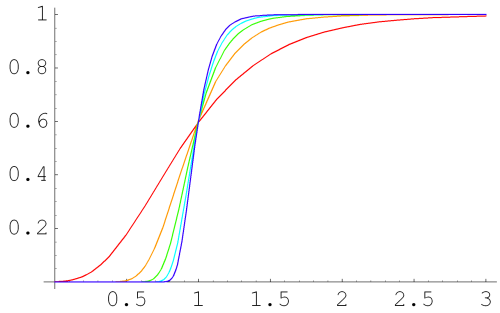


Fig. 2.2: Plots of the function $F_n(c\tau_n)$ over the range $c \in (0, 3)$ for n =10 (red), 10^2, 10^3, 10^4, 10^5 (blue).

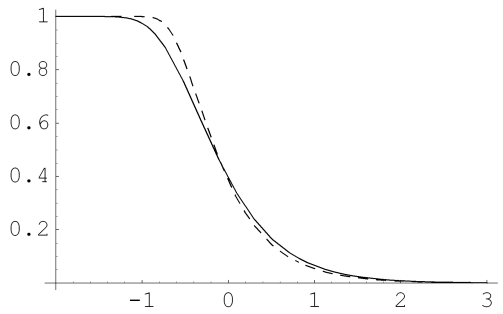


Fig. 2.3: 'Shape', as a function of c, of the coupling-cutoff and the total variation cutoff about their critical times: solid line shows $1 - \exp\left(-e^{-2c}/2\right)$ (coupling-cutoff); dashed line shows $\mathrm{Erf}\left(e^{-2c}/\sqrt{8}\right)$ (total variation cutoff).

variation distance behaves like an error function around $\tau_n/2$, whereas the proof of Proposition 2.7 shows that the distribution of the coupling time behaves like an extreme value function. The difference between these two types of behaviour is shown in Figure 2.3: it is evident that the true cutoff is slightly sharper than the coupling-cutoff.

2.1.2 Breaking the symmetry

The relatively simple example of the symmetric random walk indicates the potential for gain from studying coupling-cutoffs. Although a coupling-cutoff only gives an upper bound on the true mixing time of a chain (via the coupling inequality), in some cases the coupling construction is much easier to study. This is certainly the case when the equality of the λ_i's is broken. Indeed, equation (2.3) remains true no matter what values of $\lambda_i > 0$ are used, whereas there is no longer a simple expression for the total variation distance between $\mathcal{L}\left(X^{(n)}\right)$ and U_n when the rates are not identical (see, for example, Barrera et al. (2006)).

Diaconis (1996) considered the specific example where $\lambda_i \propto i^{-\alpha}$ for $\alpha \geq 0$, with $\sum \lambda_i = 1$. He proved that a (τ_n, b_n)-cutoff is exhibited for all such α, with the cutoff parameters as given in Figure 2.4. This result, sketched in Diaconis (1996), is based on eigen-analysis. It shows that a cutoff persists when symmetry is broken, but is unsatisfactory in that it relies on the sequence of rates decreasing in a very specific way. In particular, it does not take into account the commutativity of $\mathbb{Z}_2^n$ and the fact that the convergence behaviour is independent of coordinate permutation.

The result was recently improved upon by Barrera et al. (2006). They consider

	τ_n	b_n
$\alpha = 0$	$\frac{1}{4} n \log n$	n
$0 < \alpha < 1$	$\frac{1}{4(1-\alpha)} n(\log n - \log\log n)$	n
$\alpha = 1$	$\frac{1}{4} n \log n(\log n - \log\log n)$	$n \log n$
$1 < \alpha < \infty$	$\frac{\zeta(\alpha)}{4} n^{\alpha}(\log n - \log\log n)$	n^{α}

Fig. 2.4: Cutoff time τ_n and window b_n for the random walk on $\mathbb{Z}_2^n$ with rates $\lambda_i \propto i^{-\alpha}$. Here $\zeta(\alpha) = \sum_{s=1}^{\infty} s^{-\alpha}$ is the Riemann zeta function.

n-tuples of independent, exponentially converging processes. This class of processes includes, but is not limited to, the random walks on $\mathbb{Z}_2^n$ considered in this chapter. They give conditions under which the n-tuple exhibits a cutoff, with the convergence to equilibrium being measured by a number of distances including total variation. The following theorem states the restriction of their result to random walks on $\mathbb{Z}_2^n$.

Theorem 2.8 (Barrera et al. (2006))**.** *For $n \geq 1$, denote by $\lambda_{(1,n)}, \lambda_{(2,n)}, \ldots, \lambda_{(n,n)}$ the values of $\lambda_1, \ldots, \lambda_n$ ranked in increasing order. Define*

$$\tau_n = \max_{1 \leq i \leq n} \left\{ \frac{\log i}{4\lambda_{(i,n)}} \right\}. \tag{2.5}$$

If

$$\lim_{n \to \infty} \lambda_{(1,n)} \tau_n = \infty$$

then the random walk on $\mathbb{Z}_2^n$ with rates $\{\lambda_i\}$ exhibits a τ_n-cutoff.

It has recently come to our attention that the thesis of Chen (2006) shows that the value of τ_n in equation (2.5) is also the ℓ^2-cutoff time for this random walk. Furthermore, since this chain is reversible, Corollary 2.1 of Chen (2006) implies that the walk exhibits an ℓ^p-cutoff for all $1 < p \leq \infty$, with the ℓ^p-mixing time being of the same order as the ℓ^2-mixing time.

The fact that a total variation cutoff was shown to hold in more generality than the symmetric case by Diaconis (1996) prompted the present investigation into coupling-cutoffs. (During the process of this investigation the results of Barrera et al. (2006) and Chen (2006) were published.) The following two simple examples show

that a coupling-cutoff can hold when the rates λ_i are not identical, and provide some motivation for the results of Section 2.1.3.

Example 2.9. Consider a random walk on $\mathbb{Z}_2^n$ with $\sqrt{n}$ rates equal to 1 and the remaining rates equal to 3. Theorem 2.8 shows that the walk exhibits a cutoff at

$$\tau_n = \max\left\{\frac{\log\sqrt{n}}{4}, \frac{\log n}{12}\right\} = \frac{\log n}{8}\,.$$

Now consider the coupling time distribution:

$$F_n\left(\frac{\log n}{4} + c\right) = \left(1 - \frac{1}{2}\frac{e^{-2c}}{\sqrt{n}}\right)^{\sqrt{n}}\left(1 - \frac{1}{2}\frac{e^{-6c}}{n^{3/2}}\right)^{n-\sqrt{n}}$$
$$\xrightarrow[n\to\infty]{} \exp\left(-\frac{1}{2}e^{-2c}\right)\,.$$

Since this final expression tends to zero as $c \to -\infty$ and to one as $c \to \infty$, this random walk exhibits a $((1/4)\log n, 1)$-coupling-cutoff.

Example 2.10. For a second example, consider the rates

$$\lambda_i = \begin{cases} 1 & i = 1 \\ \log\log n & i \geq 2\,. \end{cases}$$

By Theorem 2.8, this walk has a cutoff at time

$$\tau_n = \frac{\log n}{4\log\log n}\,.$$

Furthermore, it also exhibits a $(\log n/(2\log\log n), 1/\log\log n)$-coupling-cutoff, since

$$F_n\left(\frac{\log n}{2\log\log n} + \frac{c}{\log\log n}\right) = \left(1 - \frac{1}{2}\exp\left(\frac{-(\log n + 2c)}{\log\log n}\right)\right)\left(1 - \frac{1}{2}\frac{e^{-2c}}{n}\right)^{n-1}$$
$$\xrightarrow[n\to\infty]{} \exp\left(-\frac{1}{2}e^{-2c}\right)\,.$$

These examples show that a coupling-cutoff is exhibited in at least two cases where a cutoff is known to occur, with the coupling-cutoff time being twice that of the total variation cutoff. Example 2.10 shows that a coupling-cutoff can occur even when most of the rates tend to infinity with n. In the next section, we investigate general conditions under which a coupling-cutoff occurs for a random walk on the hypercube. As remarked above, it is desirable to approach the problem from a perspective which takes into account the commutativity of $\mathbb{Z}_2^n$. We therefore choose to work with discrete probability measures μ_n on $[1, \infty)$, rather than a set of rates $\{\lambda_i\}$: the result of this will be that the existence of a coupling-cutoff is directly related to the convergence of appropriately scaled versions of $\{\mu_n\}$ as n tends to infinity.

2.1.3 A measure-based approach

Let $\{\mu_n\}$ be a sequence of probability measures on $[1, \infty)$, where μ_n is the sum of n point masses, each of weight $1/n$, whose locations may or may not be distinct. We shall assume throughout that μ_n has been scaled so that $\mu_n(\{1\}) \geq 1/n$ for all n.

Consider a random walk $X^{(n)}$ on $\mathbb{Z}_2^n$, with rates governed by the measure μ_n. The partial-independence coupling time distribution for this walk satisfies the natural generalisation of equation (2.3):

$$F_n(t) = \exp\left(n \int_1^\infty \log\left(1 - \frac{1}{2}e^{-2t\lambda}\right) \mu_n(d\lambda)\right) . \tag{2.6}$$

We now find conditions on $\{\mu_n\}$ such that the sequence $\left\{X^{(n)}\right\}$ of random walks exhibits coupling-cutoff behaviour.

Remark 2.11. The assumption that $\mu_n(\{1\}) > 0$ is not restrictive. If the sequence $\{\mu_n\}$ instead satisfies

$$\sigma(n) = \inf_{\lambda \geq 1} \{\mu_n[1, \lambda] > 0\} > 1$$

for some n, then it suffices to study the measures $\{\hat{\mu}_n\}$, where

$$\hat{\mu}_n(\{x\}) = \mu_n\left(\{x/\sigma(n)\}\right) .$$

From equation (2.6) it then follows by a simple change of variables that if $\{\hat{\mu}_n\}$ exhibits a $\hat{\tau}_n$-coupling-cutoff, the sequence $\{\mu_n\}$ will exhibit a coupling-cutoff at time

$$\tau_n = \frac{\hat{\tau}_n}{\sigma(n)} .$$

The following proposition provides a sufficient condition for the chains $\left\{X^{(n)}\right\}$ to exhibit a coupling-pre-cutoff.

Proposition 2.12. *For $n \geq 1$ and $\varepsilon > 0$, let $\lambda_n(\varepsilon)$ be defined by*

$$\lambda_n(\varepsilon) = \inf\{\lambda \geq 1 \,:\, \mu_n[1, \lambda] \geq \varepsilon\} .$$

If there exists $\varepsilon > 0$ such that $\lambda_n(\varepsilon)$ is eventually bounded above, then the sequence $\left\{X^{(n)}\right\}$ exhibits a $(\log n)/2$-coupling-pre-cutoff.

Proof. Let $\tau_n = (\log n)/2$. Since μ_n is supported on $[1, \infty)$, it follows that the coupling time T_n is stochastically dominated by the coupling time for the walk with all rates equal to 1. Thus, for any fixed $b > 1$,

$$F_n(b\tau_n) \geq \left(1 - \frac{1}{2}e^{-2b\tau_n}\right)^n = \left(1 - \frac{n^{-b}}{2}\right)^n ,$$

and so $F_n(b\tau_n) \to 1$ as $n \to \infty$.

Now let $\varepsilon > 0$ be such that $\lambda_n(\varepsilon) \leq C < \infty$ for all large n. Then F_n may be bounded above as follows:

$$\begin{aligned} F_n(t) &\leq \exp\left(n \int_1^C \log\left(1 - \frac{1}{2}e^{-2t\lambda}\right) \mu_n(d\lambda)\right) \\ &\leq \exp\left(n\mu_n[1, C] \log\left(1 - \frac{1}{2}e^{-2tC}\right)\right) \\ &\leq \left(1 - \frac{1}{2}e^{-2tC}\right)^{n\varepsilon} . \end{aligned}$$

Therefore, for fixed $b > 1$,

$$F_n\left(\frac{\tau_n}{bC}\right) \leq \left(1 - \frac{n^{-1/b}}{2}\right)^{n\varepsilon} \xrightarrow[n\to\infty]{} 0 .$$

Hence Definition 2.5(1) is satisfied for any $b > 1$, with $a = (bC)^{-1} > 0$, and so these chains exhibit a $(\log n)/2$-coupling-pre-cutoff. □

The sufficient condition of this proposition is rather restrictive: Example 2.10 shows that both a total variation and coupling-cutoff can exist even when μ_n converges vaguely to the zero measure on $[1, \infty)$.

Given a measure μ_n, define τ_n by

$$\tau_n = \max_{\lambda \geq 1} \left\{ \frac{\log(n\mu_n[1, \lambda])}{2\lambda} \right\} = \frac{\log(n\mu_n[1, \lambda_n^*])}{2\lambda_n^*} , \tag{2.7}$$

where $\lambda_n^* \in [1, \infty)$ is defined by this last equality. (If there are two or more values of λ achieving the maximum in equation (2.7) then we shall (arbitrarily) always take λ_n^* to be the minimum of these values.) Note the similarity between this definition and that given in equation (2.5) for the total variation cutoff time.

Given λ_n^*, we may define a new measure ν_n on $(0, \infty)$ as follows:

$$\nu_n(\{x\}) = \frac{\mu_n(\{\lambda_n^* x\})}{\mu_n[1, \lambda_n^*]} . \tag{2.8}$$

This measure has total mass $(\mu_n[1,\lambda_n^*])^{-1} \in [1,\infty)$ and satisfies $\nu_n(0,1] = 1$. The idea behind this scaling is as follows. λ_n^* describes in some sense the 'critical point' of μ_n: it will be shown that under a certain condition, any mass μ_n places to the left of λ_n^* will not influence the coupling-cutoff time. For ease of notation we define

$$\beta_n = n\mu_n[1,\lambda_n^*] \in [1,n]\,.$$

Lemma 2.13. *If $\beta_n \to \infty$ as $n \to \infty$ then $\nu_n(0,1] \xrightarrow{w} \delta_1$.*

Proof. By definition of τ_n (equation (2.7)),

$$\frac{\log\left(n\mu_n[1,\lambda]\right)}{\lambda} \leq \frac{\log\beta_n}{\lambda_n^*} \quad \text{for all } \lambda \geq 1.$$

Thus for all $x \geq 1/\lambda_n^*$,

$$\frac{\log\left(n\mu_n[1,x\lambda_n^*]\right)}{x} \leq \log\beta_n\,.$$

This yields

$$n\mu_n[1,x\lambda_n^*] \leq \beta_n^x \quad \text{for all } x \geq 1/\lambda_n^*. \tag{2.9}$$

Hence

$$\nu_n(0,x] = \frac{\mu_n[1,x\lambda_n^*]}{\mu_n[1,\lambda_n^*]} = \frac{n\mu_n[1,x\lambda_n^*]}{\beta_n} \leq \beta_n^{x-1}\,, \tag{2.10}$$

where the inequality follows from (2.9). Thus for all $\varepsilon \in (0,1)$,

$$\nu_n(0,1-\varepsilon] \leq \beta_n^{-\varepsilon} \xrightarrow[n\to\infty]{} 0$$

because $\beta_n \to \infty$ as $n \to \infty$. Since $\nu_n(0,1] = 1$ for all n, this proves the required convergence. □

This makes more precise what is meant by λ_n^* describing the 'critical point' of μ_n. Under the assumption that $\beta_n \to \infty$, the measures ν_n converge weakly to δ_1 on $(0,1]$: this is exactly the sort of behaviour to be expected if the sequence $\{\lambda_n^*\}$ captures information about the coupling-cutoff time. Theorem 2.15 makes this observation exact: its proof makes use of a simple inequality, which for ease of reference is produced here as a proposition.

Proposition 2.14. *For $0 \leq x \leq 1/2$,*

$$-x - x^2 \leq \log(1-x) \leq -x\,. \tag{2.11}$$

Proof. Observe that for $0 \leq u \leq 1/2$,

$$-(1+2u) \leq \frac{-1}{1-u} \leq -1\,.$$

Integrating each term in this sequence of inequalities with respect to u, over the range $0 \leq u \leq x \leq 1/2$, completes the proof. □

Theorem 2.15. *The sequence of random walks $\left\{X^{(n)}\right\}$ exhibits a τ_n-coupling-cutoff if and only if $\tau_n \to \infty$, where τ_n is defined in equation* (2.7).

Proof. First suppose that $\tau_n \not\to \infty$. The standing assumption concerning the measures $\{\mu_n\}$ is that $\mu_n(\{1\}) \geq 1/n$ for all n. Thus, for fixed $c > 0$, it follows from equation (2.6) that

$$F_n(c\tau_n) \leq \exp\left(n\mu_n(\{1\})\log\left(1-\frac{1}{2}e^{-2c\tau_n}\right)\right) \leq 1-\frac{1}{2}e^{-2c\tau_n}\,.$$

But this value is bounded away from one for all $c > 0$ if $\tau_n \not\to \infty$. Therefore $\tau_n \to \infty$ is a necessary condition for a τ_n-coupling-cutoff to exist.

Now suppose that $\tau_n \to \infty$: this implies that $\beta_n \to \infty$, since $\lambda_n^* \geq 1$. Using the measure ν_n the coupling time distribution F_n may be rewritten as follows:

$$F_n(t) = \exp\left(\beta_n \int_{1/\lambda_n^*}^{\infty} \log\left(1-\frac{1}{2}e^{-2\lambda_n^* t\lambda}\right)\nu_n(d\lambda)\right)\,. \tag{2.12}$$

For $t \in \mathbb{R}$ define

$$\theta_n(t) = \beta_n \int_{1/\lambda_n^*}^{\infty} \exp\left(-2\lambda_n^* t\lambda\right)\nu_n(d\lambda)\,. \tag{2.13}$$

This expression may be used to bound the distribution function F_n as follows. Application of Proposition 2.14 to the log term in equation (2.12) shows that, for all $t \geq 0$, the following inequalities hold:

$$-\frac{1}{2}\theta_n(t) - \frac{1}{4}\theta_n(2t) \leq \log F_n(t) \leq -\frac{1}{2}\theta_n(t)\,. \tag{2.14}$$

Furthermore, for $t \geq 0$, $\theta_n(2t) \leq \theta_n(t)$ by definition, and so

$$-\frac{3}{4}\theta_n(t) \leq \log F_n(t) \leq -\frac{1}{2}\theta_n(t)\,. \tag{2.15}$$

Thus the behaviour of F_n is determined by that of θ_n.

Now consider $\theta_n(c\tau_n)$, for fixed $c > 0$. By definition of τ_n,

$$\begin{aligned}\theta_n(c\tau_n) &= \beta_n \int_{1/\lambda_n^*}^{\infty} \exp\left(-2\lambda_n^* c \left[\frac{\log \beta_n}{2\lambda_n^*}\right] \lambda\right) \nu_n(d\lambda) \\ &= \int_{1/\lambda_n^*}^{\infty} \beta_n^{1-c\lambda} \nu_n(d\lambda)\,. \end{aligned} \tag{2.16}$$

We now search for bounds on θ_n.

Firstly, for $c \in (0,1)$,

$$\begin{aligned}\theta_n(c\tau_n) &\geq \int_{1/\lambda_n^*}^{1} \beta_n^{1-c\lambda} \nu_n(d\lambda) \\ &\geq \beta_n^{1-c} \nu_n(0,1] = \beta_n^{1-c}\,. \end{aligned} \tag{2.17}$$

For $c > 1$, integration by parts yields:

$$\theta_n(c\tau_n) = \left[\beta_n^{1-c\lambda} \nu_n(0,\lambda]\right]_{1/\lambda_n^*}^{\infty} - \log\left(\beta_n^{-c}\right) \int_{1/\lambda_n^*}^{\infty} \beta_n^{1-c\lambda} \nu_n(0,\lambda] d\lambda\,. \tag{2.18}$$

The first term of this expression is non-positive, and so using inequality (2.10), we see that

$$\begin{aligned}\theta_n(c\tau_n) &\leq c \log \beta_n \int_{1/\lambda_n^*}^{\infty} \beta_n^{1-c\lambda} \nu_n(0,\lambda] d\lambda \\ &\leq c \log \beta_n \int_{1/\lambda_n^*}^{\infty} \beta_n^{1-c\lambda} \beta_n^{\lambda-1} d\lambda \\ &= c \log \beta_n \left[-\frac{\beta_n^{-(c-1)\lambda}}{\log\left(\beta_n^{c-1}\right)}\right]_{1/\lambda_n^*}^{\infty} \\ &= \left(\frac{c}{c-1}\right) \beta_n^{-(c-1)/\lambda_n^*}\,. \end{aligned} \tag{2.19}$$

Inequalities (2.14) and (2.17) together show that for $c \in (0,1)$,

$$F_n(c\tau_n) \leq \exp\left(-\frac{1}{2}\beta_n^{1-c}\right) \xrightarrow[n\to\infty]{} 0\,,$$

since $\beta_n \to \infty$ by assumption. Furthermore, combining inequalities (2.14) and (2.19) for $c > 1$ yields

$$F_n(c\tau_n) \geq \exp\left(-\frac{3}{4}\left(\frac{c}{c-1}\right)\beta_n^{-(c-1)/\lambda_n^*}\right) \xrightarrow[n\to\infty]{} 1\,.$$

Thus there is a coupling-cutoff at time τ_n, as claimed. □

The coupling time of Proposition 2.7 may be obtained as a special case of Theorem 2.15. Let $\tilde{\mu}_n = \delta_{1/n}$ be the measure corresponding to the case where $\lambda_i = 1/n$ for all $i = 1, \ldots, n$. Similarly, let $\mu_n = \delta_1$ for all n. Then

$$\tau_n = \max_{\lambda \geq 1} \left\{\frac{\log\left(n\mu_n[1,\lambda]\right)}{2\lambda}\right\} = \frac{\log n}{2}\,,$$

with $\lambda_n^* = 1$ for all n. By Theorem 2.15, the random walk generated by $\{\mu_n\}$ exhibits a τ_n-coupling-cutoff. This then implies that the walk generated by $\{\tilde{\mu}_n\}$ exhibits a $n\tau_n$-coupling-cutoff, as proved in Proposition 2.7.

The result of Theorem 2.15 provides a coupling version of Theorem 2.8, with the coupling-cutoff being a factor of two out from the total variation cutoff time. The form of τ_n in equation (2.7) shows that $\lambda_n^* = o(\log n)$ is a necessary condition for $\tau_n \to \infty$ (and hence for a coupling-cutoff to exist). Thus although a coupling-cutoff can exist even if $\mu_n \xrightarrow{v} 0$, this will not be the case if the vague convergence takes place too quickly. In other words, the position of the 'critical mass' $\mu_n[1, \lambda_n^*]$ determining the cutoff cannot be allowed to escape to infinity as fast as $\log n$.

Theorem 2.15 also has a nice interpretation in terms of 'mini-cutoffs' (this observation was made in the case of total variation cutoffs in Barrera et al. (2006)). Note that the distribution of the i^{th} coordinate of $X^{(n)}$ at time t satisfies

$$\mathbb{P}\left(X_t^{(n)}(i) = X_0^{(n)}(i)\right) = \frac{1}{2}\left(1 + e^{-2\lambda_i t}\right).$$

It follows that the total variation distance between $X_t^{(n)}(i)$ and its stationary distribution is equal to

$$\frac{1}{2}e^{-2\lambda_i t}.$$

Thus $\mu_n[1, \lambda]$ is the proportion of coordinates of $X^{(n)}$ which converge slower than at rate $e^{-2\lambda t}$. If this proportion is sufficiently large (so that $n\mu_n[1, \lambda] \to \infty$) then this sub-tuple of coordinates will not converge before time

$$\frac{\log\left(n\mu_n[1, \lambda]\right)}{2\lambda}.$$

Theorem 2.15 shows that this will in fact be the coupling-cutoff time for $X^{(n)}$, so long as it is the latest convergence time of all such sub-tuples.

2.1.4 Window size calculations

Although Theorems 2.15 and 2.8 provide precise values of the (coupling-) cutoff times, neither result gives any information about the size of the cutoff window: information which was easy to obtain in Examples 2.9 and 2.10. Chen (2006) does provide some results for ℓ^2-cutoff window sizes for a general class of Markov chains: however,

these involve careful partition of eigenvalues into subsets of the real line, and seem less intuitive than the measure-based approach considered here for coupling-cutoffs.

Recall from Definition 2.5(3) that the sequence $X^{(n)}$ is said to exhibit a (τ_n, b_n)-coupling-cutoff if $b_n = o(\tau_n)$ and

$$F_+(c) = \limsup_{n\to\infty} F_n(\tau_n + cb_n) \quad \text{satisfies } \lim_{c\to-\infty} F_+(c) = 0\,, \tag{2.20}$$

$$F_-(c) = \liminf_{n\to\infty} F_n(\tau_n + cb_n) \quad \text{satisfies } \lim_{c\to\infty} F_-(c) = 1\,. \tag{2.21}$$

Note that this definition specifies the asymptotic behaviour of the mixing time, but gives no details about the distribution function F_n at time $\tau_n + cb_n$, nor about the functions $F_\pm$. All discussion so far has spoken of 'the cutoff window' b_n, when it is clear that if the sequence $\{G_n, P_n, T_n\}$ exhibits a (τ_n, b_n)-coupling-cutoff then it also exhibits a (τ_n, b'_n)-coupling-cutoff, where $\{b'_n\}$ is any sequence satisfying $O(b_n) \le b'_n = o(\tau_n)$. Chen (2006) distinguishes between different window sizes by defining three types of optimality for the sequence $\{b_n\}$. This distinction is more than is needed in what follows, for which a single definition of optimality will suffice:

Definition 2.16. Suppose the sequence $\{G_n, P_n, T_n\}$ exhibits a (τ_n, b_n)-coupling-cutoff. The window b_n will be said to be *optimal* if, whenever $\{G_n, P_n, T_n\}$ also exhibits a (τ_n, b'_n)-coupling-cutoff, $b_n \le O(b'_n)$.

As an example of an optimal window, consider the simple symmetric random walk on $\mathbb{Z}_2^n$, with all rates equal to $1/n$. Proposition 2.7 showed that this walk exhibits a $((n/2)\log n, n)$-coupling-cutoff. Consideration of the proof of this result shows that a $((n/2)\log n, b_n)$-coupling-cutoff does not exist for any sequence $b_n < O(n)$, and so this is indeed an optimal window.

It is possible to further analyse the optimality of the window by considering separately the windows either side of the cutoff time τ_n. That is, instead of using a single sequence $\{b_n\}$ to establish convergence in equations (2.20) and (2.21), we can consider each convergence statement separately:

Definition 2.17. Suppose the sequence $\{G_n, P_n, T_n\}$ exhibits a τ_n-coupling-cutoff. If there exists a sequence $\{b_n^L\}$ with $b_n^L = o(\tau_n)$, such that

$$F_+^L(c) = \limsup_{n\to\infty} F_n(\tau_n + cb_n^L) \quad \text{satisfies } \lim_{c\to-\infty} F_+^L(c) = 0,$$

then b_n^L will be called a *left-window* of the coupling-cutoff.

Similarly, if there exists a sequence $\{b_n^R\}$ with $b_n^R = o(\tau_n)$, such that

$$F_-^R(c) = \liminf_{n\to\infty} F_n(\tau_n + cb_n^R) \quad \text{satisfies } \lim_{c\to\infty} F_-^R(c) = 1,$$

then b_n^R will be called a *right-window* of the coupling-cutoff.

Remark 2.18. Note that if b_n^L and b_n^R satisfy Definition 2.17, then the sequence $\{G_n, P_n, T_n\}$ exhibits a (τ_n, b_n)-coupling-cutoff, where

$$b_n = \max\left\{b_n^L, b_n^R\right\} .$$

With these final general definitions in place, we now return to the analysis of the partial-independence coupling for random walks on $\mathbb{Z}_2^n$. The following two results provide general upper bounds on the optimal values of $\{b_n^L\}$ and $\{b_n^R\}$ for this walk, both of which are determined by the sequence $\{\lambda_n^*\}$.

Lemma 2.19. *Suppose there is a coupling-cutoff at time τ_n, with τ_n defined by equation* (2.7)*. Then the optimal left-window of the coupling-cutoff is bounded above by $1/\lambda_n^*$.*

Note that, since $\lambda_n^* \geq 1$, this result shows that the optimal left-window is actually bounded above by a constant.

Proof. Recall from equation (2.13) the definition of θ_n:

$$\theta_n(t) = \beta_n \int_{1/\lambda_n^*}^{\infty} \exp\left(-2\lambda_n^* t\lambda\right) \nu_n(d\lambda) .$$

Now consider $\theta_n(\tau_n + c/\lambda_n^*)$, for fixed $c \in \mathbb{R}$. Note that $\tau_n \to \infty$ by Theorem 2.15, and so for any $c \in \mathbb{R}$ it follows that $\tau_n + c/\lambda_n^* \geq 0$ for large enough n. By definition of τ_n, with $\tau_n + c/\lambda_n^* \geq 0$:

$$\begin{aligned}
\theta_n(\tau_n + c/\lambda_n^*) &= \beta_n \int_{1/\lambda_n^*}^{\infty} \exp\left(-2\lambda_n^* \left[\tau_n + c/\lambda_n^*\right] \lambda\right) \nu_n(d\lambda) \\
&\geq \beta_n \int_{1/\lambda_n^*}^{1} \exp\left(-2\lambda_n^* \left[\tau_n + c/\lambda_n^*\right] \lambda\right) \nu_n(d\lambda) \\
&\geq \beta_n \nu_n(0,1] \left(\frac{e^{-2c}}{\beta_n}\right) = e^{-2c} . \qquad (2.22)
\end{aligned}$$

Combining inequalities (2.15) and (2.22) shows that for all $c \in \mathbb{R}$, with $b_n^L = 1/\lambda_n^*$:

$$\begin{aligned} F_+^L(c) &= \limsup_{n\to\infty} F_n(\tau_n + c/\lambda_n^*) \\ &\leq \limsup_{n\to\infty} \exp\left(-\frac{1}{2}\theta_n(\tau_n + c/\lambda_n^*)\right) \\ &\leq \exp\left(-\frac{1}{2}e^{-2c}\right). \end{aligned} \tag{2.23}$$

Thus

$$\lim_{c\to-\infty} F_+^L(c) = 0\,,$$

and so $1/\lambda_n^*$ is a left-window for the coupling-cutoff. □

This result shows that $b_n^L = 1$ is a left-window for the simple symmetric random walk with rates all equal to one. Thus the optimal left-window for the random walk with rates all equal to $1/n$ is bounded above by n (and is actually equal to n by the discussion following Definition 2.16).

Lemma 2.19 provides a (perhaps surprisingly) small bound on the optimal size of the left-window. However, it turns out that the general upper bound for the optimal right-window of the coupling-cutoff is significantly larger than that for the left. This result, stated as Theorem 2.21, makes use of the Lambert W-function. This is the function defined by

$$W(x)e^{W(x)} = x\,.$$

(For details of this function, see Corless et al. (1996).) The asymptotic behaviour of $W(x)$ for large x is described in Proposition 2.20.

Proposition 2.20. *For $x \geq e$, the function W satisfies*

$$\log x - \log\log x \leq W(x) \leq \log x - \log\left(\log x - \log\log x\right). \tag{2.24}$$

In particular,

$$W(x) \sim \log x - \log\log x \quad \text{as } x \to \infty. \tag{2.25}$$

Proof. By definition,

$$W(x) = \log x - \log W(x) \tag{2.26}$$

$$= \log x - \log\left(\log x - \log W(x)\right). \tag{2.27}$$

Now simply observe that $W(x) \leq \log x$ for $x \geq e$, and insert this bound into equations (2.26) and (2.27). □

Figure 2.5 shows a graph of W with the bounds of inequality (2.24) superimposed.

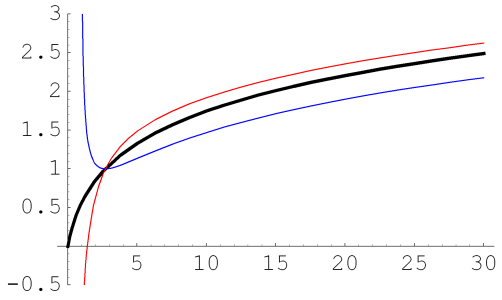


Fig. 2.5: Graph of the function W (black). For $x \geq e$, $W(x)$ is bounded below by $\log x - \log\log x$ (blue) and above by $\log x - \log(\log x - \log\log x)$ (blue).

We are now ready to state the final theorem of this section:

Theorem 2.21. *Suppose there is a coupling-cutoff at time τ_n, with τ_n defined by equation (2.7). Then the optimal right-window of the coupling-cutoff is bounded above by $W(\tau_n)$.*

Proof. In order for b_n^R to be a right-window for the coupling-cutoff, it is sufficient to show that $\theta_n\left(\tau_n + cb_n^R\right) \leq g(c)$ for sufficiently large n, where $g(c) \to 0$ as $c \to \infty$. For then, using inequality (2.15) it follows that

$$\begin{aligned} F_-^R(c) &= \liminf_{n\to\infty} F_n(\tau_n + cb_n^R) \\ &\geq \liminf_{n\to\infty} \exp\left(-\frac{3}{4}\theta_n\left(\tau_n + cb_n^R\right)\right) \\ &\geq \exp\left(-\frac{3}{4}g(c)\right) \xrightarrow[c\to\infty]{} 1\,. \end{aligned}$$

We therefore search for an upper bound on the function $\theta_n\left(\tau_n + cb_n^R\right)$ for fixed $c > 0$. Integration by parts, as in equation (2.18), yields the following:

$$\begin{aligned} \theta_n\left(\tau_n + cb_n^R\right) &= \beta_n \int_{1/\lambda_n^*}^{\infty} \left(\frac{e^{-2cb_n^R\lambda_n^*}}{\beta_n}\right)^{\lambda} \nu_n(d\lambda) \\ &= \beta_n \left[\left(\frac{e^{-2cb_n^R\lambda_n^*}}{\beta_n}\right)^{\lambda} \nu_n(0,\lambda]\right]_{1/\lambda_n^*}^{\infty} \\ &\quad + \beta_n \log(\beta_n e^{2cb_n^R\lambda_n^*}) \int_{1/\lambda_n^*}^{\infty} \left(\frac{e^{-2cb_n^R\lambda_n^*}}{\beta_n}\right)^{\lambda} \nu_n(0,\lambda]\, d\lambda\,. \end{aligned} \tag{2.28}$$

Now, for $c > 0$, this first term is negative for all n. Discarding this term, and using

inequality (2.10) to bound $\nu_n(0,\lambda]$ in the second term, we see that

$$\begin{aligned}\theta_n\left(\tau_n + cb_n^R\right) &\leq \beta_n \log(\beta_n e^{2cb_n^R\lambda_n^*}) \int_{1/\lambda_n^*}^{\infty} \left(\frac{e^{-2cb_n^R\lambda_n^*}}{\beta_n}\right)^{\lambda} \beta_n^{\lambda-1}\, d\lambda \\ &= \log(\beta_n e^{2cb_n^R\lambda_n^*}) \int_{1/\lambda_n^*}^{\infty} e^{-2cb_n^R\lambda_n^*\lambda}\, d\lambda \\ &= \log(\beta_n e^{2cb_n^R\lambda_n^*}) \frac{e^{-2cb_n^R}}{2cb_n^R\lambda_n^*} \\ &= e^{-2cb_n^R}\left(\frac{\tau_n}{cb_n^R} + 1\right).\end{aligned}$$

This upper bound blows up as n tends to infinity unless $b_n^R > O(1)$. Assuming that $b_n^R \to \infty$ therefore, with $b_n^R = o(\tau_n)$ (which necessarily holds for any window by definition), it follows that for large enough n:

$$\theta_n\left(\tau_n + cb_n^R\right) \leq e^{-c}\left(e^{-cb_n^R}\frac{\tau_n}{cb_n^R}\right) \tag{2.29}$$

for any fixed $c > 0$.

Now, by definition of Lambert's W-function, it follows that the right hand side of inequality (2.29) tends to infinity as $n \to \infty$ unless $cb_n^R \geq W(\tau_n)$. Thus, with $b_n^R = kW(\tau_n)$ for some constant $k > 0$, the right hand side of inequality (2.29) satisfies

$$e^{-c}\left(e^{-cb_n^R}\frac{\tau_n}{cb_n^R}\right) \xrightarrow[n\to\infty]{} \begin{cases}\infty & c < k^{-1} \\ e^{-c} & c = k^{-1} \\ 0 & c > k^{-1}\end{cases}.$$

It follows that for $c > k^{-1}$, $\theta_n(\tau_n + ckW(\tau_n)) \to 0$ as $n \to \infty$, and so

$$F_-^R(c) = \liminf_{n\to\infty} F_n\left(\tau_n + ckW(\tau_n)\right) = 1\,.$$

Therefore $b_n^R = W(\tau_n)$ is a right-window of the coupling-cutoff, as claimed. □

This bound on the right-window is significantly larger than that for the left-window. Since τ_n necessarily tends to infinity when a coupling-cutoff is exhibited, it follows that the right-window $W(\tau_n)$ also tends to infinity, whereas the left-window was shown to be $O(1)$. Application of Theorem 2.21 to the symmetric random walk with rates all equal to 1 shows that $b_n^R = W(\log n)$ is a bound on the optimal right-window. However, as mentioned after Definition 2.16, the optimal size of both left and right-windows is constant for this walk. Thus the bound arising from Theorem 2.21 is very conservative in this case. This is also true for the random walks

presented in Examples 2.9 and 2.10, where the optimal window size is equal to $1/\lambda_n^*$ in both cases.

However, the following example shows that the bound of Theorem 2.21 can be achieved, and so cannot be improved upon in general.

Example 2.22. Consider the random walk on $\mathbb{Z}_2^n$ governed by the probability measures

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(2\log_n(i)\vee 1)}\,.$$

(Here and throughout, for $a, b \in \mathbb{R}$, we write $a\vee b = \max\{a, b\}$ and $a\wedge b = \min\{a, b\}$.) The measure μ_n places all its mass in the interval $[1, 2]$, with

$$\mu_n[1,\lambda] = \frac{\lfloor n^{\lambda/2}\rfloor}{n} \sim n^{\lambda/2-1}, \quad \text{for all } \lambda \in [1,2] \text{ (see Figure 2.6).}$$

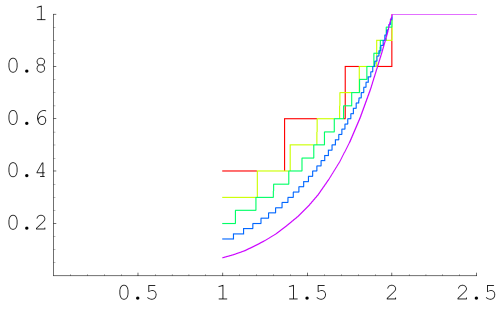


Fig. 2.6: Graph of the distribution functions of the measures μ_n for $n = 5$ (red), 10, 20, 50, 200 (purple).

For this sequence,

$$\tau_n = \max_{1\le\lambda\le 2}\left\{\frac{\log\left(n\mu_n[1,\lambda]\right)}{2\lambda}\right\} = \max_{1\le\lambda\le 2}\left\{\frac{\log\left(n^{\lambda/2}\right)}{2\lambda}\right\} = \frac{\log n}{4}\,.$$

Note that this maximum is attained at all $\lambda \in [1, 2]$: as usual we take $\lambda_n^* = 1$ to be the minimum of these values. This gives $\beta_n = \sqrt{n}$, and hence $\nu_n[1,\lambda] = \sqrt{n}\,\mu_n[1,\lambda]$. Since $\tau_n \to \infty$ as $n \to \infty$, this random walk exhibits a τ_n-coupling-cutoff.

Now, by Lemma 2.19, the optimal left-window of this coupling-cutoff is bounded above by $1/\lambda_n^*$, which is of course constant in this example. Figure 2.7 shows plots of the function

$$\exp\left(-\frac{1}{2}\theta_n\left(\frac{\log n}{4} + c\right)\right) \tag{2.30}$$

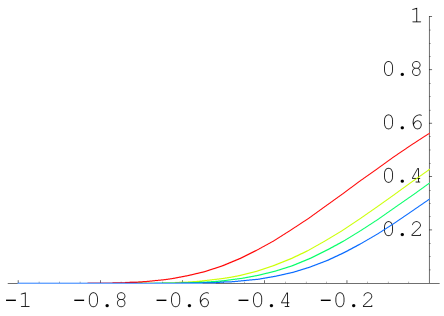


Fig. 2.7: Demonstration of the convergence of $F_n(\tau_n + c)$ to zero for fixed $c < 0$. Graphs of the function in (2.30) are shown for $n = 10$ (red), 30, 50, 100 (blue).

as a function of $c < 0$, for a range of values of n. It is apparent that this function, which is an upper bound for $F_n(\tau_n+c)$ for fixed $c < 0$ (by inequality (2.15)), converges to zero as $n \to \infty$.

However, the following calculations show that the optimal right-window for this example is given by $W(\tau_n)$. Consider once again the function $\theta_n(\tau_n + cb_n^R)$, for fixed $c > 0$ and some sequence $b_n^R = o(\tau_n)$. Since $\nu_n[1, \lambda] \sim n^{(\lambda-1)/2}$, integration by parts as in equation (2.28) yields the following:

$$\begin{aligned}
\theta_n\left(\frac{\log n}{4} + cb_n^R\right) &= \beta_n \left[\left(\frac{e^{-2cb_n^R}}{\beta_n}\right)^{\lambda} \nu_n[1,\lambda]\right]_1^2 \\
&\quad + \beta_n \log(\beta_n e^{2cb_n^R}) \int_1^2 \left(\frac{e^{-2cb_n^R}}{\beta_n}\right)^{\lambda} \nu_n[1,\lambda]\, d\lambda \\
&\sim \left(e^{-4cb_n^R} - e^{-2cb_n^R}\right) + \sqrt{n}\log(\sqrt{n}\, e^{2cb_n^R}) \int_1^2 \left(\frac{e^{-2cb_n^R}}{\sqrt{n}}\right)^{\lambda} n^{(\lambda-1)/2}\, d\lambda \\
&= \left(e^{-4cb_n^R} - e^{-2cb_n^R}\right) + \log(\sqrt{n}\, e^{2cb_n^R}) \int_1^2 e^{-2cb_n^R\lambda}\, d\lambda \\
&= \log(\sqrt{n}) \left(\frac{e^{-2cb_n^R} - e^{-4cb_n^R}}{2cb_n^R}\right) = \tau_n \left(\frac{e^{-2cb_n^R} - e^{-4cb_n^R}}{cb_n^R}\right).
\end{aligned}$$

It follows that the right-window of this coupling-cutoff cannot be bounded above by a constant, since $\theta_n(\tau_n + c) \to \infty$ as $n \to \infty$, and so

$$\limsup_{n\to\infty} F_n(\tau_n + c) \le \exp\left(-\frac{1}{2}\theta_n\left(\frac{\log n}{4} + c\right)\right) \xrightarrow[n\to\infty]{} 0\,.$$

Indeed, as in the proof of Theorem 2.21, any sequence $b_n^R = o(W(\tau_n))$ will force $\theta_n\left(\tau_n + cb_n^R\right) \to \infty$. It follows that

$$b_n^R = W(\tau_n) \sim \log\log n - \log\log\log n$$

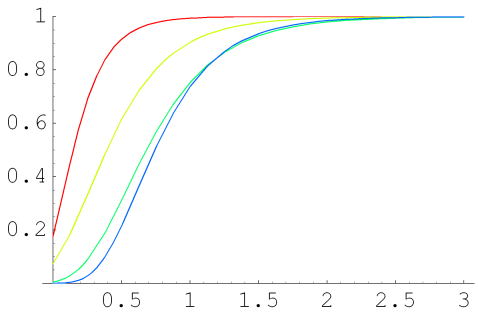


Fig. 2.8: Demonstration of the convergence of $F_n(\tau_n + cW(\tau_n))$ to one for fixed large c. Graphs of the function in (2.31) are shown for $n = 10^2$ (red), 10^3, 10^6, 10^{10} (blue).

is the optimal right-window for this coupling-cutoff. Figure 2.8 shows graphs of

$$\exp\left(-\frac{3}{4}\theta_n\left(\frac{\log n}{4} + cW\left(\frac{\log n}{4}\right)\right)\right) \tag{2.31}$$

as a function of $c > 0$ for a range of values of n. This function converges to one for large c, as expected. Note that the rate of convergence is much slower than that for the left-window.

We end this section on coupling-cutoffs for the hypercube by returning to the motivating example of Diaconis (1996).

Example 2.23. Let us reconsider the random walk on $\mathbb{Z}_2^n$ with transition rates $\lambda_i \propto i^{-\alpha}$, for $\alpha \geq 0$, and for which the total variation cutoff times are presented in Figure 2.4. We have already seen (Proposition 2.7) that when $\alpha = 0$ this walk exhibits a $((n/2)\log n, n)$-coupling-cutoff: that is, the coupling-cutoff time is twice that of the total variation cutoff. We shall now show that similar results hold for $\alpha > 0$.

The random walk in question has rates $\lambda_i = i^{-\alpha}/Z_n(\alpha)$, where $Z_n(\alpha)$ is the normalising constant. Proceeding as in Remark 2.11 it suffices to study the measures $\{\mu_n\}$ defined by

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(n/i)^\alpha}\,.$$

These are supported on $[1, \infty)$ and satisfy $\mu_n(\{1\}) = 1/n$ for all n. It follows that if the random walks driven by $\{\mu_n\}$ exhibit a (τ_n, b_n)-coupling-cutoff then the walks driven by the rates $\{\lambda_i\}_1^n$ will exhibit a $(\tau_n n^\alpha Z_n(\alpha), b_n n^\alpha Z_n(\alpha))$-coupling-cutoff.

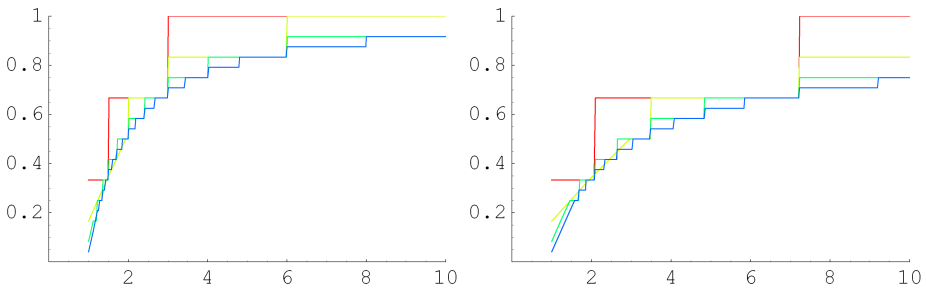


Fig. 2.9: Graph of the distribution functions of the measures μ_n when $\alpha = 1$ (left) and $\alpha = 1.8$ (right), for $n = 3$ (red), 6, 12, 24 (blue).

The measure μ_n satisfies

$$\begin{aligned}\mu_n[1,\lambda] &= \frac{1}{n}\left|\left\{1 \le i \le n \,:\, (n/i)^\alpha \le \lambda\right\}\right| \\ &\sim 1 - \lambda^{-1/\alpha} \quad \text{(see Figure 2.9).}\end{aligned} \tag{2.32}$$

Thus the prospective coupling-cutoff time τ_n satisfies

$$\tau_n = \max_{\lambda \ge 1}\left\{\frac{\log\left[n\left(1-\lambda^{-1/\alpha}\right)\right]}{2\lambda}\right\}.$$

We now claim that this maximum is achieved at

$$\lambda_n^* = \left(\frac{\alpha\log(n/\alpha)}{\alpha\log(n/\alpha)-1}\right)^\alpha. \tag{2.33}$$

Before proving this, let us consider its consequences. Plugging λ_n^* into the definition of τ_n yields

$$\tau_n \sim \frac{\log(n/\alpha) - \log\log(n/\alpha)}{2}.$$

Since this value tends to infinity with n, the condition of Theorem 2.15 is satisfied, and so the walks driven by $\{\mu_n\}$ exhibit a τ_n-coupling-cutoff. By Lemma 2.19, the optimal left-window of this coupling-cutoff is bounded above by $b_n^L = 1/\lambda_n^*$, and the right-window by $b_n^R = W(\tau_n)$. Re-scaling by $n^\alpha Z_n(\alpha)$ shows that the original random walks exhibit a $\tau_n n^\alpha Z_n(\alpha)$-coupling-cutoff, with left-window $n^\alpha Z_n(\alpha)/\lambda_n^*$ and right-window $n^\alpha Z_n(\alpha)W(\tau_n)$. By definition of $Z_n(\alpha)$,

$$n^\alpha Z_n(\alpha) \sim \begin{cases} \frac{n}{1-\alpha} & \text{for } 0 < \alpha < 1 \\ n\log n & \text{for } \alpha = 1 \\ n^\alpha\zeta(\alpha) & \text{for } 1 < \alpha\,, \end{cases}$$

where $\zeta(\alpha)$ is the Riemann zeta function. Since $\lambda_n^* \to 1$ as $n \to \infty$, the coupling-cutoffs for the original walk take the form presented in Figure 2.10.

	τ_n	b_n^L	b_n^R
$\alpha = 0$	$\frac{1}{2} n \log n$	n	n
$0 < \alpha < 1$	$\frac{1}{2(1-\alpha)} n[\log(n/\alpha) - \log\log(n/\alpha)]$	n	$n \log\log(n/\alpha)$
$\alpha = 1$	$\frac{1}{2} n \log n[\log(n/\alpha) - \log\log(n/\alpha)]$	$n \log n$	$n \log n \log\log(n/\alpha)$
$1 < \alpha < \infty$	$\frac{\zeta(\alpha)}{2} n^\alpha[\log(n/\alpha) - \log\log(n/\alpha)]$	n^α	$n^\alpha \log\log(n/\alpha)$

Fig. 2.10: Coupling-cutoff time τ_n, left-window b_n^L and right-window b_n^R for the random walk on $\mathbb{Z}_2^n$ with rates $\lambda_i \propto i^{-\alpha}$.

Comparison of Figures 2.4 and 2.10 shows that the coupling-cutoff time is asymptotically twice that of the total variation cutoff. The left-window b_n^L is the same for both types of cutoff.

All that remains is to prove that λ_n^* satisfies equation (2.33). Define for $\alpha > 0$ and $\lambda > 1$ the function ρ_n by

$$\rho_n(\lambda) = \frac{\log\left[n\left(1 - \lambda^{-1/\alpha}\right)\right]}{2\lambda}\,.$$

(So $\tau_n = \max_\lambda \rho_n(\lambda)$.) Now consider its derivative:

$$\begin{aligned}\rho_n'(\lambda) &= \frac{1}{2\lambda^2}\left(\frac{1}{\alpha(\lambda^{1/\alpha} - 1)} - \log\left[n\left(1 - \lambda^{-1/\alpha}\right)\right]\right) \\ &\propto \frac{1}{\alpha(\lambda^{1/\alpha} - 1)} - \log\left[n\left(1 - \lambda^{-1/\alpha}\right)\right]\,.\end{aligned}$$

This final expression is the difference of two continuous, strictly monotonic functions of λ (for fixed $\alpha > 0$). The first of these is decreasing whereas the second is increasing: it follows that $\rho_n(\lambda)$ has at most one turning point. We now show that $\rho_n'(\lambda_n^*) > 0$ and $\rho_n'(c\lambda_n^*) < 0$ for any fixed $c > 1$, when n is large.

First consider $\rho_n'(\lambda_n^*)$:

$$\begin{aligned}\rho_n'(\lambda_n^*) &\propto \frac{1}{\alpha({\lambda_n^*}^{1/\alpha} - 1)} - \log\left[n\left(1 - {\lambda_n^*}^{-1/\alpha}\right)\right] \\ &= \log\log(n/\alpha) - 1/\alpha\,, \quad \text{by definition of } \lambda_n^*.\end{aligned}$$

Hence $\rho_n'(\lambda_n^*) > 0$ for large enough n.

On the other hand, for fixed $c > 1$:

$$\begin{aligned}\rho_n'(c\lambda_n^*) &\propto \frac{1}{\alpha((c\lambda_n^*)^{1/\alpha}-1)} - \log\left[n\left(1-(c\lambda_n^*)^{-1/\alpha}\right)\right] \\ &= \frac{\alpha\log(n/\alpha)-1}{\alpha\left[(c^{1/\alpha}-1)\alpha\log(n/\alpha)+1\right]} - \log\left(n\left[\frac{(1-c^{-1/\alpha})\alpha\log(n/\alpha)+c^{-1/\alpha}}{\alpha\log(n/\alpha)}\right]\right) \\ &\sim \frac{1}{\alpha(c^{1/\alpha}-1)} - \log\left(n\left[1-c^{-1/\alpha}\right]\right) \quad \text{as } n\to\infty.\end{aligned}$$

Therefore there exists $N_c \in \mathbb{N}$ for all $c > 1$, such that $\rho_n'(c\lambda_n^*) < 0$ whenever $n \geq N_c$.

Now define $\tilde{\lambda}_n$ by $\rho_n'(\tilde{\lambda}_n) = 0$. Since ρ' is continuous on $[\lambda_n^*, c\lambda_n^*]$, by the above calculations it must be the case that

$$\tilde{\lambda}_n \in (\lambda_n^*, c\lambda_n^*), \quad \text{whenever } n \geq N_c.$$

Thus

$$0 < \frac{\tilde{\lambda}_n - \lambda_n^*}{\lambda_n^*} < c-1 \quad \text{for } n \geq N_c.$$

But since c can be made arbitrarily close to 1, this shows that $\tilde{\lambda}_n \sim \lambda_n^*$, and therefore that the claimed coupling-cutoffs hold.

Remark 2.24. It is actually possible to show that, for this example, the values of b_n^L in Figure 2.10 also provide an upper bound on the right-window of the coupling-cutoff.

2.2 Coupling-cutoffs for the random-to-top shuffle

A second class of random walks to which the above analysis may be applied is that of random-to-top card shuffles. These are random walks on the symmetric group S_n, where every transition is such that a card is moved to the top of the pack, with the relative positions of the other cards remaining unchanged. For the ordinary random-to-top shuffle, each card is selected with probability $1/n$. If the card currently at the top of the pack is chosen, the order of the cards remains unchanged (ensuring aperiodicity). This random walk is relatively simple to analyse, and a strong uniform time argument (as given in Diaconis (1996)) shows that a $(n\log n, n)$-total variation cutoff occurs.

A biased random-to-top shuffle may be defined by departing from the uniformity with which cards are chosen in the above setup. There are two common ways of doing this:

1) each *card* i is assigned once and for all a probability p_i, and the card to be moved to the top of the pack is chosen using these probabilities;

2) each *position* k in the deck is assigned a probability p_k, and the card in position k is then selected with probability p_k.

The scheme described in case 1 is often referred to as the *move-to-front* scheme. Whether it should strictly be called a 'shuffle' is a matter for debate, since it requires the 'shuffler' to observe the face values of the cards throughout: something that is unlikely to be acceptable in a casino! A better real-life application for this chain is a library or list management problem: imagine that n books (or computer files) are used over time with different frequencies; when a book has been used it is returned to the top of the pile.

The stationary distribution π of this random walk is in general not uniform, but satisfies

$$\pi\left(\text{card } c_k \text{ in position } k,\, 1 \le k \le n\right) = p_{c_1}\left(\frac{p_{c_2}}{1-p_{c_1}}\right)\ldots\left(\frac{p_{c_n}}{1-\sum_{k=1}^{n-1} p_{c_k}}\right).$$

Jonasson (2006) studies the time $\tau_n^{mix}(1/4)$ for this chain. He proves the following theorem:

Theorem 2.25 (Jonasson (2006))**.** *Consider the move-to-front scheme with card probabilities p_i, $i = 1, \ldots, n$, with $p_i \le 1/3$ for all i. Put*

$$t_u = \min\left\{t \;:\; \sum_{i=1}^{n-1}(1-p_i)^t \le \frac{1}{4}\right\}.$$

Then

$$\frac{1}{25}t_u - 1 \le \tau_n^{mix}(1/4) \le t_u.$$

This theorem is proved using a simple coupling argument for the upper bound, and a variant on the eigenvector technique of Wilson (2004) for the lower bound.

Now consider the following coupling scheme for this random walk. Let X and Y be two such walks, with $X_0 \sim \mathit{Uniform}(S_n)$ and $Y_0 \sim \pi$. Let Λ_i $(i = 1, \ldots, n)$ be independent Poisson processes, with the rate of Λ_i equal to p_i. X and Y may be coupled by moving card i to the top of both packs whenever an incident occurs on Λ_i: the packs will be coupled when at least one incident has occurred on all but one

of the Poisson processes. Let T_n be the time taken for at least one event to happen on all of the Λ_i. Then

$$F_n(t) = \mathbb{P}(T_n \leq t) = \prod_{i=1}^{n} \left(1 - e^{-tp_i}\right). \tag{2.34}$$

Remark 2.26. Although T_n is an upper bound on the coupling time for the random-to-top shuffle, the two times are asymptotically equal. This follows from the observation that the number of cards on which X_0 and Y_0 agree follows a $Poisson(1)$ distribution as $n \to \infty$ (see, for example, Feller (1968)).

Note the obvious similarity between equation (2.34) and that for the distribution function of the partial-independence coupling for the random walk on $\mathbb{Z}_2^n$ (equation (2.3)). This means that almost exactly the same analysis as in the previous section may be applied to this problem. In particular, the analogous version of Proposition 2.6 holds: the coupling time for this shuffle is minimised when $p_i = 1/n$ for all i.

We now proceed as before and move from a set of rates to a sequence of measures $\{\mu_n\}$ on $[1, \infty)$ (re-scaling if necessary), and define

$$\tau_n = \max_{\lambda \geq 1} \left\{ \frac{\log(n\mu_n[1,\lambda])}{\lambda} \right\} = \frac{\log(n\mu_n[1,\lambda_n^*])}{\lambda_n^*}. \tag{2.35}$$

(Note that we do not require a factor of 2 in the denominator of τ_n for this random walk, due to the missing 2 in the exponential term of equation (2.34).) With this setup, the following result follows directly from Theorem 2.15, Lemma 2.19 and Theorem 2.21:

Theorem 2.27. *The move-to-front scheme driven by the sequence of measures $\{\mu_n\}$ exhibits a τ_n-coupling-cutoff (where τ_n is defined in equation (2.35)) if and only if $\tau_n \to \infty$ as $n \to \infty$. Furthermore, $b_n^L = 1/\lambda_n^*$ is a left-window and $b_n^R = W(\tau_n)$ is a right-window of this coupling-cutoff.*

The following four examples all appear in Jonasson (2006); a coupling-cutoff holds for the first three.

Example 2.28. Let $p_i = 1/n$ for all i. Here a $(n \log n, n)$-cutoff is known to occur. Scaling the rates by n gives $\mu_n = \delta_1$ for all n: direct consideration of equation (2.34) shows that this walk exhibits a $(\log n, 1)$-coupling-cutoff, and so the scaled walk exhibits a $(n \log n, n)$-coupling-cutoff.

Example 2.29. Let

$$p_i = \begin{cases} \frac{2}{n+1}, & 1 \leq i \leq \frac{n}{2} \\ \frac{2}{n(n+1)}, & \frac{n}{2}+1 \leq i \leq n. \end{cases}$$

Jonasson proves that $\tau_n^{mix}(1/4) = O(n^2 \log n)$ for this chain. Scaling by $n(n+1)/2$ gives $\mu_n = (1/2)\delta_1 + (1/2)\delta_n$, and so $\tau_n = \log(n/2)$ and $\lambda_n^* = 1$. Undoing the scaling shows that the random walk exhibits a $(n/2)(n+1)\log(n/2)$-coupling-cutoff.

Example 2.30. Let $p_i \propto i^{-1}$: Jonasson shows that $\tau_n^{mix}(1/4) = O(n(\log n)^2)$. This example uses the same set of rates as the random walk on $\mathbb{Z}_2^n$ in Example 2.23 (with $\alpha = 1$), and so it is immediate that a coupling-cutoff occurs here too. Furthermore, Example 2.23 shows that the move-to-front scheme with $p_i \propto i^{-\alpha}$, for *any* $\alpha > 0$ will exhibit such behaviour, with the coupling-cutoff parameters being twice those given in Figure 2.10.

Example 2.31. Let $p_i = 2i/(n(n+1))$. In this case Jonasson shows that $\tau_n^{mix}(1/4) = O(n^2)$. Here though, a coupling-cutoff does not hold. To see this, scale the rates by $n(n+1)/2$ to yield

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_i \,.$$

Then

$$\tau_n = \max_{1 \leq i \leq n} \left\{ \frac{\log i}{i} \right\} \nrightarrow \infty \,,$$

and so there is no coupling-cutoff.

Following the analysis of random walks on $\mathbb{Z}_2^n$, the move-to-front scheme was simple to analyse. Note that this is not the case for the second biased random-to-top shuffling scheme (defined on page 45): the partial-independence coupling for two random walks evolving according to these dynamics is much harder to analyse in general.

2.3 Discussion and future work

The work in this chapter has provided an initial insight into a previously unreported phenomenon: that the coupling time for two random walks can exhibit similar cutoff behaviour to the distance from stationarity. In the next chapter the idea of maximal couplings will be introduced: these couplings achieve equality in the coupling

inequality (1.4). It follows that a maximal coupling for a random walk with a total variation cutoff will by definition exhibit a coupling-cutoff. What makes the results of Section 2.1.3 so interesting is the fact that the partial-independence coupling is not maximal (as can be seen by the difference between the coupling and total variation cutoff times in equations (2.5) and (2.7)). This will be discussed further below.

Theorem 2.15 shows that a random walk on $\mathbb{Z}_2^n$ exhibits a coupling-cutoff at a time which is double the instant of its total variation cutoff. The measure-based approach of Section 2.1.3 provides a more intuitive understanding of the important role played by the sequence $\{\lambda_n^*\}$ in this result than is evident from the proof of Theorem 2.8. It should be possible to extend Theorem 2.15 to produce a result similar to the more general form of Theorem 2.8 (Barrera et al. 2006). Consider a general n-tuple of independent Markov chains, $X^{(n)}$, and suppose that each coordinate $X_i^{(n)}$ may be coupled to a copy of itself in such a way that the coupling occurs (approximately) exponentially fast, with rate λ_i. The partial-independence coupling time for the sequence $\left\{X^{(n)}\right\}$ will then exhibit similar coupling-cutoff behaviour as the sequence of random walks on $\mathbb{Z}_2^n$ driven by the rates $\{\lambda_i\}$.

The results of Section 2.1.4 also provide bounds on the optimal left and right window size of the coupling-cutoff. To the best of our knowledge, little work has been carried out to date which deals with these window sizes separately. In particular, Example 2.22 appears to be a new example of a random walk which exhibits significantly different behaviour either side of a cutoff time. The general bounds on window sizes provided by Lemma 2.19 and Theorem 2.21 have very different orders of magnitude, whereas all the examples of random walks on $\mathbb{Z}_2^n$ presented in this chapter, with the exception of Example 2.22, can be shown to have their optimal right-windows also bounded above by $1/\lambda_n^*$. It would be of use to establish a good sufficient condition on $\{\mu_n\}$ for this to be true.

Recall that the existence of a coupling-cutoff is determined completely by the behaviour of the function θ_n defined in equation (2.13). This function may be interpreted as follows. For $i = 1, \ldots n$, let Z_i^n be independent, identically distributed random variables, whose distribution is a mixture of $Exp(2\lambda)$ distributions, with

mixture probability distribution μ_n. Then, for $t \geq 0$,

$$\mathbb{P}\left(Z_i^n > t\right) = \int_1^\infty e^{-2\lambda t} \mu_n(d\lambda)$$

and so

$$\mathbb{E}\left[\sum_{i=1}^n \mathbf{1}_{\left[Z_i^n > t\right]}\right] = n \int_1^\infty e^{-2\lambda t} \mu_n(d\lambda) = \theta_n(t)\,.$$

Thus θ_n describes the mean number of exceedances of t by the set of random variables $\{Z_i^n\}$. In particular, the set of random walks driven by $\{\mu_n\}$ exhibits a τ_n-coupling-cutoff if and only if

$$\mathbb{E}\left[\sum_{i=1}^n \mathbf{1}_{\left[Z_i^n > c\tau_n\right]}\right] \xrightarrow[n\to\infty]{} \begin{cases} \infty & 0 < c < 1 \\ 0 & c > 1 \end{cases}.$$

This suggests that it may be possible to link the work in this chapter to extreme value theory. This possibility is further motivated by the form taken by the 'shape' of the coupling-cutoff in many examples. For instance, it was shown in Proposition 2.7 and Examples 2.9 and 2.10 that

$$F_n\left(\tau_n + cb_n\right) \xrightarrow[n\to\infty]{} \phi(c) = \exp\left(-\frac{1}{2}e^{-2c}\right). \tag{2.36}$$

That is, ϕ describes the shape of the coupling-cutoff over the window b_n. The function ϕ belongs to one of the three classes of possible limiting distributions for the maximum of n i.i.d. random variables. (This is the Fisher-Tippett theorem: see Bingham et al. (1987).) Of course, for random walks on $\mathbb{Z}_2^n$ the distribution μ_n varies with n, and so the Fisher-Tippett theorem cannot be applied directly. However, the above observations seem to imply that linking the two areas of theory may prove fruitful.

When the shape of the cutoff is known, another direction for future research is an investigation into the *speed* at which the convergence in equation (2.36) takes place. The interest in the cutoff phenomenon arises from the desire for a quantitative answer to the question "How many steps of the random walk are needed for it to be close to equilibrium?" At first sight, proof of a cutoff seemingly yields a sharp answer to this question. However, by definition a cutoff phenomenon involves limiting behaviour! In order to restore some level of quantitativeness to examples where a cutoff has been shown to occur, it would be interesting to investigate 'how far' $X^{(n)}$ is from the cutoff (as a function of n). The speed at which the convergence in equation (2.36) occurs

is then a natural way to measure this. For example, consider the symmetric random walk with all rates equal to $1/n$. Proposition 2.7 showed that this walk exhibits a $((n/2)\log n, n)$-coupling-cutoff with shape ϕ. Inequality (2.14) shows that

$$\phi(c)\exp\left(-\frac{e^{-4c}}{4n}\right) \leq F_n\left(\frac{n\log n}{2}+cn\right) \leq \phi(c)\,.$$

It follows that

$$\left|F_n\left(\frac{n\log n}{2}+cn\right)-\phi(c)\right| \leq \phi(c)\left(1-\exp\left(-\frac{e^{-4c}}{4n}\right)\right) = O(1/n)\,.$$

Finally, it would be of real interest if a better understanding of the relationship between coupling-cutoffs and total variation cutoffs could be obtained. Based on the work in this chapter, one might be tempted to conjecture that if $\left\{X^{(n)}\right\}$ exhibits a total variation cutoff, then the 'best possible' co-adapted coupling for these random walks will exhibit a coupling-cutoff. This is not true however: from Figure 2.1 we see that the transposition shuffle on S_n exhibits a cutoff, but it follows from a simple argument (see Huber (2004)) that the distribution function for the optimal co-adapted coupling for this chain will have a tail which is too fat to produce a coupling-cutoff.

Considering instead the reverse implication, an answer to the following questions would prove very useful:

> Let $\left\{X^{(n)}\right\}$ be a sequence of random walks on $\{G_n\}$. Suppose there exists a time-homogeneous co-adapted coupling for two copies of these walks, for which a τ_n-coupling-cutoff is exhibited. Does this imply that $\left\{X^{(n)}\right\}$ also exhibits a $\hat{\tau}_n$-total variation cutoff? If so, is it then the case that $\hat{\tau}_n = O(\tau_n)$?

It is evident from some of the examples presented in this chapter that there is often a price to be paid for considering only co-adapted couplings. This varies between chains of course, but one possible method of categorising the cost is as follows:

(1) *No cost*: as the name suggests, this is when nothing is lost by restricting to co-adapted couplings. Example: the move-to-front scheme of Section 2.2 - the total-variation cutoff and partial-independence coupling-cutoff times are equal;

(2) *Constant cost*: this is when the smallest expected co-adapted coupling time and the mixing time are out by at most a constant factor. Example: random walk on $\mathbb{Z}_2^n$ - the partial-independence coupling-cutoff time is a factor of two out from the mixing time;

(3) *Growing cost*: this is when the ratio of the smallest expected co-adapted coupling time to the mixing time tends to infinity. Example: the transposition shuffle on S_n - this has a cutoff at time $(n/2)\log n$, but the expectation of any co-adapted coupling time is bounded below by $O(n^2)$ (Huber 2004).

Investigation into this categorisation scheme would be extremely interesting, since it is unclear *a priori* into which category any given random walk should fall. The difference between optimal co-adapted couplings and total variation distance is investigated further in Chapter 3, where it is shown among other things that the optimal co-adapted coupling for the symmetric random walk on $\mathbb{Z}_2^n$ has an expected coupling time of $(n/2)\log n$: it follows that this walk cannot be moved into category (1) above.

"Come together, right now."

Come together, by John Lennon

3. MAXIMAL COUPLING

In the past two chapters it has been demonstrated that coupling may be used to bound the rate of convergence of some random walks on groups. However, although the coupling approach is often much simpler to use than the more analytic options of eigen-analysis and representation theory, the bounds obtained from coupling often turn out to be poorer. For example, consider the random walk on $\mathbb{Z}_2^n$, for which the couplings described in Section 1.2 give a mixing time of $(n/2)\log n$, whereas the true mixing time is known to be $(n/4)\log n$. Although the coupling approach is only out by a factor of two in this case, things can be much worse: the best co-adapted coupling for the transposition shuffle on S_n gives a mixing time of order n^2 (Huber 2004), whereas a cutoff is known to occur at time $(n/2)\log n$.

Of course, the coupling inequality (Lemma 1.8) provides a bound on how good any coupling can possibly be:

$$\left\|\mathbb{P}\left(X_n \in \cdot\right) - \mathbb{P}\left(X_n' \in \cdot\right)\right\| \leq \mathbb{P}\left(T > n\right), \tag{3.1}$$

where T is a random time such that $X_n = X_n'$ for $n \geq T$. In this chapter we consider the existence of *maximal* couplings: that is, couplings which achieve equality in (3.1).

3.1 Existence of maximal couplings

3.1.1 Maximal coupling of measures

We start by considering couplings of measures (rather than sequences of measures) on $(E, \mathcal{E})$. The following theorem (Thorisson 2000; Lindvall 2002) proves the existence of a coupling (often called the Vasershtein coupling) that is maximal in the sense that it attains equality in the coupling inequality (1.3).

Theorem 3.1 (Maximal coupling of two probability measures)**.** *Let μ and μ' be two probability measures on the space $(E, \mathcal{E})$. Then there exists a coupling (X, X') such that*

(i) $\|\mu - \mu'\| = \mathbb{P}(X \neq X')$;

(ii) conditional on the event $\{X \neq X'\}$, X *and* X' *are independent.*

Proof. Recall that $\mu \wedge \mu'$ is the greatest common component of μ and μ'. Define

$$c = (\mu \wedge \mu')(E)\,.$$

Then equation (1.1) shows that $\|\mu - \mu'\| = 1 - c$, and thus we need to produce a coupling (X, X') such that $\mathbb{P}(X = X') = c$. Clearly if $c = 0$ then μ and μ' have disjoint support, and so it suffices to take X and X' to be independent. Similarly, if $c = 1$ then $\mu = \mu'$ and so we may define $X' = X$.

Now suppose that $0 < c < 1$. Define probability measures ν and ν' by

$$\nu = \frac{\mu - (\mu \wedge \mu')}{1 - c}\,, \qquad \nu' = \frac{\mu' - (\mu \wedge \mu')}{1 - c}\,.$$

Let I, V, W and W' be independent random variables such that

- I is a Bernoulli random variable with success probability c,
- V has law $(\mu \wedge \mu')/c$ on E,
- W has law ν, and W' has law ν', on E.

Finally, define X and X' by

$$X = \begin{cases} V & \text{if } I = 1 \\ W & \text{if } I = 0 \end{cases}, \qquad \text{and} \quad X' = \begin{cases} V & \text{if } I = 1 \\ W' & \text{if } I = 0 \end{cases}.$$

To see that this is a coupling, observe that

$$\begin{aligned} \mathbb{P}(X \in A) &= \mathbb{P}(I = 1)\,\mathbb{P}(V \in A) + \mathbb{P}(I = 0)\,\mathbb{P}(W \in A) \\ &= c\,\frac{(\mu \wedge \mu')(A)}{c} + (1 - c)\nu(A) \\ &= \mu(A), \end{aligned}$$

and $\mathbb{P}(X' \in A) = \mu'(A)$. Finally, by construction, $\mathbb{P}(X = X') = \mathbb{P}(I = 1) = c$, and conditional upon the event $\{X \neq X'\}$, X and X' are independent by the independence of W and W'. □

The intuition behind this maximal coupling is simple: X and X' are made to agree with as large a probability as possible, $(\mu \wedge \mu')(E)$, else are drawn independently from

the residual measures ν and ν'. This leads to a method of simulating a maximally-coupled pair (X, X') (see Figure 3.1). Draw a point (x, y) uniformly at random from the area lying between the x-axis and the density f of μ (with respect to some dominating measure λ), and let $X = x$. If (x, y) also lies beneath f' (the density of μ' w.r.t. λ) then set $X' = x$. If not, draw a new point (x', y') uniformly at random from beneath f' but above f, and set $X' = x'$.

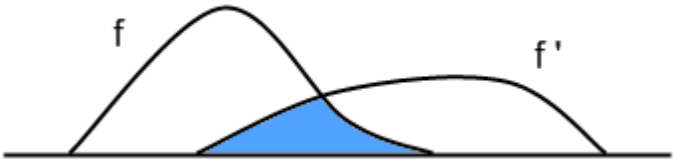


Fig. 3.1: Maximal coupling probability.

This construction can be extended to a countable collection of random variables $X_1, X_2, \ldots$, resulting in a coupling such that all subsets $\{X_{n_1}, \ldots, X_{n_k}\}$ are maximally coupled (see Thorisson (2000) for details).

Maximal coupling problems can also be viewed from the perspective of optimal transport theory. In the notation of Definition 1.2, let $\phi(x, x')$ be a *cost function* defined on $\Omega \times \Omega'$. Optimal transport problems generally involve trying to solve the minimisation problem

$$\inf_{(\hat{X}, \hat{X}')} \hat{\mathbb{E}} \left[\phi \left(\hat{X}, \hat{X}' \right) \right] ,$$

or equivalently,

$$\inf_{\hat{\mathbb{P}}} \int_{\Omega \times \Omega'} \phi(x, x') \, d\hat{\mathbb{P}} \left(x, x' \right) .$$

Here the function $\phi(x, x')$ represents the amount of work needed to move one unit of mass from x to x': the special case of $\phi(x, x') = \mathbf{1}_{x \neq x'}$ yields

$$\inf_{(\hat{X}, \hat{X}')} \hat{\mathbb{P}} \left(\hat{X} \neq \hat{X}' \right) ,$$

which is simply the problem of finding a maximal coupling, by Theorem 3.1. For further information on this area of theory, see Villani (2005).

3.1.2 Maximal coupling for stochastic processes

For the remainder of this chapter, $X = \{X_n\}$ will be a discrete time, homogeneous Markov chain defined on $(\Omega, \mathcal{F}, \mathbb{P})$, taking values in a countable state space S: let μ be its initial distribution, and P its transition matrix. We shall write $\mu_n = \mu P^n$ for the n-step distribution of X. Recall from Chapter 1 that X is said to be weakly ergodic if

$$\lim_{n\to\infty} \sum_{k\in S} |\delta_x P^n(k) - \delta_y P^n(k)| = 0 \quad \text{for all } x, y \in S\,.$$

The proof of Theorem 3.1 shows that it is easy to construct a maximal coupling of two random variables (the Vasershtein coupling). In order to maximally couple two Markov chains X and X' however, it is necessary to define *simultaneously for all* n a maximal coupling of the random variables X_n and X'_n, such that the sequences $\{X_n\}$ and $\{X'_n\}$ (considered separately) evolve as Markov chains. As such, the existence of a maximal coupling for Markov chains is harder to prove than the equivalent result for a pair of random variables. Nevertheless, such a result does exist:

Theorem 3.2 (Griffeath (1975))**.** *Any Markov chain X has a maximal coupling (achieving equality in* (3.1)*). Thus there exists a successful maximal coupling for X if and only if X is weakly ergodic.*

This result is profound, but the proof as given in Griffeath (1975) is rather tedious (as Griffeath himself admits!). The maximal coupling constructed is necessarily non-Markovian, and this makes the proof rather technical and somewhat unintuitive.

However, the beautiful paper of Pitman (1976) approaches the construction of Griffeath's maximal coupling in a far more intuitive manner, via the use of randomised stopping times (recall Definition 1.16). RSTs are particularly useful when coupling Markov chains, since they can be used to bound the total variation distance in a manner similar to the distributional coupling method of Definition 1.11. Thus, if T is a RST of a chain X defined on $(\Omega, \mathcal{F}, \mathbb{P})$, and T' is a RST of X' defined on $(\Omega', \mathcal{F}', \mathbb{P}')$, such that

$$\mathbb{P}\left(T = n, X_n = y\right) = \mathbb{P}'(T' = n, X'_n = y)\,, \quad n \in \mathbb{N},\, y \in S\,, \tag{3.2}$$

then it follows that

$$\left\|\mu_n - \mu'_n\right\| \leq \mathbb{P}\left(T > n\right) = \mathbb{P}'(T' > n)\,.$$

Of course, if the random time T is a RST of both chains, then we recover the normal coupling inequality. If not, then Pitman (1976) makes the following point: given the matching distributions in equation (3.2), it is possible to define copies of X and X', say $\hat{X}$ and $\hat{X}'$, on a probability space $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}})$ which also supports a random time $\hat{T}$ and an S-valued random variable $\hat{Y}$ satisfying:

(i) the $\hat{\mathbb{P}}$ distribution of $(\hat{X}, \hat{T}, \hat{Y})$ equals the $\mathbb{P}$ distribution of (X, T, X_T) (with the same statement holding for the appropriate primed distributions);

(ii) under $\hat{\mathbb{P}}$ the Markov chains $\hat{X}$ and $\hat{X}'$ are conditionally independent given $\hat{T}$ and $\hat{Y}$.

The first of these conditions ensures that $\hat{Y} = \hat{X}_{\hat{T}} = \hat{X}'_{\hat{T}}$ $\hat{\mathbb{P}}$-almost surely. This means that it is possible to define a new process $\hat{X}''$ by crossing over from $\hat{X}'$ to $\hat{X}$ at time $\hat{T}$:

$$\hat{X}'' = \begin{cases} \hat{X}'_n & \text{on } \{\hat{T} > n\} \\ \hat{X}_n & \text{on } \{\hat{T} \leq n\}. \end{cases}$$

The conditional independence in (ii) and the strong Markov property of $\hat{X}$ and $\hat{X}'$ at the RST $\hat{T}$ then ensures that $\hat{X}''$ is a Markov chain with initial distribution μ' and transition kernel P. The result is therefore a coupling $(\hat{X}, \hat{X}'')$ of (X, X') with coupling time $\hat{T}$. Thus RSTs enable us to 'glue' distributions together and hence move from a distributional (weak) coupling to a non-distributional (strong) coupling.

Pitman's coupling construction proceeds by identifying a specific time T which is a RST of both X and X': this T will be the coupling time. The maximal coupling is achieved by firstly specifying the space-time distribution of (T, X_T), and then the conditional laws (given T and X_T) of the two pre-T processes and the single post-T process, with the requirement that these three fragments be conditionally independent given T and X_T. We now state his result.

Let $\tilde{\Omega}$ be the space of all sequences

$$\tilde{\omega} = ((\omega_0, \omega'_0), (\omega_1, \omega'_1), \ldots)$$

of pairs of points in S, and equip $\tilde{\Omega}$ with the product σ-algebra $\tilde{\mathcal{F}}$ generated by the coordinate maps $X_0, X_1, \ldots, X'_0, X'_1, \ldots$, where $X_n(\tilde{\omega}) = \omega_n$. Define

$$T(\tilde{\omega}) = \min\left\{n \,:\, X_n(\tilde{\omega}) = X'_n(\tilde{\omega})\right\},$$

with $T(\tilde{\omega}) = \infty$ if there is no such n. Finally recall that, for a general σ-finite signed measure ν,

$$\nu^+ = \max\{\nu, 0\} \quad \text{and} \quad \nu^- = \max\{0, -\nu\}$$

are both positive measures, and $\nu = \nu^+ - \nu^-$.

Theorem 3.3 (Pitman (1976)). *Let μ and μ' be two mutually singular probabilities on S, let $\alpha_n = \mu_n - \mu'_n$, and suppose that $\lim_{n\to\infty} \|\alpha_n\| = 0$. Then there exists a unique probability measure $\tilde{\mathbb{P}}$ on $(\tilde{\Omega}, \tilde{\mathcal{F}})$ such that*

1. *$\tilde{\mathbb{P}}(T = n, X_n = z) = (\alpha^+_{n-1}P - \alpha^+_n)(z)$, $n \geq 1, z \in S$, and*
2. *under $\tilde{\mathbb{P}}$ conditional on $(T = n, X_n = z)$, for each $n \geq 1, z \in S$,*
   (a) *the two pre-T processes $(X_0, \ldots, X_n)$ and $(X'_0, \ldots, X'_n)$ are inhomogeneous Markov chains with reverse transition probabilities from y at time m to x at time $m-1$ given respectively by*

$$\alpha^+_{m-1}(x)P(x,y)/\alpha^+_{m-1}P(y) \quad \text{and} \quad \alpha^-_{m-1}(x)P(x,y)/\alpha^-_{m-1}P(y)\,;$$

   (b) *the post-T processes $(X_n, X_{n+1}, \ldots)$ and $(X'_n, X'_{n+1}, \ldots)$ are a.s. identical, forming a single homogeneous Markov chain with transition probabilities P starting at z;*
   (c) *the two pre-T processes and the single post-T process are mutually independent. Under this probability $\tilde{\mathbb{P}}$ the two marginal processes X and X' are Markov (with transition kernel P, and initial distributions μ and μ' respectively); these chains agree $\tilde{\mathbb{P}}$-a.s. after the random time T when they first meet,*

$$\tilde{\mathbb{P}}(T > n) = \|\alpha_n\|\,,$$

   and the maximal coupling thus provided is the maximal coupling of Griffeath (1975).

This construction can also be made to work when X takes values in a general measurable state space, or when X is inhomogeneous: the Markov property is the only essential requirement (Pitman 1976). In Section 3.2 we will use Pitman's construction to extend the idea of a maximal coupling to that of a *maximal coalescent*

coupling, whereby we aim to minimise the time taken for N chains to couple. Further discussion of this construction is thus postponed until that time.

Following on from the result of Griffeath (1975), Goldstein (1979) considered the question of the existence of maximal couplings for two (not necessarily Markovian) processes X^1 and X^2, taking values in a standard Borel space. He proved the following theorem:

Theorem 3.4 (Goldstein (1979))**.** *The following are equivalent:*

(i) There exists a successful coupling of X^1 and X^2;

(ii) $\mu^1_\infty = \mu^2_\infty$ (agreement on the tail σ-algebra);

(iii) $\lim_{n\to\infty} \left\|\mu^1_n - \mu^2_n\right\| = \lim_{n\to\infty} \left\|\mu^1 - \mu^2 \,|\, \mathcal{F}_n\right\| = 0\,.$

The proof that (iii) implies (i) follows directly from the existence of a maximal coupling, which Goldstein constructs by piecing together a sequence of measures. If X^1 and X^2 are copies of the same Markov chain (with countable state space), then this coupling turns out to be exactly that of Griffeath (1975).

The idea of piecing together measures was subsequently used by Thorisson (1986), who proved the existence of a maximal distributional coupling for processes on a general state space. Recall that $(\hat{X}, \hat{X}')$ is a distributional coupling of X and X' with coupling times T and T' if both of the following hold:

(a) $\hat{X} \stackrel{\mathcal{D}}{=} X$ and $\hat{X}' \stackrel{\mathcal{D}}{=} X'$;

(b) $(\theta_T \hat{X}, T) \stackrel{\mathcal{D}}{=} (\theta_{T'} \hat{X}', T')\,.$

Theorem 3.5 (Thorisson (1986))**.** *Let X and X' be discrete time stochastic processes on a general state space $(E, \mathcal{E})$. Then there exists a (not necessarily successful) maximal distributional coupling of X and X'.*

Of course, if there exists a regular version of the conditional distribution of $\hat{X}$ given $(\theta_T \hat{X}, T)$, then $\hat{X}$ and $\hat{X}'$ can be glued together using part (b) of the above definition. This 'glueing together' is the same as used by Pitman (1976), and in this way the non-distributional maximal coupling of Griffeath may once again be recovered.

The remainder of this chapter considers some extensions and applications of maximal couplings. In Section 3.2 the idea of a maximal coupling of two chains is extended to that of a maximal coalescent coupling for $N \geq 2$ chains. This construction will be further used in Chapter 4 when we examine the efficiency of a perfect simulation algorithm known as CFTP. The final two sections of this chapter consider maximal couplings for specific examples of random processes (namely the random walk on $\mathbb{Z}_2^n$ introduced in Example 1.12, Brownian motion in $\mathbb{R}^d$ and the Ornstein-Uhlenbeck process).

3.2 Maximal coalescent coupling

Suppose now that, instead of simply coupling two chains X and X', we wish to couple N chains $\{X^i,\ i = 1, \ldots, N\}$, for some $2 \leq N \leq \infty$. Can we find a 'maximal coupling' for these chains? This is a natural question to ask, given the existence of a maximal (Vasershtein) coupling for a countable number of probability distributions, as mentioned in Section 3.1.1. The answer is yes, and in this section we explicitly describe such a coalescent coupling. Recall that Pitman's description of a maximal coupling of two Markov chains is very transparent: Theorem 3.3 explicitly describes the behaviour of each chain both before and after the coupling time. It therefore seems natural to use this result as a basis for producing a maximal coalescent coupling: the following construction is based upon the method of Pitman (1976), and includes a proof of his theorem as a special case (when $N = 2$).

Maintaining the notation of Section 3.1.2, we again write P for the (common) transition matrix of the chains, and μ_n^i for the distribution of X_n^i. To ease notation we define the measure

$$\lambda_n = \mu_n^1 \wedge \mu_n^2 \wedge \cdots \wedge \mu_n^N$$

to be the greatest common component of the measures $\{\mu_n^i\}$, and write

$$d_n = 1 - \sum_{y \in S} \lambda_n(y).$$

d_n is simply a generalisation of total variation distance to the case of $N \geq 2$ measures (recall Equation (1.1)). To make this comment precise, suppose that $\{\hat{X}^i\}$ is a coupling of $\{X^i\}$ on some space $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}})$, such that $\hat{X}_n^i(\hat{\omega}) = \hat{X}_n^j(\hat{\omega})$ for all i, j

whenever $n \geq \hat{T}(\hat{\omega})$, for some random time $\hat{T}$. Then

$$\hat{\mathbb{P}}\left(\hat{X}_n^i = y, \hat{T} \leq n\right) = \hat{\mathbb{P}}\left(\hat{X}_n^j = y, \hat{T} \leq n\right) \leq \mu_n^j(y) \quad \text{for all } j.$$

Thus

$$\hat{\mathbb{P}}\left(\hat{X}_n^i = y, \hat{T} \leq n\right) \leq \lambda_n(y)$$

and so

$$\hat{\mathbb{P}}\left(\hat{T} > n\right) \geq d_n\,. \tag{3.3}$$

Inequality (3.3) is therefore a natural generalisation of the coupling inequality, and we shall call a coupling of the N chains $\left\{X^i\right\}$ a *maximal coalescent coupling* if its coupling time achieves equality in (3.3).

The following notation directly extends that used in Theorem 3.3. Let $\tilde{\Omega}$ be the space of all sequences

$$\tilde{\omega} = ((\omega_0^1, \ldots, \omega_0^N), (\omega_1^1, \ldots, \omega_1^N), \ldots)$$

of N-tuples of points in S, and we equip $\tilde{\Omega}$ with the product σ-field $\tilde{\mathcal{F}}$ generated by the coordinate maps $X_0^0, X_1^0, \ldots, X_0^N, X_1^N, \ldots$ where, for instance, $X_n^i(\tilde{\omega}) = \omega_n^i$. Let

$$T(\tilde{\omega}) = \min\left\{n \,:\, X_n^i(\tilde{\omega}) = X_n^1(\tilde{\omega}) \text{ for all } i \geq 1\right\}\,,$$

with $T(\tilde{\omega}) = \infty$ if there is no such n. Finally, let

$$\beta_n^i = \mu_n^i - \lambda_n, \quad i = 1, \ldots, N.$$

(Note that β_n^i is a non-negative measure for all i and n, and that if $N = 2$ then $\beta_n^1 = \alpha_n^+$ and $\beta_n^2 = \alpha_n^-$, where α_n is defined in Theorem 3.3.)

The main result of this section is that it is possible to construct a maximal coalescent coupling. We begin the proof by extending the proposition of Pitman (1976):

Proposition 3.6. *For $i = 1, \ldots, N$, let T^i be a randomised stopping time (RST) of the chain X^i defined on $(\Omega^i, \mathcal{F}^i, \mathbb{P}^i)$. If the $\mathbb{P}^i$ distribution of (T^i, X_{T^i}) is identical for all i, that is if*

$$\mathbb{P}^i\left(T^i = n, X_n^i = y\right) = \mathbb{P}^1\left(T^1 = n, X_n^1 = y\right), \quad \textit{for all } i \leq N,\ n \in \mathbb{N},\ y \in S, \tag{3.4}$$

then

$$d_n \leq \mathbb{P}^1\left(T^1 > n\right) = \mathbb{P}^i\left(T^i > n\right), \quad i = 2, \ldots, N,\ n \in \mathbb{N}. \tag{3.5}$$

Proof. This follows from the argument leading to inequality (3.3), using the strong Markov property of RSTs and equation (3.4). □

We now use this proposition and follow Pitman's proof of Theorem 3.3 to prove the main result of this section.

Theorem 3.7. *Let μ_0^i, λ_n and β_n^i $(i = 1, \ldots, N)$ be as above, and suppose that $\lim_{n\to\infty} d_n = 0$. Then there exists a unique probability $\tilde{\mathbb{P}}$ on $(\tilde{\Omega}, \tilde{\mathcal{F}})$ such that*

1. *$\tilde{\mathbb{P}}\left(T = n, X_n^1 = z\right) = (\lambda_n - \lambda_{n-1}P)(z)$, $n \geq 1, z \in S$, and*
2. *under $\tilde{\mathbb{P}}$ conditional on $(T = n, X_n^1 = z)$, for each $n \geq 1, z \in S$,*
   (a) *the N pre-T processes $(X_0^i, \ldots, X_n^i)$ $(i = 1, \ldots, N)$ are inhomogeneous Markov chains with reverse transition probabilities from y at time m to x at time $m-1$ given by $\beta_{m-1}^i(x)P(x,y)/\beta_{m-1}^i P(y)$ respectively;*
   (b) *the post-T processes $(X_n^i, X_{n+1}^i, \ldots)$ $(i = 1, \ldots, N)$ are a.s. identical, forming a single homogeneous Markov chain starting at z, with transition kernel P;*
   (c) *the N pre-T processes and the single post-T process are mutually independent. Under this probability $\tilde{\mathbb{P}}$ the N marginal processes X^i are Markov (with transition kernel P, and initial distributions μ_0^i respectively); these chains agree $\tilde{\mathbb{P}}$-a.s. after the random time T when they first meet, and*

$$\tilde{\mathbb{P}}\left(T > n\right) = d_n\,. \tag{3.6}$$

Proof. First observe that to obtain equality in (3.5) it suffices to construct randomised stopping times T^i such that

$$\mathbb{P}^i\left(T^i > n, X_n^i = y\right) = \beta_n^i(y), \quad 1 \leq i \leq N,\ y \in S, \tag{3.7}$$

since then

$$\begin{aligned}\mathbb{P}^i\left(T^i > n\right) &= \sum_{y\in S} \beta_n^i(y) = \sum_{y\in S} \left[\mu_n^i(y) - \lambda_n(y)\right] \\ &= 1 - \sum_{y\in S} \lambda_n(y) = d_n\,.\end{aligned}$$

Equation (3.7), along with the strong Markov property of RSTs, implies that

$$\begin{aligned}\mathbb{P}^i\left(T^i = n, X_n^i = y\right) &= \beta_{n-1}^i P(y) - \beta_n^i(y) \\ &= (\lambda_n - \lambda_{n-1}P)(y), \quad 1 \le i \le N. \end{aligned} \tag{3.8}$$

We thus attempt to satisfy equation (3.7), since this will yield RSTs $\left\{T^i\right\}$ satisfying equation (3.4) and such that equality is achieved in (3.5). When this has been completed, the rest of the theorem follows simply by moving from a distributional to a non-distributional coupling, as in the discussion immediately after Definition 1.16.

Following Pitman's proof, the simplest thing to try is to make the conditional distribution of $\left\{T^i > n\right\}$ given $\left(X_0^i, \ldots, X_n^i\right)$ and $\left\{T^i \ge n\right\}$ equal to $r_n^i(X_n^i)$ for some functions $r_n^i : S \to [0,1]$. This then yields

$$\begin{aligned}\mathbb{P}^i\left(T^i > n \,|\, X_0^i = x_0^i, \ldots, X_n^i = x_n^i\right) &= r_0^i(x_0^i) r_1^i(x_1^i) \ldots r_n^i(x_n^i) \\ &= f_n^i(x_0^i, \ldots, x_n^i)\,, \end{aligned} \tag{3.9}$$

where the functions f_n^i are defined by this last equation: note that they are by definition decreasing in n. This may be achieved by letting $(\Omega^i, \mathcal{F}^i, \mathbb{P}^i)$ support a random variable $U^i \sim \mathit{Uniform}[0,1]$ which is independent of $\left\{X_n^i\right\}_n$, and setting

$$T^i = \inf\left\{n \,:\, f_n^i(X_0^i, \ldots, X_n^i) \le U^i\right\}\,.$$

We now claim that (3.7) may be achieved in this way by taking

$$r_0^i = \beta_0^i/\mu_0^i\,; \qquad r_n^i = \beta_n^i/\beta_{n-1}^i P, \quad n \ge 1, \tag{3.10}$$

where for two measures ϕ and ψ on S with $\phi \le \psi$, ϕ/ψ denotes the density of ϕ with respect to ψ. This follows by a simple induction argument. For $n = 0$,

$$\begin{aligned}\mathbb{P}^i\left(T^i > 0, X_0^i = y\right) &= \mathbb{P}^i\left(T^i > 0 \,|\, X_0^i = y, T^i \ge 0\right) \mathbb{P}^i\left(X_0^i = y\right) \\ &= r_0^i(y)\mu_0^i(y) = \beta_0^i(y)\,. \end{aligned}$$

Now assume that equation (3.7) holds for $n = k \ge 0$. It is then the case that

$$\begin{aligned}\mathbb{P}^i\left(T^i > k+1, X_{k+1}^i = y\right) &= r_{k+1}^i(y)\mathbb{P}^i\left(T^i > k, X_{k+1}^i = y\right) \\ &= \left(\beta_{k+1}^i(y)/\beta_k^i P(y)\right)\beta_k^i P(y) \\ &= \beta_{k+1}^i(y)\,, \end{aligned}$$

as required.

Finally, note that the construction outlined above has the feature that for each i, conditional on either $\{T^i = n\}$ or $\{T^i \geq n\}$ for each $n \geq 1$, the pre-n process $(X_0^i, \ldots, X_n^i)$ is an inhomogeneous Markov chain (this follows from the multiplicative form of equation (3.9)). The reverse transition probabilities of this chain (whatever the value of $n \geq 1$) are as follows: the transition probability from y at time m to x at time $m-1$ is given by

$$\beta_{m-1}^i(x)P(x,y)/\beta_{m-1}^i P(y), \quad i = 1, \ldots, N\,,$$

as claimed in the statement of the theorem. This completes the proof.

□

Note that the result of Theorem 3.7 can be generalised in two directions. Firstly, there is no need for the state space S to be countable: any Polish space will suffice. Secondly, the result also holds for an uncountable number of chains (if, for example, S is continuous).

3.3 Maximal coupling for the simple random walk on $\mathbb{Z}_2^n$

Although a maximal coupling of Markov chains is known to exist, such a coupling is typically (at best) very difficult to compute explicitly. This usually arises because the maximal coupling is often not co-adapted, and thus an intuitive description of the joint evolution of X and X' may be difficult, if not impossible. Until recently, for example, the best known explicit coupling for the transposition shuffle on S_n was the co-adapted coupling described in Aldous (1983). Instead of giving the correct mixing time of $(n/2)\log n$ however, this strategy takes $O(n^2)$ steps to couple. Indeed, it is quite simple to show that *any* co-adapted coupling for this random walk will take at least $O(n^2)$ steps. A (non-co-adapted) coupling with a coupling time of $O(n \log n)$ was finally produced by Burton and Kovchegov in 2006.

However, for the running example of a random walk on the hypercube (introduced in Example 1.12), an almost-maximal coupling due to Matthews (1987) has been known for some time. This is a non-co-adapted coupling based on the Aldous coupling of Example 1.13: we now briefly describe this construction.

3.3.1 An almost-maximal (non-co-adapted) coupling

Let X and Y be the two discrete-time random walks to be coupled, with $X_0 = 0$ and $Y_0 \sim \mathit{Uniform}(\mathbb{Z}_2^n)$. Matthews' coupling proceeds as follows. Define a 'mythical' $(n+1)^{\text{th}}$ coordinate, which moves whenever $Y_k = Y_{k-1}$. This yields a new process Y^* on $\mathbb{Z}_2^{n+1}$: set

$$Y_0^*(n+1) = \begin{cases} 1 & \text{if } |Y_0| \text{ is odd} \\ 0 & \text{otherwise.} \end{cases}$$

We similarly define a process X^*, with $X_0^*(n+1) = 0$: this of course makes $|Y_0^*| = |Y_0^* - X_0^*|$ even. The coupling strategy actually couples the processes X^* and Y^*, but this of course ensures that X and Y are also coupled.

The coupling time will now be defined by running the process Y^* until some stopping time T. The sample path of X^* over the interval $[0, T]$ will then be constructed from the *entire* path of Y^* over $[0, T]$. That is, what X^* does at time k, given $k \leq T$, will depend upon what Y^* does up to time T: thus the coupling will certainly not be co-adapted.

With this in mind, define

$$U_0 = \{1 \leq i \leq n+1 \,:\, Y_0^*(i) = 1\} = \{1 \leq i \leq n+1 \,:\, X_0^*(i) \neq Y_0^*(i)\} \,,$$

and, for $k \geq 0$, let

$$U_k = \{i \in U_0 \,:\, Y_k^*(i) = 1\} \,.$$

The stopping time T (which will turn out to be the coupling time) is then given by:

$$T = \min\left\{k \geq 0 \,:\, |U_k| = \frac{1}{2}\,|U_0|\right\} \,. \tag{3.11}$$

Thus T is the time taken for half of the originally unmatched coordinates to have changed to their opposite value.

Given T and $Y_0^*, \ldots, Y_T^*$, we now construct $X_1^*, \ldots, X_T^*$ such that X^* evolves as a simple random walk on $\mathbb{Z}_2^{n+1}$ and $X_T^* = Y_T^*$. This is done as follows: for $k \geq 1$ let i_k be the coordinate on which Y_k^* and Y_{k-1}^* differ. Since $|U_T| = |U_0 \backslash U_T|$, we can define a bijection

$$\rho : U_0 \backslash U_T \to U_T \,.$$

Then at step k, X^*_k is obtained from X^*_{k-1} by flipping coordinate j_k, where

$$j_k = \begin{cases} i_k & \text{if } k > T\,; \\ i_k & \text{if } k \le T \text{ and } i_k \notin U_0\,; \\ \rho(i_k) & \text{if } k \le T \text{ and } i_k \in U_0 \backslash U_T\,; \\ \rho^{-1}(i_k) & \text{if } k \le T \text{ and } i_k \in U_T\,. \end{cases}$$

The similarity between this coupling and the Aldous coupling (in Example 1.13) is clear: matched coordinates are made to move synchronously, whereas unmatched coordinates are matched in pairs. The difference between the two is that this coupling 'cheats' by looking into the future, and then using this future information about Y to produce a bijection ρ which makes the pairwise coupling happen as fast as possible. Matthews (1987) shows explicitly that X^* does have the correct transition kernel to be a simple random walk, and also analyses the distribution of T. He proves that, with $\tau_n = (n/4)\log n$,

$$\mathbb{P}\left(T > \tau_n + cn\right) = 2\Phi\left(\frac{e^{-2c}}{\sqrt{2}}\right) - 1 + o(1)\,, \tag{3.12}$$

showing the mixing time to be $(n/4)\log n$. This improves the coupling time of the Aldous coupling by a factor of $1/2$, and achieves the time at which there is a known cutoff (recall Theorem 2.3). Equation (3.12) is very similar to the result of Diaconis et al. (1990) (Theorem 2.4 of this thesis), which showed that

$$\|\mathcal{L}(X_{\tau_n+cn}) - \mathcal{L}(Y_{\tau_n+cn})\| = 2\Phi\left(\frac{e^{-2c}}{2}\right) - 1 + o(1)\,. \tag{3.13}$$

Matthews' coupling is therefore extremely close to being maximal: it does not strictly qualify as maximal due to the slight difference between equations (3.12) and (3.13). This coupling does, however, demonstrate the potential gain to be made by considering non-co-adapted couplings.

3.3.2 An optimal co-adapted coupling

We have seen in Chapter 2 that the partial-independence coupling for the continuous-time random walk on $\mathbb{Z}_2^n$ (with all rates equal to $1/n$) has coupling time of order $n\log n/2$. Matthew's maximal coupling for this random walk does give the correct order of $n\log n/4$, but the coupling is clearly not co-adapted. This therefore suggests the following question: what is the best possible co-adapted coupling for this walk,

and is it unique? In this section we answer both of these questions for the continuous-time version of X.

Remark 3.8. Before proceeding further, it is important to clarify some terminology. In what follows, any coupling achieving equality in the coupling inequality (3.1) will continue to be called a *maximal* coupling. Thus a co-adapted maximal coupling is a coupling which is both maximal and co-adapted. We will also be considering 'best possible' couplings (which may or may not be maximal) in the class of co-adapted couplings - these will be referred to throughout as *optimal co-adapted* couplings.

The issue of finding an optimal co-adapted coupling can be viewed as an optimal control problem: we aim to find the best way of controlling the chains X and Y (in a co-adapted manner) so as to minimise the expected coupling time. (Note that we could equally reasonably wish to consider optimality with respect to some other function of the coupling time - this will be discussed in Section 3.5.)

Let $\mathcal{C}$ be the class of all successful co-adapted couplings for two random walks on $\mathbb{Z}_2^n$. That is, $\mathcal{C}$ contains all co-adapted couplings c for which the coupling time T^c satisfies $\mathbb{P}(T^c < \infty) = 1$. We now describe a framework for a general co-adapted coupling, and introduce the notation to be used in this section. As usual, X and Y are the two random walks that we wish to couple. To simplify the algebra, we shall assume in what follows that each coordinate of X (and of Y) moves at rate 1 (rather than rate $1/n$), independently of all other coordinates.

For $0 \leq i, j \leq n$ let Λ_{ij} be independent unit-rate marked Poisson processes, with marks U_{ij} chosen uniformly on the interval $(0, 1)$. The transitions of X and Y will be driven by these processes, and controlled by a co-adapted process $\{Q(t)\}_{t\geq 0}$, where $Q(t) = \{q_{ij}(t) : 1 \leq i, j, \leq n\}$ is a $n \times n$ doubly sub-stochastic matrix. Such a matrix implicitly defines terms $\{q_{0j}(t) : 1 \leq j \leq n\}$ and $\{q_{i0}(t) : 1 \leq i \leq n\}$ satisfying

$$\sum_{i=0}^{n} q_{ij}(t) = 1 \quad \text{for all } 1 \leq j \leq n \text{ and } t \geq 0\,, \tag{3.14}$$

$$\text{and} \quad \sum_{j=0}^{n} q_{ij}(t) = 1 \quad \text{for all } 1 \leq i \leq n \text{ and } t \geq 0\,. \tag{3.15}$$

For convenience we also define $q_{00}(t) = 0$ for all $t \geq 0$.

A general co-adapted coupling for X and Y may now be defined as follows:

if there is an incident on the process Λ_{ij} at time $t \geq 0$, *and* the mark $U_{ij}(t) \leq q_{ij}(t)$, then set $X_t(i) = 1 - X_{t-}(i)$ and $Y_t(j) = 1 - Y_{t-}(j)$.

From this construction it follows directly that X and Y each have the correct marginal transition rates for a continuous-time simple random walk on $\mathbb{Z}_2^n$. Furthermore, there is a one-to-one correspondence between co-adapted couplings $c \in \mathcal{C}$ and matrix processes $\{Q(t)\}_{t\geq 0}$ satisfying the above conditions. Finally, since a co-adapted coupling strategy is allowed to depend at time t upon

$$\mathcal{F}_t = \sigma\left\{ \bigcup_{i,j} \Lambda_{ij}(s), \bigcup_{i,j} U_{ij}(s), Q(s) \,:\, s \leq t \right\},$$

we shall write $Q(t, A_t)$ whenever we wish to emphasise that $Q(t)$ depends not only upon t but also upon some event $A_t \in \mathcal{F}_t$.

Before proceeding to a discussion of optimal couplings, we introduce here the last of the notation to be used in this section. At time $t \geq 0$ define

$$U_t = \{1 \leq i \leq n \,:\, X_t(i) \neq Y_t(i)\}$$

to be the set of unmatched coordinates, and let M_t be the complement of U_t (that is, the set of matched coordinates at time t). Finally, for subsets V and W of $[0, n] = \{0, \ldots, n\}$, we define the following three functions:

$$\begin{aligned} \varphi_t(V, W) &= \sum_{i\in V} \sum_{j\in W} q_{ij}(t)\,, \\ \psi_t(V, W) &= \varphi_t(V, W) + \varphi_t(W, V)\,, \\ \text{and} \quad \Delta_t(V) &= \sum_{i\in V} q_{ii}(t)\,. \end{aligned}$$

When a subset $V \subset [1, n]$ is a single state x, we shall simply write $V = x$, rather than the more cumbersome $V = \{x\}$.

Note that ψ_t is symmetric in its two arguments, and that the identities

$$\varphi_t(0 \cup M_t, U_t) = |U_t| - \varphi_t(U_t, U_t) = \varphi_t(U_t, 0 \cup M_t) \tag{3.16}$$

$$\varphi_t(0 \cup U_t, M_t) = |M_t| - \varphi_t(M_t, M_t) = \varphi_t(M_t, 0 \cup U_t) \tag{3.17}$$

follow as a consequence of equations (3.14) and (3.15). This notation allows us to define concisely the partial-independence coupling of Example 1.14 by the two statements:

$$\varphi_t(0, U_t) = |U_t| \quad \text{and} \quad \Delta_t(M_t) = |M_t|\,, \quad \text{for all } t \geq 0\,.$$

(This forces $\varphi_t(U_t, 0) = |U_t|$, and so unmatched coordinates evolve independently while matched coordinates move synchronously.)

Now let us consider a general co-adapted coupling $c \in \mathcal{C}$. Due to the symmetry of $\mathbb{Z}_2^n$, any coupling scheme should clearly be invariant under permutation of the coordinates of the hypercube, and so we need only consider the hitting time at zero of the counting process $N_t = |U_t|$. Note that, since the Poisson processes Λ_{ij} are independent, the probability of two or more events occurring on the superimposed Poisson process $\bigcup \Lambda_{ij}$ in a time interval of length δ is $O(\delta^2)$.

Let N_t^c denote the state of the chain N at time t when coupling strategy $c \in \mathcal{C}$ is used over the period $[0, t]$. The hitting time of N^c at zero (which will be the coupling time of X and Y) is denoted τ^c. Since $\mathcal{C}$ only contains successful couplings,

$$\mathbb{P}(\tau^c < \infty) = \mathbb{P}(N_t^c = 0 \text{ for sufficiently large } t) = 1 \quad \text{for all } c \in \mathcal{C}.$$

The setup described above makes it possible to write down the infinitesimal transition rates for the chain N^c: for small $\delta > 0$,

$$\begin{aligned}
\mathbb{P}\left(N_{t+\delta}^c = N_t^c + 2\right) &= \delta\,(\varphi_t(M_t, M_t) - \Delta_t(M_t)) + o(\delta) \\
\mathbb{P}\left(N_{t+\delta}^c = N_t^c + 1\right) &= \delta\,\psi_t(0, M_t) + o(\delta) \\
\mathbb{P}\left(N_{t+\delta}^c = N_t^c - 1\right) &= \delta\,\psi_t(0, U_t) + o(\delta) \\
\mathbb{P}\left(N_{t+\delta}^c = N_t^c - 2\right) &= \delta\,(\varphi_t(U_t, U_t) - \Delta_t(U_t)) + o(\delta) \\
\mathbb{P}\left(N_{t+\delta}^c = N_t^c\right) &= 1 - \delta\,(\psi_t(0, [1, n]) + \varphi_t(U_t, U_t) + \varphi_t(M_t, M_t) - \Delta_t([1, n])) + o(\delta).
\end{aligned} \tag{3.18}$$

Suppose that $N_t^c = k$. We now argue that any coupling $c \in \mathcal{C}$ which has a positive chance of breaking two of the $n - k$ already-matched coordinates cannot be optimal. To see this, note that the only way in which two matches can be broken is if there is an event on one of the $n - k$ matched coordinates of X_t in time $[t, t + \delta)$. Thus the expected time taken by any coupling to break two matches is at least $(n - k)^{-1}$. This is greater than the time taken to break one match, which can be achieved in time $(n - k)^{-1}/2$ (by allowing matched coordinates to evolve independently). If two matches are broken then $N_{t+\delta}^c = k + 2$, and in order for the process N to reach zero it must almost surely pass through at least one of the states $\{k, k + 1\}$. However, both of these states can be reached faster from state k than by going via state $k + 2$, and so any coupling strategy c which allows for two matches to be broken cannot be

optimal. Therefore any candidate optimal coupling $c \in \mathcal{C}$ must satisfy

$$\varphi_t(M_t, M_t) = \Delta_t(M_t) \quad \text{for all } t \geq 0\,. \tag{3.19}$$

Let $\mathcal{C}' \subset \mathcal{C}$ be the set of successful co-adapted couplings satisfying equation (3.19).

It is not the case however that breaking single matches is necessarily a bad idea. Indeed, consider the following coupling strategy: if N is odd, do independence (not partial-independence) coupling until N becomes even, and then couple in pairs in the manner of the Aldous coupling of Example 1.12 ("when an unmatched bit moves on X, move a different unmatched bit on Y"). In terms of the coupling matrix description above, this is achieved by setting (for $m \geq 1$):

$$\varphi_t(U_t, U_t) - \Delta_t(U_t) = |U_t| \quad \text{and} \quad \Delta_t(M_t) = |M_t| \quad \text{when } N_t = 2m \tag{3.20}$$

$$\varphi_t(0, [1, n]) = \varphi_t([1, n], 0) = n \qquad \text{when } N_t = 2m - 1. \tag{3.21}$$

An alternative version of this coupling is the following: when N is even, proceed exactly as above (couple unmatched coordinates in pairs); but when N is odd, employ the partial-independence coupling, i.e. carry out independence coupling for the *unmatched* coordinates only. In our matrix notation:

$$\varphi_t(U_t, U_t) - \Delta_t(U_t) = |U_t| \quad \text{and} \quad \Delta_t(M_t) = |M_t| \quad \text{when } N_t = 2m \tag{3.22}$$

$$\varphi_t(0, U_t) = \varphi_t(U_t, 0) = |U_t| \quad \text{and} \quad \Delta_t(M_t) = |M_t| \quad \text{when } N_t = 2m - 1. \tag{3.23}$$

Unlike the coupling described above, this second scheme never breaks any matches (since $\Delta_t(M_t) = |M_t|$ for all t). However, a simple calculation shows that if $N_t = 1$ the coupling time distribution is identical under either strategy.

The main theorem of this section is the following:

Theorem 3.9. *A necessary and sufficient condition for a coupling $\hat{c}$ to be an optimal co-adapted coupling (with respect to minimising the expected coupling time) is that $\hat{c}$ corresponds to a matrix process $\{\hat{Q}(t)\}_{t\geq 0}$ satisfying equations* (3.22) *and* (3.23) *whenever $N_t \geq 2$. When $N_t = 1$ only the first part of* (3.23) *is necessary.*

Proof. By the discussion above, it suffices to restrict the search for an optimal co-adapted coupling to the set $\mathcal{C}'$. (The proof below can be modified to include all couplings in $\mathcal{C}$, but the observation that we can restrict to $\mathcal{C}'$ greatly simplifies the

computations.) First suppose that a coupling $\hat{c} \in \mathcal{C}'$ satisfies equations (3.22) and (3.23). For such a coupling it is simple to calculate the expected coupling time given the starting state N_0. Recall that if $N_0 = 2m$ is even then a coupling satisfying equations (3.22) and (3.23) proceeds by coupling unmatched coordinates in pairs. The coupling time in this instance is then given by

$$\tau^{\hat{c}} = \sum_{k=1}^{m} S_k,$$

where S_k is the time taken for an incident to occur on one of the Poisson processes Λ_{ij} with i, j both unmatched, given that $N_0 = 2k$. Such an incident occurs at rate $2k$ (using equation (3.22)) and so $\{S_k\}$ form a set of independent $Exp(2k)$ random variables. This yields

$$\mathbb{E}\left[\tau^{\hat{c}} \,|\, N_0 = 2m\right] = \sum_{k=1}^{m} \frac{1}{2k} \qquad (m \geq 1)\,. \tag{3.24}$$

Similarly, when $N_0 = 2m+1$ is odd, the coupling proceeds to couple an unmatched coordinate as fast as possible, and to then proceed as for $N_0 = 2m$. Thus

$$\mathbb{E}\left[\tau^{\hat{c}} \,|\, N_0 = 2m+1\right] = \frac{1}{2(2m+1)} + \sum_{k=1}^{m} \frac{1}{2k} \qquad (m \geq 0)\,. \tag{3.25}$$

Now define the value function v^* by

$$v^*(x) = \inf_{c \in \mathcal{C}'} \mathbb{E}\left[\tau^c \,|\, N_0 = x\right] = \inf_{c \in \mathcal{C}'} \mathbb{E}_x\left[\tau^c\right]\,. \tag{3.26}$$

The optimality of $\hat{c}$ will follow if it can be shown that $\hat{v} = v^*$, where $\hat{v}$ is defined (using equations (3.24) and (3.25)) by

$$\hat{v}(x) = \begin{cases} 0 & x = 0 \\ \sum_{k=1}^{m} \frac{1}{2k} & x = 2m\,(m \geq 1) \\ \frac{1}{2(2m+1)} + \sum_{k=1}^{m} \frac{1}{2k} & x = 2m+1\,(m \geq 0)\,. \end{cases} \tag{3.27}$$

With this aim in mind, define for a coupling $c \in \mathcal{C}'$:

$$\hat{V}_t^c = \int_0^t \mathbf{1}[N_s^c > 0]\; ds + \hat{v}(N_t^c)\,.$$

Since $\hat{v}(x)$ is non-negative and increasing in x, and $\mathcal{C}'$ contains only successful co-adapted couplings, it follows that

$$0 \leq \lim_{t\to\infty} \mathbb{E}_x\left[\hat{v}(N_t^c)\right] \leq \lim_{t\to\infty} \hat{v}(n)\mathbb{P}_x\left(N_t^c > 0\right) = 0\,,$$

and so

$$\lim_{t\to\infty} \mathbb{E}_x\left[\hat{V}_t^c\right] = \lim_{t\to\infty} \mathbb{E}_x\left[\int_0^t \mathbf{1}[N_s^c > 0]\; ds\right] + \lim_{t\to\infty} \mathbb{E}_x\left[\hat{v}(N_t^c)\right] = \mathbb{E}_x\left[\tau^c\right]. \tag{3.28}$$

Now suppose that it is possible to show that $\hat{V}^c$ is a submartingale for all $c \in \mathcal{C}'$ and starting states $x \in [0, n]$, and furthermore that $\hat{V}^{\hat{c}}$ is a martingale. It would then follow that

$$\hat{v}(x) = \hat{V}_0^c \leq \mathbb{E}_x\left[\hat{V}_t^c\right] \to \mathbb{E}_x\left[\tau^c\right] \quad \text{as } t \to \infty\,.$$

Since this holds for all c and x, we see that $\hat{v} \leq v^*$. Moreover, since $\hat{V}^{\hat{c}}$ is a martingale, it follows that $\hat{v}$ can be attained (using the coupling $\hat{c}$), and thus $\hat{v} \geq v^*$. This optimisation argument is known as Bellman's Principle (Krylov 1980), and yields the required proof of the optimality of $\hat{c}$. We therefore aim to prove that $\hat{V}^c$ is a submartingale for all $c \in \mathcal{C}'$ and starting states $x \in [0, n]$, and that $\hat{V}^{\hat{c}}$ is a martingale.

Consider then the following expected change in $\hat{V}^c$ over a small interval $[t, t+\delta)$, given that $\tau^c > t$:

$$\mathbb{E}_x\left[\hat{V}_{t+\delta}^c \,|\, N_t^c = k > 0, \mathcal{F}_t\right] - \hat{V}_t^c = \mathbb{E}_k\left[\int_0^\delta \mathbf{1}[N_s^c > 0]\; ds\right] + \mathbb{E}_x\left[\hat{v}(N_{t+\delta}^c) - \hat{v}(N_t^c) \,|\, N_t^c = k, \mathcal{F}_t\right]. \tag{3.29}$$

Now, using the transition rates in (3.18), along with the condition in equation (3.19) that is satisfied by any coupling $c \in \mathcal{C}'$, it follows that (ignoring all $o(\delta)$ terms)

$$\begin{aligned}\mathbb{E}_x\left[\hat{v}(N_{t+\delta}^c) - \hat{v}(N_t^c) \,|\, N_t^c = k, \mathcal{F}_t\right] = {}& \delta\,\psi_t(0, U_t)\hat{v}(k-1) \\ &+ \delta\,\psi_t(0, M_t)\hat{v}(k+1) \\ &+ \delta\,[\varphi_t(U_t, U_t) - \Delta_t(U_t)]\,\hat{v}(k-2) \\ &- \delta\,[\,\psi_t(0, [1, n]) + \varphi_t(U_t, U_t) - \Delta_t(U_t)]\,\hat{v}(k)\,.\end{aligned} \tag{3.30}$$

Combining equations (3.29) and (3.30) results in:

$$\begin{aligned}\lim_{\delta\downarrow 0} \frac{\mathbb{E}_x\left[\hat{V}_{t+\delta}^c \,|\, N_t^c = k > 0, \mathcal{F}_t\right] - \hat{V}_t^c}{\delta} = {}& 1 + \psi_t(0, U_t)\hat{v}(k-1) \\ &+ \psi_t(0, M_t)\hat{v}(k+1) \\ &+ [\varphi_t(U_t, U_t) - \Delta_t(U_t)]\,\hat{v}(k-2) \\ &- [\,\psi_t(0, [1, n]) + \varphi_t(U_t, U_t) - \Delta_t(U_t)]\,\hat{v}(k)\,.\end{aligned}$$

In order to prove that $\hat{V}^c$ is a submartingale for all $c \in \mathcal{C}'$ and a martingale for $\hat{c}$, it therefore suffices to show that

$$\begin{aligned}\inf_{c\in\mathcal{C}'}\{1 + \ &\psi_t(0,U_t)\hat{v}(k-1) \\ &+ \ \psi_t(0,M_t)\hat{v}(k+1) + [\varphi_t(U_t,U_t) - \Delta_t(U_t)]\,\hat{v}(k-2) \\ &- [\,\psi_t(0,[1,n]) + \varphi_t(U_t,U_t) - \Delta_t(U_t)]\,\hat{v}(k)\} = 0\,,\end{aligned}$$

with the infimum being attained at $c = \hat{c}$. That is,

$$\inf_{c\in\mathcal{C}'} L_t^c(\hat{v}) = 0 \quad \text{for all } t \geq 0\,, \tag{3.31}$$

where (rearranging some terms)

$$\begin{aligned}L_t^c(\hat{v}) = 1 &- [\varphi_t(U_t,U_t) - \Delta_t(U_t)]\,(\hat{v}(k) - \hat{v}(k-2)) \\ &+ \psi_t(0,M_t)\,(\hat{v}(k+1) - \hat{v}(k)) \\ &- \psi_t(0,U_t)\,(\hat{v}(k) - \hat{v}(k-1))\ .\end{aligned} \tag{3.32}$$

Since the function $\hat{v}$ depends upon the parity of N (recall equations (3.24) and (3.25)), we consider two different situations in order to show this.

Case 1: $\mathbf{N_t^c = 2m}$ $(m \geq 1)$

In this case we obtain the following expression:

$$\begin{aligned}L_t^c(\hat{v}) &= 1 - [\varphi_t(U_t,U_t) - \Delta_t(U_t)]\,(\hat{v}(2m) - \hat{v}(2m-2)) \\ &\quad + \psi_t(0,M_t)\,(\hat{v}(2m+1) - \hat{v}(2m)) \\ &\quad - \psi_t(0,U_t)\,(\hat{v}(2m) - \hat{v}(2m-1)) \\ &= 1 - [\varphi_t(U_t,U_t) - \Delta_t(U_t)]\,\frac{1}{2m} + \psi_t(0,M_t)\frac{1}{2(2m+1)} \\ &\quad - \psi_t(0,U_t)\frac{(m-1)}{2m(2m-1)}\,,\end{aligned} \tag{3.33}$$

by definition of $\hat{v}$.

Now, from equation (3.16) it follows that

$$\begin{aligned}\varphi_t(U_t,U_t) - \Delta_t(U_t) &\leq \varphi_t(U_t,U_t) \\ &= N_t^c - \varphi_t(U_t, 0 \cup M_t) \\ &\leq N_t^c - \frac{\psi_t(0,U_t)}{2}\,.\end{aligned} \tag{3.34}$$

Since $N_t^c = 2m$ and ψ_t is non-negative by definition, using inequality (3.34) to bound the second term of equation (3.33) yields

$$\begin{aligned} L_t^c(\hat{v}) &\geq \psi_t(0, U_t) \left(\frac{1}{4m} - \frac{(m-1)}{2m(2m-1)} \right) \\ &= \frac{\psi_t(0, U_t)}{4m(2m-1)} \geq 0 \,. \end{aligned} \tag{3.35}$$

Furthermore, it follows from equation (3.22) that the $\hat{Q}(t, N_t^{\hat{c}} = 2m)$ matrix for $\hat{c}$ satisfies the following:

$$\varphi_t(0, [1, n]) = \varphi_t([1, n], 0) = \Delta_t(U_t) = 0 \quad \text{and} \quad \varphi_t(U_t, U_t) = 2m \,.$$

Inserting these values into equation (3.33) we see that $L_t^{\hat{c}}(\hat{v}) = 0$, as required.

Case 2: $\mathbf{N_t^c = 2m + 1}$

We first consider the situation $m \geq 1$. In this case we obtain

$$\begin{aligned} L_t^c(\hat{v}) = 1 &- \left[\varphi_t(U_t, U_t) - \Delta_t(U_t) \right] \frac{4m^2 - 2m - 1}{2m(2m-1)(2m+1)} \\ &+ \psi_t(0, M_t) \frac{m}{2(m+1)(2m+1)} - \frac{\psi_t(0, U_t)}{2(2m+1)} \,. \end{aligned} \tag{3.36}$$

Proceeding as in Case 1, inequality (3.34) may be used to bound the second term of equation (3.36):

$$\begin{aligned} L_t^c(\hat{v}) &\geq \frac{1}{2m(2m-1)} - \psi_t(0, U_t) \left(\frac{1}{2(2m+1)} - \frac{(4m^2 - 2m - 1)}{4m(2m-1)(2m+1)} \right) \\ &= \frac{1}{2m(2m-1)} - \frac{\psi_t(0, U_t)}{4m(2m-1)(2m+1)} \geq 0 \,, \end{aligned}$$

since $\psi_t(0, U_t) = \varphi_t(0, U_t) + \varphi_t(U_t, 0) \leq 2 \left| U_t \right| = 2(2m+1)$.

When $m = 0$, equation (3.32) must be modified slightly, to allow for the fact that it is no longer possible to make two new matches. The appropriate equation reads

$$\begin{aligned} L_t^c(\hat{v}) &= 1 - \psi_t(0, U_t) \left(\hat{v}(1) - \hat{v}(0) \right) + \psi_t(0, M_t) \left(\hat{v}(2) - \hat{v}(1) \right) \\ &= 1 - \frac{\psi_t(0, U_t)}{2} \\ &\geq 0 \,, \end{aligned} \tag{3.37}$$

since $\psi_t(0, U_t) \leq 2(2m+1) = 2$, by definition.

Finally, recall from equation (3.23) that the $\hat{Q}(t, N_t^{\hat{c}} = 2m+1)$ matrix for $\hat{c}$ satisfies the following:

$$\varphi_t(0, U_t) = \varphi_t(U_t, 0) = 2m+1$$
$$\text{and} \quad \varphi_t(0, M_t) = \varphi_t(M_t, 0) = \varphi_t(U_t, U_t) = 0\,.$$

Inserting these values into equations (3.36) and (3.37) we see that $L_t^{\hat{c}}(\hat{v}) = 0$ for all $m \geq 0$, as required.

It has therefore been shown that $\hat{V}^c$ is a submartingale for all couplings $c \in \mathcal{C}'$, and that $\hat{V}^{\hat{c}}$ is a martingale. This proves that any coupling $\hat{c}$ satisfying equations (3.22) and (3.23) is an optimal co-adapted coupling.

The proof of the other half of the theorem now follows simply. It has been proved above that $\hat{v} = v^*$, and so for a coupling $c \in \mathcal{C}'$ to be optimal it must satisfy $L_t^c(\hat{v}) = 0$: consideration of the calculations in Cases 1 and 2 shows that equations (3.22) and (3.23) are necessary for this to hold. For example, consider equation (3.33), which holds when $N_t^c = 2m \geq 2$:

$$\begin{aligned} L_t^c(\hat{v}) = 1 - [\varphi_t(U_t, U_t) - \Delta_t(U_t)]\,\frac{1}{2m} + \psi_t(0, M_t)\frac{1}{2(2m+1)} \\ - \psi_t(0, U_t)\frac{(m-1)}{2m(2m-1)}\,. \end{aligned}$$

Firstly, inequality (3.35) shows that $\psi_t(0, U_t) = 0$ is necessary for $L_t^c(\hat{v}) = 0$. For such a coupling c, the above expression reduces to

$$L_t^c(\hat{v}) = 1 - [\varphi_t(U_t, U_t) - \Delta_t(U_t)]\,\frac{1}{2m} + \psi_t(0, M_t)\frac{1}{2(2m+1)}\,.$$

But since Δ_t and ψ_t are non-negative, and $\varphi_t(U_t, U_t) \leq |U_t| = N_t^c = 2m$, this means that

$$\varphi_t(U_t, U_t) - \Delta_t(U_t) = 2m \quad \text{and} \quad \psi_t(0, M_t) = 0$$

are also both necessary. These conditions necessarily force $\varphi_t(U_t, M_t) = \varphi_t(M_t, U_t) = 0$. Combining all these necessary conditions yields

$$\varphi_t(U_t, U_t) - \Delta_t(U_t) = |U_t| \quad \text{and} \quad \Delta_t(M_t) = |M_t| \quad \text{when } N_t^c = 2m,$$

which is exactly the condition presented in equation (3.22). A similar argument shows that equation (3.23) is necessary when $N_t^c = 2m+1$ for $m \geq 1$. Note however

that, by equation (3.37), only the first half of (3.23) is necessary when $N_t^c = 1$. This confirms the comment made before this theorem, where it was remarked that the same expected coupling time could be obtained by allowing matched coordinates to become unmatched when $N_t = 1$. □

A direct consequence of this theorem is the following:

Corollary 3.10. *The expected coupling time of any co-adapted coupling for the simple random walk on $\mathbb{Z}_2^n$, when each coordinate moves at rate $1/n$, is asymptotically bounded below by $(n/2)\log n$.*

Proof. The optimality of the coupling $\hat{c}$ with respect to minimisation of the expected coupling time was proved in Theorem 3.9. Equations (3.24) and (3.25) show that, when each coordinate moves at rate 1, $\mathbb{E}\left[\tau^{\hat{c}}\right] \sim (\log n)/2$ for large n. □

Note that, in the proof of Theorem 3.9, only Markovian couplings $c \in \mathcal{C}'$ are considered. That is, the control $Q(t)$ depends only upon the value of N_t^c, even though as a co-adapted coupling it is allowed to depend upon any events in $\mathcal{F}_t$. However, this apparent restriction does not affect the validity of the proof, since the driving processes Λ_{ij} are Markovian. Furthermore, the cost function in question (the expected remaining time to couple) is also independent of the past, given N_t^c. It is therefore sufficient to consider only Markovian couplings in the search for maximality (see Krylov (1980), Chapter 1).

This concludes the current investigation into random walks on the hypercube. Directions for further research in this area will be discussed in Section 3.5. In the next section, we turn our attention to two diffusions on $\mathbb{R}^d$, and investigate how good co-adapted couplings can possibly be for these continuous-time processes.

3.4 Maximal coupling for Brownian motion and the O-U process

Consider a random walk $S = \{S_n\}$ on $\mathbb{R}$, with step distribution F. Rogers (1999) considers the problem of coupling a pair of these walks S and S', with $S_0 = 0$ and $S'_0 = a > 0$. He first constructs a pair of random variables (X, Y) such that

(i) $X \leq Y$ almost surely;

(ii) X has law F and Y has law $F_a = \delta_a * F$;

(iii) $\mathbb{E}[\varphi(X-Y)]$ is maximal for any non-negative decreasing convex function φ.

(Note the similarity of item (iii) to the optimal transport problem outlined at the end of Section 3.1.) This provides a maximal coupling of (S_1, S_1'). The monotonicity of the coupling in item (i) is essential for the optimality in item (iii) to hold for all suitable functions φ. Iteration of the above construction then provides a co-adapted coupling of the random walks S and S', with each step being locally maximal. However, in the case where F is unimodal this local optimality turns out to hold globally, and so the coupling is both co-adapted and maximal. The random walks so constructed satisfy $S_n \leq S_n'$ for all $n \geq 0$.

The proof of Rogers' theorem provides an explicit form for the joint law of the random variables (X, Y) above. In the case where F is unimodal, with a density f which is symmetric about zero, this joint law takes on a particularly simple form:

$$\mathbb{P}(X \in dx, Y = X) = (f(x) \wedge f(x-a))\,dx \qquad \text{for } x \in \mathbb{R}, \tag{3.38}$$

$$\mathbb{P}(X \in dx, Y = a - X) = (f(x) - f(x-a))\,dx \qquad \text{for } x \leq a/2. \tag{3.39}$$

Thus we see that the one-step coupling in this case is simply a particular instance of the Vasershtein coupling: X and Y are made to agree with as large a probability as possible, else they are chosen from the residual densities, but now with Y being the reflection of X about $a/2$.

As a scaling limit of a symmetric unimodal random walk, it is to be expected that a similar result should hold for the coupling of Brownian motions on $\mathbb{R}$. A discussion of this problem, including its application to the coupling of other diffusions such as the Ornstein-Uhlenbeck (O-U) process, is the main subject of this section.

3.4.1 Maximal coupling for Brownian motion

Define

$$p_t(x,y) = \frac{e^{-|x-y|^2/2t}}{\sqrt{2\pi t}}$$

to be the Gaussian heat kernel on $\mathbb{R}$. Let X and Y be Brownian motions on $\mathbb{R}$, with $(X_0, Y_0) = (x, 0)$ for some fixed state $x \in \mathbb{R}$. The following results all hold for

d-dimensional Euclidean Brownian motion, $1 \leq d < \infty$, but for simplicity of notation we shall always set $d = 1$ throughout this section.

The symmetry of the heat kernel means that the reflection of Y around zero, $-Y$, is also a Brownian motion of course, and this suggests a simple method for coupling the two processes: simply consider the process $(x - B, B)$, where B is a standard Brownian motion started at zero. This process is equal in distribution to (X, Y), and by the path continuity of B the coupling time is equal to the first hitting time of B at the level $x/2$. It is a standard result that this hitting time, $\tau_{x/2}$, satisfies

$$\mathbb{P}\left(\tau_{x/2} \geq t\right) = \sqrt{\frac{2}{\pi t}} \int_0^{x/2} e^{-u^2/2t} du = \mathrm{Erf}\left(\frac{x/2}{\sqrt{2t}}\right), \tag{3.40}$$

and a simple calculation shows that this tail distribution agrees with the total variation distance between the laws of X_t and Y_t. This reflection coupling is therefore a co-adapted maximal coupling. There do exist other maximal couplings for Euclidean Brownian motion, but these all turn out to be non-co-adapted:

Theorem 3.11 (Hsu and Sturm)**.** *Let $x, y \in \mathbb{R}^d$. The reflection coupling is the unique co-adapted maximal coupling of d-dimensional Brownian motions X and Y, where $(X_0, Y_0) = (x, y)$.*

The proof of this theorem proceeds by showing that the reflection coupling of equations (3.38) and (3.39) applied to $f(z) = p_t(x, z)$, with $a = |x - y|/2$, is the *unique* maximal coupling of $p_t(x, z)$ and $p_t(y, z)$ with respect to the Vasershtein distance. The result then follows from the path continuity of Brownian motion.

Theorem 3.11 was recently generalised a little in Kuwada (2006). This paper considers diffusions on more general spaces, but only ones for which there is a specific 'reflection structure'. More precisely, the requirement is that Z is a diffusion on a space $\mathcal{X}$ such that the following two properties hold for any fixed $x, y \in \mathcal{X}$:

1. There is a continuous map $R : \mathcal{X} \to \mathcal{X}$ with $R \circ R = id$, such that $\mathcal{L}(Z^x) \circ R^{-1} = \mathcal{L}(Z^y)$;

2. The set of fixed points $H = \{x \in \mathcal{X} : R(x) = x\}$ separates $\mathcal{X}$ into two disjoint open sets $\mathcal{X}_1$ and $\mathcal{X}_2$, with $R(\mathcal{X}_1) = \mathcal{X}_2$.

This structure means that, as for Brownian motion in $\mathbb{R}^d$, a natural reflection coupling exists, and this is again a co-adapted maximal coupling. The main purpose of

the paper by Kuwada (2006) is to investigate when this reflection coupling is unique among co-adapted maximal couplings. A sufficient (but unnecessary) condition for this uniqueness to hold is given: this condition is not particularly pleasant, but the result does show that the uniqueness (among co-adapted couplings) of reflection coupling for Brownian motions also holds when the processes are defined on a complete Riemannian manifold.

Let us return now to the case of Brownian motion on $\mathbb{R}$. Suppose that, instead of fixing starting states x and y for X and Y, we let $X_0 \sim \mu$ and $Y_0 \sim \mu'$. Does there exist a co-adapted maximal coupling for these processes? As pointed out by Hsu and Sturm, it is not clear that such a coupling does always exist, but the answer *is* positive in certain situations where the unique minimisers of the Vasershtein distance

$$d_V^{\phi}(\mu, \mu') = \inf_{(\hat{\mu}, \hat{\mu}')} \int_{\mathbb{R}\times\mathbb{R}} \phi\left(|x-y|\right) \hat{\mu}(dx)\hat{\mu}'(dy)$$

(where $(\hat{\mu}, \hat{\mu}')$ is a coupling of μ and μ') are independent of the choice of strictly concave function ϕ. This holds, for example, if $(\mu - \mu')^+$ is supported on a half space and $(\mu - \mu')^-$ is the reflection of $(\mu - \mu')^+$ in the other half space (Hsu and Sturm). Thus if μ and μ' are Dirac point masses, or two Gaussian distributions of equal variance, it is possible to produce a co-adapted maximal coupling by drawing (X_0, Y_0) using the Vasershtein coupling of (μ, μ') and then using the reflection coupling as before.

What happens, however, if we now mix these two examples and fix $X_0 = x$ but draw $Y_0 \sim N(0, \sigma^2)$? The following (new) result shows that reflection coupling for these two processes is no longer maximal.

Lemma 3.12. *Fix $X_0 = x \in \mathbb{R}$ and let $Y_0 \sim N(0, \sigma^2)$. Then the reflection coupling for the pair of Brownian motions (X, Y) is not maximal.*

Proof. Let T^R and T^* be the coupling times under reflection and maximal couplings respectively. The tail distribution for T^R is given by averaging the tail distribution in equation (3.40) over the possible values of Y_0:

$$\mathbb{P}\left(T^R > t\right) = \int_{-\infty}^{\infty} p_{\sigma^2}(0, y) \int_{-\infty}^{\infty} \left(\left(p_t(x, z) - p_t(y, z)\right) \vee 0\right) \, dzdy \,. \tag{3.41}$$

On the other hand, the tail distribution for T^* can be calculated using the formula

for total variation distance given in equation (1.1):

$$\mathbb{P}\left(T^* > t\right) = \int_{-\infty}^{\infty} \left(\left(p_t(x,z) - p_{\sigma^2+t}(0,z)\right) \vee 0\right)\, dz\,. \tag{3.42}$$

Now condition upon the state z at which coupling takes place, and consider the difference between equations (3.42) and (3.41) under this conditioning:

$$\begin{aligned}
&\left(\left(p_t(x,z) - p_{\sigma^2+t}(0,z)\right) \vee 0\right) - \int_{-\infty}^{\infty} p_{\sigma^2}(0,y)\left(\left(p_t(x,z) - p_t(y,z)\right) \vee 0\right)\, dy \\
&\quad = \left(\int_{-\infty}^{\infty} p_{\sigma^2}(0,y)\left(p_t(x,z) - p_t(y,z)\right)\, dy\right) \vee 0 \\
&\qquad\qquad - \int_{-\infty}^{\infty} p_{\sigma^2}(0,y)\left(\left(p_t(x,z) - p_t(y,z)\right) \vee 0\right)\, dy \\
&\quad = \left(\mathbb{E}\left[f_{t,z}(Y_0)\right] \vee 0\right) - \mathbb{E}\left[f_{t,z}(Y_0) \vee 0\right]\,,
\end{aligned} \tag{3.43}$$

where $f_{t,z}(y) = p_t(x,z) - p_t(y,z)$,

$$\begin{aligned}
&= \Big(\left(\mathbb{E}\left[f_{t,z}(Y_0) \vee 0\right] - \mathbb{E}\left[(-f_{t,z}(Y_0)) \vee 0\right]\right) \vee 0\Big) - \mathbb{E}\left[f_{t,z}(Y_0) \vee 0\right] \\
&< 0
\end{aligned} \tag{3.44}$$

since the random variable $\operatorname{sgn}(f_{t,z}(Y_0))$ is not constant.

Thus we see that there is a conditional deficit at location z between the tail distributions of the two coupling times. Since this deficit holds for all $z \in \mathbb{R}$ and $t \geq 0$, it follows that reflection coupling is not maximal. □

Of course, this proof breaks down when Y_0 is deterministic, since then the random variable $\operatorname{sgn}(f_{t,z}(Y_0))$ is constant.

Although reflection coupling is not maximal when the starting state for Y is randomised as above, it is clear that reflection is an optimal co-adapted coupling for X and Y when Y_0 is randomised using *any* distribution. This follows from the observation that any co-adapted coupling must be conditioned at time zero upon the σ-algebra $\mathcal{F}_0 = \sigma\left\{X_s, Y_s \,:\, s \leq 0\right\}$. In particular, the coupling scheme at time zero is conditioned on the event $\{Y_0 = y_0\}$. So the best that any co-adapted coupling can do is to match the coupling time of a maximal coupling between X and Y when $(X_0, Y_0) = (x_0, y_0)$, integrated over the distribution of Y_0. This bound is achieved by the reflection coupling (recall equation (3.41)), making it an optimal co-adapted coupling, as claimed.

Furthermore, conditional on $(X_0, Y_0) = (x_0, y_0)$, we have seen that reflection coupling is the *unique* optimal co-adapted coupling of X and Y (Hsu and Sturm). This observation, combined with the above argument, proves the following corollary:

Corollary 3.13. *The reflection coupling is the unique optimal co-adapted coupling for a pair of d-dimensional Brownian motions (X, Y). In particular, this result holds whatever the initial distributions of X_0 and Y_0.*

This brief investigation into reflection coupling for Euclidean Brownian motions has highlighted some, perhaps surprising, results concerning the differences between co-adapted and non-co-adapted couplings. In particular, we have seen that, depending upon the distribution of (X_0, Y_0), it is possible for an optimal co-adapted, but not a co-adapted maximal coupling to exist. The original motivation for this study was provided by an interest in the coupling time for two Ornstein-Uhlenbeck processes, and it is this topic which forms the focus of the final section of this chapter.

3.4.2 Maximal coupling for the O-U process

Consider an Ornstein-Uhlenbeck process on $\mathbb{R}$. This is the unique solution of the following stochastic differential equation:

$$dX_t = -\alpha X_t \, dt + \sigma\sqrt{2\alpha} \, dB_t, \quad X_0 = x_0 \, . \tag{3.45}$$

Here α and σ are positive constants, and $(B_t)_{t\geq 0}$ is a standard Brownian motion. It is simple to show that the distribution of X_t is Gaussian, with

$$\mathbb{E}\left[X_t\right] = e^{-\alpha t}x \quad \text{and} \quad Var(X_t) = \sigma^2\left(1 - e^{-2\alpha t}\right) \, .$$

This process converges to its stationary distribution, $N(0, \sigma^2)$, as $t \to \infty$.

Now suppose that X is an O-U process starting at x and Y is an O-U process with the same parameters α and σ but started at some fixed value $y \in \mathbb{R}$. These may be coupled as follows. Define the process $\hat{Y}$ by

$$d\hat{Y}_t = -\alpha \hat{Y}_t \, dt - \sigma\sqrt{2\alpha} \, dB_t, \quad \hat{Y}_0 = y,$$

where the noise component B_t is the same as that used in the definition of X. Clearly $\hat{Y}$ has the correct transition kernel to be an O-U process started at y.

Lemma 3.14. *For fixed starting values $x, y \in \mathbb{R}$, the reflection coupling defined above is a maximal coupling. That is,*

$$\|\mathcal{L}(X_t) - \mathcal{L}(Y_t)\| = \mathbb{P}\left(T^R(x,y) > t\right), \tag{3.46}$$

where $T^R(x,y)$ is the reflection coupling time of X and Y.

This (unsurprisingly) is the same result as that which holds for Brownian motion (contained within Theorem 3.11). A direct proof is easy, since the distribution of the hitting time at zero of the process $X - \hat{Y}$ is well known. However, in order to apply the other results of the previous section to O-U processes, it is more convenient to rewrite the O-U process X as a time- and space-change of a Brownian motion. This may be done as follows (Chaumont and Yor 2003). Let $\tilde{B}_x = \left\{\tilde{B}_x(t)\right\}_t$ be a Brownian motion on $\mathbb{R}$ with $\tilde{B}_x(0) = x$. Then it is simple to check that the following process has the same distribution as X:

$$\tilde{X}(t) = e^{-\alpha t}\tilde{B}_x\left(\sigma^2\left(e^{2\alpha t} - 1\right)\right), \quad \tilde{X}(0) = x\,. \tag{3.47}$$

In order to couple $\tilde{X}$ with another O-U process $\tilde{Y}$, it is thus sufficient to consider the difference process

$$\tilde{X}(t) - \tilde{Y}(t) = e^{-\alpha t}\left[\tilde{B}_x\left(\sigma^2\left(e^{2\alpha t} - 1\right)\right) - \tilde{\tilde{B}}_y\left(\sigma^2\left(e^{2\alpha t} - 1\right)\right)\right], \tag{3.48}$$

where $\tilde{\tilde{B}}$ is the Brownian motion driving $\tilde{Y}$. The fundamental point to note in the above expression is that the time- and space-change is identical for $\tilde{X}$ and $\tilde{Y}$. Therefore in order to couple these two processes it is sufficient to consider couplings of the time changed Brownian motions $\tilde{B}_x\left(\sigma^2\left(e^{2\alpha t} - 1\right)\right)$ and $\tilde{\tilde{B}}_y\left(\sigma^2\left(e^{2\alpha t} - 1\right)\right)$. But this, of course, is equivalent to coupling two un-time-changed Brownian motions (since the time-change is the same for both processes), and this is exactly the problem studied in the previous section. After making this sequence of observations, the following collection of results follows easily.

Theorem 3.15. *Let X and Y be two Ornstein-Uhlenbeck processes in $\mathbb{R}^d$. Then the following results all hold.*

1. *if $(X_0, Y_0) = (x, y)$ is deterministic:*

   (i) reflection coupling is the unique co-adapted maximal coupling of X and Y;

(ii) the reflection coupling time satisfies

$$T^R(x,y) = \inf\left\{t \geq 0 \ : \ X_t = e^{-\alpha t}\left(\frac{x+y}{2}\right)\right\}.$$

2. *if $X_0 = x$ and $Y_0 \sim N(0,\sigma^2)$ (the equilibrium distribution of X):*

   (i) reflection coupling is no longer maximal;

   (ii) however, reflection coupling is *the unique optimal co-adapted coupling of X and Y.*

3.5 Future work

The construction of a maximal coupling for two Markov chains due to Pitman (Section 3.1.2) was based on the idea of stitching the chains together using randomised stopping times. RSTs are also used in Greven (1987): given initial distributions μ and μ' for two Markov chains X and X' with common transition kernel, this paper constructs RSTs S of X and T of X' such that $\mathcal{L}(X_S) = \mathcal{L}(X'_T)$, with S and T both finite. Furthermore, S and T may be constructed such that the chains $\{X_n \ : \ n < S\}$ and $\{X'_n \ : \ n < T\}$ almost surely live in disjoint regions of the state space. Such a coupling clearly exhibits nice spatial behaviour: it would be of interest to investigate when this construction also yields a maximal coupling.

More generally, a better understanding of the relationship between maximal coupling and any structural properties of the state space would be useful. For example, consider the maximal coalescent coupling for N chains, produced in the proof of Theorem 3.7, for a finite state space $\mathcal{X}$ with $|\mathcal{X}| = N$. Suppose that this has the property that the distribution of the coupling time is equal to that of the maximal coupling for chains started from two specific states x and y. Does this imply the existence of a (partial) ordering on $\mathcal{X}$ (with x and y the extremal states)?

In Section 3.3 an optimal control approach was used to prove the optimality of a co-adapted coupling for the random walk on $\mathbb{Z}_2^n$. This was carried out with respect to minimising the expected coupling time. It would be of interest to develop this analysis further by considering other optimality criteria. For example, it is not too hard to show that the coupling $\hat{c}$ defined by equations (3.22) and (3.23) also maximises the function

$$\mathbb{E}_x\left[e^{-\lambda T^c}\right], \quad \text{for all } \lambda > 0\,.$$

This observation motivates the following conjecture:

Conjecture 3.16. *The coupling $\hat{c}$ defined by equations* (3.22) *and* (3.23) *results in a coupling time $\hat{T}$ which is stochastically smaller than that produced by any other coupling $c \in \mathcal{C}$.*

Note that proof of this conjecture would also prove the existence of a coupling-pre-cutoff for $\hat{c}$ when each coordinate moves at rate $1/n$: this follows from the existence of a total variation cutoff at time $(n/4)\log n$, and Proposition 2.7 which proved the existence of a $(n/2)\log n$-coupling-cutoff for the partial-independence coupling. It would also be interesting to investigate the existence and properties of an optimal co-adapted coupling for the random walk on $\mathbb{Z}_2^n$ where each coordinate moves at a different rate (as studied in Chapter 2). This is likely to be much harder than a proof of the above conjecture however.

Finally, it was shown in Section 3.4.2 that some interesting results concerning maximal couplings of O-U processes follow easily from similar work on Brownian motion. In particular, reflection coupling is not maximal when the starting state of one process is randomised. Consider now a general diffusion X on $\mathbb{R}^d$ driven by a d-dimensional Brownian motion B:

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dB_t\,.$$

It would be interesting to investigate the range of such processes to which a similar analysis may be applied. Reflection coupling of multidimensional diffusions was studied by Lindvall and Rogers (1986): they give conditions on the drift and volatility which ensure that a reflection coupling is successful, but do not address the issue of maximality. The proof of Theorem 3.15 relies on the representation of an O-U process as a time- and space- change of a Brownian motion (equation (3.47)), where these changes are independent of the starting state. Such a representation will not be available for a general diffusion X, and so an alternative approach will be needed. In the light of the results in Section 3.4, it seems natural to conjecture that the reflection coupling will not be maximal for any diffusion driven by a Brownian motion when the starting state of one process is non-deterministic.

Dark Helmet: *What happened to then?*
Colonel Sandurz: *We passed then.*
Dark Helmet: *When?*
Colonel Sandurz: *Just now. We're at now, now.*
Dark Helmet: *Go back to then!*
Colonel Sandurz: *When?*
Dark Helmet: *Now.*
Colonel Sandurz: *Now?*
Dark Helmet: *Now!*
Colonel Sandurz: *I can't.*
Dark Helmet: *Why?*
Colonel Sandurz: *We missed it.*
Dark Helmet: *When?*
Colonel Sandurz: *Just now.*
Dark Helmet: *When will then be now?*
Colonel Sandurz: *Soon.*

Spaceballs

4. AN INTRODUCTION TO PERFECT SIMULATION

A common requirement of many problems in a variety of disciplines (such as stochastic geometry, Bayesian inference, statistical physics and computer science) is for a sample to be drawn from some probability distribution π. It is often the case that direct methods for doing this do not apply, or are infeasible: for example, the normalisation constant may be inaccessible. One possible solution to this problem is to use a Markov chain Monte Carlo (MCMC) algorithm. Such an approach involves the design of an ergodic Markov chain X which has π as its stationary distribution, and then running a simulation of X until it is near equilibrium.

Since its introduction in Metropolis et al. (1953), MCMC has proved to be an area of much interest, both theoretically and practically, and it is now a routine simulation technique in the researcher's toolbox. However, an obvious drawback with MCMC is that the user running the simulation does not know how many steps of the algorithm are needed for X to be close to equilibrium, and so this decision is ultimately up to the user. Of course, if the wrong choice is made then the algorithm may return a sample from a distribution that is far from π. A number of methods have been proposed for determining this *burn-in* time. These range from simple observation of algorithm output (such as autocovariance plots), to more analytical techniques involving bounds obtained from coupling, eigenvalue analysis or Foster-Lyapunov drift conditions.

An attractive alternative to these solutions, however, is to adapt the MCMC algorithm to form what is known as a *perfect simulation* algorithm. Such a procedure has two very desirable features:

1. the algorithm determines for itself when it should stop;
2. if the algorithm is successful then it returns a sample drawn *exactly* from the stationary distribution π.

The first paper showing how the above can be done in practice was that of Propp

and Wilson (1996). They introduced an algorithm known as Coupling from the Past (CFTP), and showed how to use such an algorithm in practice by drawing from the exact equilibrium distribution of the critical Ising model. In this chapter we describe the CFTP algorithm and note some of the common problems associated with its implementation. Some variations on this algorithm are then discussed. In section 4.2 the ideas of read-once CFTP, small-set CFTP and dominated CFTP will be introduced: all of these are necessary background material for the subsequent work of this and the final chapter. Finally, Sections 4.3 and 4.4 will discuss the efficiency and theoretical limitations of CFTP. Much of the discussion of the classic and dominated CFTP algorithms is based on the material in Connor (2007).

4.1 Coupling from the Past (CFTP)

Before launching into a description of the CFTP algorithm, we begin with some background information regarding stochastic recursive sequences and coupling.

4.1.1 Stochastic recursive sequences

Let X be a Markov chain with state space $\mathcal{X}$ and transition kernel P. The idea of a stochastic recursive sequence (SRS) (also called a randomising operation (Wilson 2000b)) is that the transitions of X can be defined using an i.i.d sequence of uniform random variables and a deterministic function f. Thus all the structure of the transition kernel P can be placed in f, making analysis much simpler. More specifically, it is possible to construct a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, an i.i.d. sequence $\{\xi_n\}_{n=-\infty}^{\infty}$ of $\mathit{Uniform}[0,1]$ random variables, and a measurable function $f : \mathcal{X} \times [0,1] \to \mathcal{X}$ such that X satisfies the recursion

$$X_0 = x_0, \quad X_{n+1} = f(X_n, \xi_n), \quad n \geq 0\,,$$

and X has transition kernel $P(x, \cdot)$ (Borovkov and Foss 1993). There are of course many different possible SRS constructions for any given chain X depending, for example, on the joint specification of $f(x,u)$ and $f(y,u)$ for different x and y.

The SRS representation provides a useful way to couple chains X and X' such that the two chains stay together for all time once they agree. Define updates for X

and X' by

$$X_{n+1} = f(X_n, \xi_n), \quad X'_{n+1} = f(X'_n, \xi_n)\,,$$

and let $T = \inf\{n \geq 0 \,:\, X_n = X'_n\}$. Then it is clear that $X_n = X'_n$ for all $n \geq T$, and that (assuming X is ergodic) T is almost surely finite. The beauty of the SRS representation however, is that it enables the coupling of any number of chains in this way, using the same f and the same randomness $\{\xi_n\}$ for all the chains. For this reason, stochastic recursive sequences are also called *grand couplings*. As we shall see, these grand couplings are exactly what is needed in the context of CFTP.

4.1.2 The CFTP algorithm

Coupling from the Past is now ten years old, and so many descriptions of the algorithm have been published in this period that the concept is now well known. In this section we give a short description of the algorithm - the reader is referred to any of the following references for further discussion: Propp and Wilson (1996), Wilson (2000a), Häggström (2002), Kendall (2005).

The basic concept behind CFTP is the following: consider a (hypothetical) copy of the chain of interest, $\tilde{X}$, which has been running since time $-\infty$, and which is in equilibrium at time zero, i.e. $\tilde{X}_0 \sim \pi$. This would be of obvious use if we could determine $\tilde{X}_0$, but clearly we cannot run a chain from time $-\infty$ in practice! However, it may be possible to determine $\tilde{X}_0$ by looking back only a *finite* number of steps into the past of $\tilde{X}$: it is this idea that is fundamental to CFTP.

To fully describe the algorithm, recall the SRS representation above. Suppose that we have available to us an i.i.d. sequence $\{\xi_n\}_{-\infty}^{0}$ of *Uniform*$[0,1]$ random variables and a deterministic update function f. For $v \geq -u$, define the random *input-output* map $F_{(-u,v]}(x) : \mathcal{X} \to \mathcal{X}$ by

$$F_{(-u,v]}(x) = f\left(f\left(\ldots f(f(x, \xi_{-u}), \xi_{-u+1})\ldots, \xi_{v-2}\right), \xi_{v-1}\right)\,. \tag{4.1}$$

Thus if X is begun at time $-u$ with the value $X_{-u} = x$, then we may set $X_v = F_{(-u,v]}(x)$. We also write $X_v^{x,-u}$ for the value of the chain X at time v, when X is started at time $-u$ from state x.

Definition 4.1. For a given update function f, the *backwards coalescence time* T^* of the chain X is defined by

$$\begin{aligned} T^* &= \min\left\{n \geq 0 \ : \ F_{(-t,0]}(x) = F_{(-s,0]}(y), \text{ for all } x, y \in \mathcal{X} \text{ and for all } s, t \geq n\right\} \\ &= \min\left\{n \geq 0 \ : \ X_0^{x,-t} = X_0^{y,-s}, \text{ for all } x, y \in \mathcal{X} \text{ and for all } s, t \geq n\right\}. \end{aligned}$$

Note that T^* is a random variable since its value depends upon the sequence $\{\xi_n\}$.

Now consider starting chains $X^{x,-n}$ from *all* states $x \in \mathcal{X}$ at time $-n$, and recall that the chain $\tilde{X}$ is a copy of X which has been running since time $-\infty$. Using the SRS representation all of these chains may be coupled using the same update function f and the same source of randomness $\{\xi_n\}_{-\infty}^{0}$. The big idea behind CFTP is that if n is large enough such that $n \geq T^*$, then $X_0^{x,-n} = X_0$ (say) is the same for all x, and so $X_0 = \tilde{X}_0 \sim \pi$. That is, if the chains $X^{x,-n}$ are started far enough back into the past that they have all coalesced by time 0, then their common value at time 0 is an exact draw from π, as required. Of course, we do not know the value of T^*, and so the CFTP algorithm simply repeats the above procedure for larger and larger n until $n \geq T^*$ is achieved:

Algorithm 4.2 (CFTP)**.**

- `set` $n \leftarrow 1$
- `while` $F_{(-n,0]}(\mathcal{X})$ `not constant`
  $n \leftarrow 2n$
- `return` $F_{(-n,0]}(\mathcal{X})$

The most important aspect of the algorithm to note here is that the random sequence $\{\xi_n\}$ is *re-used* in each execution of the `while` loop above. This implies a possible issue with computer memory in practice, but there are tricks to avoid this, such as the *read-once* CFTP algorithm (Section 4.2). The other comment to make here is the use of the binary search for T^* contained within the `while` loop: we are free to increase n in any way we like, but the binary search means that the total number of Markov chain steps simulated is linear in T^*, whereas if we used $n \leftarrow n+1$

(for example) it would grow quadratically. Furthermore, this strategy means that the number of simulation steps comes within a factor of 4 of the true value of T^* (see Propp and Wilson (1996) and Wilson (2000b) for more details).

The proof that the CFTP algorithm really does return a draw from π is very simple, and so we include it here for completeness. This was first proved of course by Propp and Wilson (1996), but here we repeat the version of the proof given in Kendall (2005).

Theorem 4.3 (Propp and Wilson (1996))**.** *If the backwards coalescence time T^* is almost surely finite then CFTP samples from equilibrium.*

Proof. Assume that $T^* < \infty$. Use the input-output map F to couple chains X^x started from all possible starting states $x \in \mathcal{X}$: that is, define

$$X_v^{x,-u} = F_{(-u,v]}(x) \quad \text{for } -u \leq v\,.$$

By definition of T^*, and the time-homogeneity of X,

$$\begin{aligned} X_0^{x,-n} &= X_0^{x,-T^*} \quad \text{whenever } n \geq T^*, \\ \text{and} \quad \mathcal{L}\left(X_0^{x,-n}\right) &= \mathcal{L}\left(X_n^{x,0}\right)\,, \end{aligned}$$

for all x. Now, since X converges to its equilibrium distribution π in total variation,

$$\begin{aligned} \left\|\mathcal{L}\left(X_0^{x,-T^*}\right) - \pi\right\| &= \lim_{n\to\infty} \left\|\mathcal{L}\left(X_0^{x,-n}\right) - \pi\right\| \\ &= \lim_{n\to\infty} \left\|\mathcal{L}\left(X_n^{x,0}\right) - \pi\right\| = 0\,. \end{aligned}$$

□

Note that if $\mathcal{X}$ is finite then T^* is almost surely finite (but very large) for the independence coupling (where chains evolve independently until they meet, after which they agree forever). In fact, $\mathbb{P}\left(T^* < \infty\right)$ is always either zero or one, whatever the choice of update function f (this follows from a tail σ-algebra argument: see Foss and Tweedie (1998) for details). A good CFTP algorithm uses a function f which has a high probability of making target chains coalesce quickly.

4.1.3 A simple example

The following example was considered in Thönnes (2000). Consider a symmetric reflecting random walk X on $\mathcal{X} = \{1, 2, 3, 4\}$, satisfying $\pi(i) = 1/4$ for all $i \in \mathcal{X}$:

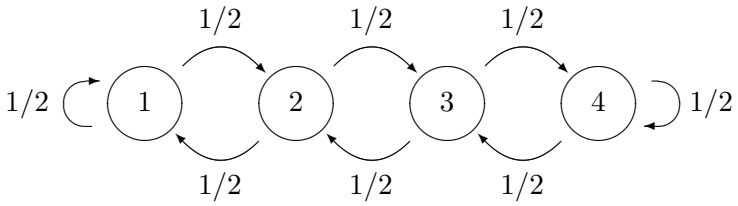


Figure 4.1 shows a realisation of Algorithm 4.2 applied to this random walk. The update function used is

$$f(x,u) = \begin{cases} (x+1) \wedge 4 & \text{if } u \leq 1/2 \\ (x-1) \vee 1 & \text{if } u > 1/2\,. \end{cases} \tag{4.2}$$

The grey arrows indicate the transitions determined by f and the sequence $\{\xi_n\}$. The algorithm runs chains $X^{x,-n}$ started from all states x, with $n = 1, 2, 4, \ldots$. When $n = 8$ is reached, all of the target chains have the same value at time zero: $X_0^{x,-8} = 2$ for this realisation (for which $T^* = 7$). The red circles highlight the possible values of the chains $X^{x,-8}$ at each step: coalescence occurs the first time that $X^{1,-n}$ hits 4 or $X^{4,-n}$ hits 1.

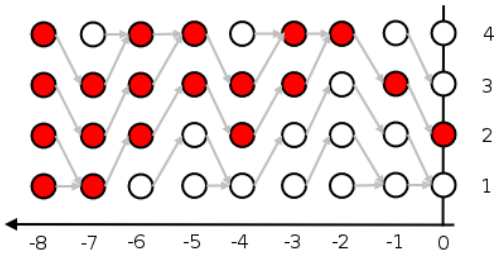


Fig. 4.1: CFTP for a simple symmetric reflecting random walk.

Implementation of CFTP for this example is trivial, and Figure 4.2(a) shows a histogram of the results of 10,000 runs of the algorithm. A χ^2-test to compare this output to π yields a value of 1.039 on 3 degrees of freedom, resulting in a p-value of 0.79: it is clear that the algorithm is drawing from the correct distribution. In contrast, Figure 4.2(b) shows the output of a wrongly implemented CFTP algorithm, in which the random sequence $\{\xi_n\}$ is not re-used when the chains are restarted

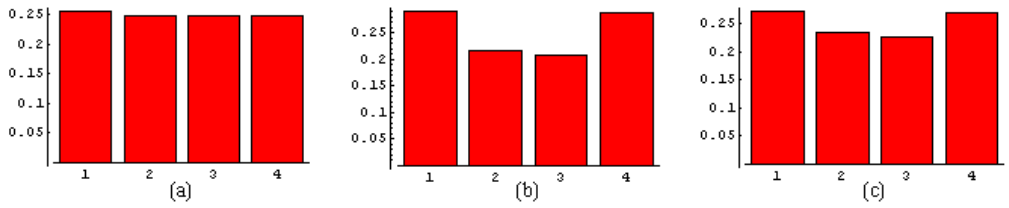


Fig. 4.2: Output of 10,000 runs of the CFTP algorithm for the random walk X on $\{1, 2, 3, 4\}$ described above: (a) a proper implementation, re-using randomness; (b) not re-using randomness; (c) re-using randomness, but with simulations with 'long' run-times interrupted and discarded.

further into the past. Here a definite departure from uniformity is visible, and indeed a χ^2-test gives a tiny p-value of 2×10^{-16}. For more examples of the bias that is introduced when not re-using randomness, see Propp and Wilson (1996), Wilson (2000b), Häggström (2002). Figure 4.2(c) shows another possible source of bias: that which is introduced by *user-impatience*. This happens when simulations with long run-times are interrupted and discarded (whether by an impatient user, or due to computer breakdown etc.). Since the output of the CFTP algorithm is in general not independent of the algorithm run time, this introduces a bias. The p-value for this sample is again very small: $p = 2.6 \times 10^{-11}$.

Although this is just a toy example, it does highlight the ease with which CFTP may be applied to state spaces with a partial order. More specifically, suppose that the state space $\mathcal{X}$ admits a partial order $\preceq$ which is respected by the update function f. That is,

$$x \preceq y \quad \Rightarrow \quad f(x, u) \preceq f(y, u)$$

for all $x, y \in \mathcal{X}$. Furthermore, suppose that $\mathcal{X}$ contains maximal and minimal elements, x^{max} and x^{min}, satisfying $x^{min} \preceq x \preceq x^{max}$ for all $x \in \mathcal{X}$.

With this setup it becomes simple to check for coalescence in Algorithm 4.2, since the monotonicity of f guarantees that

$$F_{(-n,0]}(x^{min}) = F_{(-n,0]}(x^{max}) \quad \Rightarrow \quad F_{(-n,0]}(\mathcal{X}) \text{ is constant}\,.$$

This means that, instead of running target chains $X^{x,-n}$ starting from all states $x \in \mathcal{X}$ at time $-n$, we now only need to simulate chains started from x^{min} and x^{max}. Note that this is the case in the toy example above, where f is clearly monotonic (using the normal ordering on the integers): coalescence occurs exactly when the two

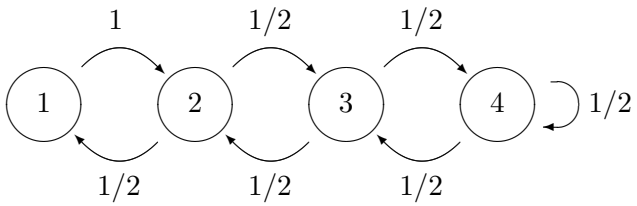
chains X^1 and X^4 meet.

The idea of *monotone CFTP* was used by Propp and Wilson (1996) to produce a perfect draw from the equilibrium distribution of the critical Ising model on a finite lattice. This model has probability mass function proportional to

$$\exp\left(\frac{1}{kT}\sum\sum_{i\sim j}\sigma_i\sigma_j\right)$$

with indices i, j running through the vertices of a (large, but finite) square lattice. Here k is Boltzmann's constant, T is the temperature of the system, $i \sim j$ indicates that sites i and j are neighbours, and $\sigma_i = \pm 1$ is the spin at site i. Propp and Wilson (1996) used the single-bond heat-bath algorithm (Sweeny 1983) to produce a Markov chain with the correct stationary distribution. This algorithm treats the Ising model as a random cluster model (Fortuin and Kasteleyn 1972). The updates of the algorithm respect a partial order on the state space, and so the monotone CFTP algorithm is easy to implement (and converges fast - Propp and Wilson claim that convergence is generally achieved when starting at time -30 for a 512×512 toroidal grid at critical temperature). Furthermore, similar perfect simulation algorithms may be used to draw from the equilibrium of variations on the Ising model, such as when an external magnetic field is applied to the system: this results in the *conditioned* Ising model, which is used in image analysis (see, for example, Besag (1986)).

Note however that monotonicity, whilst useful for CFTP, is not essential. Consider the following modification of the random walk X considered above:



This chain, say $\hat{X}$, is no longer reversible, and there is no monotonic update function under the usual ordering on the integers (although a different ordering, or the use of subsampling, may reintroduce monotonicity). However, a CFTP algorithm for this chain can be constructed using the *crossover trick* (Thönnes 2000): this first appeared in Kendall (1998) to deal with *anti-monotone chains*, and is also used in Häggström and Nelander (1998).

A variant of this approach is to use the original monotonic random walk X as an *envelope process* for $\hat{X}$. Suppose we define an update function $\hat{f}$ as follows:

$$\hat{f}(1,u) = 2\,, \quad \text{and} \quad \hat{f}(x,u) = \begin{cases} (x+1) \wedge 4 & \text{if } u \leq 1/2 \\ (x-1) \vee 1 & \text{if } u > 1/2\,. \end{cases} \qquad \text{for } x = 2,3,4\,. \quad (4.3)$$

Note that $\hat{f}$ is a valid update function for $\hat{X}$, and that $f(x,u) \leq \hat{f}(x,u)$ for all $x \in \mathcal{X}$ and $u \in [0,1]$ (where f is defined in equation (4.2)). This suggests the following perfect simulation algorithm:

- run a target chain $X^{1,-n}$ using the update function f and random sequence $\{\xi_i\}$, until the first time $S \leq 0$ that state 4 is hit (if $S \not\leq 0$, increase n and repeat, reusing $\{\xi_i\}$);
- due to the ordering of f and $\hat{f}$, $X_S^{1,-n} = \hat{X}_S^{x,-n} = 4$ for all x, and so all target chains $\hat{X}^{x,-n}$ have coalesced by time S;
- run the chain $\hat{X}^{4,S}$ up to time zero, using $\hat{f}$ and the same $\{\xi_i\}$. Return $\hat{X}_0^{4,S}$.

The chain $\hat{X}$ still has a simple equilibrium distribution of course:

$$\pi = (1/7, 2/7, 2/7, 2/7)\,.$$

The sample distribution obtained from 10,000 runs of the perfect simulation algorithm just described is shown in Figure 4.3. A χ^2-test for the output from this algorithm gave a p-value of 0.68, providing no evidence against the algorithm sampling from the correct distribution.

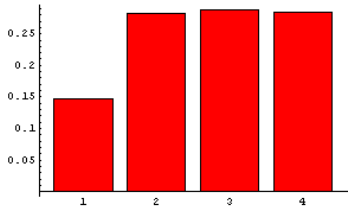


Fig. 4.3: Output of 10,000 runs of the CFTP algorithm for the random walk $\hat{X}$ on $\{1,2,3,4\}$.

The idea of using an envelope process was studied in more depth, under the name *bounding chains*, in Huber (2004): the technique can be used in situations which are neither monotonic nor anti-monotonic, and can also give bounds on the expected run-time of CFTP. This trick, along with the crossover trick mentioned

above, increases the possibility of efficient implementation of CFTP for a variety of chains. Furthermore, a different kind of simulation algorithm, known as the FMMR method (after Fill et al. 2000) neither assumes monotonicity nor has (under suitable implementation) the user-impatience bias observed for CFTP above (Thönnes 1999). This algorithm, based on Fill's method (Fill 1999), is based on strong uniform times (encountered in Chapter 1) and set-valued duals. Although a useful simulation algorithm (see Thönnes 1999 for its application to point process simulation, for example), FMMR is harder to describe than CFTP, and we do not go into a detailed exposition of the method here. The reader is referred to any of the following references for further information: Fill et al. (2000), Dobrow and Fill (2003), Kendall (2005).

4.2 Variants of CFTP

The CFTP algorithm works by ensuring that chains started in the past from all possible starting states $x \in \mathcal{X}$ have coalesced by time zero, but in practice CFTP is unlikely to be as simple to implement as in the above examples. We have seen how to check for coalescence when $|\mathcal{X}|$ is finite, and how this task is simplified when there is a partial order on $\mathcal{X}$ (respected by the update function), but what if $\mathcal{X}$ is continuous? The algorithm also requires the re-use of randomness whenever coalescence at zero is not achieved and the target chains are started further into the past: this can quickly lead to issues with computer memory. Furthermore, it turns out that CFTP can only be applied to chains which converge *uniformly* to their equilibrium distribution (Section 4.4): many Markov chains of course do not satisfy this criterion.

In this section we introduce three variants on Algorithm 4.2, which can potentially be used when faced with one of the problems outlined above. There do exist further variants on CFTP and other simulation algorithms, which we do not go into here: a discussion of the following three algorithms is, however, essential preparation for the rest of the work in this thesis.

4.2.1 Small-set CFTP

We begin with the issue of a continuous state space, with no assumptions about the existence of a useful partial ordering.

Definition 4.4. A subset $C \subseteq \mathcal{X}$ is a *small set (of order m)* for the Markov chain X if the following *minorisation condition* holds: for some $\varepsilon \in (0, 1]$ and a probability measure ν,

$$\mathbb{P}_x (X_m \in E) \geq \varepsilon \nu(E), \quad \text{for all } x \in C \text{ and measurable } E \subset \mathcal{X}. \tag{4.4}$$

In this case we say that C is m-small. If X hits the small set C at time n then with probability ε it can be made to regenerate at time $n + m$ (using the measure ν).

This is similar to the setup for the Vasershtein coupling of Theorem 3.1: if two chains X and X' both belong to C at time n, then the distributions of X_{n+m} and X'_{n+m} have common component $\varepsilon\nu$. With probability ε we may therefore set $X_{n+m} = X'_{n+m}$ (using a draw from ν), and with probability $1 - \varepsilon$ draw X_{n+m} and X'_{n+m} independently from their residual kernels. Furthermore, if regeneration does occur then a single draw from ν may be used for any number of copies of X belonging to C at time n, resulting in their coalescence at time $n + m$.

This construction is rendered more useful by the fact that small sets (of non-trivial measure) exist for any ψ-irreducible chain (Meyn and Tweedie 1993). Moreover, under the assumption of ψ-irreducibility, it is possible to cover the whole state space with a countable collection of small sets (Meyn and Tweedie (1993), Proposition 5.2.4).

If the whole state space is small, then target chains started from all possible states can therefore be made to regenerate (and thus coalesce) at the same time. This leads to the following perfect simulation algorithm (Murdoch and Green 1998). Starting at time $-n$, draw i.i.d. random variables $U_{-n}, U_{-(n-1)}, \ldots \sim \mathit{Uniform}[0, 1]$ until the earliest k for which $U_k \leq \varepsilon$. Draw X_{k+m} from the distribution ν, and follow a trajectory of X from this value forwards to time zero: $X_0 \sim \pi$, as required. As usual, if when starting at time $-n$ there does not exist a $k \in [-n, 0]$ such that $U_k \leq \varepsilon$, then set $n \leftarrow 2n$ and repeat, ensuring that the path of X over $[-n, 0]$ is constructed using the residual kernel at each step (i.e. respecting the fact that $U_{-n}, \ldots, U_0 > \varepsilon$).

Note that we don't actually need to simulate chains from all starting states in this algorithm: it is only necessary to follow the trajectory of one chain after a coalescent event has occurred. However, since ε can typically be very small indeed (and m

very large), the coupling time as constructed above may be rather large: Green and Murdoch (1998) discuss ways of producing a faster backward coupling time in practice, including a method whereby $\mathcal{X}$ is partitioned into a number of small sets, each with a different minorisation condition.

4.2.2 Read-once CFTP

So far all of the perfect simulation algorithms described have required the re-use of randomness each time chains are started further into the past. This can be costly when considering computer memory capacity for complicated implementations. However, Wilson (2000a) noted that CFTP *can* be formulated as a forwards time algorithm.

As a simple example of how this may be achieved, consider once again the small set CFTP algorithm of the last section, in the case when $\mathcal{X}$ is a small set. This algorithm could equally well be implemented *forwards* in time as follows. Draw $X_0 \sim \nu$, and $V \sim Geom(\varepsilon)$. Then run X forwards in time until time $V-1$, using the residual kernel at each step, returning $X_{V-1} \sim \pi$.

This construction may be generalised. Recall that for CFTP to work, we need to be able to identify when the input-output map $F_{(-n,0]}$ of Algorithm 4.2 is coalescent. We can view the map $F_{(-nt,0]}$ as the composition of smaller i.i.d. blocks of length t:

$$F_{(-nt,0]} = F_{(-t,0]} \circ F_{(-2t,-t]} \circ \ldots \circ F_{(-nt,-(n-1)t]} \,.$$

Now suppose that t is chosen large enough such that there is a positive probability p_t that the map $F = F_{(-t,0]}$ is coalescent. As for the example above, the following algorithm can be shown to be equivalent to the classic CFTP algorithm:

Algorithm 4.5 (Read-once CFTP)**.**

- draw independent realisations of F until F coalescent
- set $x = F(\cdot)$
- draw independent realisations of F
  - while F not coalescent
    $x \leftarrow F(x)$
- return x

Note that the third step of this algorithm is equivalent to composing a $Geom(p_t)$ number of non-coalescent blocks (as for the small-set CFTP example), but using this construction avoids the problem of calculating the maps F conditioned on coalescence or non-coalescence. Since this algorithm starts at time zero and runs into the future, any randomness used does not need to be stored for future use. The read-once algorithm is also well suited to producing a number of draws from π: as soon as a coalescent block F is detected in the final step, the result x of the previous (non-coalescent) block is returned and then $F(x)$ may be used as the starting point for another run of the algorithm, starting at the second step.

Of course, the time taken for this algorithm to return a draw from the stationary distribution is highly dependent upon the length t of the block used. Wilson (2000a) compares the performance of read-once CFTP with Algorithm 4.2. The expected running time is within a constant factor of that of CFTP, and in some situations (especially if $|\mathcal{X}|$ is continuous) the gains to be made from using Algorithm 4.5 can be significant.

4.2.3 Dominated CFTP

As mentioned earlier, the classic CFTP algorithm of Propp and Wilson (1996) has a major drawback: a successful CFTP algorithm for X exists if and only if X is *uniformly ergodic* (see Section 4.4.2). However, there does exist a major extension of CFTP, known as *dominated CFTP* (domCFTP) or *Coupling into and from the Past*

(CIAFTP), which can be applied to chains not satisfying this restriction (Kendall 1997; Kendall and Møller 2000).

Whereas CFTP works by checking for *vertical* coalescence (target chains started from all states have coalesced by time zero), domCFTP checks for *horizontal* coalescence (all sufficiently early starts from a specific state lead to the same result at time zero). A second chain Y is used to identify how far into the past one has to go to determine that this coalescence has occurred. To describe the algorithm in the simplest possible way, we consider here only a monotonic chain X on $\mathcal{X} = [0, \infty)$: see Cai and Kendall (2002) for a much more general formulation.

Suppose that copies of X can be coupled such that, for each $x, t, u \geq 0$, and $s \geq -t$, we can construct $X^{x,-t}$ (begun at state x at time $-t$) satisfying

$$X_s^{x,-t} \leq X_s^{x,-u} \qquad \Rightarrow \qquad X_{s+1}^{x,-t} \leq X_{s+1}^{x,-u} \,. \tag{4.5}$$

Suppose too that we can construct a dominating process Y on $\mathcal{X}$ which is stationary, defined for all time, and may be coupled to the target chains X such that

$$X_s^{x,-t} \leq Y_s \qquad \Rightarrow \qquad X_{s+1}^{x,-t} \leq Y_{s+1} \,. \tag{4.6}$$

The domCFTP algorithm then proceeds as follows:

1. Draw Y_0 from its stationary distribution;
2. Simulate Y backwards to time $-n$;
3. Set $y = Y_{-n}$. Simulate $X^{y,-n}$ and $X^{0,-n}$ forwards to time zero (coupled to each other and to Y so that (4.5) and (4.6) are satisfied);
4. If $X_0^{y,-n} = X_0^{0,-n}$ then return this value as a perfect draw from π. If not, extend the realisation of Y back to time $-2n$, set $n \leftarrow 2n$, and go to step 3.

A proof that this algorithm returns a perfect draw from π (so long as it terminates almost surely) may be found in Kendall and Møller (2000). A consequence of the coupling in equation (4.5) is that the target processes are *funnelled*: the earlier the two target chains (X^0 and X^y) are started, the closer they will be at time zero. As a simple example of the algorithm in practice, consider a birth-death process X with transitions $x \to x + 1$ at rate $\alpha_x \leq \alpha < \infty$, and $x \to x - 1$ at rate μx. This chain

is clearly monotonic, and may be dominated by the birth-death chain Y which has births at rate α and deaths at rate μ. It is easy to see that Y is reversible and has a $Poisson(\alpha/\mu)$ equilibrium distribution. A realisation of the domCFTP algorithm for this chain when $\alpha_x = 10 - \log(x+1)/(x+1)$ and $\mu = 1$ is shown in Figure 4.4.

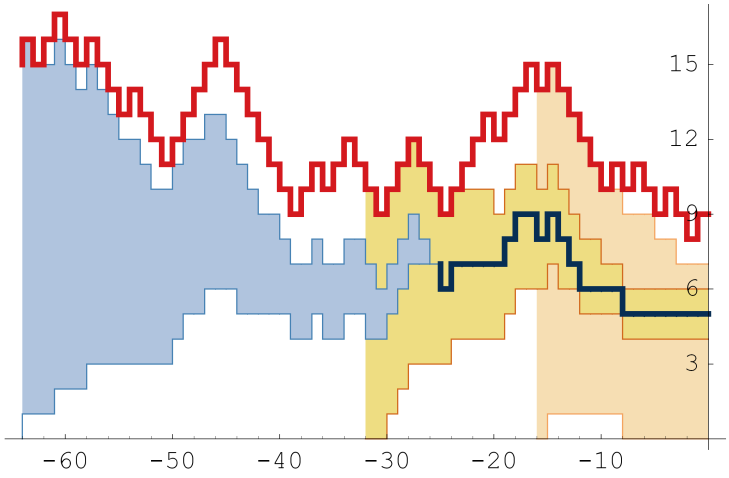


Fig. 4.4: Implementation of domCFTP for a birth-death process. The topmost line shows the evolution of the dominating process into the past, and the shaded regions demonstrate the funnelling of target processes. In this realisation, all target chains started beneath the dominating process at time -64 have coalesced by time zero.

This algorithm can also be made to work if multiple simultaneous deaths are allowed to occur, or if the birth rate μ varies with x, as is the case with many problems in stochastic geometry (Kendall 1998; Kendall and Møller 2000).

4.2.4 *Extended state-space CFTP*

A more general theorem for domCFTP is presented in Cai and Kendall (2002). This involves further abstraction of the original CFTP idea, but results in an algorithm which has practical applications: Cai and Kendall (2002) use this setup to carry out perfect simulation for correlated Poisson random variables conditioned to be positive. We now summarise this abstraction, since it will be useful to us in the following chapter.

As usual, the target chain X lives on the space $\mathcal{X}$. The main idea of Cai and Kendall (2002) is to embed $\mathcal{X}$ into a partially ordered space $(\mathcal{Y}, \preceq)$ such that $\mathcal{X}$ is at the bottom of $\mathcal{Y}$. That is, such that for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$y \preceq x \quad \text{implies} \quad y = x\,. \tag{4.7}$$

We then identify a suitable process Y on $\mathcal{Y}$ which is eventually absorbed in $\mathcal{X}$, and which follows the stochastic dynamics of X once this occurs. More specifically, we actually work with a sequence of processes

$$Y^{(n)} = \left\{ Y_t^{(n)} \,:\, -n \leq t \leq \infty \right\},$$

for $n = 1, 2, 4, \ldots$, which are identically distributed up to a shift in time. These processes $Y^{(n)}$ all live on $\mathcal{Y}$ and satisfy the following three conditions:

1. $Y^{(n)}$ is eventually absorbed in $\mathcal{X}$: for any fixed t, as $n \to \infty$,

$$\mathbb{P}\left(Y_t^{(n)} \in \mathcal{X} \right) \to 1\,; \tag{4.8}$$

2. $Y^{(n)}$ evolves using the dynamics of X once it hits $\mathcal{X}$;

3. The processes $Y^{(n)}$ obey the following *funnelling* condition:

$$Y_t^{(m)} \preceq Y_t^{(n)}, \quad \text{for all } -m \leq -n \leq t \leq 0. \tag{4.9}$$

With these conditions in place we are ready to state the theorem of Cai and Kendall (2002):

Theorem 4.6 (Cai and Kendall (2002))**.** *Let X be an ergodic Markov chain living on the state space $\mathcal{X}$ and with equilibrium distribution π. Suppose that $\mathcal{X}$ may be embedded in the partially ordered space $\mathcal{Y}$ such that $\mathcal{X}$ is at the bottom of $\mathcal{Y}$, as in* (4.7)*. Suppose further that the processes $Y^{(n)}$ live on $\mathcal{Y}$ and satisfy conditions 1-3 above. Define*

$$T = \inf\left\{ n \geq 1 \,:\, Y_0^{(n)} \in \mathcal{X} \right\}. \tag{4.10}$$

Then T is almost surely finite, and $Y_0^{(T)} \sim \pi$.

Proof. Condition 1 above implies that $T < \infty$ almost surely, and so it remains to show that $Y_0^{(T)} \sim \pi$. By the funnelling in equation (4.9), if $n > T$ then

$$Y_0^{(n)} \preceq Y_0^{(T)}.$$

Therefore, since $Y_0^{(T)} \in \mathcal{X}$, the embedding of $\mathcal{X}$ at the bottom of $\mathcal{Y}$ implies that $Y_0^{(n)} = Y_0^{(T)}$ for all $n > T$, and so

$$Y_0^{(T)} = \lim_{n \to \infty} Y_0^{(n)}$$

exists almost surely. Conditions 1 and 2, along with the fact that the $Y^{(n)}$ are identically distributed up to time shifts, imply that $Y^{(n)}$ has the same equilibrium distribution as X, and so

$$\begin{aligned}\left\|\mathcal{L}\left(Y_0^{(T)}\right)-\pi\right\| &= \lim_{n\to\infty}\left\|\mathcal{L}\left(Y_0^{(n)}\right)-\pi\right\| \\ &= \lim_{n\to\infty}\left\|\mathcal{L}\left(X_n\right)-\pi\right\| ,\end{aligned}$$

where X is started at time 0 with the common hitting distribution of the $Y^{(n)}$ on $\mathcal{X}$. But X has π as its unique equilibrium distribution, and so this final limit is equal to 0. Thus $Y_0^{(T)} \sim \pi$, as required. □

For obvious reasons, Cai and Kendall (2002) call this algorithm *extended state-space CFTP*. Although it appears to be rather more abstract than the perfect simulation algorithms we have encountered above, it has the nice property that it assumes very little about the state space of X. For example, nothing is assumed about the existence of a partial order on $\mathcal{X}$, nor about the presence of monotonicity or maximal and minimal elements if such a partial order does exist.

4.3 Efficiency considerations

In this section we consider the efficiency of CFTP. There are, of course, a number of ways to define 'efficiency', and we begin by considering a couple of these that have already received attention in the literature.

Propp and Wilson (1996), in their original paper on CFTP, consider the tail distribution of the backwards coalescence time T^* (Definition 4.1). Suppose that $\mathcal{X}$ is a partially ordered space, and let ℓ be the length of the longest totally ordered subset of $\mathcal{X}$. Let

$$\bar{d}(k) = \max_{x,y}\left\|\delta_x P^k - \delta_y P^k\right\| .$$

Then, for the monotonic CFTP algorithm, Propp and Wilson (1996) show that

$$\frac{\mathbb{P}\left(T^* > k\right)}{\ell} \leq \bar{d}(k) \leq \mathbb{P}\left(T^* > k\right) .$$

and

$$\mathbb{E}\left[T^*\right] \leq 2\tau^{mix}(1+\log\ell) .$$

Thus monotonic CFTP is within a constant multiplicative factor of being as good as possible.

Propp and Wilson also consider different possible methods of recursively 'searching' for T^* in the CFTP algorithm. They prove that the binary search method used in Algorithm 4.2 minimises the worst-case number of steps, as well as almost minimising the expected number of steps required.

A different approach to considering the efficiency of CFTP comes from the paper of Burdzy and Kendall (2000). They consider the 'gap' between the rate at which a Markov chain approaches equilibrium, and the rate at which co-adapted coupling can happen for two such chains. Of course, the coupling rate has to be slower than the convergence rate (Lemma 1.8). The result of Burdzy and Kendall (2000) however, shows that co-adapted coupling is strictly slower than convergence to stationarity when it is possible for a pair of co-adapted chains to transpose before coupling. For such a chain X, a CFTP algorithm using co-adapted coupling will converge at an exponential rate slower than X converges to equilibrium. (Of course, this does not apply for monotone chains, where the transposition of coupled chains will not occur.)

This result is not too surprising, given our earlier work on co-adapted and maximal coupling: we have seen instances where the optimal co-adapted coupling is not maximal (e.g. for a random walk on $\mathbb{Z}_2^n$, Section 3.3). With this in mind, we now ask the following question:

Can CFTP be as fast as convergence to equilibrium?

In other words, what is the smallest possible value of T^* for a given Markov chain, and how can we design a CFTP algorithm using this knowledge? This question motivates the work in the following section.

4.3.1 Impractical CFTP

We will assume throughout this section that the chain of interest X takes values in a finite state space $\mathcal{X} = \{1, 2, \ldots, N\}$. The run-time of the CFTP algorithm is of course dependent upon the distribution of the time taken for chains started from each state $1 \leq i \leq N$ to coalesce, which in turn depends upon the input-output maps F used in the algorithm. This suggests that a fast (but impractical, see Remark 4.9) CFTP algorithm could be designed using the maximal coalescent coupling of Section 3.1.

Recall that this coupling works by drawing the coupling time and place (T, X_T) from a given distribution, and then constructing the N pre-T processes and the

single post-T process according to the prescription in Theorem 3.7. We will not be concerned with the inhomogeneous pre-T chains in this application, since the output of the CFTP algorithm will depend only on what happens after coalescence. The apparent problem with trying to use a maximal coupling with CFTP is that the first of these techniques concerns forwards coupling while the second relies on a backwards construction. Thus the following (perhaps at first sight reasonable) algorithm will not return a draw from equilibrium:

Algorithm 4.7 (Incorrect CFTP)**.**

- `set` $n \leftarrow 1$
- `draw` (T, X_T) `from maximal coupling distribution`
- `while` $T > n$
  - $n \leftarrow 2n$
  - `draw` (T, X_T) `from maximal coupling distribution`
- `run` X `from` $(-(n-T), X_T)$ `to time zero`
- `return` X_0

The problem with this algorithm of course is the independence of the draws from the maximal coupling distribution. In order for the algorithm to return a draw from equilibrium, it would be necessary for the draw of the pair (T, X_T) at time $-2n$ to be conditioned on coalescence not having occurred in the interval $[-n, 0]$.

However, we *can* produce a perfect simulation algorithm employing the maximal coalescent coupling by using Wilson's read-once trick (Section 4.2.2). Recall that this algorithm circumvents the issue of re-using randomness by running from time zero into the future, composing i.i.d. blocks $F = F_{(-t,0]}$.

Using the same notation as in Section 4.2.2, equation (3.6) means that the coalescence time T for the maximal coalescent coupling has the following distribution:

$$\mathbb{P}(T \leq n) = \sum_{y \in S} \lambda_n(y) = 1 - d_n. \tag{4.11}$$

We also have an explicit form for the distribution of a candidate chain X at time n conditional upon the event $\{T \leq n\}$, $\mathcal{L}(X_n \mid T \leq n)$, which is of course independent of X_0 (since T is the coalescence time). Indeed, from equation (3.8) and the fact that the post-T process is a version of the original homogeneous Markov chain begun at X_T, we see that:

$$\begin{aligned}\mathbb{P}(X_n = x \mid T \leq n) &= \frac{\sum_{y=1}^{N}\sum_{k=1}^{n}\mathbb{P}(T=k, X_k = y)\,\mathbb{P}(X_n = x \,|\, X_k = y, T = k)}{\mathbb{P}(T \leq n)} \\ &= (1-d_n)^{-1}\sum_{y=1}^{N}\sum_{k=1}^{n}(\lambda_k - \lambda_{k-1}P)(y)P^{n-k}(y,x)\,. \end{aligned} \tag{4.12}$$

The (impractical) read-once CFTP algorithm is therefore as follows.

Algorithm 4.8 (Read-once CFTP using maximal coalescent coupling)**.**

- `fix` $t \in \mathbb{N}$ `and define` $p_t^* = \mathbb{P}(T \leq t) = 1 - d_t$
- `draw` $X_0 = x$ `from` $\mathcal{L}(X_t \,|\, T \leq t)$
- `draw` $V \sim Geom(p_t^*)$
- `run` X `forwards until time` $(V-1)t$`, conditional on the event that none of the` $(V-1)$ `blocks of length` t `are coalescent blocks`
- `return` $X_{(V-1)t}$

Note that in the penultimate step of this algorithm, there is a closed form for the distribution $\mathcal{L}(X_t \mid T > t, X_0 = x)$:

$$\begin{aligned}\mathbb{P}(X_t = y \,|\, T > t, X_0 = x) &= \mathbb{P}(X_t = y, T > t \,|\, X_0 = x)\,\mathbb{P}(T > t \,|\, X_0 = x)^{-1} \\ &= \beta_t^x(y)(1-p_t^*)^{-1}.\end{aligned}$$

(This follows from equation (3.7).) Thus the algorithm can draw values for the chain at times $t, 2t, \ldots, (V-1)t$ using this distribution iteratively.

Remark 4.9. This explains the use of the word 'impractical' in the title of this section: although it is possible to write out closed form expressions for each of the distributions used in Algorithm 4.8, it would not be possible in general to sample from these distributions in practice.

In order to compare the expected cost of this algorithm with that of the read-once CFTP algorithm of Section 4.2.2 we make the assumption that, for fixed $t \in \mathbb{N}$, a draw from either of the distributions $\mathcal{L}(X_t \mid T \leq t)$ and $\mathcal{L}(X_t \mid T > t)$ is about as expensive as checking to see if the map $F = F_{(-t,0]}$ used in Algorithm 4.5 is coalescent. It is also assumed that this cost is linear in t.

The 'standard' read-once CFTP algorithm proceeds as follows (recall Algorithm 4.5). First of all a block length t is chosen: for a given input-output map $F = F_{(-t,0]}$ any such block has probability p_t of being coalescent, independent of all other blocks. A sequence of blocks is then generated: after $k+1$ coalescent blocks have been observed the algorithm will have returned k draws from equilibrium. Since the number of blocks needed to obtain a coalescent block is distributed as $Geom(p_t)$, the expected cost of this algorithm is proportional to

$$\frac{(k+1)t}{p_t}\,.$$

Similarly, Algorithm 4.8 draws once from $\mathcal{L}(X_t \mid T \leq t)$ and $V-1$ times from $\mathcal{L}(X_t \mid T > t)$ for each required draw from equilibrium. Since $V \sim Geom(p_t^*)$, it follows that the expected cost for k such draws is proportional to

$$\frac{kt}{p_t^*}\,.$$

Now, for any fixed $t \in \mathbb{N}$, $p_t^* \geq p_t$ for any input-output map F. Therefore, the expected cost of Algorithm 4.8 is a lower bound on the cost of Algorithm 4.5. If the maximal coalescent coupling is equivalent to a co-adapted coalescent coupling, then it is possible for p_t to equal p_t^* for all t, and hence for the 'standard' read-once algorithm to perform as well as Algorithm 4.8. Furthermore, although the value of p_t^* may be hard to calculate, it can be bounded above by

$$\min_{i,j \in \mathcal{X}} q_t^*(i,j)\,,$$

where $q_t^*(i,j)$ is the probability that maximally-coupled chains started at i and j have coupled by time t. (This maximal pairwise coupling is of course not necessarily the same as the maximal coalescent coupling.) This may provide a more tractable lower bound on the cost of any read-once algorithm. These observations show that the future research identified at the end of Chapter 3 (regarding the relationships

between co-adapted, coalescent and maximal couplings) may have consequences for our understanding of the limitations of perfect simulation algorithms.

Algorithm 4.8 is near-optimal, in the sense described above, but it is not necessarily optimal since the cost of the algorithm depends upon the balance between t and p_t^*. Careful choice of the block size t would be required in order to minimise the ratio t/p_t^* and thereby optimise the expected cost of the algorithm. Since p_t and p_t^* are of course specific to the Markov chain of interest (and, in the case of p_t, to the choice of F), it is not possible to determine an optimal value of t which holds in generality.

Finally note that, since Theorem 3.7 holds for an uncountable number of chains, Algorithm 4.8 can also be applied to chains on continuous state-spaces.

4.4 Ergodicity considerations

The way in which an ergodic Markov chain approaches its stationary distribution is a topic of great interest: although Markov chains have been studied for over eighty years now, this is still an area of active research. We have already seen (in Chapter 2) that the cutoff phenomenon is one type of possible behaviour which is particularly interesting when studying random walks on groups. For Markov chains without so much structure though, we cannot hope to establish such startling results. However, the speed at which a chain approaches equilibrium is still a very interesting question, and one which has important consequences for MCMC and perfect simulation algorithms.

In this section we present a couple of striking results (due to Foss and Tweedie (1998) and Kendall (2004)) linking the possibility of perfect simulation for a Markov chain with the rate at which the chain converges to its equilibrium distribution.

4.4.1 Definitions and notation

Let $X = (X_0, X_1, \ldots)$ be a discrete-time Markov chain on a Polish state space $\mathcal{X}$. The Markov transition kernel for X is denoted by P, and the n-step kernel by P^n;

$$P^n(x, E) = \mathbb{P}_x \left(X_n \in E \right),$$

where $\mathbb{P}_x$ is the conditional distribution of the chain given $X_0 = x$. The corresponding expectation operator will be denoted $\mathbb{E}_x$. If g is a non-negative function then we write $Pg(x)$ for the function $\int g(y)P(x, dy)$, and for a signed measure μ we write $\mu(g)$ for $\int g(y)\mu(dy)$. The f-norm is defined as $\|\mu\|_f = \sup_{g:|g|\leq f} |\mu(g)|$; taking $f \equiv 1$ yields the usual total variation norm, for which we continue to write $\|\mu\|$.

We assume throughout that X is aperiodic (in the sense of Meyn and Tweedie 1993) and Harris-recurrent. The stationary distribution of X shall be denoted by π, and the *first hitting time* of a measurable set $A \subseteq \mathcal{X}$ by $\tau_A = \min\{n \geq 1 \,:\, X_n \in A\}$. Being Harris-recurrent, X is ψ-recurrent (for some measure ψ): a set A is called *full* if $\psi(A^c) = 0$ and *absorbing* if $P(x, A) = 1$ for all $x \in A$.

Recall from Definition 4.4 the notion of small sets, which will feature heavily throughout the remainder of this thesis:

> A subset $C \subseteq \mathcal{X}$ is a *small set (of order m)* for the Markov chain X if the following *minorisation condition* holds: for some $\varepsilon \in (0, 1]$ and a probability measure ν,
>
> $$\mathbb{P}_x(X_m \in E) \geq \varepsilon\nu(E), \quad \text{for all } x \in C \text{ and measurable } E \subset \mathcal{X}. \tag{4.13}$$

Many results in the literature are couched in terms of the more general idea of *petite sets*; however for aperiodic ψ-irreducible chains the two notions are equivalent (Meyn and Tweedie 1993, Theorem 5.5.7). Small sets belong to a larger class of *pseudo-small* sets, as introduced by Roberts and Rosenthal (2001), but such sets only allow for the coupling of pairs of chains. Implementation of domCFTP (to be considered below) requires a positive chance of a continuum of chains coalescing when belonging to a given set C, and so henceforth we shall deal solely with small sets.

4.4.2 Uniform ergodicity and CFTP

Let us now consider two particular types of convergence: those of geometric and uniform ergodicity.

Definition 4.10. The chain X is said to be *geometrically ergodic* if there exists a constant $\gamma \in (0, 1)$ and some function $R : \mathcal{X} \to [0, \infty)$ such that, for all x in a full and absorbing set,

$$\|P^n(x, \cdot) - \pi(\cdot)\| \leq R(x)\gamma^n. \tag{4.14}$$

If R can be chosen to be bounded then X is said to be *uniformly ergodic*.

Thus a chain is uniformly ergodic if its convergence rate does not depend upon its starting state. Uniform ergodicity of X can be shown to be equivalent to the whole state space $\mathcal{X}$ being a small set (Meyn and Tweedie (1993), Theorem 16.2.2). This fact may be used to prove the following theorem:

Theorem 4.11 (Foss and Tweedie (1998))**.** *There exists a successful (that is, almost surely finite) backwards coalescence time T^* for X if and only if X is uniformly ergodic.*

In other words, there exists a CFTP algorithm for the chain X (in the sense of Propp and Wilson (1996)) if and only if X is uniformly ergodic.

Proof. (Sketch.) Theorem 6 of Propp and Wilson (1996) shows that T^* is submultiplicative:

$$\mathbb{P}(T^* > m+n) \leq \mathbb{P}(T^* > m)\,\mathbb{P}(T^* > n) \quad \text{for any } m, n \in \mathbb{N}. \tag{4.15}$$

Suppose first that T^* is successful. Inequality (4.15) implies that

$$\mathbb{P}(T^* > n) \leq c\gamma^n\,,$$

for some $c < \infty$ and $\gamma < 1$. Now define T to be the *forward* coupling time for the family of chains $\{X^x\}$. Then the time-homogeneity of X implies that

$$\mathbb{P}(T^* > n) = \mathbb{P}(T > n)\,.$$

Therefore the distribution of T also has a geometric tail. But since T is a coupling time we can now use the coupling inequality (1.4) to deduce that

$$\|P^n(x,\cdot) - \pi\| \leq \mathbb{P}(T > n) \leq c\gamma^n\,. \tag{4.16}$$

Therefore X is indeed uniformly ergodic.

Now suppose that X is uniformly ergodic. As mentioned above, this is equivalent to the whole state space $\mathcal{X}$ being a small set. Thus the small-set CFTP construction of Section 4.2.1 can be applied to produce a perfect simulation algorithm for X. □

4.4.3 Geometric ergodicity and domCFTP

It may be thought that this result constrains the possible applicability of perfect simulation. We have seen above however that the classic CFTP of Propp and Wilson is not the only perfect simulation technique available, and indeed it turns out that the domCFTP method of Kendall (1998) (Section 4.2.3) can be used for chains which converge geometrically (but not necessarily uniformly) fast. In fact, Kendall (2004) proves that there exists a domCFTP algorithm for *all* geometrically ergodic chains.

We do not reproduce the proof of this result here, but it is helpful to examine the basis for the construction of a dominating process for a general geometrically ergodic chain (this is essential preparation for the work in the following chapter in fact). Before doing this, we need to present a little background theory concerning geometrically ergodic chains.

The most common way to establish the rate of ergodicity of a chain X is to check for the existence of a *drift* and a *minorization* condition satisfied by X. For example, the following condition is equivalent to X being positive recurrent:

Condition **PR**:

There exists a positive constant $b < \infty$, a small set C and a scale function $V : \mathcal{X} \to [1, \infty)$, bounded on C, such that

$$\mathbb{E}\left[V(X_{n+1}) \,|\, X_n = x\right] \leq V(x) - 1 + b\mathbf{1}_C(x)\,. \tag{4.17}$$

We shall usually refer to such a condition simply as a drift condition - the minorization component is implied by the fact that C is a small set. For simplicity we also often write inequality (4.17) as $PV \leq V - 1 + b\mathbf{1}_C$. Condition PR states that the chain $V(X)$ behaves as a supermartingale before X hits C. When the chain hits C then it can increase in expectation, but only by a bounded amount.

The first hitting time of C is related to drift conditions in the following way (extracted from Meyn and Tweedie (1993), Theorem 11.3.5):

Theorem 4.12. *For an ergodic chain X, the function $V_C(x) = \mathbb{E}_x\left[\tau_C\right]$ is the pointwise minimal solution to the inequality*

$$PV(x) \leq V(x) - 1, \quad x \notin C\,. \tag{4.18}$$

(Note that (4.18) is simply the drift condition PR for $x \notin C$.) This result can be shown to imply that all sub-level sets are small (Meyn and Tweedie (1993), Lemma 11.3.7), and since V is bounded on C we will henceforth always take C to be a sub-level set of the form $\{x \in \mathcal{X} \,:\, V(x) \leq d\}$.

The following geometric *Foster-Lyapunov* condition (Foster 1953) is stronger than the drift condition PR:

Condition **GE**:

> There exist positive constants $\beta < 1$ and $b < \infty$, a small set C and a scale function $V : \mathcal{X} \to [1, \infty)$, bounded on C, such that
>
> $$PV \leq \beta V + b\mathbf{1}_C \,. \tag{4.19}$$

Inequality (4.19) will be referred to as GE(V, β, b, C) when we need to be explicit about the scale function and constants. As with condition PR, a chain X satisfying GE behaves as a supermartingale (under the scale function V), but now the drift towards C is geometric. Under our global assumptions on X, this drift condition is actually equivalent to X being geometrically ergodic (Meyn and Tweedie 1993, Theorem 15.0.1). Furthermore, if X satisfies (4.19) then we can take $R = V$ in equation (4.14).

The following result can be extracted from Meyn and Tweedie (1993), Theorems 15.0.1 and 16.0.1.

Theorem 4.13. *Suppose X is ψ-irreducible and aperiodic. Then X is geometrically ergodic if and only if there exists $\kappa > 1$ such that the corresponding geometric moment of the first return time to C is bounded:*

$$\sup_{x \in C} \mathbb{E}_x \left[\kappa^{\tau_C}\right] < \infty. \tag{4.20}$$

Indeed, given the drift condition (4.19), Theorem 15.2.5 of Meyn and Tweedie (1993) shows that (4.20) holds for any $\kappa \in (1, \beta^{-1})$. Roberts and Tweedie (1999) also show that under condition GE, the time until X first regenerates according to ν in (4.13) has an exponential moment, and use this to find bounds on γ in (4.14).

With this theory to hand, we can now properly state the result of Kendall (2004):

Theorem 4.14 (Kendall 2004)**.** *If X satisfies the drift condition*

$$PV \leq \beta V + b\mathbf{1}_C$$

for $0 < \beta < 1$, then there exists a domCFTP *algorithm for X (possibly subject to sub-sampling) using a dominating process based on the scale V.*

To begin to understand this result, it is first necessary to define what is meant by 'a dominating process based on the scale V'. The remainder of this section is a summary of Kendall (2004).

Definition 4.15. Suppose that V is a scale function for a Harris-recurrent Markov chain X. We say that the stationary ergodic random process Y on $[1, \infty)$ is a *dominating process for X based on the scale function V* (with *threshold* h and *coalescence probability* ε) if it can be coupled co-adaptively to realisations of $X^{x,-t}$ (the Markov chain X begun at x at time $-t$) as follows:

(a) for all $x \in \mathcal{X}$, $n > 0$, and $-t \leq 0$, almost surely

$$V(X^{x,-t}_{-t+n}) \;\leq\; Y_{-t+n} \;\;\Rightarrow\;\; V(X^{x,-t}_{-t+n+1}) \;\leq\; Y_{-t+n+1}; \tag{4.21}$$

(b) if $Y_n \leq h$ then the probability of *coalescence* at time $n+1$ is at least ε, where coalescence at time $n+1$ means that the set

$$\left\{ X^{x,-t}_{n+1} \;:\; -t \leq n \text{ and } V(X^{x,-t}_{n}) \leq Y_n \right\} \tag{4.22}$$

is a singleton set;

(c) and finally, $\mathbb{P}(Y_n \leq h)$ must be positive.

The most important component of the domCFTP algorithm described in Kendall (2004) is the construction of a stationary process Y which satisfies equation (4.21). Since we only have knowledge of the dynamics of $V(X)$ through its moments (via the drift condition GE), it is natural to ask that

$$\mathbb{P}_z(Y_1 \geq \beta z y) \;\geq\; \sup_{x:V(x)\leq z} \frac{\mathbb{E}_x[V(X_1)]}{\beta z y}, \tag{4.23}$$

and then Markov's inequality provides the domination required in Definition 4.15(a) (see chapter IV of Lindvall 2002, for example). It has already been remarked that it is no restriction to set $C = \{x \,:\, V(x) \leq d\}$, and this yields

$$\begin{aligned} \sup_{x:V(x)\leq z} \frac{\mathbb{E}_x\left[V(X_1)\right]}{\beta z y} &\leq \sup_{x:V(x)\leq z} \frac{\beta V(x) + b\mathbf{1}_{[V(x)\leq d]}}{\beta z y} \\ &\leq \frac{1}{y} \qquad \text{if } z \geq d + \frac{b}{\beta}\,. \end{aligned}$$

Define U to be the system workload of a $D/M/1$ queue, sampled just before arrivals, with arrivals every $\log(1/\beta)$ units of time, and service times being independent and of unit rate Exponential distribution. If $Y = (d + b/\beta)\exp(U)$ and $y \geq 1$, then

$$\mathbb{P}_z\left(Y_1 \geq \beta z y\right) = \frac{1}{y}\,, \quad \text{if } z \geq d + \frac{b}{\beta},$$

and so (4.23) is satisfied. U is positive recurrent only if $\beta < e^{-1}$, but a new geometric drift condition with β replaced by β^{k-1} can be produced by subsampling X with a fixed subsampling period k. If k is chosen large enough to fix $\beta^{k-1} < e^{-1}$ then the above argument produces a stationary dominating process for the subsampled chain. There is, of course, more to the proof of Theorem 4.14: an explicit coupling between Y and target chains X which satisfies the regeneration requirement (4.22) must be constructed, and Y must also be shown to satisfy part (c) of Definition 4.15. It must also be explained why and how Definition 4.15 delivers a *domCFTP* algorithm. Details are provided in (Kendall 2004).

Note that Y is easy both to sample from in equilibrium and to run in reversed-time, which is essential for implementation of domCFTP. Note too that Y belongs to a family of *universal* dominating processes for geometrically ergodic chains, although this dominator need not generally lead to a practical simulation algorithm. The main difficulties in application are in implementing practical domination derived from (4.23), and in determining whether or not regeneration has occurred when Y visits the set $\{Y \leq h\}$. This task is rendered even less practical if subsampling has taken place, since then detailed knowledge of convolutions of the transition kernel for X is required.

Theorem 4.14 leads to an obvious question: does there exist a similar domCFTP algorithm for chains which converge at a subgeometric rate? This question forms the

starting point for the work in the next chapter, where the problem of moving from geometric to subgeometric ergodicity is investigated extensively.

Les hommes ont oublié cette vérité, dit le renard. Mais tu ne dois pas l'oublier. Tu deviens responsable pour toujours de ce que tu as apprivoisé.

(*"Men have forgotten this truth," said the fox. "But you must not forget it. You become responsible, forever, for what you have tamed."*)

Le Petit Prince, by Antoine de Saint-Exupéry

5. PERFECT SIMULATION FOR SLOW MARKOV CHAINS

We have seen in Section 4.4.3 that the existence of a (possibly impractical) perfect simulation algorithm is guaranteed for a Markov chain if it converges at a geometric rate. In this chapter we extend this result by introducing a new class of positive-recurrent chains (*tame chains*) for which domCFTP is possible in principle. Most of the content of this chapter is based upon the paper by Connor and Kendall (2007) and the associated research report (Connor and Kendall 2006). (It should be noted that these articles have recently been found to contain an error: this mistake has been rectified in what follows. In particular, Lemma 5.8 and consequently the proof of Theorem 5.17 have been corrected.)

Before we begin, we should state clearly what we are permitting to be a part of an 'impractical algorithm'. The Foss and Tweedie (1998) algorithm for uniformly ergodic chains (Theorem 4.11 of this thesis) requires us to be able to identify when regeneration occurs for the target Markov chain X sub-sampled every k time-steps: here k is the order of the whole state-space considered as a small set for X. It also assumes that it is then possible to draw from the regeneration distribution. For the geometric ergodicity result of Kendall (2004) (reviewed in Section 4.4.3), a little more is required: namely that it is possible to couple the target chain X and the dominating chain Y when sub-sampled every k time-steps, and that this domination is preserved while so doing. Furthermore, it also assumes that we can implement the coupling between X and Y in a monotonic fashion even when conditioning on small-set regeneration occurring or not occurring. For the extension to tame chains presented in this chapter, it turns out that we do not need to assume any more than for the geometrically ergodic case, except that now the sub-sampling order k is not fixed for all time, but can vary according to the current value of the dominating process.

Finally, note that such an 'impractical algorithm' is really more of a *strategy*

than an algorithm: we are simply concerned with describing a theoretical method for producing a perfect sample from a distribution, without any consideration of the run-time of such an approach.

5.1 Preliminaries

We begin this chapter with a review of the relevant literature, and also present some drift condition results which will prove useful in the work that follows.

5.1.1 Past research into subgeometrically ergodic chains

Throughout this chapter, X will be a discrete-time Markov chain on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, satisfying the same assumptions as in Section 4.4.1. In particular, X is aperiodic and Harris-recurrent, with stationary distribution π. X is ψ-recurrent, and we write $\mathcal{B}^+(\mathcal{X}) = \{A \in \mathcal{B}(\mathcal{X}) : \psi(A) > 0\}$ for the set of *accessible* sets. A set A is full if $\psi(A^c) = 0$ and absorbing if $P(x, A) = 1$ for all $x \in A$.

In the last chapter we met the definition of a geometrically ergodic chain: this definition was given for the rate of convergence of X in total variation norm. In this chapter we will consider a more general form of convergence, that of (f, r)-ergodicity:

Definition 5.1. The chain X is said to be (f, r)-*ergodic* if there exists a rate function $r : \mathbb{N} \to [1, \infty)$ and a function $f : \mathcal{X} \to [1, \infty)$ such that, for all x in a full and absorbing set $S(f, r)$,

$$r(n) \left\| P^n(x, \cdot) - \pi(\cdot) \right\|_f \to 0 \quad \text{as } n \to \infty. \tag{5.1}$$

A set $A \in \mathcal{B}(\mathcal{X})$ is said to be (f, r)-*regular* if for every $B \in \mathcal{B}^+(\mathcal{X})$,

$$\sup_{x \in A} \mathbb{E}_x \left[\sum_{n=0}^{\tau_B - 1} r(n) f(X_n) \right] < \infty \,. \tag{5.2}$$

A point $x \in \mathcal{X}$ is called (f, r)-regular if $A = \{x\}$ is (f, r)-regular.

If $f \equiv 1$ then equation (5.1) gives the rate of convergence of X in total variation, and equation (5.2) provides information about the moments of τ_B. If $r \equiv 1$ then it is usual practice to call X simply f-regular.

The first major piece of work on general (f, r)-ergodicity when r is a subgeometric rate function was that of Tuominen and Tweedie (1994). (Previous work

concentrated on the cases when $f \equiv 1$, $r \equiv 1$ or the geometrically ergodic case, when $r(n) = \kappa^n$.) The following classes of subgeometric rate function were originally defined in Nummelin and Tuominen (1983):

$$\Lambda_0 = \left\{ r \,:\, r \text{ is a positive increasing function, with } \frac{\log r(n)}{n} \downarrow 0 \text{ as } n \to \infty \right\};$$

$$\Lambda = \left\{ r \,:\, \text{there exists } r_0 \in \Lambda_0 \text{ such that } \liminf_{n\to\infty} \frac{r(n)}{r_0(n)} > 0 \text{ and } \limsup_{n\to\infty} \frac{r(n)}{r_0(n)} < \infty \right\}.$$

The class Λ includes polynomial rate functions (where $r_0(n) = (1+n)^\beta$ for some $\beta > 0$) and some functions which increase faster than polynomially, for example, functions r for which

$$r_0(n) = (1 + \log(n))^\alpha (n+1)^\beta e^{cn^\gamma}$$

for $\alpha, \beta \in \mathbb{R}$, $\gamma \in (0,1)$ and $c > 0$.

It will be convenient for the later work in this chapter to define subsets of Λ as follows:

$$\Lambda(\delta) = \left\{ r \in \Lambda \,:\, r(n) \le O(n^\delta) \text{ as } n \to \infty \right\} \quad \text{and} \quad \Lambda^* = \bigcup_{0<\delta<1} \Lambda(\delta)\,.$$

Thus Λ^* contains, for example, all rate functions $r \in \Lambda$ with

$$r_0(n) = (1 + \log(n))^\alpha (n+1)^\beta$$

for $\alpha \in \mathbb{R}$ and $\beta < 1$. In particular, if $r \in \Lambda^*$ then $r(n)/n \to 0$ as $n \to \infty$.

With this notation in place, we can now state the main result of Tuominen and Tweedie (1994):

Theorem 5.2 (Tuominen and Tweedie (1994))**.** *Suppose that X is ψ-irreducible and aperiodic, and let $f : \mathcal{X} \to [1,\infty)$ and $r \in \Lambda$ be given. The following conditions are equivalent:*

(i) there exists a small set $C \in \mathcal{B}(\mathcal{X})$ such that

$$\sup_{x\in C} \mathbb{E}_x \left[\sum_{n=0}^{\tau_C - 1} r(n) f(X_n) \right] < \infty\,; \tag{5.3}$$

(ii) there exists a sequence $\{V_n\}$ of functions $V_n : \mathcal{X} \to [0,\infty]$, a small set $C \in \mathcal{B}(\mathcal{X})$ and $b \in \mathbb{R}_+$, such that V_0 is bounded on C,

$$V_0(x) = \infty \Rightarrow V_1(x) = \infty\,,$$

and

$$PV_{n+1} \leq V_n - r(n)f + br(n)\mathbf{1}_C, \quad n = 0, 1, 2, \ldots ; \tag{5.4}$$

(iii) there exists an (f, r)-regular set $A \in \mathcal{B}^+(\mathcal{X})$;

Any of these conditions implies that for all (f, r)-regular points x,

$$r(n) \left\| P^n(x, \cdot) - \pi(\cdot) \right\|_f \to 0 \quad \text{as } n \to \infty,$$

and the set of all (f, r)-regular points is full, absorbing, and contains the set $\{V_0 < \infty\}$.

In fact, more can be said here. Careful examination of the proof of Theorem 5.2 shows that the following bound holds (G. Fort, personal communication):

Corollary 5.3. *Suppose X is ψ-irreducible and aperiodic. Let $f : \mathcal{X} \to [1, \infty)$ and $r \in \Lambda$ be such that*

$$\sup_{x \in C} \mathbb{E}_x \left[\sum_{n=0}^{\tau_C - 1} r(n) f(X_n) \right] < \infty .$$

Then there exists $M < \infty$ such that

$$r(n) \left\| P^n(x, \cdot) - \pi(\cdot) \right\|_f \leq M \mathbb{E}_x \left[\sum_{n=0}^{\tau_C - 1} r(n) f(X_n) \right] . \tag{5.5}$$

The chain X is called (f, r)-regular if the conditions of Theorem 5.2 are satisfied and every point is (f, r)-regular. From these definitions it follows that

$$X \text{ is } (f, r)\text{-regular} \Rightarrow X \text{ is } (f, r)\text{-ergodic.}$$

Although this theorem provides a way of determining (f, r)-ergodicity, the sequence of drift conditions contained in inequality (5.4) are extremely hard to check in practice. As such, these multiple drift conditions are hardly ever used directly. However, a single drift condition was introduced by Jarner and Roberts (2002) and shown to imply the existence of the multiple drift conditions when X is polynomially ergodic. Their drift condition is as follows:

Condition **PE**:

There exist constants $0 < b, c < \infty$ and $\alpha \in (0, 1)$, a small set C, and a scale function $V : \mathcal{X} \to [1, \infty)$ which is bounded on C, such that

$$PV \leq V - cV^{\alpha} + b\mathbf{1}_C \,. \tag{5.6}$$

We will refer to condition PE as PE(V, c, α, b, C) when we need to be explicit about the scale function and constants. As with condition GE, this drift condition tells us that $V(X)$ behaves as a supermartingale before X hits C, but now the drift towards the small set occurs at a subgeometric rate (and hence τ_C has no exponential moment).

Condition PE is much easier to check in practice than the multiple drift conditions: at the end of this section an example of how such a drift condition may be established is provided. Jarner and Roberts (2002) use this condition to prove the following theorem:

Theorem 5.4 (Jarner and Roberts 2002)**.** *Suppose X is ψ-irreducible, aperiodic, and satisfies drift condition PE. Then X is (V_ρ, r_ρ)-regular for each $1 \leq \rho \leq 1/(1-\alpha)$, where*

$$V_\rho(x) = V^{1-\rho(1-\alpha)}(x), \quad \textit{and} \, r_\rho(n) = (n+1)^{\rho-1} \,. \tag{5.7}$$

In particular, the following polynomial convergence statements hold for all x:

$$(n+1)^{\rho-1} \left\| P^n(x, \cdot) - \pi(\cdot) \right\|_{V_\rho} \to 0, \ \textit{as } n \to \infty.$$

Note that if $\alpha = 1$ then we regain the geometric drift condition GE, whilst $\alpha = 0$ leads to condition PR. This result shows the trade-off between the rate of convergence and the norm: the larger the latter, the slower the former. The gap $(1 - \alpha)$ is the power lost for each order of convergence gained, with the fastest rate $(r(n) = n^{\alpha/(1-\alpha)})$ corresponding to the total variation norm. The V^α norm is the largest norm for which we can establish convergence (at rate $r \equiv 1$): indeed, given only the drift condition PE, V^α is the largest function that can be guaranteed to have a finite moment under π (Meyn and Tweedie (1993), Theorem 14.0.1).

The result of Jarner and Roberts (2002) was generalised by Douc et al. (2004), who introduced the following drift condition:

Condition **SGE**:

There exist a constant $b < \infty$, a small set C, a scale function $V : \mathcal{X} \to [1, \infty)$ which is bounded on C, and a concave, non-decreasing, differentiable function $\phi : [1, \infty) \to (0, \infty)$ with $\phi'(t) \to 0$ as $t \to \infty$, such that

$$PV \leq V - \phi \circ V + b\mathbf{1}_C \,. \tag{5.8}$$

We will refer to condition SGE as SGE(V, ϕ, b, C) when we need to be explicit about the scale function and constants. Condition PE can easily be recovered from SGE by taking $\phi(x) = cx^\alpha$. Also, as ϕ' is non-increasing, if $\phi'(t) \nrightarrow 0$ then $\phi \circ V \geq (1 - \beta)V$ for sufficiently large V, and so SGE reduces to condition GE.

Under condition SGE, the $(\phi \circ V)$-norm is the largest norm for which convergence can be proved, and this again corresponds to the slowest rate of convergence ($r \equiv 1$). Douc et al. (2004) show that condition SGE again implies the existence of the multiple drift conditions in inequality (5.4). They also prove a generalised form of Theorem 5.4, showing how the convergence rate varies with the norm. In order to state this result, we need a little more notation.

For a concave, non-decreasing, differentiable function $\phi : [1, \infty) \to (0, \infty)$, define

$$H_\phi(v) = \int_1^v \frac{du}{\phi(u)} \,. \tag{5.9}$$

H_ϕ is then another concave, non-decreasing, differentiable function on $[1, \infty)$, which increases to infinity. Its inverse therefore exists and is differentiable and, keeping to the notation of Douc et al. (2004), we define

$$r_\phi(v) = (H_\phi^{-1})'(v) = \phi \circ H_\phi^{-1}(v) \,. \tag{5.10}$$

Thus, for example, if $\phi(x) = cx^\alpha$ (as in the case of a polynomially ergodic chain),

$$H_\phi(v) = \frac{v^{1-\alpha} - 1}{1 - \alpha} \quad \text{and } r_\phi(v) = [(1 - \alpha)v + 1]^{\alpha/(1-\alpha)} \,.$$

Finally, we introduce a useful set of functions, Υ. This is the set of pairs of ultimately non-decreasing functions (Ψ_1, Ψ_2) defined on $[1, \infty)$, such that $\lim_{x\to\infty} \Psi_1(x) = \infty$ or $\lim_{x\to\infty} \Psi_2(x) = \infty$, and

$$\Psi_1(x)\Psi_2(y) \leq x + y \quad \text{for all } x, y \geq 1.$$

Υ includes the pair $(x, 1)$ of course, but also more interesting examples such as $((px)^{1/p}, (qy)^{1/q})$ where $1/p + 1/q = 1$.

The main result of Douc et al. (2004) is then as follows:

Theorem 5.5 (Douc et al. (2004))**.** *Suppose X is ψ-irreducible, aperiodic, and satisfies drift condition SGE(V, ϕ, b, C). Let $(\Psi_1, \Psi_2) \in \Upsilon$. Then X is $(\Psi_2(\phi \circ V), \Psi_1(r_\phi))$-regular:*

$$\mathbb{E}_x \left[\sum_{n=0}^{\tau_C - 1} \Psi_1(r_\phi(n)) \Psi_2(\phi \circ V(X_n)) \right] \leq MV(x)\,, \tag{5.11}$$

for some constant $M < \infty$. In particular, the following convergence statements hold for all x in the full set $\{V < \infty\}$:

$$\Psi_1(r_\phi(n)) \left\| P^n(x, \cdot) - \pi(\cdot) \right\|_{\Psi_2(\phi \circ V)} \to 0, \text{ as } n \to \infty.$$

Once again, this reduces to the result of Jarner and Roberts (2002) when $\phi(x) = cx^\alpha$ and $(\Psi_1, \Psi_2)(x) = (((1-p)x)^{1-p}, (px)^p)$ for some $p \in (0, 1)$.

This concludes the review of past research into subgeometrically ergodic chains. All work on these chains has to date been concerned with three general notions: establishing precise rates of ergodicity under different norms (Tuominen and Tweedie (1994), Jarner and Roberts (2002), Douc et al. (2002), Jarner and Tweedie (2003), Douc et al. (2004)); showing the drift conditions PE and SGE hold for different examples of chains (those papers just cited, as well as Fort and Moulines (2000), Fort and Moulines (2003)); or proving the existence of central limit theorem results arising from the drift conditions (Jarner and Roberts (2002)). All such work, by definition, deals solely with establishing the long-term behaviour of a chain, given knowledge of its one-step transitions via a drift condition. In the subsequent work of this chapter, we will primarily be using drift conditions to study the behaviour of a chain X over a fixed *finite* period. Therefore previous research involving asymptotic behaviour, other than the results reviewed in this section, will not in general be of great use in what follows.

We end this introduction to subgeometric drift conditions with an example of how drift conditions are proved to hold in practice. The following example is probably the simplest possible polynomially ergodic chain.

Example 5.6 (Forward recurrence time chain). Let $\{Y_0, Y_1, \dots\}$ be a sequence of i.i.d. random variables, with common distribution p, taking values on $\mathbb{Z}_+$. Define

$$Z_n = \sum_{i=0}^{n} Y_i \,,$$

so that $\{Z_n\}$ forms a discrete-time renewal process. The *forward recurrence time chain* X is defined by

$$X_n = \inf\left(Z_m - n \;:\; Z_m > n\right), \quad n \geq 0\,. \tag{5.12}$$

(See Meyn and Tweedie (1993), page 44.) The dynamics of X are extremely simple. If $X_n = k > 1$ then at time $n+1$ the forward recurrence time to the next renewal of Z has come down to $k-1$, and thus $X_{n+1} = k-1$. If $X_n = 1$ then a renewal occurs at time $n+1$, and so X_{n+1} is distributed exactly according to p.

$C = \{1\}$ is a regenerative atom for this chain and X is δ_1-irreducible. By construction,

$$\mathbb{E}_1\left[\kappa^{\tau_1}\right] = \sum_n \kappa^n \mathbb{P}_1\left(\tau_1 = n\right) = \sum_n \kappa^n p(n)\,. \tag{5.13}$$

Therefore, by Theorem 4.13, X is geometrically ergodic if and only if the distribution $p(n)$ has geometrically decreasing tails. Let us now assume that this is not the case, and that $\sum_n \kappa^n p(n) = \infty$ for all $\kappa > 1$, but that

$$b = \sum_n n^{1+\varepsilon} p(n) < \infty \tag{5.14}$$

for some $0 < \varepsilon < 1$. Under this assumption, X satisfies condition PE: for $x > 2$,

$$\begin{aligned}
\mathbb{E}_x\left[X_1^{1+\varepsilon}\right] &= (x-1)^{1+\varepsilon} \\
&\leq x^{1+\varepsilon} - (1+\varepsilon)\,(x-1)^{\varepsilon} \\
&\leq x^{1+\varepsilon} - (1+\varepsilon)\,(x^{\varepsilon} - 1) \\
&\leq x^{1+\varepsilon} - \varepsilon x^{\varepsilon}\,.
\end{aligned}$$

Combining this bound with that in equation (5.14) we obtain the PE drift condition

$$\mathbb{E}_x\left[V(X_1)\right] \leq V(x) - \varepsilon V(x)^{\alpha} + b\mathbf{1}_{[x \leq 2]}\,, \tag{5.15}$$

where $V(x) = x^{1+\varepsilon}$, $\alpha = \varepsilon/(1+\varepsilon)$, and b is defined in equation (5.14). Of course, a variant of this example satisfying the more general drift condition SGE can be produced simply by changing the moment condition in equation (5.14) appropriately.

5.1.2 Useful drift condition results

We now present a few simple results arising from the drift conditions for geometrically and subgeometrically ergodic chains, which for ease of reference are repeated here:

Condition **GE**:

There exist positive constants $\beta < 1$ and $b < \infty$, a small set C and a scale function $V : \mathcal{X} \to [1, \infty)$, bounded on C, such that

$$PV \leq \beta V + b\mathbf{1}_C \,. \tag{4.19}$$

Condition **SGE**:

There exist a constant $b < \infty$, a small set C, a scale function $V : \mathcal{X} \to [1, \infty)$ which is bounded on C, and a concave, non-decreasing, differentiable function $\phi : [1, \infty) \to (0, \infty)$ with $\phi'(t) \to 0$ as $t \to \infty$, such that

$$PV \leq V - \phi \circ V + b\mathbf{1}_C \,. \tag{5.8}$$

Recall that the drift condition PE for polynomially ergodic chains is simply condition SGE with $\phi(x) = cx^\alpha$ for some $\alpha \in (0, 1)$.

The first result demonstrates how the scale function V in inequality (4.19) may be changed to obtain a new drift condition using the same small set:

Lemma 5.7. *If the chain X satisfies condition GE(V, β, b, C), then for any $\xi \in (0, 1]$,*

$$PV^\xi \leq (\beta V)^\xi + b^\xi \mathbf{1}_C \,.$$

Thus GE(V, β, b, C) implies GE$(V^\xi, \beta^\xi, b^\xi, C)$.

Proof. Calculus shows that $(x + y)^\xi \leq x^\xi + y^\xi$ for $x, y \geq 0$ and $0 < \xi \leq 1$. The result follows by Jensen's inequality for $(PV)^\xi$, using inequality (4.19) above. □

The next Lemma shows that the geometric drift condition for the chain subsampled at time $n \in \mathbb{N}$ may be made independent of n.

Lemma 5.8. *Suppose X satisfies condition GE(V, β, b, C). Then for any positive time $n \in \mathbb{N}$:*

$$\mathbb{E}_x\left[V(X_n)\right] \leq \beta V(x) + b_1 \mathbf{1}_{C_1}(x)\,,$$

where $b_1 = b/(1-\beta)$ and $C_1 = \left\{x : V(x) \leq b/(\beta(1-\beta)^2)\right\} \cup C$.

The same β, b_1 and C_1 work for all values of n, since the constant b_1 swallows up the higher order terms in β below.

Proof. Iterate the drift condition (4.19) and treat the cases $\{n = 1\}$ and $\{n > 1\}$ separately:

$$\begin{aligned}
\mathbb{E}_x\left[V(X_n)\right] &\leq \mathbb{E}_x\left[\beta^n V(x) + b\sum_{j=1}^{n}\beta^{j-1}\mathbf{1}_C(X_{n-j})\right] \\
&\leq \left(\beta V(x) + b\mathbf{1}_C(x)\right)\mathbf{1}_{[n=1]} + \left(\beta^2 V(x) + \frac{b}{1-\beta}\right)\mathbf{1}_{[n>1]} \\
&\leq \left(\beta V(x) + b\mathbf{1}_C(x)\right)\mathbf{1}_{[n=1]} + \left(\beta V(x) + b_1\mathbf{1}_{C_1}(x)\right)\mathbf{1}_{[n>1]} \\
&\leq \beta V(x) + b_1\mathbf{1}_{C_1}(x)\,.
\end{aligned}$$

□

The following result shows that subgeometrically ergodic chains satisfy a result analogous to Lemma 5.7. However, there is no analogue to Lemma 5.8 when X is subgeometrically ergodic, since the geometric ergodicity case makes essential use of the convergence of the series $\sum \beta^j$.

Lemma 5.9. *If the chain X satisfies condition SGE, then for any concave, non-decreasing, differentiable function $\varphi : [1, \infty) \to (0, \infty)$, there exists $0 < b_1 < \infty$ such that*

$$P(\varphi \circ V) \leq \varphi \circ V - (\phi \circ V)(\varphi' \circ V) + b_1\mathbf{1}_C\,. \tag{5.16}$$

Proof. Since φ' is necessarily non-increasing, it follows that

$$\varphi(z - y) \leq \varphi(z) - y\varphi'(z) \tag{5.17}$$

for all $0 \leq y \leq z$. As usual we write $C = \{x \,:\, V(x) \leq d\}$.

First consider the case when $x \notin C$. By Jensen's inequality:

$$\begin{aligned}
P(\varphi \circ V) &\leq \varphi(PV) \leq \varphi(V - \phi \circ V) \\
&\leq \varphi \circ V - (\phi \circ V)(\varphi' \circ V)\,,
\end{aligned}$$

using inequality (5.17), since $V - \phi \circ V \geq 0$ by condition SGE.

Secondly, for $x \in C$:

$$\begin{aligned} P(\varphi \circ V) &\leq \varphi(PV) \leq \varphi(V - \phi \circ V + b) \\ &= \varphi \circ V - (\phi \circ V)(\varphi' \circ V) + \big(\varphi(V - \phi \circ V + b) - \varphi \circ V + (\phi \circ V)(\varphi' \circ V)\big) \\ &\leq \varphi \circ V - (\phi \circ V)(\varphi' \circ V) + \big(\varphi(d + b) + \phi(d)\varphi'(1)\big) \,. \end{aligned}$$

Therefore inequality (5.16) is satisfied, with $b_1 = \varphi(d + b) + \phi(d)\varphi'(1) < \infty$. □

Note that, as in Lemma 5.7, the same small set C appears in the new drift condition when the scale function is changed in this way. When $\phi(x) = cx^\alpha$ and $\varphi(x) = x^\xi$ for some $\alpha, \xi \in (0, 1)$, this result reduces to Lemma 3.5 of Jarner and Roberts (2002).

Recall the definition of the function H_ϕ, introduced in Section 5.1.1:

$$H_\phi(x) = \int_1^x \frac{du}{\phi(u)} \,.$$

Corollary 5.10. *Suppose X satisfies condition SGE. Then, for $x \notin C$,*

$$\mathbb{E}_x \left[\tau_C\right] \leq H_\phi \circ V(x).$$

Proof. By definition of H_ϕ, $(\phi \circ V)(H'_\phi \circ V) = 1$. Furthermore, H_ϕ is a non-decreasing concave differentiable function on $[1, \infty)$. Lemma 5.9 yields

$$P(H_\phi \circ V) \leq H_\phi \circ V - 1$$

for $x \notin C$, and the result then follows by Theorem 4.12. □

Finally, we have the following Lemma, which follows from the results of the last section and which will prove useful in Section 5.2.5.

Lemma 5.11. *Suppose X is ψ-irreducible, aperiodic and satisfies condition SGE. Let $(\Psi_1, \Psi_2) \in \Upsilon$. Then for any fixed $n \in \mathbb{N}$, there exist constants $c, M < \infty$ such that*

$$\mathbb{E}_x \left[\Psi_2(\phi \circ V(X_n))\right] \leq \frac{MV(x)}{\Psi_1(r_\phi(n))} + c \,. \tag{5.18}$$

Proof. Since V is bounded on the small set C, Theorem 5.5 asserts that

$$\sup_{x\in C} \mathbb{E}_x \left[\sum_{k=0}^{\tau_C - 1} \Psi_1(r_\phi(k)) \Psi_2(\phi \circ V(X_k)) \right] < \infty \,.$$

It follows that, for some constants $R, M < \infty$,

$$\begin{aligned} \Psi_1(r_\phi(n)) \left\| P^n(x,\cdot) - \pi(\cdot) \right\|_{\Psi_2(\phi\circ V)} &\leq R\mathbb{E}_x \left[\sum_{k=0}^{\tau_C - 1} \Psi_1(r_\phi(k)) \Psi_2(\phi \circ V(X_k)) \right] \\ &\leq MV(x) \,, \end{aligned}$$

where the first inequality follows from Corollary 5.3 and the second by Theorem 5.5 once again. Finally, since X satisfies condition SGE, Theorem 14.3.7 of Meyn and Tweedie (1993) confirms that $\pi(\phi \circ V) < \infty$, and so $c = \pi(\Psi_2(\phi \circ V)) < \infty$. This completes the proof. □

5.2 Tame chains

We now turn our attention to the question posed at the end of the last chapter: what can be said about the existence of a perfect simulation algorithm for a chain that converges at a subgeometric rate? Note that if we try to directly produce a dominating process for a subgeometrically ergodic chain by replacing drift condition GE with SGE in the proof of Theorem 4.14, then the resulting process is transient. (In fact, this process is a D/M/1 queue with unit rate Exponential service times again, but now with the arrival rate increasing (and unbounded) as the number of people in the queue increases!) Therefore another approach is needed.

The principal idea behind the subsequent work is to investigate when it is possible to subsample a subgeometrically ergodic chain X to produce a geometrically ergodic chain. For non-geometrically ergodic chains a fixed subsampling interval will not work and so we seek an appropriate simple adaptive subsampling scheme. A similar scheme can then be used to *delay* the dominating process Y for geometrically ergodic chains (constructed in Section 4.4.3), and to show that this new process D dominates the chain $V(X)$ at the times when D moves.

Several issues must be addressed in order to derive a perfect simulation algorithm using this idea:

1. what is an appropriate adaptive subsampling scheme?

2. when does such a scheme exist?

3. how does the dominating process D dominate $V(X)$ when D moves?

4. can we simulate D in equilibrium, and in reversed-time?

The answers to these questions are quite subtle.

5.2.1 Adaptive subsampling

We begin by defining more carefully what we mean by an adaptive subsampling scheme.

Definition 5.12. An *adaptive subsampling scheme* for the chain X, with respect to a scale function V, is a sequence of stopping times $\{\theta_n\}$ defined recursively by

$$\theta_0 = 0; \quad \theta_{n+1} = \theta_n + F(V(X_{\theta_n})), \tag{5.19}$$

where $F : [1, \infty) \to \{1, 2, \ldots\}$ is a deterministic function.

Remark 5.13. Note that a set of stopping times $\{\theta_n\}$ such that $\{X_{\theta_n}\}$ is *uniformly* ergodic can be produced as follows. Using the Athreya-Nummelin split-chain construction (Meyn and Tweedie 1993) we may suppose there is a state ω with $\pi(\omega) > 0$. Define

$$F(V(x)) = \min \left\{ m > 0 : \mathbb{P}_x \left(X_m = \omega \right) > \frac{\pi(\omega)}{2} \right\}. \tag{5.20}$$

Then the time until $\{X_{\theta_n}\}$ hits ω from any starting state x is majorised by a Geometric random variable with success probability $\pi(\omega)/2$. This implies that the subsampled chain is uniformly ergodic, as claimed. F as defined in equation (5.20) depends upon knowledge of π however, and this is obviously unavailable (it is the distribution from which we are trying to sample!). This example shows that adaptive subsampling can have drastic effects on X. However, construction of a domCFTP algorithm for X using this subsampling scheme (in the manner to be described in Section 5.2.3) turns out to be impossible unless X is itself uniformly ergodic.

Reverting to the previous discussion, suppose that there is an explicit adaptive subsampling scheme such that the chain $X' = \{X_{\theta_n}\}$ satisfies condition GE with

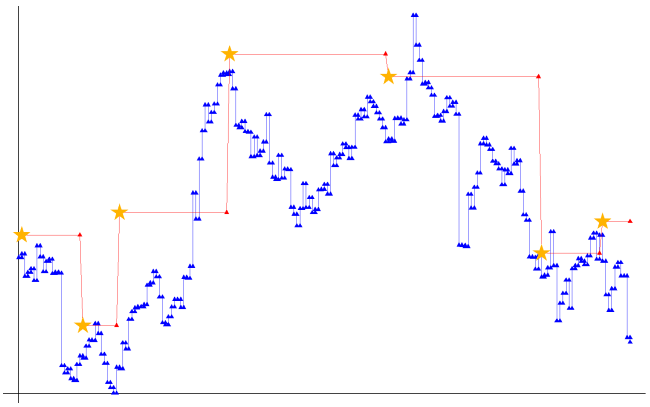


Fig. 5.1: Construction of the delayed dominating process. The chain $V(X)$ (blue) is dominated by Y at times $\{\sigma_n\}$ (marked with stars). The process D (red) uses the jumps of Y (thus maintaining domination at times $\{\sigma_n\}$) and is defined to be constant on the intervals $[\sigma_n, \sigma_{n+1})$.

drift parameter $\beta < e^{-1}$. Then a candidate dominating process D can be produced for $V(X)$ in the following way. Begin with an exponential queue workload process Y that dominates $V(X')$ (as in Section 4.4.3). Then *slow down* Y by generating *pauses* using the function F to produce the process D. That is, given $D_0 = Y_0 = z$, pause D by setting

$$D_1 = D_2 = \ldots = D_{F(z)-1} = z.$$

Then define the law of $D_{F(z)}$ by $\mathcal{L}(D_{F(z)} \mid D_{F(z)-1} = z) = \mathcal{L}(Y_1 \mid Y_0 = z)$. Iteration of this construction leads to a sequence of times $\{\sigma_n\}$ at which D moves, defined recursively by

$$\sigma_{n+1} = \sigma_n + F(D_{\sigma_n}),$$

with D constant on each interval of the form $[\sigma_n, \sigma_{n+1})$ (see Figure 5.1).

Such a process D is a plausible candidate for a dominating process. To be suitable for use in a domCFTP algorithm however, it must be possible to compute its equilibrium distribution. Now, D as we have just defined it is only a semi-Markov process: it is Markovian at the times $\{\sigma_n\}$, but not during the delays between jumps. To remedy this, augment the chain by adding a second coordinate N that measures the time until the next jump of D. (In fact, N is an example of a forward recurrence time chain - recall Example 5.6.) This yields the Markov chain $\{(D_n, N_n)\}$ on

$[0,\infty) \times \{1,2,\ldots\}$ with transitions controlled by:

$$\begin{aligned}
&\mathbb{P}\left(D_{n+1} = D_n,\, N_{n+1} = N_n - 1 \mid D_n,\, N_n\right) = 1\,, \quad \text{if } N_n \geq 2;\\
&\mathbb{P}\left(D_{n+1} \in E \mid D_n = z,\, N_n = 1\right) = \mathbb{P}\left(Y_1 \in E \mid Y_0 = z\right)\,,\\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{for all measurable } E \subseteq [1,\infty);\\
&\mathbb{P}\left(N_{n+1} = F(D_{n+1}) \mid D_n,\, N_n = 1,\, D_{n+1}\right) = 1.
\end{aligned}$$

Using the standard equilibrium equations, if $\tilde{\pi}$ is the equilibrium distribution of (D, N) then

$$\tilde{\pi}(z,1) = \tilde{\pi}(z,2) = \ldots = \tilde{\pi}(z,F(z)),$$

and thus $\pi_D(z) = \tilde{\pi}(z,\cdot) \propto \pi_Y(z)F(z)$. Hence the equilibrium distribution of D is the equilibrium of Y re-weighted using F. It is a classical probability result (see Appendix) that the equilibrium distribution of the queue workload U is a mixture of an atom at zero with an Exponential distribution of rate $(1-\eta)$, where η is the smallest positive root of

$$\eta = \beta^{1-\eta}.$$

(Note that $0 < \eta < 1$ since $\beta < e^{-1}$.) Since $Y \propto \exp(U)$, the equilibrium density of Y, π_Y, satisfies

$$\pi_Y(z) \propto z^{-(2-\eta)}. \tag{5.21}$$

Re-weighting Y using F yields the equilibrium density of D:

$$\pi_D(z) \propto F(z) z^{-(2-\eta)}. \tag{5.22}$$

A suitable function F must therefore satisfy $F(z) < z^{1-\eta}$ in order to obtain a probability density in (5.22): in particular, this means that $F(z)/z \to 0$ as $z \to \infty$.

5.2.2 Tame and wild chains

The above discussion motivates the following definition of a *tame* chain.

Definition 5.14. A Markov chain X is *tame with respect to a scale function* V if the following two conditions hold:

(a) there exists a small set $C' = \{x : V(x) \leq d'\}$, and a non-decreasing *taming function* $F : [1, \infty) \to \{1, 2, \ldots\}$ of the form

$$F(z) = \begin{cases} \lceil g(z) \rceil & z > d' \\ 1 & z \leq d' \end{cases} \tag{5.23}$$

for some increasing function $g \in \Lambda(\delta)$, $\delta \in (0, 1)$, such that the chain $X' = \{X_{\theta_n}\}$ possesses the drift condition

$$PV \leq \beta V + b' \mathbf{1}_{C'}, \tag{5.24}$$

where $\{\theta_n\}$ is an adaptive sampling scheme defined using F, as in (5.19);

(b) the constants δ and β above satisfy

$$\log \beta < \delta^{-1} \log(1 - \delta). \tag{5.25}$$

We say that X is *tamed (with respect to V) by the function F*. We may also simply say that X is *tame*, without mention of a specific scale function. A chain that is not tame is said to be *wild*.

Thus a tame chain is one for which it is possible to exhibit an explicit adaptive subsampling scheme using a function F of the form in equation (5.23), and for which the subsampled chain so produced is geometrically ergodic with sufficiently small β. Note that all geometrically ergodic chains are trivially tame: if X satisfies condition GE(V, β, b, C) then X is tamed by the function

$$F(z) = k \qquad \text{for } z > \sup_{y \in C} V(y),$$

for any integer $k > 1 - 1/\log \beta$.

Definition 5.14 is strongly motivated by the discussion in Section 5.2.1. From equation (5.23) we see that F produces a simple adaptive subsampling scheme as in Definition 5.12. F is also a non-decreasing function, which accords with our

intuition: if $V(X)$ is large then we expect to wait longer before subsampling again, to create enough drift in the chain to produce a geometric Foster-Lyapunov condition. Requirement (b) of Definition 5.14 is made for two reasons. Firstly, it ensures that $\beta < e^{-1}$, and so delivers ergodicity of the $D/M/1$ queue workload U used in the construction of Y. Secondly, it ensures that the weighted equilibrium distribution of Y using F (as described at the end of Section 5.2.1), is a proper distribution; this will be shown in the proof of Theorem 5.17.

Kendall (2004) shows that a dominating process exists for $V(X')$ even if $\beta > e^{-1}$, but recall that this involves a further subsampling of X' with a fixed period k. Here $\beta < e^{-1}$ is made a requirement of the adaptive subsampling process to avoid this situation, since further subsampling of X' would result in a composite non-deterministic subsampling scheme.

Example 5.15 (Example 5.6 revisited). Recall the definition of the forward recurrence time chain:

$$X_n = \inf\left(Z_m - n \ : \ Z_m > n\right), \quad n \geq 0\,,$$

where $\{Z_n\}$ is a discrete-time renewal process. In equation (5.15) it was shown that, under a mild condition, X satisfies the PE drift condition

$$\mathbb{E}_x\left[V(X_1)\right] \leq V(x) - \varepsilon V(x)^{\alpha} + b\mathbf{1}_{[x\leq 2]}\,, \tag{5.26}$$

where $0 < \varepsilon < 1$, $V(x) = x^{1+\varepsilon}$, $\alpha = \varepsilon/(1+\varepsilon)$, and $b < \infty$.

Due to its very simple dynamics, it is trivial to show that X is tame. Define $\delta = (1+\varepsilon)^{-1}$, and fix β such that $\log\beta < \delta^{-1}\log(1-\delta)$. Now let

$$\lambda = 1 - \beta^{1/(1+\varepsilon)}\,,$$

and define the taming function F by

$$F(z) = \begin{cases} \lceil \lambda\, z^{\delta} \rceil & z > 2 \\ 1 & z \leq 2\,. \end{cases} \tag{5.27}$$

Due to the choice of δ and β above, part (b) of Definition 5.14 is automatically satisfied.

Suppose that $X_0 = x > 2$. Using F to construct an adaptive sampling scheme as in Definition 5.12 we see that

$$\theta_1 = F(V(x)) = \lceil \lambda x \rceil\ ,$$

and so

$$\begin{aligned}\mathbb{E}_x\left[V(X_{\theta_1})\right] &= (x-\theta_1)^{1+\varepsilon} \\ &\leq (1-\lambda)^{1+\varepsilon}V(x) \\ &= \beta V(x)\,.\end{aligned}$$

Finally, if $x \leq 2$,

$$\mathbb{E}_x\left[V(X_1)\right] \leq \beta V(x) + b'\,,$$

for $b' = 2(1-\beta) + b < \infty$. Therefore X is tame, as claimed. Note that, in proving this chain to be tame, the behaviour of X_{θ_1} when $x \in C' = [1,2]$ was trivial to deal with. This will always be the case, since the small set C' is a sub-level set (and so it is always possible to deal with the drift condition on C' by making the constant b' in (5.24) large enough). In the future therefore, we may simply restrict attention to proving the geometric drift condition (5.24) for the subsampled chain X' when $X'_0 \notin C'$.

The main theorem of this chapter is the following:

Theorem 5.16. *Suppose X is tame with respect to a scale function V. Then there exists a domCFTP algorithm for X using a dominating process based on V.*

Theorem 5.16 is true for all geometrically ergodic chains by the result of Kendall (2004). As with the results of Foss and Tweedie (1998) and Kendall (2004), there is no reason to suppose this algorithm to be implementable in practice. The proof of Theorem 5.16 results directly from Theorem 5.17 and the discussion in Section 5.2.3 below, where a description of the algorithm is given.

Theorem 5.17. *Suppose X satisfies the weak drift condition $PV \leq V + b\mathbf{1}_C$, and that X is tamed with respect to V by the function*

$$F(z) = \begin{cases} \lceil g(z) \rceil & z > d' \\ 1 & z \leq d', \end{cases}$$

for some increasing function $g \in \Lambda(\delta)$ with the resulting subsampled chain X' satisfying a drift condition $PV \leq \beta V + b'\mathbf{1}_{[V \leq d']}$, with $\log\beta < \delta^{-1}\log(1-\delta)$. Then there exists a stationary ergodic process D which dominates $V(X)$ at the times $\{\sigma_n\}$ when D moves.

Proof. Since $g \in \Lambda(\delta)$, we may prove this result assuming that $g(z) = \lambda z^{\delta}$: the more general result follows by increasing the size of the small set C^* that is constructed below.

As mentioned earlier, D will be constructed by starting with a process Y and pausing it using F. First choose $\beta^* > \beta$ such that

$$\log \beta < \log \beta^* < \delta^{-1} \log(1-\delta). \tag{5.28}$$

(That this is possible is a result of the definition of tameness.)

Suppose that $D_{\sigma_n} = z$, and that $V(X_{\sigma_n}) = V(x) \leq z$. We wish to show that $D_{\sigma_{n+1}}$ can be made to dominate $V(X_{\sigma_{n+1}})$, where $\sigma_{n+1} = \sigma_n + F(z)$ is the time at which D next moves. Domination at successive times $\{\sigma_j\}$ at which D moves then follows inductively. For simplicity in the calculations below we set $\sigma_n = 0$.

Our aim is to control $\mathbb{E}_x\left[V(X_{F(z)})\right]$, recalling that $F(z)$ is deterministic and that $F(V(x)) \leq F(z)$.

$$\begin{aligned}
\mathbb{E}_x\left[V(X_{F(z)})\right] &= \mathbb{E}_x\left[V(X_{F(V(x))})\right] + \mathbb{E}_x\left[V(X_{F(z)}) - V(X_{F(V(x))})\right] \\
&= \mathbb{E}_x\left[V(X'_1)\right] + \mathbb{E}_x\left[V(X_{F(z)}) - V(X_{F(V(x))})\right] \\
&\leq \beta V(x) + b' \mathbf{1}_{[V(x) \leq d']} + b\left[F(z) - F(V(x))\right] \\
&\leq \beta z + b' + b(\lambda+1) z^{\delta} \\
&\leq \beta^* z, \quad \text{for } z \geq h^*,
\end{aligned} \tag{5.29}$$

where $h^* < \infty$ is a constant chosen sufficiently large for inequality (5.29) to hold. The first inequality in this sequence holds due to the drift conditions satisfied by X' and X. The second follows from the definition of F and the assumption that $V(x) \leq z$.

Now define the process $Y = h^* \exp(U)$, where U is the system workload of a $D/M/1$ queue with arrivals every $\log(1/\beta^*)$ time units and service times being independent and of unit Exponential distribution. Positive recurrence of U follows from inequality (5.28). Pause Y using F (as described on page 130) and call the resulting process D. The stationary distribution of D, as shown at the end of Section 5.2.1, is given by

$$\begin{aligned}
\pi_D(z) &\propto F(z) z^{-(2-\eta)} \\
&\propto z^{-(2-\eta-\delta)} \quad (\text{for } z > h^*),
\end{aligned} \tag{5.30}$$

where $\eta < 1$ is the smallest positive solution to the equation

$$\eta = \beta^{*(1-\eta)}.$$

Now, by the choice of β^* above,

$$(1-\eta)^{-1}\log\eta = \log\beta^* < \delta^{-1}\log(1-\delta),$$

and so $\eta < 1-\delta$. Hence $2-\eta-\delta > 1$, and so it is evident from equation (5.30) that π_D is a proper density.

Finally, observe that D takes values in $[h^*, \infty)$. Inequality (5.29) therefore shows that if $D_0 = z$ and $X_0 = x$ with $V(x) \leq z$, then

$$\mathbb{E}_x\left[V(X_{F(z)})\right] \leq \beta^* z\,.$$

As in the proof of Theorem 4.14, it follows that $V(X_{F(z)})$ can be dominated by $D_{F(z)}$ (Lindvall 2002), as required. □

Note that questions 1 and 3 of page 129 have now been answered: we have defined what is meant by an adaptive subsampling scheme and shown that, if this takes a particular (power function) form, a stationary process D that dominates $V(X)$ at times $\{\sigma_n\}$ can be produced.

5.2.3 The domCFTP algorithm for tame chains

In this section we describe the domCFTP algorithm for tame chains, and hence complete the proof of Theorem 5.16. We begin this by answering question 4 of page 129, by showing how to simulate (D, N) in equilibrium, and in reversed-time. Furthermore this simulation is quite easy to implement when $F = \lceil \lambda z^\delta \rceil$, which we shall assume in what follows for simplicity.

The first point to make here is that one can simulate easily from π_D using rejection sampling (Robert and Casella 2004): using equation (5.22), for some constant $\gamma > 0$,

$$\begin{aligned}\pi_D(z) &= \gamma\left(\frac{1}{2}\frac{\lceil \lambda\, z^\delta\rceil}{\lambda\, z^\delta}\right)\frac{1}{z^{2-\eta-\delta}}\\ &= \gamma\, p(z) g(z),\end{aligned}$$

where $p(z) \in [1/2, 1]$, and $g(z)$ is a Pareto density (since $2-\eta-\delta > 1$, as in the proof of Theorem 5.17). Now, given $D_0 = z_0$ as a draw from π_D, set $N_0 = n_0$,

where $n_0 \sim \text{Uniform}\{1, 2, \ldots, F(z_0)\}$. It follows from the construction of (D, N) in Section 5.2.1 that $(D_0, N_0) \sim \tilde{\pi}$, as required. The chain (D, N) may then be run in reversed-time using the following algorithm (see Figure 5.2):

Algorithm 5.18 (Dominating process (D, N) in reversed-time)**.**

set $j \leftarrow -1$

for $i = 0, -1, -2, \ldots$:

  if $N_i < F(D_i)$:

    $D_{i-1} = D_i$;

    $N_{i-1} = N_i + 1$;

  else

    set $\sigma_j = i$ and $j \leftarrow j - 1$;

    draw D_{i-1} from the reverse kernel $q(D_i\,;\, dz)$, where

$$q(z'\,;\, dz)\pi_Y(z')dz' = p(z\,;\, dz')\pi_Y(z)dz,$$

    and $p(z\,;\, dz')$ is the transition kernel for $Y = \exp(U)$;

    $N_{i-1} \leftarrow 1$

We now show that D is a dominating process for X (at the times when D moves) based on the scale function V, with threshold h^* (recall Definition 4.15). We define the sub-level set C^* by $C^* = \{x \,:\, V(x) \leq h^*\}$: this set is m-small (say).

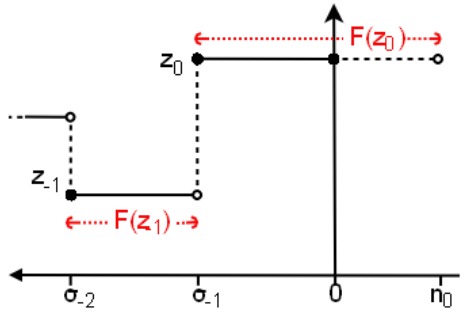


Fig. 5.2: Construction of D in reversed-time.

Firstly, the proof of Theorem 5.17 shows that the link between stochastic domination and coupling (Lindvall 2002) may be exploited to couple the various $X^{x,\sigma_{-M}}$ with D such that for all $n \leq M$,

$$V(X^{x,\sigma_{-M}}_{\sigma_{-n}}) \;\leq\; D_{\sigma_{-n}} \;\;\Rightarrow\;\; V(X^{x,\sigma_{-M}}_{\sigma_{-(n-1)}}) \;\leq\; D_{\sigma_{-(n-1)}}. \tag{5.31}$$

We now turn to part (b) of Definition 4.15. Since C^* is m-small, there exists a probability measure ν and a scalar $\varepsilon \in (0,1)$ such that for all Borel sets $B \subset [1,\infty)$, whenever $V(x) \leq h^*$

$$\mathbb{P}\left(V(X_m) \in B \mid X_0 = x\right) \geq \varepsilon\nu(B).$$

Suppose first of all that $F(h^*) \geq m$. In this case,

$$\mathbb{P}\left(V(X_{F(h^*)}) \in B \mid X_0 = x\right) \geq \varepsilon P_\nu^{F(h^*)-m}(B), \tag{5.32}$$

and so C^* is $F(h^*)$-small. Furthermore, the stochastic domination which has been arranged in the construction of D means that for all $u \geq 1$, whenever $V(x) \leq y$,

$$\mathbb{P}\left(V(X_{F(y)}) > u \mid X_0 = x\right) \leq \mathbb{P}\left(Y_1 > u \mid Y_0 = y\right).$$

We can couple in order to arrange for regeneration if a probability measure $\tilde{\nu}$ can be identified, defined solely in terms of $P_\nu^{F(h^*)-m}$ and the dominating jump distribution $\mathbb{P}\left(Y_1 \geq u \mid Y_0 = y\right)$, such that for all $u \geq 1$, whenever $V(x) \leq y$:

$$\begin{aligned}
\mathbb{P}_x\left(V(X_{F(y)}) > u\right) - \varepsilon P_\nu^{F(h^*)-m}((u,\infty)) &\leq \mathbb{P}_y\left(Y_1 > u\right) - \varepsilon\tilde{\nu}((u,\infty)); \\
P_\nu^{F(h^*)-m}((u,\infty)) &\leq \tilde{\nu}((u,\infty)); \\
\text{and} \quad \mathbb{P}_y\left(Y_1 \in E\right) &\geq \varepsilon\tilde{\nu}(E),
\end{aligned} \tag{5.33}$$

for all measurable $E \subseteq [1,\infty)$.

Recall the following result, a proof of which is provided in Kendall (2004):

Lemma 5.19. *Suppose U, V are two random variables defined on $[1,\infty)$ such that*

(a) The distribution $\mathcal{L}(U)$ is stochastically dominated by the distribution $\mathcal{L}(V)$:

$$\mathbb{P}\left(U > u\right) \leq \mathbb{P}\left(V > u\right) \quad \textit{for all } u \geq 1;$$

(b) U satisfies a minorization condition: for some $\beta \in (0,1)$ and probability measure ψ,

$$\mathbb{P}(U \in E) \geq \beta\psi(E) \quad \textit{for all Borel sets } E \subseteq [1,\infty).$$

Then there is a probability measure μ stochastically dominating ψ and such that $\beta\mu$ is minorised by $\mathcal{L}(V)$. Moreover, μ depends only on $\beta\psi$ and $\mathcal{L}(V)$.

Therefore, using Lemma 5.19, $\mathcal{L}(X_{\sigma_{-(n-1)}} \mid X_{\sigma_{-n}} = x)$ may be coupled to $\mathcal{L}(D_{\sigma_{-(n-1)}} \mid D_{\sigma_{-n}} = y)$ whenever $V(x) \leq y$, in a way that implements stochastic domination and ensures that all the $X_{\sigma_{-(n-1)}}$ can regenerate simultaneously whenever $D_{\sigma_{-n}} \leq h^*$.

If $F(h^*) < m$ however, part (b) of Definition 4.15 is harder to satisfy. It then becomes necessary, when $D_0 = h^*$, for D to dominate $V(X)$ not at time $\sigma_1 = F(h^*)$ but at time

$$\sigma_k = \inf_{j \geq 2} \{\sigma_j \ : \ \sigma_j \geq m\} \ .$$

This involves running the chain D from time zero until the first time after m that it jumps (σ_k): since $\sigma_k \geq m$ it follows as for inequality (5.32) that C^* is σ_k-small. In order to couple the chains in a way that implements domination and which allows all the target chains to regenerate, it then becomes necessary to satisfy the three inequalities in (5.33), but now with $F(y) = F(h^*) = \sigma_k$ and $Y_1 = Y_k$, for any $k \geq 1$. This unfortunately renders the algorithm less practical, which is an issue that we are currently trying to resolve. For simplicity of exposition in what follows, it is assumed henceforth that $F(h^*) \geq m$.

Finally, it is easy to see that part (c) of Definition 4.15 is satisfied: the system workload U of the queue will hit zero infinitely often and therefore D will hit level h^* infinitely often.

We can now describe a perfect simulation algorithm based on X which yields a draw from the equilibrium distribution, the final step of which is depicted in Figure 5.3.

Algorithm 5.20 (Perfect simulation algorithm for tame chains)**.**

`simulate` D `backwards in time` (as a component of the stationary process (D, N), using Algorithm 5.18) `until the most recent` $\sigma_{-M} < 0$ `for which` $D_{\sigma_{-M}} \leq h^*$;

`while coalescence does not occur at time` σ_{-M}`:`

`extend` D `backwards till the most recent` $\sigma_{-M'} < \sigma_{-M}$ `for which` $D_{\sigma_{-M'}} \leq h^*$;

$M \leftarrow M'$;

`simulate the coupled` X `forwards at times` $\sigma_{-M}, \sigma_{-(M-1)}, \ldots, \sigma_{-1}$, `starting with the unique state produced by the coalescence event at time` σ_{-M}; (see Figure 5.3)

`run` X `forward (from its unique state) from time` σ_{-1} `to time 0 (without reference to` D`);`

`return` X_0.

Lemma 5.21. *The output of the above algorithm is a draw from the stationary distribution of the target chain X.*

Proof. The stochastic domination of equation (5.31) and Theorem 2.4, Ch. IV of Lindvall (2002) guarantee the existence of a joint transition kernel $P_{X,D}$ that provides

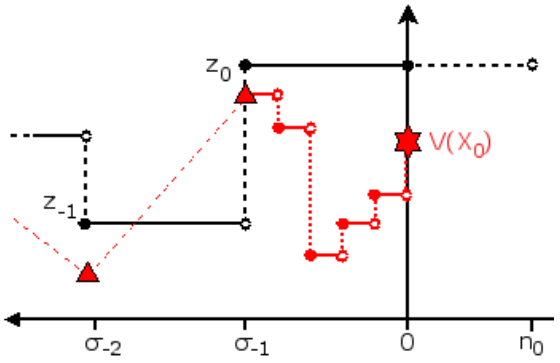


Fig. 5.3: Final stage of the domCFTP algorithm: D (black circles •) dominates $V(X)$ (red triangles ▲) at times $\{\sigma_{-n}\}$. To obtain the draw from equilibrium, X_0, X can be run from time σ_{-1} to 0 without reference to D after time σ_{-1}.

domination of X by D when D moves, and such that the marginal distributions of X and D are correct. That is, for $x \leq y$, with $n = F(y)$, for all $z \geq 1$:

$$\begin{aligned} P^n_{X,\,D}\left(x, y\,;\; V^{-1}((z,\infty)), [1,z]\right) &= 0\,; \\ \int_{V^{-1}([1,z])} \int_1^\infty P^n_{X,\,D}\,(x, y\,;\; du, dv) &= P^n_X\left(x\,;\; V^{-1}([1,z])\right)\,; \\ \int_{\mathcal{X}} \int_1^z P^n_{X,\,D}\,(x, y\,;\; du, dv) &= P^n_D\,(y\,;\; [1,z])\,. \end{aligned}$$

The chains X and D (run forwards) may therefore be constructed in either of two ways.

1. Given $D_{\sigma_{-m}}$ and $X_{\sigma_{-m}} \leq D_{\sigma_{-m}}$, with $n = F(D_{\sigma_{-m}})$:

   - draw $D_{\sigma_{-(m-1)}}$ from the probability kernel

   $$P^n_D\left(D_{\sigma_{-m}}\,;\,\cdot\,\right);$$

   - draw $X_{\sigma_{-(m-1)}}$ from the regular conditional probability

   $$\frac{P^n_{X,\,D}\left(X_{\sigma_{-m}}, D_{\sigma_{-m}}\,;\,\cdot\,, D_{\sigma_{-(m-1)}}\right)}{P^n_D\left(D_{\sigma_{-m}}\,;\, D_{\sigma_{-(m-1)}}\right)}\,;$$

   - draw $X_{\sigma_{-m}+1}, X_{\sigma_{-m}+2}, \ldots, X_{\sigma_{-(m-1)}-1}$ as a realisation of X conditioned on the values of $X_{\sigma_{-m}}$ and $X_{\sigma_{-(m-1)}}$ (that is, as a *Markov bridge* between $X_{\sigma_{-m}}$ and $X_{\sigma_{-(m-1)}}$).

2. Given $D_{\sigma_{-m}}$ and $X_{\sigma_{-m}} \leq D_{\sigma_{-m}}$, with $n = F(D_{\sigma_{-m}})$:

   - draw $X_{\sigma_{-m}+1}, X_{\sigma_{-m}+2}, \ldots, X_{\sigma_{-(m-1)}}$ using the normal transition kernel for X. Note that the distribution of $X_{\sigma_{-(m-1)}}$ is exactly the same as if it were drawn directly from $P^n_X\left(X_{\sigma_{-m}}\,;\,\cdot\,\right)$;

   - draw $D_{\sigma_{-(m-1)}}$ from the regular conditional probability

   $$\begin{aligned} &P^n_{D|\{X\}}\left(\,\cdot\mid D_{\sigma_{-m}}, X_{\sigma_{-m}}, X_{\sigma_{-m}+1}, \ldots, X_{\sigma_{-(m-1)}}\right) \\ &\qquad = \frac{P^n_{X,\,D}\left(X_{\sigma_{-m}}, D_{\sigma_{-m}}\,;\, X_{\sigma_{-(m-1)}},\,\cdot\,\right)}{P^n_X\left(X_{\sigma_{-m}}\,;\, X_{\sigma_{-(m-1)}}\right)}\,. \end{aligned}$$

Each of these two methods produces chains X and D which satisfy the stochastic domination of equation (5.31). Method 1 is that which is effectively used by the algorithm, although there is no need for the final superfluous step (the Markov bridge) when implementing the algorithm. Method 2, however, makes it clear that X has the correct Markov transition kernel to be the required target chain. Furthermore, the equivalence of the two schemes proves the validity of the final step of the algorithm, where the chain X is run from time σ_{-1} to 0 without reference to D.

Finally, the proof that the algorithm returns a draw from equilibrium follows a common renewal theory argument. Consider a stationary version of the chain X, $\hat{X}$ say, run from time $-\infty$ to 0. The regenerations of $\hat{X}$ (when it visits the small set C^*), and those of D (when it hits level h^*) form two positive recurrent renewal processes (with that of $\hat{X}$ being aperiodic). Therefore, if D is started far enough in the past, there will be a time $-T$ at which both $\hat{X}$ and D regenerate simultaneously. Now consider the process $\tilde{X}_n = \hat{X}_n \mathbf{1}_{[n<-T]} + X_n \mathbf{1}_{[n\geq -T]}$. Clearly, $\tilde{X}$ is stationary and follows the same transitions of X from time $-T$ to 0. Thus $X_0 = \tilde{X}_0 \sim \pi$, and so the output of the algorithm is indeed a draw from the required equilibrium distribution. □

This concludes the proof of Theorem 5.16: we have produced a perfect simulation algorithm based on the scale function V for the tame chain X.

5.2.4 Extended state-space CFTP for tame chains

In the previous section, a perfect simulation algorithm for tame chains was described, and shown to sample from the correct distribution. This algorithm was based upon the domCFTP algorithm of Kendall (2004) for geometrically ergodic chains. However, the algorithm for tame chains as presented is not strictly a domCFTP algorithm. Recall that the idea behind domCFTP is that of *horizontal coupling*: all sufficiently early starts (i.e. at times $-n, -(n+1), \ldots$, for large enough n) from a given state x lead to the same result at time 0. The construction of Algorithm 5.20 only insists upon sufficiently early starts *at times σ_{-n} when the dominating process D moves* leading to the same state at time 0. Thus chains started within a time interval of the form $(\sigma_{-n}, \sigma_{-(n-1)})$ are not guaranteed to lead to the same state at time 0, no matter how large n is, since such chains will not be dominated by D. This is why

it was necessary to explicitly prove the correctness of the algorithm in Lemma 5.21, rather than by simply appealing to usual domCFTP arguments.

However, it is possible to modify Algorithm 5.20 so that it does fit into the more normal dominated CFTP framework, by using the extended state-space CFTP construction of Cai and Kendall (2002), presented in Section 4.2.3. In this section we show how to do this.

In order to use the extended state-space CFTP construction we need to define an embedding of the space of interest within another, partially ordered, space. To do this, we shall not work directly with $\mathcal{X}$, but with the space $\tilde{\mathcal{X}}$, defined below. In the following we shall write V_t for $V(X_t)$. Recall that F is the (deterministic) function used in the construction of the dominating process (D, N) in the last section.

Definition 5.22. Define $\tilde{\mathcal{X}}$ and $\tilde{\mathcal{Y}}$, with $\tilde{\mathcal{X}} \subset \tilde{\mathcal{Y}}$, as follows:

$$\tilde{\mathcal{X}} = \{(d, n, 2, z) \,:\, 1 \leq d < \infty,\ n \in \mathbb{N},\ \text{and } z \text{ is a } (F(d) + 1 - n)\text{-tuple}\},$$
$$\tilde{\mathcal{Y}} = \{(d, n, w, z) \,:\, 1 \leq d < \infty,\ n \in \mathbb{N},\ w \in \{0, 1, 2\},\ \text{and } z \text{ is a } (F(d) + 1 - n)\text{-tuple}\}.$$

The partial ordering $\preceq$ on $\tilde{\mathcal{Y}}$ is defined by

$$x = \left(d^{(1)}, n^{(1)}, w^{(1)}, z^{(1)}\right) \preceq \left(d^{(2)}, n^{(2)}, w^{(2)}, z^{(2)}\right) = y$$

if and only if one of the following two conditions holds:

- either $x = y$;
- or $d^{(1)} = d^{(2)}$, $n^{(1)} = n^{(2)}$, and $w^{(1)} > w^{(2)}$.

Note that $\tilde{\mathcal{X}}$ is embedded at the bottom of $\tilde{\mathcal{Y}}$ in the manner of section 4.2.3, since if $x \in \tilde{\mathcal{X}}$, $y \in \tilde{\mathcal{Y}}$ and $y \preceq x$ then $x = y$.

We now identify chains $\tilde{Y}^{(n)}$ on $\tilde{\mathcal{Y}}$ which are identically distributed up to a shift in time (recall from Section 4.2.3 that $\tilde{Y}^{(n)}$ is a chain started at time $-n < 0$). To do this we use the dominating process (D, N): let $\{(D_t, N_t), -\infty < t \leq 0\}$ be a realisation of this chain constructed by drawing (D_0, N_0) from its equilibrium distribution $\tilde{\pi}$ (see page 131) and running (D, N) backwards in time (using Algorithm 5.18). Then initially, at time $-n$, we set

$$\tilde{Y}^{(n)}_{-n} = \left(D_{-n}, N_{-n}, 0, \overrightarrow{D_{-n}}\right), \tag{5.34}$$

where $\overrightarrow{D_{-n}}$ is a vector of length $(F(D_{-n}) + 1 - N_{-n})$, with each element equal to D_{-n}.

Then, for $-n < t \leq 0$ we define

$$\tilde{Y}_t^{(n)} = (D_t, N_t, W_t, Z_t) \ ,$$

where D_t and N_t are just the components of (D, N) at time t, and the processes W and Z are defined together (in a Markovian way) according to the following set of transition rules:

1) if $W_{t-1} \neq 2$ and $N_{t-1} > 1$:

$$(W_t, Z_t) = \left(W_{t-1}, \overrightarrow{D_t}\right) \ ,$$

where $\overrightarrow{D_t}$ is of length $(F(D_t) + 1 - N_t)$;

2) if $W_{t-1} = 0$ and $N_{t-1} = 1$:

- if $D_{t-1} > h^*$:

$$(W_t, Z_t) = (0, (D_t)) \ ;$$

- else, if $D_{t-1} \leq h^*$:
  - with probability $\varepsilon > 0$,

$$(W_t, Z_t) = (1, (D_t)) \ ,$$

where $\varepsilon > 0$ is the constant in the minorisation condition (5.32) satisfied by X on the small set $C^* = \{x \ : \ V(x) \leq h^*\}$;

  - with probability $1 - \varepsilon$,

$$(W_t, Z_t) = (0, (D_t)) \ ;$$

3) if $W_{t-1} = 1$, $N_{t-1} = 1$:

$$(W_t, Z_t) = (2, (V_t)) \ ,$$

where V_t is the unique state produced by the coalescent event for a set of chains $V(X)$, which necessarily occurs at time t (see below);

4) else, if $W_{t-1} = 2$:

$$(W_t, Z_t) = \left(2, \left(V_{t-(F(D_t)-N_t)}, \ldots, V_t\right)\right) \ .$$

Although these updates look complicated, they can be described simply as follows:

- $W = 0$ until the first time after $-n$, say σ_{-k}, that D drops down to level h^* and a regeneration is approved (this occurs with probability ε): W then equals 1 until time $\sigma_{-(k-1)}$ when D next moves, at which point W is set equal to 2. W_t is then equal to 2 for all $t \geq \sigma_{-(k-1)}$;

- Z_t is equal to a vector of length $F(D_t) + 1 - N_t$, with each element equal to D_t, until time $\sigma_{-(k-1)}$: at this time all chains $V(X)$ currently dominated by $D_{\sigma_{-(k-1)}}$ regenerate (as identified by $W_{\sigma_{-(k-1)}} = 2$) using the measure on the right-hand-side of inequality (5.32), and then $Z_{\sigma_{-(k-1)}} = V_{\sigma_{-(k-1)}}$ (the unique state produced by the coalescent event). At this point the chain $\tilde{Y}^{(n)}$ enters $\tilde{\mathcal{X}}$. Z_t is then equal to a vector of length $F(D_t) + 1 - N_t$, recording the trajectory of the coalesced chain $V(X)$ over the time interval $[t - (F(D_t) - N_t), t]$;

From equation (5.34) it is evident that the chains $\tilde{Y}^{(n)}$ are identically distributed up to a shift in time. Furthermore, $\tilde{Y}^{(n)}$ hits $\tilde{\mathcal{X}}$ when all the target chains lying below D regenerate: this happens in finite time almost surely. Once $\tilde{Y}^{(n)}$ hits $\tilde{\mathcal{X}}$ it stays within this space until time 0, with its final component following the trajectory of $V(X)$ between the jump times of D. Finally, we see that the funnelling requirement of Theorem 4.6 (condition 3) is satisfied for the chains $\tilde{Y}^{(n)}$, as the following statements hold for $-m \leq -n \leq t \leq 0$:

- by construction, $D_t^{(m)} = D_t^{(n)}$ and $N_t^{(m)} = N_t^{(n)}$ for all $-n \leq t$;

- if $Y_t^{(m)} \preceq Y_t^{(n)}$ for some $t \geq -n$ then this equality persists up until time 0;

- if there are no coalescent events in the interval $[-m, -n)$ then $Y_{-n}^{(m)} = Y_{-n}^{(n)}$;

- if there is a coalescent event in the interval $[-m, -n)$ then $W_{-n}^{(m)} > 0 = W_{-n}^{(n)}$, and so $Y_{-n}^{(m)} \preceq Y_{-n}^{(n)}$.

Accordingly, $Y_t^{(m)} \preceq Y_t^{(n)}$ for all $-n \leq t \leq 0$.

Thus, by Theorem 4.6, setting

$$T = \inf \left\{ n \geq 1 \ : \ Y_0^{(n)} \in \mathcal{X} \right\},$$

we see that $T < \infty$ almost surely, and that

$$\tilde{Y}_0^{(T)} = \left(D_0, N_0, 2, \left(V_{-(F(D_0)-N_0)}, \ldots, V_0\right)\right)$$

is a draw from the joint equilibrium distribution of the dominating process D, the time N since the last jump of D, and the trajectory of $V(X)$ from σ_{-1} (the time of the last jump of D before 0) until time 0. This is much more than we were actually trying to achieve! A draw from the equilibrium distribution of X (our original aim) can of course be recovered from $\tilde{Y}_0^{(T)}$ by margining out the distribution of V_0.

5.2.5 When is a chain tame?

As a consequence of Theorem 5.16, question 2 of page 129 can be rephrased as: when is a chain tame? Note that a tame chain will not necessarily be tamable with respect to all scale functions.

In this section we present an equivalent definition of tameness, and prove some sufficient conditions for a chain to be tame. The following theorem shows that tameness is determined precisely by the behaviour of the chain until the time that it first hits the small set C.

Theorem 5.23. *Suppose X satisfies the weak drift condition $PV \leq V + b\mathbf{1}_C$. Then, for $n(x) = o(V(x))$, the following two conditions are equivalent:*

(i) there exists $\beta \in (0,1)$ such that $\mathbb{E}_x\left[V(X_{n(x)})\right] \leq \beta V(x)$, for $V(x)$ sufficiently large;

(ii) there exists $\beta' \in (0,1)$ such that $\mathbb{E}_x\left[V(X_{n(x)\wedge\tau_C})\right] \leq \beta' V(x)$, for $V(x)$ sufficiently large.

Furthermore, if $V(x)$ is large enough, we may take $|\beta - \beta'| < \varepsilon$ for any $\varepsilon > 0$.

Proof. Since $C = \{x \,:\, V(x) \leq d\}$ is a sub-level set, it is possible to split the expectation of $V(X_{n(x)\wedge\tau_C})$ according to whether $\tau_C \leq n(x)$ or not, to show

$$\begin{aligned}\mathbb{E}_x\left[V(X_{n(x)\wedge\tau_C})\right] &\leq \sup_{y\in C} V(y) + \mathbb{E}_x\left[V(X_{n(x)})\,;\, \tau_C > n(x)\right] \\ &\leq \sup_{y\in C} V(y) + \mathbb{E}_x\left[V(X_{n(x)})\right],\end{aligned}$$

and so (i) implies (ii).

Now consider the reverse implication. Using the weak drift condition for X, and recalling that $n(x)$ is deterministic:

$$\begin{aligned}
\mathbb{E}_x\left[V(X_{n(x)})\,;\,\tau_C \le n(x)\right] &= \sum_{k=1}^{n(x)} \mathbb{E}_x\left[\mathbb{E}_{X_k}\left[V(X_{n(x)-k})\right]\,;\,\tau_C = k\right] \\
&\le \sum_{k=1}^{n(x)} \sup_{y\in C} \mathbb{E}_y\left[V(X_{n(x)-k}) \,|\, X_k = y\right] \mathbb{P}_x\left(\tau_C = k\right) \\
&\le \sum_{k=1}^{n(x)} \sup_{y\in C} \left(V(y) + b(n(x)-k)\right) \mathbb{P}_x\left(\tau_C = k\right) \\
&\le d + n(x)b.
\end{aligned}$$

Assuming (ii), it follows that

$$\begin{aligned}
\mathbb{E}_x\left[V(X_{n(x)})\right] &\le \mathbb{E}_x\left[V(X_{n(x)\wedge\tau_C})\right] + \mathbb{E}_x\left[V(X_{n(x)})\,;\,\tau_C \le n(x)\right] \\
&\le \beta' V(x) + d + n(x)b \\
&\le \beta V(x),
\end{aligned}$$

for all large enough $V(x)$, since $n(x) = o(V(x))$.

Finally, due to the restriction upon the size of $n(x)$, it is clear that β and β' may be made arbitrarily close simply by restricting attention to x for which $V(x)$ is sufficiently large. □

Example 5.24 (Epoch chain). Consider the Markov chain X on $\{0, 1, 2, \ldots\}$ with the following transition kernel: for all $x \in \{0, 1, 2, \ldots\}$,

$$\begin{aligned}
&P(x, x) = \theta_x; \qquad\qquad P(0, x) = \zeta_x; \\
&P(x, 0) = 1 - \theta_x.
\end{aligned}$$

Thus X spends a random length of time (an epoch) at level x before jumping to 0 and regenerating. Meyn and Tweedie (1993) (page 362) show that this chain is ergodic if $\zeta_x > 0$ for all x, and

$$\sum_x \zeta_x (1 - \theta_x)^{-1} < \infty. \tag{5.35}$$

Furthermore, they show that the chain is not geometrically ergodic if $\theta_x \to 1$ as $x \to \infty$, no matter how fast $\zeta_x \to 0$.

Suppose that $\theta_x = 1 - \kappa(x + 1)^{-\gamma}$, for some suitable $\kappa, \gamma > 0$. We now slightly strengthen condition (5.35) on $\{\zeta_x\}$ to obtain a polynomial drift condition: we require

that there exists $\varepsilon > 0$ such that $\sum_x \zeta_x x^{(1+\varepsilon)\gamma} < \infty$. Under this assumption, with $C = \left[0, \kappa^{1/\gamma}\right]$, the following drift condition holds:

$$\mathbb{E}_x \left[V(X_1)\right] \leq V(x) - \kappa V^{\alpha}(x) + b\mathbf{1}_C(x), \tag{5.36}$$

where $V(x) = (x+1)^m$, $m = (1+\varepsilon)\gamma$, and $\alpha = \varepsilon/(1+\varepsilon)$.

We now show that X is tame, using Theorem 5.29. Define the taming function F by

$$F(z) = \left\lceil \lambda z^{1/(1+\varepsilon)} \right\rceil$$

for some λ satisfying

$$\lambda > \frac{1+\varepsilon}{\kappa} \log\left(\frac{1+\varepsilon}{\varepsilon}\right). \tag{5.37}$$

Write $F_x = F(V(x))$ for simplicity. Then,

$$\begin{aligned}
\mathbb{E}_x \left[V(X_{F_x \wedge \tau_0})\right] &\leq V(x)\theta_x^{F_x} + 1 \\
&\sim V(x)\left(1 - \frac{\kappa}{(x+1)^\gamma}\right)^{F_x} \\
&\leq V(x)\left(1 - \frac{\kappa}{(x+1)^\gamma}\right)^{\lambda(x+1)^\gamma} \\
&\leq \beta V(x),
\end{aligned}$$

where $\beta = e^{-\kappa\lambda}$ satisfies

$$\log\beta < (1+\varepsilon)\log\left(\frac{\varepsilon}{1+\varepsilon}\right),$$

by inequality (5.37). Therefore X is tame, by Theorem 5.29.

Suppose that we now modify the behaviour of a tame chain X when it is in the small set C. The following simple corollary of Theorem 5.23 shows that, so long as the resulting chain still satisfies a weak drift condition, tameness is preserved under such modification.

Corollary 5.25. *Suppose X satisfies the weak drift condition $PV \leq V + b\mathbf{1}_C$ and that X is tamed by the function F, to produce a chain X' satisfying GE(V, β, b', C'). Let $\hat{X}$ be a new chain produced by modifying the behaviour of X when in C, such that $\hat{X}$ satisfies $PV \leq V + \hat{b}\mathbf{1}_C$. Then F also tames $\hat{X}$, and the resulting chain $\hat{X}'$ satisfies GE$(V, \hat{\beta}, \hat{b}', \hat{C}')$, for any $\hat{\beta}' \in (\beta, 1)$.*

Proof. Write $F_x = F(V(x))$. Since X is tame, Theorem 5.23 states that for $V(x)$ large enough,

$$\mathbb{E}_x \left[V(X_{F_x \wedge \tau_C}) \right] \leq \tilde{\beta} V(x),$$

for any $\tilde{\beta} \in (\beta, 1)$. Now, since

$$\hat{X} \mathbf{1}_{[\hat{\tau}_C \geq F_x]} \stackrel{\mathcal{D}}{=} X \mathbf{1}_{[\tau_C \geq F_x]}$$

by definition,

$$\mathbb{E}_x \left[V(\hat{X}_{F_x \wedge \hat{\tau}_C}) \right] \leq \tilde{\beta} V(x).$$

Furthermore, since $\hat{X}$ satisfies the drift condition $PV \leq V + \hat{b} \mathbf{1}_C$, a second application of Theorem 5.23 yields

$$\mathbb{E}_x \left[V(\hat{X}_{F_x}) \right] \leq \hat{\beta} V(x),$$

where $\hat{\beta} \in (\tilde{\beta}, 1)$ may be chosen arbitrarily close to $\tilde{\beta}$ (and hence to β). Thus the same function F also tames $\hat{X}$. □

It has already been remarked that all geometrically ergodic chains are tame: the rest of this section investigates conditions which imply that a subgeometrically ergodic chain is tame.

Theorem 5.26. *Any chain satisfying drift condition PE may be adaptively subsampled, using a taming function F of the form*

$$F(z) = \left\lceil \lambda z^{\delta} \right\rceil, \quad \delta > 0,$$

to produce a chain X' which is geometrically ergodic.

A similar result (with a different form for the function F) holds for chains satisfying condition SGE.

Proof. Recall from Lemma 5.11 that for fixed $n \in \mathbb{N}$,

$$\mathbb{E}_x \left[V_\rho(X_n) \right] \leq \frac{MV(x)}{n^{\rho - 1}} + c, \tag{5.38}$$

for some constants $c, M < \infty$, where $1 < \rho < (1 - \alpha)^{-1}$ and $V_\rho = V^{1 - \rho(1-\alpha)}$. By Lemma 5.9,

$$PV_\rho \leq V_\rho - V_\rho^{\alpha'} + b_1 \mathbf{1}_C,$$

for some $\alpha' < 1$ and $b_1 < \infty$. We shall seek a time change that produces a geometric Foster-Lyapunov condition on this scale, V_ρ.

We wish to control $\mathbb{E}_x [V_\rho(X_{F_x})]$ where $F_x = F(V_\rho(x))$. By inequality (5.38),

$$\mathbb{E}_x [V_\rho(X_{F_x})] \leq \frac{MV(x)}{F_x^{\rho-1}} + c. \tag{5.39}$$

Therefore if F is defined such that

$$F_x \geq (\lambda V(x))^\delta \tag{5.40}$$

where

$$\delta = \frac{\rho(1-\alpha)}{\rho - 1} \in \left(\frac{1-\alpha}{\alpha}, \infty \right), \tag{5.41}$$

we obtain

$$\mathbb{E}_x [V_\rho(X_{F_x})] \leq \frac{MV(x)}{(\lambda V(x))^{\delta(\rho-1)}} + c \leq \beta V_\rho(x) + c\,, \tag{5.42}$$

where β may be made as small as desired simply by increasing λ. Since $c < \infty$ and V_ρ is bounded on small sets, inequality (5.42) is equivalent to condition GE, as required. □

This theorem highlights the importance of part (b) in the definition of tameness. The above result shows that *any* subgeometrically ergodic chain can be subsampled using a taming function F to produce a geometrically ergodic chain. However, this is not enough for our needs in Section 5.2.2, where tight control over the form of F is necessary in order for the constructed dominating process to have a proper equilibrium distribution. For example, from equation (5.41) it is evident that polynomially ergodic chains with a small drift exponent α need a large amount of time between subsampling instants, and this is where part (b) of Definition 5.14 may fail.

In the following two results attention is restricted to polynomially ergodic chains, and the range of α for which the definition of tameness is completely satisfied is investigated. Note that tameness is monotonic in the drift exponent α, because chains satisfying PE(V, c, α, b, C) also satisfy PE(V, c, α', b, C) for all $\alpha' \leq \alpha$. Similarly, any chain satisfying SGE(V, ϕ, b, C) with drift $\phi(x) \geq cx^\alpha$ also satisfies condition PE(V, c, α, b, C). The first result follows directly from the proof of Theorem 5.26.

Theorem 5.27. *Let X be a chain satisfying the drift condition $PV \leq V - cV^\alpha + b\mathbf{1}_C$, with $\alpha > 3/4$. Then X is tame.*

Proof. Let V_ρ be as in the proof of the last theorem. There we saw that X can be tamed by the function F, where

$$F_x = F(V_\rho(x)) \propto V(x)^\delta, \qquad \text{and} \quad \delta = \frac{\rho(1-\alpha)}{\rho - 1}\,.$$

In other words, F must satisfy

$$F(z) \propto z^{\frac{\delta}{1-\rho(1-\alpha)}}\,.$$

In order to completely satisfy Definition 5.14 therefore, we require

$$\frac{\delta}{1-\rho(1-\alpha)} < 1\,,$$

so that $F(z)/z \to 0$ as $z \to \infty$. Writing $\rho = \varepsilon/(1-\alpha)$, for some $1-\alpha < \varepsilon < 1$, this reduces to the requirement that

$$\varepsilon^2 - \varepsilon + (1-\alpha) < 0\,. \tag{5.43}$$

This inequality does not hold for the full range of $\varepsilon \in (1-\alpha, 1)$, and so it is necessary for the quadratic to have a positive discriminant if inequality (5.43) is to hold for some choice of ε. Since the discriminant is equal to $1 - 4(1-\alpha)$, this means that inequality (5.43) can only be satisfied for $\alpha > 3/4$. This completes the proof.

□

The proof of Theorem 5.27 cannot be improved upon directly: no information is lost through the use of loose inequalities. The only place where an improvement could possibly be found is in inequality (5.39), which was the starting point for the proof of the last result. The following theorem uses a different approach and provides a slightly better bound than inequality (5.39): this in turn leads to an improved bound for α.

Theorem 5.28. *Let X be a chain satisfying the drift condition $PV \leq V - cV^\alpha + b\mathbf{1}_C$ with $\alpha > \alpha^*$, where $\alpha^* = 0.704...$ is the only real root of the cubic equation $a^3 - 4a^2 + 8a - 4 = 0$. Then X is tame.*

Proof. Recall Holder's inequality: if $p, q > 0$ satisfy $1/p + 1/q = 1$, then for non-negative random variables Y and Z,

$$\mathbb{E}\,[YZ] \leq \mathbb{E}\,[Y^p]^{\frac{1}{p}}\,\mathbb{E}\,[Z^q]^{\frac{1}{q}}\,.$$

Let $V_\rho = V^{1-\rho(1-\alpha)}$ as before, where $1 < \rho < (1-\alpha)^{-1}$. As in the last two proofs, we seek a function F which tames X with respect to V_ρ, and by Theorem 5.23 it suffices to control

$$\mathbb{E}_x \left[V_\rho(X_{F_x}) \mathbf{1}_{[F_x < \tau_C]} \right] ,$$

where $F_x = F(V_\rho(x))$ again.

We now apply Holder's inequality to this expression, with

$$p = \frac{\alpha}{1 - \rho(1-\alpha)} > 1 ,$$

and writing c for a generic constant (which may change between successive lines). This yields (see the following notes for a step-by-step argument):

$$\begin{aligned}
\mathbb{E}_x \left[V_\rho(X_{F_x}) \mathbf{1}_{[F_x < \tau_C]} \right] &\leq \mathbb{E}_x \left[V^\alpha(X_{F_x}) \right]^{\frac{1}{p}} \mathbb{P}_x \left(F_x < \tau_C \right)^{1-\frac{1}{p}} \\
&\leq (cV(x))^{\frac{1}{p}} \left(\frac{\mathbb{E}_x \left[\tau_C^{1/(1-\alpha)} \right]}{F_x^{1/(1-\alpha)}} \right)^{1-\frac{1}{p}} \qquad (5.44) \\
&\leq cV(x) \left(\frac{1}{F_x^{1/(1-\alpha)}} \right)^{1-\frac{1}{p}} \qquad (5.45) \\
&= \frac{cV(x)}{F_x^{(\rho-1)/\alpha}} . \qquad (5.46)
\end{aligned}$$

Inequality (5.44) follows from Markov's inequality and Lemma 5.11: with $\Psi_1(x) = 1$ and $\Psi_2(x) = x$, this Lemma yields:

$$\mathbb{E}_x \left[V^\alpha(X_{F_x}) \right] \leq cV(x) .$$

The next line follows from Theorem 5.5 with $\Psi_1(x) = x$ and $\Psi_2(x) = 1$: for any $1 \leq \gamma \leq 1/(1-\alpha)$,

$$\mathbb{E}_x \left[\sum_{k=0}^{\tau_C - 1} k^{\gamma - 1} \right] \leq MV(x)$$

for some $M < \infty$. Taking $\gamma = 1/(1-\alpha)$ results in

$$\mathbb{E}_x \left[\tau_C^{1/(1-\alpha)} \right] \leq MV(x) ,$$

which proves the validity of inequality (5.45). Finally, equation (5.46) simply follows by definition of p.

Note that inequality (5.46) is an improvement upon the bound in inequality (5.39), since $\alpha < 1$. Therefore a candidate taming function for X is given by a function F

satisfying

$$F_x = F(V_\rho(x)) \propto V^\delta(x), \qquad \text{where} \quad \delta = \frac{\rho\alpha(1-\alpha)}{\rho - 1}\,.$$

As in the proof of Theorem 5.27 however, F must also satisfy $F(z)/z \to 0$ as $z \to \infty$, and this is equivalent to showing that the following holds for some value of $\rho \in (1, (1-\alpha)^{-1})$:

$$(1-\alpha)\rho^2 - ((1-\alpha)^2 + 1)\rho + 1 < 0\,.$$

As before, we need this quadratic to have a positive discriminant, and this in turn requires α to solve the following inequality:

$$\alpha^3 - 4\alpha^2 + 8\alpha - 4 > 0\,.$$

This concludes the proof. □

Although this result only appears to be a slight improvement upon that of Theorem 5.27, it is possible that the proof could be improved upon to further lower the bound on α. This possibility follows from two observations:

1) the final bound obtained in inequality (5.46) is still not tight: by Theorem 5.5 we know that
$$\sum_{n=0}^{\infty} n^{\rho-1}\mathbb{E}_x\left[V_\rho(X_n)\mathbf{1}_{[n<\tau_C]}\right] < \infty\,.$$
However, the bound in (5.46) only tells us that
$$\begin{aligned}\sum_{n=0}^{\infty} n^{\rho-1}\mathbb{E}_x\left[V_\rho(X_n)\mathbf{1}_{[n<\tau_C]}\right] &\leq \sum_{n=0}^{\infty} n^{(\rho-1)(1-1/\alpha)}V(x)\\ &= \infty, \qquad \text{since } \rho \leq \frac{1}{1-\alpha}\,.\end{aligned}$$

2) unlike in the proof of Theorem 5.27, there are a number of places in the proof of Theorem 5.28 (inequalities (5.44) to (5.46)) where information is lost (and thus where improvements might possibly be made).

The sufficient condition of Theorem 5.28 is far from necessary, as shall be demonstrated by example in the next section. Showing tameness for chains with $\alpha < 0.704$ is therefore still an open question, and one which is the subject of current research.

We conclude this discussion with a simple, but slightly restrictive, sufficient condition for a subgeometrically ergodic chain to be tame.

Theorem 5.29. *Let X be a chain satisfying a drift condition $PV \le V - \phi \circ V + b\mathbf{1}_C$ for which $H_\phi \in \Lambda^*$ and for which $V(X)$ has bounded upward jumps whenever $X \notin C$. That is, $V(X_1) \le V(X_0) + K$ whenever $X_0 \notin C$, for some constant $K < \infty$. Then X is tame.*

Note that if X satisfies condition PE, that is if $\phi(x) \propto x^\alpha$ for some $\alpha \in (0,1)$, then

$$H_\phi(x) = \int_1^x \frac{du}{\phi(u)} \propto x^{1-\alpha} \in \Lambda(1-\alpha)\,.$$

Thus this theorem applies to all chains satisfying condition PE.

Proof. From Theorem 5.23 we see that it is sufficient to show that by choosing an appropriate taming function F it is possible to obtain the bound

$$\mathbb{E}_x\left[V(X_{F_x})\;;\; F_x < \tau_C\right] \le \beta V(x)\,, \tag{5.47}$$

for all sufficiently large $V(x)$, where $F_x = F(V(x))$.

Fix $\alpha \in (0,1)$ such that $H_\phi \in \Lambda(1-\alpha)$. Thus there exists $d_\alpha < \infty$ such that $H_\phi(z) \le cz^{1-\alpha}$ for all $z > d_\alpha$. Choose β sufficiently small to satisfy

$$\log\beta < (1-\alpha)^{-1}\log\alpha\,, \tag{5.48}$$

and then choose $0 < \lambda < \beta$. Define the constant d_λ by

$$d_\lambda = \max\left\{y \;:\; y < \frac{K}{\beta-\lambda}H_\phi(y)\right\} < \infty\,,$$

and let $C_\lambda = \{x \;:\; V(x) \le d_\lambda\}$. Note that, if $x \notin C_\lambda$,

$$(\beta-\lambda)\,V(x) \ge K H_\phi \circ V(x)\,. \tag{5.49}$$

Finally, set $d' = \max\{d, d_\alpha, d_\lambda\}$, and let $C' = \{x \;:\; V(x) \le d'\}$.

Now define the taming function F by

$$F(z) = \begin{cases} \lceil \lambda^{-1} H_\phi(z) \rceil & \text{for } z > d' \\ 1 & \text{for } z \le d', \end{cases} \tag{5.50}$$

and note that our choice of β and F satisfy part (b) of the definition of tame chains. Then, for $x \notin C'$, since the upward jumps of $V(X)$ before time τ_C are bounded above

by K:

$$\begin{aligned}
\mathbb{E}_x\left[V(X_{F_x})\,;\,F_x<\tau_C\right] &\leq (V(x)+KF_x)\mathbb{P}_x\left(\tau_C>F_x\right)\\
&\leq (V(x)+KF_x)\frac{\mathbb{E}_x\left[\tau_C\right]}{F_x}, \quad \text{by Markov's inequality,}\\
&\leq (V(x)+KF_x)\frac{H_\phi\circ V(x)}{F_x}, \quad \text{by Corollary 5.10,}\\
&\leq \lambda V(x)+KH_\phi\circ V(x) \quad \text{using equation (5.50),}\\
&\leq \beta V(x), \quad \text{by inequality (5.49).}
\end{aligned}$$

Finally, for $x\in C'$, we have

$$\begin{aligned}
\mathbb{E}_x\left[V(X_{F_x})\right]=\mathbb{E}_x\left[V(X_1)\right] &\leq V(x)+b\\
&\leq \beta V(x)+(1-\beta)d'+b\\
&= \beta V(x)+b',
\end{aligned}$$

where $b'=(1-\beta)d'+b<\infty$. Hence (5.47) is satisfied and X is tame. □

Neither of the sufficient conditions presented above are necessary for a subgeometrically ergodic chain to be tame: in the next section we include an example of a chain that satisfies condition PE with drift coefficient $\alpha=1/2$, and which does not have bounded jumps for $X\notin C$, and show explicitly that it is tame.

5.3 Examples

We have already met two examples of polynomially ergodic chains (the Forward recurrence time chain of Examples 5.6 and 5.15 and the Epoch chain of Example 5.24) that have been shown to be tame. We now present four more examples of polynomially ergodic chains, and show that they are tame. The first of these is tame by Theorem 5.29, and the next two by Theorem 5.27. The final example shows that the sufficient conditions of Theorems 5.28 and 5.29 are not necessary for X to be tame.

We conclude this section by presenting a subgeometrically (but not polynomially) ergodic chain: we show explicitly that this chain is wild.

Example 5.30 (Delayed death process). Consider the Markov chain X on $\{0,1,\ldots\}$

with the following transition kernel:

$$\begin{aligned} P(x,x) &= \theta_x, \quad x \geq 1 \\ P(x,x-1) &= 1-\theta_x, \quad x \geq 1 \\ P(0,x) &= \zeta_x > 0, \quad x \in \{0,1,2,\ldots\}, \end{aligned}$$

where $\theta_x > 0$ for all $x \geq 1$ and $\theta_x \to 1$ as $x \to \infty$. We also assume that the mean jump from zero, μ_0, is finite. This chain is clearly aperiodic and δ_0-irreducible. The expected return time to zero is given by

$$\mathbb{E}_0[\tau_0] = 1 + \sum_{j=1}^{\infty} \zeta_j \sum_{k=1}^{j} (1-\theta_k)^{-1}, \tag{5.51}$$

and so we assume that $\zeta_j \to 0$ fast enough for (5.51) to be finite, which makes X ergodic.

X is not geometrically ergodic, however. To see this, define random variables $V_k \sim Geom(1-\theta_k)$. We aim to show that $\mathbb{E}_0[r^{\tau_0}] = \infty$ for all $r > 1$, and then appeal to Theorem 4.13. First note that for fixed $r > 1$,

$$\mathbb{E}\left[r^{V_k}\right] = \begin{cases} \frac{r(1-\theta_k)}{1-r\theta_k} & \text{if } r < \theta_k^{-1} \\ \infty & \text{otherwise.} \end{cases}$$

Therefore:

$$\begin{aligned} \mathbb{E}_0[r^{\tau_0}] &= r \sum_x \zeta_x \mathbb{E}_x[r^{\tau_0}] \\ &\geq r \sum_x \zeta_x \mathbb{E}\left[r^{V_x}\right] \\ &= \infty, \end{aligned}$$

since $\theta_x \to 1$ as $x \to \infty$, and so $r \geq \theta_x^{-1}$ for large enough x.

Now suppose that $\theta_x = 1 - \kappa(x+1)^{-\lambda}$, for some $\kappa > 0, \lambda > 1$, and that $\{\zeta_x\}$ are defined so that

$$\sum_{j=1}^{\infty} \zeta_j j^{1+\lambda+\varepsilon} < \infty, \tag{5.52}$$

for some $\varepsilon > 0$. (Note that this is more than sufficient for $\mathbb{E}_0[r^{\tau_0}]$ to be finite.) A polynomial drift condition is easy to obtain here, by letting $V(x) = (x+1)^m$, for

some $m > 1$:

$$\begin{aligned}
\mathbb{E}_x\left[V(X_1)\right] &= \mathbb{E}_x\left[(X_1+1)^m\right] \\
&= (x+1)^m\theta_x + x^m(1-\theta_x) \\
&= (x+1)^m - \frac{\kappa}{(x+1)^\lambda}\left((x+1)^m - x^m\right) \\
&\leq (x+1)^m - \kappa(x+1)^{m-1-\lambda} \\
&= V(x) - \kappa V^\alpha(x),
\end{aligned}$$

where $\alpha = (m-1-\lambda)/m$. To satisfy the drift condition (5.6) completely, it is necessary to choose m such that $\alpha > 0$ and the drift when X hits the small set C is bounded (here we can take $C = [0, \kappa^{1/(1-\alpha)}]$). Due to the assumption in (5.52), both of these requirements are met when $m = 1 + \lambda + \varepsilon$. Thus X satisfies the drift condition

$$\mathbb{E}_x\left[V(X_1)\right] \leq V(x) - \kappa V^\alpha(x) + b\mathbf{1}_C(x), \tag{5.53}$$

where $V(x) = (x+1)^{1+\lambda+\varepsilon}$, $\alpha = \varepsilon/(1+\lambda+\varepsilon)$, and $b < \infty$. Since the upward jumps of $V(X)$ when X is large are clearly bounded for this chain, X is tame by Theorem 5.29.

Example 5.31 (Delayed simple random walk). Similarly, a delayed reflected simple random walk can be defined by the following transition probabilities:

$$\begin{array}{ll}
P(x,x+1) = \kappa(1-p)(x+1)^{-1} \quad (x \geq 1), & P(0,1) = \kappa(1-p), \\
P(x,x) = 1 - \kappa(x+1)^{-1} \quad (x \geq 1), & P(0,0) = 1 - \kappa(1-p), \\
P(x,x-1) = \kappa p(x+1)^{-1} \quad (x \geq 1), &
\end{array}$$

for suitable $\kappa > 0$, and with $p > 1/2$ to ensure that X is ergodic. X is not geometrically ergodic however, since the time to hit 0 from any point x dominates that taken by the delayed death process (with $\lambda = 1$) in Example 5.30.

To find a drift condition for X, again take $V(x) = (x+1)^m$, for some $m > 2$. (Note that $\mathbb{E}_0\left[V(X_1)\right] < \infty$ for any $m > 0$.) Now,

$$\begin{aligned}
\mathbb{E}_x\left[(X_1+1)^m\right] &= (x+1)^m - \frac{\kappa}{x+1}\left((x+1)^m - px^m - (1-p)(x+2)^m\right) \\
&\leq (x+1)^m - \kappa m(2p-1)(x+1)^{m-2} + Kx^{m-3}, \\
&\qquad \text{for some constant } K > 0, \\
&\leq (x+1)^m - \kappa(2p-1)(x+1)^{m-2}, \quad \text{for large } x.
\end{aligned}$$

Thus X satisfies condition PE(V, c, α, b, C) with $V(x) = (x+1)^m$ and $\alpha = (m-2)/m$ for any integer $m > 2$. Although the bound $|X_{n+1} - X_n| \leq 1$ holds, it is not possible to apply Theorem 5.29 since that requires a uniform bound on $|V(X_{n+1}) - V(X_n)|$. However, choosing $m > 10$ means that Theorem 5.27 may be applied to show that X is tame.

Example 5.32 (Random walk Metropolis-Hastings). For a more practical example, consider a random walk Metropolis Hastings algorithm on $\mathbb{R}^d$, with proposal density q and target density p. Fort and Moulines (2000) consider the case when q is symmetric and compactly supported, and $\log p(z) \sim -|z|^s$, $0 < s < 1$ as $|z| \to \infty$. (When $d = 1$, this class of target densities includes distributions with tails typically heavier than the Exponential, such as the Weibull distributions.) They show that, under these conditions, the Metropolis-Hastings algorithm converges at *any* polynomial rate. In particular, it is possible to choose a scale function V such that the chain satisfies condition PE with $\alpha > 3/4$. Therefore, by Theorem 5.27 this chain is tame.

Example 5.33 (Random walk on a half-line). For our final example of a tame chain, consider Example 5.1 of Tuominen and Tweedie (1994). This is the random walk on $[0, \infty)$ given by

$$X_{n+1} = (X_n + Z_{n+1})^+, \tag{5.54}$$

where $\{Z_n\}$ is a sequence of i.i.d. real-valued random variables. We suppose that $\mathbb{E}[Z] = -\mu < 0$ (so $\{0\}$ is a positive-recurrent atom) and that $\mathbb{E}[(Z^+)^m] = \mu_m < \infty$ for some integer $m \geq 2$.

We also assume that $\mathbb{E}\left[r^{Z^+}\right] = \infty$ for all $r > 1$, and claim that this forces X to be subgeometrically ergodic. To see this, consider the chain $\hat{X}$ which uses the same downward jumps as X but stays still when X increases. That is,

$$\hat{X}_{n+1} = \left(\hat{X}_n - Z_{n+1}^-\right)^+.$$

Let τ_0 be the first time that X hits 0, and $\hat{\tau}_0$ be the corresponding hitting time for $\hat{X}$. Note that, for all $n > 0$,

$$\mathbb{E}_x\left[\hat{X}_{n \wedge \hat{\tau}_0}\right] \geq x - \mathbb{E}_x\left[n \wedge \hat{\tau}_0\right] \hat{\mu}, \tag{5.55}$$

where $\hat{\mu} = -\mathbb{E}[Z \,;\, Z \leq 0] > 0$. Now, the left hand side of (5.55) is dominated by x,

and $\mathbb{E}_x[\hat{\tau}_0] < \infty$, so letting $n \to \infty$ yields

$$\mathbb{E}_x[\tau_0] \geq \mathbb{E}_x[\hat{\tau}_0] \geq x/\hat{\mu}. \tag{5.56}$$

Thus, for $r > 1$:

$$\begin{aligned}
\mathbb{E}_0[r^{\tau_0}] &= r\mathbb{E}_0[\mathbb{E}_{X_1}[r^{\tau_0}]] \\
&\geq r\mathbb{E}_0\left[r^{\mathbb{E}_{X_1}[\tau_0]}\right] \\
&\geq r\mathbb{E}_0\left[r^{X_1/\hat{\mu}}\right] = \infty, \quad \text{by assumption.}
\end{aligned}$$

Therefore, by Theorem 4.13, X is not geometrically ergodic.

Now, Jarner and Roberts (2002) show that if $m \geq 2$ is an integer, then X satisfies condition PE with $V(x) = (x+1)^m$ and $\alpha = (m-1)/m$. Clearly the upward jumps of $V(X)$ when $X \notin C$ are not necessarily bounded, and so Theorem 5.29 cannot be applied. Furthermore, if $m \leq 3$ then $\alpha \leq 2/3$ and so Theorem 5.28 cannot be applied. However, we now show that X is still tame when $m = 2$ (and thus tame for all $m \geq 2$).

(i) First assume that the law of Z is concentrated on $[-z_0, \infty)$ for some $z_0 > 0$, and so $\mathbb{E}\left[Z^2\right] < \infty$. Then, if $x \geq z_0$:

$$\begin{aligned}
\mathbb{E}_x\left[(X_1+1)^2\right] &= \mathbb{E}\left[(x+1+Z)^2\right] \\
&= (x+1)^2 + 2(x+1)\mathbb{E}[Z] + \mathbb{E}\left[Z^2\right] \\
&\leq (x+1)^2 - 2\mu(x+1) + (\mu_2 + z_0^2).
\end{aligned}$$

Thus, for any $0 < \beta < 1$ there exists $z_\beta > z_0$ and $b_\beta < \infty$ such that, with $V(x) = (x+1)^2$ and $\alpha = 1/2$,

$$\mathbb{E}_x[V(X_1)] \leq V(x) - (2-\beta)\mu V^\alpha(x) + b_\beta \mathbf{1}_{[x \leq z_\beta]}. \tag{5.57}$$

Assume that $\beta < 1/4$ and a corresponding $z_\beta > z_0$ are fixed. Write $C_\beta = [0, z_\beta]$, and for $V(x) > z_\beta$ define $F(V(x)) = \left\lceil V^{1/2}(x)/\mu \right\rceil$. Iterating the drift condition

(5.57) we obtain for $x \notin C$, with $F_x = F(V(x))$:

$$\begin{aligned}
\mathbb{E}_x\left[V(X_{F_x})\right] &\leq V(x) - (2-\beta)\mu \sum_{k=0}^{F_x-1} \mathbb{E}_x\left[V^{1/2}(X_k)\right] + b_\beta F_x \\
&\leq (x+1)^2 - (2-\beta)\mu \sum_{k=0}^{F_x-1} (x+1-k\mu) + b_\beta F_x \qquad (5.58) \\
&\quad \text{since } \mathbb{E}_x\left[V^{1/2}(X_k)\right] = \mathbb{E}_x\left[(X_k+1)\right] \geq x+1-k\mu, \\
&\leq \left(1-(2-\beta)+\frac{(2-\beta)}{2}\right)(x+1)^2 + \gamma x \\
&\quad \text{for some } \gamma > 0, \\
&\leq \frac{\beta}{2} V(x) + \gamma V^{\frac{1}{2}}(x).
\end{aligned}$$

Thus there exists a sub-level set C' and a constant $b' < \infty$ such that if

$$F(x) = \begin{cases} \lceil x^{1/2}/\mu \rceil & x \notin C' \\ 1 & x \in C', \end{cases}$$

we obtain

$$\mathbb{E}_x\left[V(X_{F_x})\right] \leq \beta V(x) + b' \mathbf{1}_{C'}(x).$$

with $\beta < 1/4$. Since $\alpha = 1/2$, the requirement that $\log \beta < (1-\alpha)^{-1} \log \alpha$ is satisfied, and so this chain is indeed tame.

(ii) In the general case, we can proceed by truncating the law of Z at a level $-z_0$ so that the truncated distribution has a negative mean. The resulting chain, X^* say, is tame by the above argument. However, X^* stochastically dominates X on the whole of $[0, \infty)$, and so X must also be tame.

Remark 5.34. A polynomial drift condition for this chain can still be shown to hold when $m \in (1,2)$ (corresponding to drift $\alpha \in (0, 1/2)$). Although it is quite simple to produce an adaptive subsampling scheme in this situation that produces a chain satisfying condition GE(V, β, b, C), we have not yet been able to do this in a way such that β is small enough to satisfy part (b) of Definition 5.14. Therefore it is unclear at present whether such chains are in fact tame.

Now recall the Epoch chain of Example 5.24. This is the chain X on $\{0, 1, 2, \ldots\}$ with the following transition kernel: for all $x \in \{0, 1, 2, \ldots\}$,

$$\begin{aligned}
&P(x, x) = \theta_x\,; \qquad && P(0, x) = \zeta_x\,; \\
&P(x, 0) = 1 - \theta_x\,. &&
\end{aligned}$$

It was shown above that if $\theta_x = 1 - \kappa(x+1)^{-\gamma}$ then X satisfies a polynomial drift condition and is tame. We now consider a variant of this chain which satisfies the subgeometric drift condition SGE, but which is not polynomially ergodic: this chain turns out to be wild.

Example 5.35 (Epoch chain II). Consider the Epoch chain with

$$\theta_x = 1 - \frac{\log(x+1)}{x},$$

and with $\{\zeta_x\}$ satisfying

$$\sum_x \zeta_x x < \infty, \tag{5.59}$$

$$\text{but} \quad \sum_x \zeta_x \left(\frac{x}{\log(x+1)}\right)^{1/(1-\alpha)} = \infty \quad \text{for all } \alpha > 0. \tag{5.60}$$

(This is satisfied, for example, if $\zeta_x \propto (x\log(x+1))^{-2}$.) X is not geometrically ergodic since $\theta_x \to 1$ as $x \to \infty$ (Meyn and Tweedie 1993, page 362). Furthermore, we claim that X does not satisfy condition PE. For suppose there exists a scale function W such that X satisfies the drift condition

$$PW \leq W - cW^\alpha + b\mathbf{1}_C, \tag{5.61}$$

for some $\alpha \in (0,1)$. By design, for all $x > 0$, X satisfies

$$\begin{aligned} PW(x) &= W(x)\theta_x + W(0)(1-\theta_x) \\ &= W(x) - \frac{\log(x+1)}{x}\left(W(x) - W(0)\right). \end{aligned} \tag{5.62}$$

Combining equations (5.61) and (5.62) it is evident that, to satisfy condition PE for the scale function W, we must have

$$W(x) \geq \left(\frac{cx}{\log(x+1)}\right)^{1/(1-\alpha)}. \tag{5.63}$$

However, to satisfy inequality (5.61) completely, it is necessary that the drift from the small set C is bounded. In particular, we require $\mathbb{E}_0\left[W(X_1)\right] < \infty$. However, from the bound in inequality (5.63) it follows that

$$\mathbb{E}_0\left[W(X_1)\right] \geq \sum_x \zeta_x \left(\frac{cx}{\log(x+1)}\right)^{1/(1-\alpha)} = \infty$$

for all α, by the condition imposed upon $\{\zeta_x\}$ in (5.60). Thus X does not satisfy condition PE, as claimed.

X clearly does satisfy drift condition SGE however: with $V(x) = x + 1$,

$$\mathbb{E}_x [V(X_1)] = V(x) - \log V(x) + b\mathbf{1}_C ,$$

where the bound on the drift from $\{0\}$ is now guaranteed by equation (5.59). However,

$$H_\phi(x) = \int_1^x \frac{du}{\log u} \geq \frac{x}{\log x} \notin \Lambda^* ,$$

and so Theorem 5.29 cannot be applied.

We now show that, unlike the original epoch chain, X is wild. To see this, consider a drift function W and a taming function F:

$$\begin{aligned} \mathbb{E}_x \left[W(X_{F(W(x))}) \right] &\geq W(x) \mathbb{P}_x \left(\tau_0 > F(W(x)) \right) \\ &= W(x) \left(1 - \frac{\log(x+1)}{x} \right)^{F(W(x))} \\ &\sim W(x) \quad \text{if } F(W(x)) < \frac{x}{\log(x+1)} . \end{aligned}$$

However, if $F(W(x)) \geq x/\log(x+1)$ then by equation (5.22) the equilibrium distribution of the resulting dominating process D, π_D, satisfies (up to a normalisation constant):

$$\pi_D(x) \geq F(W(x)) x^{-(2-\eta)} \geq \frac{x^{-(1-\eta)}}{\log(x+1)} ,$$

for sufficiently large x, for some value of $\eta \in (0, 1)$. But this means that π_D is an improper density. Therefore there does not exist a taming function for X, and so this chain is wild.

5.4 Conclusions and questions

We have introduced the concept of a *tame* Markov chain, and shown that a perfect simulation algorithm exists for all such chains. We have also shown how this algorithm may be viewed in an extended state-space CFTP setting, as described in Cai and Kendall (2002). Our algorithm is not expected to be practical in general, but it directly extends the results of Foss and Tweedie (1998) and Kendall (2004): in a practical setting of course, one would use a dominating process that is better suited

to the chain of interest. The main assumption of the above work amounts to supposing that we can translate into practice the theoretical possibility of implementing various stochastic dominations as couplings. Although this is quite impractical in its most general setting, it should be noted that practical and implemented CFTP algorithms can correspond very closely to those described in this thesis. For example, we have already seen that the CFTP algorithm arising from the result of Foss and Tweedie (1998) is essentially the simplest case of the exact sampling algorithm proposed by Murdoch and Green (1998) (reviewed in Section 4.2.1); the scheme proposed in Kendall (2004) is closely related to fast domCFTP algorithms for perpetuities with sample step $k = 1$.

Aside from the development of a perfect simulation algorithm, we have proved two sufficient conditions for a subgeometrically ergodic chain to be tame, and provided an example which demonstrates that neither of these sufficient conditions are necessary. Our suspicion, which is shared by those experts with whom we have discussed this, is that the following conjecture is true:

Conjecture 5.36. *There exists a chain satisfying condition PE which is wild.*

On the other hand, we do not rule out the possibility that all polynomially ergodic chains are tame. A resolution of this conjecture would do much to further our understanding of such chains. The tame/wild classification provides some structure to the class of subgeometrically ergodic Markov chains that goes beyond the rate at which they converge to equilibrium. Although purely theoretical at present, this may prove to be important in understanding tricky MCMC implementations: for a tame chain, the existence of a time-change which produces a geometrically ergodic chain could possibly be exploited to improve the behaviour of an MCMC algorithm.

We have also given an example of a wild chain satisfying the drift condition SGE (Example 5.35), but this chain does not satisfy any polynomial drift condition. The existence of a perfect simulation algorithm for this and similar chains is another open question.

"I predict that within 100 years, computers will be twice as powerful, 10,000 times larger, and so expensive that only the five richest kings of Europe will own them."

Professor Frink, in *The Simpsons: Much Apu About Nothing*

6. CONCLUSION

The work in this thesis has hopefully shown that coupling is a beautiful and powerful technique, with many applications in both applied and theoretical probability. A number of these applications have been investigated over the course of the past five chapters, resulting in some interesting new theory and an abundance of stimulating open questions.

A major topic of interest arising from this investigation is that of the difference between co-adapted and non-co-adapted couplings. Co-adapted couplings are usually far more intuitive that their non-co-adapted counterparts, and as such are more commonly used in practice. (This is certainly true for perfect simulation algorithms such as CFTP.) The results of Chapters 2 and 3 show that for some processes there exists a co-adapted maximal coupling, whereas for others this is not the case. Examples of this first kind include the random-to-top shuffle of Section 2.2, and Brownian motions with fixed starting states (Section 3.4). Processes for which optimal co-adapted coupling is not maximal include the simple symmetric random walk on $\mathbb{Z}_2^n$ (Section 3.3.2), and Brownian motions with suitably randomised starting states (Section 3.4). To the best of our knowledge, little research has been carried out to date into the differences between optimal co-adapted and maximal couplings.

The 'price to be paid' for using co-adapted couplings of random walks on groups has already been discussed at the end of Chapter 2. There a possible three-class categorisation system for random walks was proposed, based upon the size of this cost. At present there is no obvious system by which chains can easily be classified in this way. For walks generated by the uniform probability measure on the set $H \subseteq G$, research into conditions relating the structure of H to the cost of optimal co-adapted couplings would be extremely interesting.

In the case of Brownian motion and the O-U process however, the difference between co-adapted and maximal couplings depends completely upon the starting

states of the two processes. For example, randomising the starting state of one process while fixing that of the other breaks the maximality of the reflection coupling. Generalisation of this observation to other diffusions on $\mathbb{R}^d$, and investigation into the structure of maximal couplings for these processes under initial randomisation, are also directions for future research.

Finally, Chapter 5 introduced a new class of Markov chains, named tame chains. Although this definition stemmed from the investigation into the existence of dom-CFTP algorithms, the tame/wild classification is nevertheless interesting in itself. Much is known about the asymptotic behaviour of subgeometrically ergodic chains, but surprisingly little research into their short-term properties exists. From a theoretical viewpoint, development of necessary conditions for a chain to be tame would prove valuable, as would resolution of Conjecture 5.36. For applied probabilists, it is also to be hoped that a better understanding of tame chains may lead to improvements in MCMC implementation for slowly-converging chains (as mentioned at the end of Chapter 5).

Who can say where all these paths will lead? Lines of research are of course co-adapted, so we will just have to wait and see...

APPENDIX: EQUILIBRIUM CALCULATIONS FOR A $D/M/1$ QUEUE

Recall from Section 4.4.3 that U is defined as the system workload of a $D/M/1$ queue, sampled just before arrivals, with arrivals every $\log(1/\beta)$ units of time, and service times being independent and of unit rate Exponential distribution. This satisfies the recurrence

$$U_{n+1} = \left(U_n + Exp(1) - \log \beta^{-1}\right) \vee 0.$$

The queue workload just before an arrival is equal to the sum of N independent $Exp(1)$ random variables, where N is the number of people in the queue (not including the arrival being considered). Now, from Grimmett and Stirzaker (2000) we see that the equilibrium distribution for N is Geometric, with parameter η, where η is the smallest positive root of

$$\eta = \beta^{1-\eta}. \tag{A-1}$$

Thus if $E_i \overset{\text{i.i.d}}{\sim} Exp(1)$, and $Z \sim Poisson(t)$:

$$\begin{aligned}
\mathbb{P}\left(U_n \leq t\right) &= \sum_{m=0}^{\infty} \mathbb{P}\left(\sum_{i=1}^{m} E_i \leq t\right) \eta^m (1-\eta) = \sum_{m=0}^{\infty} \eta^m (1-\eta) \mathbb{P}\left(Z \geq m\right) \\
&= \sum_{m=0}^{\infty} \eta^m (1-\eta) \sum_{k=m}^{\infty} \frac{e^{-t} t^k}{k!} \qquad = \sum_{k=0}^{\infty} (1-\eta) \frac{e^{-t} t^k}{k!} \sum_{m=0}^{k} \eta^m \\
&= \sum_{k=0}^{\infty} (1-\eta) \frac{e^{-t} t^k}{k!} \frac{(1-\eta^{k+1})}{(1-\eta)} \qquad = 1 - \eta e^{-t} \sum_{k=0}^{\infty} \frac{(t\eta)^k}{k!} \\
&= 1 - \eta e^{-t(1-\eta)}.
\end{aligned}$$

To check that this really is the equilibrium distribution of the workload, we now show that U_{n+1} has the same distribution. Let $E \sim Exp(1)$. Then:

$$\begin{aligned}
\mathbb{P}\left(U_{n+1} \leq t\right) &= \mathbb{P}\left(U_n + E - \log(1/\beta) \leq t\right) \\
&= \int_0^{t+\log(1/\beta)} e^{-s}\, \mathbb{P}\left(U_n \leq t + \log(1/\beta) - s\right)\, ds \\
&= \int_0^{t+\log(1/\beta)} e^{-s} \left(1 - \eta e^{-(1-\eta)(t+\log(1/\beta)-s)}\right)\, ds \\
&= 1 - e^{-(t+\log(1/\beta))} - \eta e^{-(1-\eta)(t+\log(1/\beta))} \int_0^{t+\log(1/\beta)} e^{-\eta s}\, ds \\
&= 1 - e^{-(1-\eta)(t+\log(1/\beta))} \\
&= 1 - \eta e^{-t(1-\eta)}\,,
\end{aligned}$$

by the definition of η in equation (A-1). Thus the equilibrium distribution of U is a mixture of an atom at zero with an $Exp(1-\eta)$ distribution, as claimed in Section 5.2.1.

REFERENCES

Aldous, D. (1983). *Random walks on finite groups and rapidly mixing Markov chains*, Volume 986 of *Lecture Notes in Math.*, pp. 243–297. Berlin: Springer.

Aldous, D. and P. Diaconis (1986). Shuffling cards and stopping times. *Amer. Math. Monthly 93*, 333–348.

Aldous, D. and P. Diaconis (1987). Strong uniform times and finite random walks. *Adv. in Appl. Math. 8*, 69–97.

Aldous, D. and J. Fill (2002). Reversible Markov chains and random walks on graphs. Unpublished Manuscript.

Barrera, J., B. Lachaud, and B. Ycart (2006). Cut-off for n-tuples of exponentially converging processes. *Stochastic Process. Appl. 116*(10), 1433–1446.

Bayer, D. and P. Diaconis (1992). Trailing the dovetail shuffle to its lair. *Ann. Appl. Probab. 2*(2), 294–313.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological) 48*(3), 259–302.

Bingham, N., C. Goldie, and J. Teugels (1987). *Regular variation*. Cambridge University Press.

Borovkov, A. A. and S. G. Foss (1993). Stochastically recursive sequences and their generalizations. In *Limit theorems for random processes and their applications (Russian)*, Volume 20 of *Trudy Inst. Mat.*, pp. 32–103, 303. Izdat. Ross. Akad. Nauk Sib. Otd. Inst. Mat., Novosibirsk.

Burdzy, K. and W. S. Kendall (2000). Efficient Markovian couplings: examples and counterexamples. *Ann. Appl. Probab. 10*(2), 362–409.

Burton, R. and Y. Kovchegov. Mixing times via super-fast coupling. Submitted.

Cai, Y. and W. S. Kendall (2002). Perfect simulation for correlated Poisson random variables conditioned to be positive. *Stat. Comput. 12*(3), 229–243.

Chaumont, L. and M. Yor (2003). *Exercises in probability*, Volume 13 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

Chen, G.-Y. (2006). *The cutoff phenomenon for finite Markov Chains*. Ph. D. thesis, Cornell University.

Connor, S. (2007). *Wiley Encyclopedia of Statistics in Quality and Reliability*, Chapter - Perfect Sampling. Wiley.

Connor, S. and W. Kendall (2006). Perfect simulation for a class of positive recurrent Markov chains. Research Report 446, University of Warwick.

Connor, S. B. and W. S. Kendall (2007). Perfect simulation for a class of positive recurrent Markov chains. *Ann. Appl. Probab. 17*(3), 781–808.

Corless, R. M., G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth (1996). On the Lambert W function. *Adv. Comput. Math. 5*(4), 329–359.

Diaconis, P. (1988). *Group Representations in Probability and Statistics*, Volume 11 of *Lecture Notes - Monograph Series*. Institute of Mathematical Statistics.

Diaconis, P. (1996). The cutoff phenomenon in finite Markov chains. *Proc. Nat. Acad. Sci. U.S.A. 93*(4), 1659–1664.

Diaconis, P., R. L. Graham, and J. A. Morrison (1990). Asymptotic analysis of a random walk on a hypercube with many dimensions. *Random Structures Algorithms 1*(1), 51–72.

Diaconis, P. and L. Saloff-Coste (2006). Separation cut-offs for birth and death chains. *Ann. Appl. Probab. 16*(4), 2098–2122.

Diaconis, P. and M. Shahshahani (1981). Generating a random permutation with random transpositions. *Z. Wahrsch. Verw. Gebiete 57*(2), 159–179.

Diaconis, P. and M. Shahshahani (1987). Time to reach stationarity in the Bernoulli-Laplace diffusion model. *SIAM J. Math. Anal. 18*(1), 208–218.

Dobrow, R. P. and J. A. Fill (2003). Speeding up the FMMR perfect sampling algorithm: a case study revisited. *Random Structures Algorithms 23*(4), 434–452.

Doeblin, W. (1938). Exposé de la theorie des chaînes constantes de Markov à un nombre fini d'états. *Rev. Math. Union Interbalkanique 2*, 77–105.

Douc, R., G. Fort, E. Moulines, and P. Soulier (2002). Computable bounds for subgeometric ergodicity. Technical Report 186, Equipe d'Analyse et Probabilités, Université d'Évry.

Douc, R., G. Fort, E. Moulines, and P. Soulier (2004). Practical drift conditions for subgeometric rates of convergence. *Ann. Appl. Probab. 14*(3), 1353–1377.

Feller, W. (1968). *An introduction to probability theory and its applications* (3rd ed.). Wiley.

Fill, J. A. (1999). An interruptible algorithm for perfect sampling via Markov chains. In *STOC '97 (El Paso, TX)*, pp. 688–695 (electronic). New York: ACM.

Fill, J. A., M. Machida, D. J. Murdoch, and J. S. Rosenthal (2000). Extension of Fill's perfect rejection sampling algorithm to general chains. In *Proceedings of the Ninth International Conference "Random Structures and Algorithms" (Poznan, 1999)*, Volume 17, pp. 290–316.

Fort, G. (2001). *Contrôle explicite d'ergodicité de chaîne de Markov : Applications à l'analyse de convergence de l'algorithme Monte-Carlo EM.* Ph. D. thesis, University Paris 6.

Fort, G. and E. Moulines (2000). V-subgeometric ergodicity for a Hastings-Metropolis algorithm. *Statist. Probab. Lett. 49*(4), 401–410.

Fort, G. and E. Moulines (2003). Polynomial ergodicity of Markov transition kernels. *Stochastic Process. Appl. 103*(1), 57–99.

Fortuin, C. M. and P. W. Kasteleyn (1972). On the random-cluster model. I. Introduction and relation to other models. *Physica 57*, 536–564.

Foss, S. G. and R. L. Tweedie (1998). Perfect simulation and backward coupling. *Stochastic Models 14*, 187–203.

Foster, F. G. (1953). On the stochastic matrices associated with certain queueing processes. *Ann. Math. Statistics 24*, 355–360.

Gibbs, A. L. and F. E. Su (2002). On choosing and bounding probability metrics. *International Statistical Review 70*, 419.

Gilbert, E. (1955). Theory of shuffling. Technical report, Bell Laboratories.

Goldstein, S. (1979). Maximal coupling. *Z. Wahrsch. Verw. Gebiete 46*(2), 193–204.

Green, P. J. and D. J. Murdoch (1998). Exact sampling for Bayesian inference: towards general purpose algorithms (with discussion). In J. Bernardo, J. Berger, A. Dawid, and A. Smith (Eds.), *Bayesian Statistics 6*, pp. 301–321. The Clarendon Press, Oxford University Press.

Greven, A. (1987). Couplings of Markov chains by randomized stopping times. I. Couplings, harmonic functions and the Poisson equation. *Probab. Theory Related Fields 75*(2), 195–212.

Griffeath, D. (1975). A maximal coupling for Markov chains. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 31*, 95–106.

Grimmett, G. R. and D. R. Stirzaker (2000). *Probability and Random Processes* (Second ed.). Oxford University Press.

Häggström, O. (2002). *Finite Markov chains and algorithmic applications*, Volume 52 of *London Mathematical Society Student Texts*. Cambridge University Press.

Häggström, O. and K. Nelander (1998). Exact sampling from anti-monotone systems. *Statist. Neerlandica 52*(3), 360–380.

Hardy, G. H., J. E. Littlewood, and G. Pólya (1952). *Inequalities*. Cambridge, at the University Press. 2d ed.

Hsu, E. and K.-T. Sturm. Maximal coupling of Euclidean Brownian motions. Preprint.

Huber, M. (2004). Perfect sampling using bounding chains. *Ann. Appl. Probab. 14*(2), 734–753.

Jarner, S. F. and G. O. Roberts (2002). Polynomial convergence rates of Markov chains. *Ann. Appl. Probab. 12*(1), 224–247.

Jarner, S. F. and R. L. Tweedie (2003). Necessary conditions for geometric and polynomial ergodicity of random-walk-type Markov chains. *Bernoulli 9*(4), 559–578.

Jonasson, J. (2006). Biased random-to-top shuffling. *Ann. Appl. Probab. 16*(2), 1034–1058.

Kendall, W. S. (1997). Perfect simulation for spatial point processes. In Proceedings of the 51st Session of the ISI, Istambul (August 1997), 1997; 3: 163–166.

Kendall, W. S. (1998). Perfect simulation for the area-interaction point process. In L. Accardi and C. Heyde (Eds.), *Probability Towards 2000*, pp. 218 – 234. Springer-Verlag.

Kendall, W. S. (2004). Geometric ergodicity and perfect simulation. *Electron. Comm. Probab. 9*, 140–151.

Kendall, W. S. (2005). *Markov chain Monte Carlo: Innovations and Applications*, Volume 7 of *Lecture Notes Series, Institute for Mathematical Sciences, National University of Singapore*, Chapter Notes on perfect simulation. World Scientific.

Kendall, W. S. and J. Møller (2000). Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes. *Adv. in Appl. Probab. 32*, 844–865.

Krylov, N. V. (1980). *Controlled diffusion processes*, Volume 14 of *Applications of Mathematics*. New York: Springer-Verlag. Translated from the Russian by A. B. Aries.

Kuwada, K. (2006). On uniqueness of maximal coupling for diffusion processes with a reflection. To appear in Journal of Theoretical Probability.

Lachaud, B. (2005). Cut-off and hitting times of a sample of Ornstein-Uhlenbeck processes and its average. *J. Appl. Probab. 42*(4), 1069–1080.

Lindvall, T. (1983). On coupling of diffusion processes. *J. Appl. Probab. 20*(1), 82–93.

Lindvall, T. (2002). *Lectures on the coupling method*. Dover.

Lindvall, T. and L. C. G. Rogers (1986). Coupling of multidimensional diffusions by reflection. *Ann. Probab. 14*(3), 860–872.

Matthews, P. (1987). Mixing rates for a random walk on the cube. *SIAM J. Algebraic Discrete Methods 8*(4), 746–752.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953). Equation of state calculation by fast computing machines. *Journal of Chemical Physics 21*(6), 1087 – 1092.

Meyn, S. and R. Tweedie (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag.

Murdoch, D. J. and P. J. Green (1998). Exact sampling from a continuous state space. *Scand. J. Statist. 25*(3), 483–502.

Nummelin, E. and P. Tuominen (1983). The rate of convergence in Orey's theorem for Harris recurrent Markov chains with applications to renewal theory. *Stochastic Process. Appl. 15*(3), 295–311.

Peres, Y. (2005). Mixing for Markov chains and spin systems. Draft Lecture notes for summer school at UBC.

Pitman, J. W. (1976). On coupling of Markov chains. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 35*(4), 315–322.

Pitman, J. W. and T. P. Speed (1973). A note on random times. *Stochastic Processes Appl. 1*, 369–374.

Propp, J. and D. Wilson (1998). Coupling from the past: a user's guide. In D. Aldous and J. Propp (Eds.), *Microsurveys in Discrete Probability*, Volume 41 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pp. 181–192. American Mathematical Society.

Propp, J. G. and D. B. Wilson (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures Algorithms 9*, 223–252.

Reeds, J. (1981). Theory of riffle shuffling.

Robert, C. P. and G. Casella (2004). *Monte Carlo statistical methods*. Springer Texts in Statistics. New York: Springer-Verlag.

Roberts, G. O. and J. S. Rosenthal (2001). Small and pseudo-small sets for Markov chains. *Stoch. Models 17*(2), 121–145.

Roberts, G. O. and R. L. Tweedie (1999). Bounds on regeneration times and convergence rates for Markov chains. *Stochastic Process. Appl. 80*, 211–229.

Rogers, L. C. G. (1999). Fastest coupling of random walks. *J. London Math. Soc. (2) 60*(2), 630–640.

Rosenthal, J. S. (1995a). Convergence rates for Markov chains. *SIAM Rev. 37*(3), 387–405.

Rosenthal, J. S. (1995b). On generalizing the cut-off phenomenon for random walks on groups. *Adv. in Appl. Math. 16*(3), 306–320.

Saloff-Coste, L. (1997). Lectures on finite Markov chains. In *Lectures on probability theory and statistics (Saint-Flour, 1996)*, Volume 1665 of *Lecture Notes in Math.*, pp. 301–413. Berlin: Springer.

Saloff-Coste, L. (2004). Random walks on finite groups. In *Probability on discrete structures*, Volume 110 of *Encyclopaedia Math. Sci.*, pp. 263–346. Berlin: Springer.

Sweeny, M. (1983, April). Monte Carlo study of weighted percolation clusters relevant to the Potts models. *Physical Review B 27*, 4445–4455.

Thönnes, E. (1999). Perfect simulation of some point processes for the impatient user. *Adv. in Appl. Probab. 31*(1), 69–87.

Thönnes, E. (2000). A primer on perfect simulation. In *Statistical physics and spatial statistics (Wuppertal, 1999)*, Volume 554 of *Lecture Notes in Phys.*, pp. 349–378. Berlin: Springer.

Thorisson, H. (1986). On maximal and distributional coupling. *Ann. Probab. 14*(3), 873–876.

Thorisson, H. (2000). *Coupling, stationarity, and regeneration*. Springer-Verlag.

Tuominen, P. and R. L. Tweedie (1994). Subgeometric rates of convergence of f-ergodic Markov chains. *Adv. in Appl. Probab. 26*, 775–798.

Villani, C. (2005). Optimal transport, old and new. Lecture notes, Saint-Flour summer school.

Wilson, D. B. (2000a). How to couple from the past using a read-once source of randomness. *Random Structures Algorithms 16*(1), 85–113.

Wilson, D. B. (2000b). *Layered multishift coupling for use in perfect sampling algorithms (with a primer on CFTP)*, Volume 26 of *Fields Inst. Commun.*, pp. 143–179. Providence, RI: Amer. Math. Soc.

Wilson, D. B. (2004). Mixing times of Lozenge tiling and card shuffling Markov chains. *Ann. Appl. Probab. 14*(1), 274–325.

Ycart, B. (1999). Cutoff for samples of Markov chains. *ESAIM Probab. Statist. 3*, 89–106 (electronic).