Mathematical foundations of the Markov chain Monte Carlo method

Mark Jerrum*

Department of Computer Science, University of Edinburgh, The King's Buildings, Edinburgh EH9 3JZ, United Kingdom.

Summary. The Markov chain Monte Carlo (MCMC) method exploits the idea that information about a set of combinatorial objects may be obtained by performing an appropriately defined random walk on those objects. In the area of statistical physics, MCMC algorithms have been in use for many years for the purpose of estimating various quantities of physical interest, often expectations of random variables on "configurations" of a statistical model. The running time of MCMC algorithms depends on the rate at which the random walk converges to equilibrium; only when a condition of near-equilibrium has been achieved can the algorithm discover what "typical" objects are like. In the past decade or so, it has become possible to derive a priori bounds on the rate of convergence to equilibrium of random walks underlying MCMC algorithms of practical interest. In cases where a priori bounds cannot be derived, it may still be possible to conduct rigorously grounded experiments. Many of the main ideas and techniques are set out here, with the recent developments being discussed at greater length.

* Supported in part by Esprit Working Group No. 21726, "RAND2."

Acknowledgements

Two colleagues merit special acknowledgement for their contribution to this survey. Some of the results presented here were products of an extended period of collaboration with Alistair Sinclair, and in describing them I have freely plundered and adapted material from our joint articles. The experiment described in Section 7.2 was jointly undertaken with Vivek Gore, and is published here for the first time.

I also thank an anonymous referee for carefully reading and providing helpful comments on a draft of this chapter.

1. Introduction

The classical *Monte Carlo method* is an approach to estimating quantities that are hard to compute exactly. The quantity z of interest is expressed as the expectation $z = \operatorname{Exp} Z$ of a random variable (r.v.) Z for which some efficient sampling procedure is available. By taking the mean of some sufficiently large set of independent samples of Z, one may obtain an approximation to z. For example, suppose

$$S = \left\{(x, y) \in [0,1]^2 : p_i(x, y) \leq 0, \text{ for all } i\right\}$$

is some region of the unit square defined by a system of polynomial inequalities $p_i(x, y) \leq 0$. Let Z be the r.v. defined by the following experiment or trial: choose a point (x, y) uniformly at random (u.a.r.) from $[0,1]^2$; let $Z = 1$ if $p_i(x, y) \leq 0$ for all i, and $Z = 0$ otherwise. Then the area a of S is equal to $\operatorname{Exp} Z$, and an estimate of a may be obtained from the sample mean of a sufficiently long sequence of trials. In this example, the use of the Monte Carlo method is perhaps avoidable, at the expense of a more complex algorithm; for more essential uses, see, for example, Knuth's proposal [48] for estimating the size of a tree by taking a random path from the root to a leaf, or Rasmussen's [54] for estimating the permanent of a 0,1-matrix.

The *Markov chain* Monte Carlo (MCMC) method is a development of the foregoing approach, which is sometimes applicable when Z cannot be sampled "directly." Computer scientists approaching this subject with only the most basic probabilistic tools can, for the moment, think of a Markov chain $\mathfrak{M}$ as being a kind of finite automaton, in which the transitions from any state are labelled, not by letters from some alphabet, but by non-negative real numbers (probabilities) summing to 1. The Markov chain $\mathfrak{M}$ starts in a distinguished state x_0 at time 0, and makes a sequence of transitions at successive time-steps, resulting in $\mathfrak{M}$ passing through a sequence of states $X_0 = x_0, X_1, X_2, \ldots$. The transitions are guided by the specified probabilities; if $X_t = x_t$, i.e., $\mathfrak{M}$ is in state x_t after the tth transition, then the probability that $X_{t+1} = x_{t+1}$ is just the number assigned to the transition from state x_t to state x_{t+1}.

Suppose Ω denotes the (finite) state space of $\mathfrak{M}$. The Markov chain $\mathfrak{M}$ will be completely specified if we give the matrix of transition probabilities $\big(P(x,y) : x, y \in \Omega\big)$, where, for all pairs of states $x, y \in \Omega$,

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

is the probability that the Markov chain is in state y at time $t+1$, conditioned on it being in state x at time t. Note the crucial "forgetting property" of Markov chains: the state at time $t+1$ depends probabilistically on the state at time t, but not on the state at any earlier time.

Provided a certain technical condition—let's call it *ergodicity*—is met, $\mathfrak{M}$ will converge to well-defined *stationary distribution* π. More precisely, there is a probability distribution π on Ω such that $\Pr(X_t = y \mid X_0 = x) \to \pi(y)$, as $t \to \infty$, for all pairs of states $x, y \in \Omega$. Note that the initial state x is "forgotten" by $\mathfrak{M}$ over a sufficiently large number of states.

So suppose we have a r.v. Z for which no obvious direct sampling procedure exists. The idea behind MCMC is to construct an ergodic Markov chain $\mathfrak{M}$ whose state space is the range of Z (or at least includes the range of Z) and whose stationary distribution matches the probability distribution of Z. Then the required samples are obtained by simulating $\mathfrak{M}$ for sufficiently many steps τ from some fixed initial state, and returning the final state. Of course, what we obtain is not a perfect sample from the probability distribution of Z, but if τ is large the error will be negligible. Naturally, the determination of a suitable τ is a significant concern in rigorous applications of MCMC.

As an example of the approach, we consider the problem of estimating the number of (vertex) q-colourings of a graph G. In Section 2. we consider how sample q-colourings of G, generated independently and u.a.r., can be used to obtain an estimate for the *number* of q-colourings of G. This step of the MCMC programme—how samples are used—is often (though not always) rather routine. We therefore leave graph colouring as our one representative example, and turn from the *use* of samples to their *generation*. In Section 3., we show how to design a simple Markov chain on colourings that, given a certain condition on the graph G and the number of colours q, is ergodic and has uniform stationary distribution. Again, this step—the design of the Markov chain—is often rather routine.

We then turn to what is the crux, often the sticking point, of the method, namely determining good upper bounds on the "mixing time" τ, i.e., the number of steps before the Markov chain is "close" to the stationary distribution. Section 4. presents three methods for bounding the mixing time in the context of a toy example, namely a Markov chain on q-colourings of the empty graph. Obviously, the toy example is of no practical value, but its very simplicity brings the various techniques into sharp relief. Section 5. applies the same three methods to some more realistic and challenging applications. Most of the material of Sections 2. to 5. can be followed in greater detail (though sometimes with different examples) in the survey article of Jerrum and Sinclair [36].

The remainder of the article deals in greater depth with a topic, namely the coupling method, which has grown in perceived importance since the survey article [36] was written. Coupling is a classical (and elementary) technique for bounding the convergence rate of a Markov chain, but some of us working in the analysis of MCMC algorithms had been guilty of thinking it too weak in practice to be applied to interesting examples. Two recent developments—"coupling from the past" and "path coupling"—are beginning to correct that perception.

2. Approximate counting, uniform sampling and their relationship

What do we mean precisely by (efficient) approximate counting and uniform sampling?

Suppose $N : \Sigma^* \to \mathbb{N}$ is a function mapping problem instances (encoded as words over some convenient alphabet Σ) to natural numbers. For example, N might map (encodings of) a graph G to the number $N(G)$ of perfect matchings in G. It should be clear that any combinatorial enumeration problem can be cast in this framework. A *randomised approximation scheme* for N is a randomised algorithm that takes as input a word (instance) $x \in \Sigma^n$ and and an error bound $\varepsilon > 0$, and produces as output a number Y (a random variable) such that[1]

$$\Pr\big((1-\varepsilon)N(x) \le Y \le (1+\varepsilon)N(x)\big) \ge \frac{3}{4}. \tag{2.1}$$

A randomised approximation scheme is said to be *fully polynomial* [43] if it runs in time polynomial in n (the input length) and ε^{-1}. We shall abbreviate the rather unwieldy phrase "fully polynomial randomised approximation scheme" to FPRAS.

Suppose now that $S \subset \Sigma^* \times \Sigma^*$ is a relation between (encodings of) problem instances and (encodings of) feasible solutions to that instance. Thus, S might assign to each graph G the set $S(G)$ of perfect matchings in G. We insist that the set $S(x)$ is finite for all x. (The relationship we envisage between S and the counting function N discussed earlier is, of course, that $N(x) = |S(x)|$ for all meaningful encodings $x \in \Sigma^*$ of problem instances.) For any two probability distributions π and π' on a countable set Ω, define the *total variation distance* between π and π' to be

[1] There is no significance in the constant $\frac{3}{4}$ appearing in the definition, beyond its lying strictly between $\frac{1}{2}$ and 1. Any success probability greater than $\frac{1}{2}$ may be boosted to $1-\delta$ for any desired $\delta > 0$ by performing a small number of trials and taking the median of the results; the number of trials required is $O(\ln \delta^{-1})$ [37].

$$D_{\mathrm{tv}}(\pi,\pi') := \max_{A\subseteq\Omega}|\pi(A)-\pi'(A)| = \frac{1}{2}\sum_{x\in\Omega}|\pi(x)-\pi'(x)|.$$

This definition extends to uncountable probability spaces with the maximum replaced by a supremum over measurable sets A, or the sum by an integral. An *almost uniform sampler* for S is a randomised algorithm that takes as input a word (instance) $x\in\Sigma^n$ and a tolerance $\delta>0$, and produces a feasible solution $Z\in S(x)$ (a random variable) such that the probability distribution of Z is within variation distance δ of the uniform distribution on $S(x)$. An almost uniform sampler is said to be *fully polynomial* if it runs in time polynomial in n (the input length) and $\log\delta^{-1}$.

There is a close connection between almost uniform sampling and approximate counting, which has been discussed at some length by Jerrum, Valiant, and Vazirani [37]. In brief, provided a certain technical condition known as *self-reducibility* is met, almost uniform sampling is possible in polynomial time if and only if approximate counting is. Here is a possible way to make the connection concrete in the case of graph colourings.

Proposition 2.1. *Suppose we have an almost uniform sampler for q-colourings of a graph, which works for graphs G with maximum degree bounded by $\Delta<q$; and suppose that the sampler has time complexity $T(n,\delta)$, where n is the number of vertices in G, and δ the allowed deviation from uniformity in the sampling distribution. Then we may construct a randomised approximation scheme for the number of q-colourings of a graph, which works for graphs G with maximum degree bounded by Δ, and which has time complexity*

$$O\left(\frac{m^2}{\varepsilon^2}\,T\Big(n,\frac{\varepsilon}{6m}\Big)\right),$$

where m is the number of edges in G, and ε the specified error bound.

At this point we merely indicate the key algorithmic technique underlying Proposition 2.1. A full proof, including a detailed statistical analysis, can be found in the Appendix.

Denote by $\Omega(G)$ the set of all q-colourings of G. Let $G=G_m>G_{m-1}>\cdots>G_1>G_0=(V,\emptyset)$ be any sequence of graphs in which each graph G_{i-1} is obtained from the previous graph G_i by removing a single edge. We may express the quantity we wish to estimate as a product of ratios:

$$|\Omega(G)| = \frac{|\Omega(G_m)|}{|\Omega(G_{m-1})|}\times\frac{|\Omega(G_{m-1})|}{|\Omega(G_{m-2})|}\times\cdots\times\frac{|\Omega(G_1)|}{|\Omega(G_0)|}\times|\Omega(G_0)|, \qquad (2.2)$$

where, it will be observed, $|\Omega(G_0)|=q^n$. Our strategy is to estimate the ratio

$$\varrho_i = \frac{|\Omega(G_i)|}{|\Omega(G_{i-1})|}$$

for each i in the range $1\le i\le m$, and by substituting these quantities into identity (2.2), obtain an estimate for the number of q-colourings of G:

$$|\Omega(G)| = q^n \varrho_1 \dots \varrho_m.$$

To estimate the ratio ϱ_i we use the almost uniform sampler to obtain a sufficiently large sample of q-colourings from $\Omega(G_{i-1})$ and compute the proportion of samples that lie in $\Omega(G_i)$. The analysis presented in the Appendix places a bound on the sample size required.

For background material on approximate counting, refer to Welsh's survey article [58].

3. Sampling by Markov chain simulation

Let G be an undirected graph on vertex set $V = [n] = \{0, 1, \dots, n-1\}$ whose maximum degree is bounded by $\Delta = \Delta(G)$, and let $Q = [q]$ be a set of q colours. Let $X_0 : V \to Q$ be a proper colouring of the vertices of G, i.e., one in which every edge has endpoints of different colours. Such a colouring always exists if $q \geq \Delta + 1$, as can be appreciated by considering a simple sequential colouring algorithm. Indeed Brooks' theorem asserts that a colouring exists when $q \geq \Delta$, provided $\Delta \geq 3$ and G does not contain $K_{\Delta+1}$ as a connected component [7, 9].

Consider the Markov chain (X_t) whose state space $\Omega = \Omega(G, q)$ is the set of all q-colourings of G, and whose transition probabilities from state (colouring) X_t are modelled by the following procedure:

(1) Select a vertex $v \in V$ uniformly at random (u.a.r.), and then a colour $c \in Q$ u.a.r. from the set of legal colours for v. (A colour is legal if it is different from the colour of any neighbour of v.)
(2) Recolour vertex v with colour c, and let the resulting colouring be X_{t+1}.

This procedure describes what would be termed, by the statistical physics community, the "heat-bath" dynamics of an antiferromagnetic q-state Potts model at zero temperature. For $t \in \mathbb{N}$, let $P^t : \Omega^2 \to [0, 1]$ denote the t-step transition probabilities[2] arising from this procedure, so that $P^t(x, y) = \Pr(X_t = y \mid X_0 = x)$ for all $x, y \in \Omega$.

Assume now that $q \geq \Delta + 2$. It is easily verified that the Markov chain (X_t)—which we refer to in the sequel as $\mathfrak{M}_{\mathrm{col}}(G, q)$ or simply $\mathfrak{M}_{\mathrm{col}}$—is (a) *irreducible*, i.e., for all $x, y \in \Omega$, there is a t such that $P^t(x, y) > 0$, and (b) *aperiodic*, i.e., $\gcd\{t : P^t(x, y) > 0\} = 1$ for all $x, y \in \Omega$. Irreducibility of $\mathfrak{M}_{\mathrm{col}}$ follows from the observation that any colouring x may be transformed to any other colouring y by sequentially assigning new colours to the vertices V in ascending sequence; before assigning a new colour c to vertex v it is necessary to recolour all neighbouring vertices $u > v$ that have colour c, but there is always at least one "free" colour to allow this to be done, provided $q \geq \Delta + 2$. Aperiodicity follows from the fact that the loop probabilities $P(x, x)$ are non-zero for all $x \in \Omega$.

[2] We drop the superscript t in the case $t = 1$.

A finite Markov chain that is irreducible and aperiodic is *ergodic*; i.e., there is a *stationary distribution* $\pi : \Omega \to [0,1]$ such that $\lim_{t\to\infty} P^t(x,y) = \pi(y)$ for all $x, y \in \Omega$. In the case of $\mathfrak{M}_{\mathrm{col}}$, this stationary distribution is actually the uniform distribution on Ω, as can be verified using the following simple but useful fact.

Lemma 3.1. *Let $\mathfrak{M}$ be an ergodic Markov chain with finite state space Ω and transition probabilities $P(\cdot\,,\cdot)$. If $\pi' : \Omega \to [0,1]$ is any function satisfying "detailed balance"*

$$\pi'(x)P(x,y) = \pi'(y)P(y,x), \quad \text{for all } x, y \in \Omega, \tag{3.1}$$

and the normalisation condition $\sum_{x\in\Omega} \pi'(x) = 1$, then π' is indeed the stationary distribution of $\mathfrak{M}$.

A Markov chain whose stationary distribution satisfies the detailed balance condition is said to be *time-reversible.*

In Section 5.3 we demonstrate that $\mathfrak{M}_{\mathrm{col}}$ is "rapidly mixing," i.e., the t-step distribution approaches close to the stationary distribution in time polynomial in n, provided $q \geq 2\Delta + 1$. To make this statement precise we need to explain what is meant by "close" here. In the light of our definition of almost uniform sampler, it seems natural to measure closeness to stationarity in terms of the variation distance. For $t \in \mathbb{N}$ define

$$\delta_x(t) := D_{\mathrm{tv}}(P^t(x,\,\cdot\,), \pi) := \max_{A\subseteq\Omega} |P^t(x,A) - \pi(A)|,$$

where x is the initial state and $P^t(x,A) = \sum_{y\in A} P^t(x,y)$. The rate of convergence to stationarity from initial state x may be measured by the *mixing time*, i.e., the function

$$\tau_x(\delta) = \min\{t : \delta_x(t') \leq \delta \text{ for all } t' \geq t\}.$$

When making statements about rate of convergence that are independent of the initial state, the appropriate version of mixing time is $\tau(\delta) = \max_x \tau_x(\delta)$, where the maximum is over all $x \in \Omega$. By *rapid mixing*, we mean that $\tau(\delta) \leq \mathrm{poly}(n, \log \delta^{-1})$.

The rapid mixing result of Section 5.3 provides us with a simple almost uniform sampler for q-colourings in G: simulate the Markov chain $\mathfrak{M}_{\mathrm{col}}$, starting at an arbitrary state, for a sufficiently large (but polynomial) number of steps, and return the current state as result. As a corollary we obtain, via Proposition 2.1, an FPRAS for the number of q-colourings of a graph in the case $q \geq 2\Delta + 1$.

As a warm up we consider first the rather trivial case of an empty graph (i.e., $\Delta = 0$).

4. A toy example: colourings of the empty graph

In this section we survey those techniques for proving rapid mixing that have shown themselves to have some degree of general applicability. The three techniques described here—which might be titled "canonical paths," "geometric" and "coupling"—cover the majority of applications. Nevertheless, some ingenious special techniques have been introduced to handle specific problems, most notably Feder and Mihail's inductive argument to demonstrate rapid mixing of the basis-exchange random walk on a "balanced" matroid [27].

The three techniques will be illustrated by applying each in turn to the graph-colourings Markov chain $\mathfrak{M}_{\mathrm{col}}(O_n, q)$ of Section 3., specialised to the empty graph $O_n := (V, \emptyset)$, where, as usual, $V = [n]$. Since the state space in this case is simply $\Omega = Q^n$, it would be a trivial matter to sample from Ω directly. On the other hand, the very triviality of the situation will allow us to concentrate on the methods without getting bogged down in calculation or technical detail. Section 5. will consider some more realistic applications.

4.1 Canonical paths

Let $\mathfrak{M}$ be an ergodic Markov chain with finite state space Ω, transition probabilities $P(\cdot, \cdot)$, and stationary distribution π. Any description of the canonical path argument is considerably simplified if we assume $\mathfrak{M}$ to be time-reversible. In the light of the detailed balance condition (3.1), we may view $\mathfrak{M}$ as an undirected graph (Ω, T) with vertex set Ω and edge set

$$T = \bigl\{\{x, y\} \in \Omega^{(2)} : \widetilde{P}(x, y) > 0\bigr\}, \tag{4.1}$$

where

$$\widetilde{P}(x, y) := \pi(x)P(x, y) = \pi(y)P(y, x). \tag{4.2}$$

For each (ordered) pair $(x, y) \in \Omega^2$, we specify a canonical path γ_{xy} from x to y in the graph (Ω, T); the canonical path γ_{xy} corresponds to a sequence of legal transitions in $\mathfrak{M}$ that leads from initial state x to final state y. Denote by $\Gamma = \{\gamma_{xy} : x, y \in \Omega\}$ the set of all canonical paths. For the method to yield good bounds, it is important to choose a set of paths Γ that avoids the creation of "hot spots:" edges of the graph that carry a particularly heavy burden of canonical paths. The degree to which an even loading has been achieved is measured by the quantity

$$\bar{\varrho} = \bar{\varrho}(\Gamma) := \max_t \frac{1}{\widetilde{P}(t)} \sum_{\gamma_{xy} \ni t} \pi(x)\pi(y)|\gamma_{xy}|,$$

where the maximum is over oriented edges (transitions) t of (Ω, T), and $|\gamma_{xy}|$ denotes the length of the path γ_{xy}.

Intuitively, we might expect a Markov chain to be rapidly mixing if it contains no "bottlenecks," i.e., if it admits a choice of paths Γ for which

$\bar{\varrho}(\Gamma)$ is not too large. This intuition is formalised in the following result derived from Sinclair [55], which is a development of a theorem of Diaconis and Stroock [19].

Theorem 4.1. *Let $\mathfrak{M}$ be a finite, time-reversible, ergodic Markov chain with loop probabilities $P(x,x) \geq \frac{1}{2}$ for all states x. Let Γ be a set of canonical paths with maximum edge loading $\bar{\varrho} = \bar{\varrho}(\Gamma)$. Then the mixing time of $\mathfrak{M}$ satisfies $\tau(\varepsilon) \leq \bar{\varrho}(\ln \pi(x)^{-1} + \ln \varepsilon^{-1})$.*[3]

Proof. Combine [55, Prop. 1] and [55, Thm 5].

We demonstrate the canonical path method by applying it to the toy example. For convenience, we shall work with a slightly modified version of the Markov chain $\mathfrak{M}_{\mathrm{col}}$ of section 3.. The transitions will be defined as before, except for the addition a preliminary step:

(0) with probability $\frac{1}{2}$ let X_{t+1} equal X_t and halt; otherwise, progress to step (1).

This modification has the effect of adding an additional loop probability $\frac{1}{2}$ to every state (and reducing all other transition probabilities by a similar factor). Let us refer to the modified Markov chain with increased loop probabilities as $\mathfrak{M}'_{\mathrm{col}}$. Note that $\mathfrak{M}'_{\mathrm{col}}(O_n, q)$ satisfies the conditions of Theorem 4.1.

Let $x = (x_0, x_1, \ldots, x_{n-1})$ and $y = (y_0, y_1, \ldots, y_{n-1})$ be arbitrary colourings in $\Omega = Q^n$. The canonical path γ_{xy} from x to y is composed of n edges, 0 to $n-1$, where edge i is simply

$$\big((y_0, \ldots, y_{i-1}, x_i, x_{i+1}, \ldots x_{n-1}), (y_0, \ldots, y_{i-1}, y_i, x_{i+1}, \ldots x_{n-1})\big),$$

i.e., we change the ith colour from x_i to y_i. Note that some of the edges may be loops (if $x_i = y_i$). To compute $\bar{\varrho}$, fix attention on a particular (oriented) edge

$$t = (w, w') = \big((w_0, \ldots, w_i, \ldots w_{n-1}), (w_0, \ldots, w'_i, \ldots w_{n-1})\big),$$

and consider the number of canonical paths γ_{xy} that include t. The number of possible choices for x is q^i, as the final $n-i$ positions are determined by $x_j = w_j$, for $j \geq i$; and by a similar argument the number of possible choices for y is q^{n-i-1}. Thus the total number of canonical paths using a particular edge t is q^{n-1}; furthermore, $\widetilde{P}(t) = \pi(w)P(w,w') \geq q^{-n}(2qn)^{-1}$, and the length of every canonical path is exactly n. Plugging all these bounds into the definition of $\bar{\varrho}$ yields $\bar{\varrho} \leq 2n^2$. Thus, by Theorem 4.1, the mixing time of $\mathfrak{M}'_{\mathrm{col}}(O_n, q)$ is $\tau(\varepsilon) \leq 2n^2(n \ln q + \ln \varepsilon^{-1})$. Note that the mixing time of $\mathfrak{M}'_{\mathrm{col}}(O_n, q)$ grows only polynomially with the input size n, even though the size of the state space is exponential in n, i.e., $\mathfrak{M}'_{\mathrm{col}}(O_n, q)$ is "rapidly mixing" in the sense of Section 3.. The bound on mixing time we have derived is some

[3] This Theorem also has a suitably stated converse; see [55, Thm 8].

way off the exact answer [17], which is $\tau(\varepsilon) = \mathrm{O}\bigl(n(\log n + \log \varepsilon^{-1})\bigr)$, and the slackness we see here is typical of the method.

On reviewing the canonical path argument, we perceive what appears to be a major weakness. In order to compute the key quantity $\bar{\varrho}$, we needed in turn to compute quantities such as $\widetilde{P}(t)$ that depend crucially on the size of the state space Ω. In the current example this does not present a problem, but in more interesting examples we do not know the size of the state space: indeed, our ultimate goal will often be to estimate this very quantity. Fortunately, it is possible to finesse this obstacle by implicit counting using a carefully constructed injective map. The idea will be illustrated by application to the Markov chain $\mathfrak{M}'_{\mathrm{col}}(O_n, q)$.

Let edge $t = (w, w')$ be as before, and denote by $\mathrm{cp}(t) = \{(x, y) : \gamma_{xy} \ni t\}$ the set of all (endpoints of) canonical paths that use edge t. Define the map $\eta_t : \mathrm{cp}(t) \to \Omega$ as follows: if $(x, y) = \bigl((x_0, \ldots, x_{n-1}), (y_0, \ldots, y_{n-1})\bigr) \in \mathrm{cp}(t)$ then

$$\eta_t(x, y) = (u_0, \ldots, u_{n-1}) := (x_0, \ldots, x_{i-1}, w_i, y_{i+1}, \ldots, y_{n-1}).$$

The crucial feature of the map η_t is that it is injective. To see this, observe that x and y may be unambiguously recovered from $(u_0, \ldots, u_{n-1}) = \eta_t(x, y)$ through the explicit expressions

$$x = (u_0, \ldots, u_{i-1}, w_i, w_{i+1}, \ldots, w_{n-1})$$

and

$$y = (w_0, \ldots, w_{i-1}, w'_i, u_{i+1}, \ldots, u_{n-1}).$$

Using the injective map η_t it is possible to evaluate $\bar{\varrho}$ without recourse to explicit counting. Noting[4] that $\pi(x)\pi(y) = \pi(w)\pi(\eta_t(x, y))$, we have

$$\begin{aligned}
\frac{1}{\widetilde{P}(t)} \sum_{\gamma_{xy} \ni t} \pi(x)\pi(y)|\gamma_{xy}| &= \frac{1}{\pi(w)P(w, w')} \sum_{\gamma_{xy} \ni t} \pi(w)\pi(\eta_t(x, y))\,|\gamma_{xy}| \\
&= \frac{n}{P(w, w')} \sum_{\gamma_{xy} \ni t} \pi(\eta_t(x, y)) \\
&\le \frac{n}{P(w, w')} \le 2qn^2,
\end{aligned}$$

where the penultimate inequality follows from the facts that η_t is injective, and that π is a probability distribution. Since the above argument is valid uniformly over the choice of t, we deduce $\bar{\varrho} \le 2qn^2$. The factor of q as compared with the direct argument was lost to redundancy in the encoding: the map η_t was not a bijection.

[4] This is a trivial observation when the stationary distribution is uniform, as it is here, but it is sometimes possible, by judicious choice of η_t, to contrive such an identity even when the stationary distribution is non-uniform. See Section 5.1 for an example.

4.2 Geometry

As before, suppose $\mathfrak{M}$ is a finite, time-reversible, ergodic Markov chain with stationary distribution π, and recall definitions (4.2) and (4.1) of $\widetilde{P}$ and T from the previous section. The *conductance* [57] of $\mathfrak{M}$ is defined by

$$\Phi = \Phi(\mathfrak{M}) := \min_{\substack{S \subset \Omega \\ 0 < \pi(S) \leq 1/2}} \frac{\widetilde{P}(S, \overline{S}\,)}{\pi(S)}, \tag{4.3}$$

where $\widetilde{P}(S, \overline{S}\,)$ denotes the sum of $\widetilde{P}(x, y)$ over edges $\{x, y\} \in T$ with $x \in S$ and $y \in \overline{S} = \Omega \setminus S$. The conductance may be viewed as a weighted version of edge expansion of the graph (Ω, T) associated with $\mathfrak{M}$. Alternatively, the quotient appearing in (4.3) can be interpreted as the conditional probability that the chain in equilibrium escapes from the subset S of the state space in one step, given that it is initially in S; thus Φ measures the readiness of $\mathfrak{M}$ to escape from any small enough region of the state space, and hence to make rapid progress towards equilibrium. This intuitive connection can be given a precise quantitative form as follows. (Related results may be found in the work of Aldous [2] and Alon [4].)

Theorem 4.2. [Sinclair.] *Let $\mathfrak{M}$ be a finite, reversible, ergodic Markov chain with loop probabilities $P(x, x) \geq \frac{1}{2}$ for all states x. Let Φ be the conductance of $\mathfrak{M}$ as defined in* (4.3). *Then the mixing time of $\mathfrak{M}$ satisfies $\tau(\varepsilon) \leq 2\Phi^{-2} \times (\ln \pi(x)^{-1} + \ln \varepsilon^{-1})$.*

Proof. Combine [55, Prop. 1] and [55, Thm 2].

Our approach in this section to bounding the conductance of a Markov chain $\mathfrak{M}$ is to give $\mathfrak{M}$ a geometric interpretation, in which states of $\mathfrak{M}$ are identified with certain polytopes, and transitions with their common facets. A lower bound on conductance then follows from an "isoperimetric inequality." This was the approach pioneered by Dyer, Frieze and Kannan in the analysis of a random walk in a convex body [22], and Karzanov and Khachiyan in the context of a Markov chain on linear extensions of a partial order [44] (see also Section 5.1). The following isoperimetric inequality of Dyer and Frieze, is particularly well suited to the purpose. To state the inequality, we need the concept of the dual of a norm. If $\|\cdot\|$ is a norm, then the norm $\|\cdot\|^*$ *dual* to $\|\cdot\|$ is defined by

$$\|x\|^* = \sup\{a \cdot x : \|a\| = 1\}.$$

The symbol ∂ denotes "boundary of."

Theorem 4.3. [Dyer and Frieze.] *Suppose $K \subseteq \mathbb{R}^n$ is a convex body and f a log-concave function on $\operatorname{int} K$. For a set $S \subseteq K$ such that $\sigma = \partial S \setminus \partial K$ is a piecewise smooth surface, define $\mu(S) = \int_S f(x)\,dx$ and $\mu'(S) = \int_\sigma f(x)\,\|u(x)\|^*\,dx$, where $u(x)$ is the Euclidean unit normal to σ at $x \in \sigma$. If*

$\mu(S) \le \frac{1}{2}\mu(K)$ then $\mu(S)/\mu'(S) \le \frac{1}{2}\operatorname{diam} K$, where the diameter $\operatorname{diam} K$ is measured with respect to the (primal) norm $\|\cdot\|$.

Proof. See [20, Thm 3] and preliminary lemmas.

We illustrate the utility of Theorem 4.3 by applying it to the toy example. We again work with the modified Markov chain $\mathfrak{M}'_{\mathrm{col}}(O_n, q)$, with inflated loop probabilities, applied to the empty graph O_n. We view states (colourings of G) as functions $V \to Q$, where $V = [n]$ and $Q = [q]$. For each colouring $c \in \Omega$, define a corresponding polytope (a closed, bounded region formed by the intersection of halfspaces) in $\mathbb{R}^{n\times q}$ by

$$R(c) = \left\{x = (x_{ij}) \in \mathbb{R}^{n\times q} : 0 \le x_{ij} \le 1 \text{ and } x_{i,c(i)} \ge x_{ij} \text{ for all } i,j\right\}.$$

For any $S \subseteq \Omega$, let $R(S) = \bigcup_{c\in S} R(c)$, and observe that $K := R(\Omega) = \frac{1}{2}B_\infty$, where $\frac{1}{2}B_\infty$ denotes the l_∞-ball of radius $\frac{1}{2}$, or unit cube. Clearly, $\operatorname{diam} K = 1$, where diameter is measured with respect to l_∞-norm. Note that, by symmetry, $\mathrm{vol}_{nq} R(c) = |\Omega|^{-1}$ for any $c \in \Omega$, and hence

$$\mathrm{vol}_{nq} R(S) = \frac{|S|}{|\Omega|}. \tag{4.4}$$

Recall the definitions of $\widetilde{P}$ (4.2) and of conductance (4.3). A transition is available between colourings c and c' (we say the colourings are *adjacent*) if they differ at exactly one vertex; equivalently, if $R(c)$ and $R(c')$ share a common facet (i.e., $(nq-1)$-dimensional face). By calculus the area (i.e., $(nq-1)$-dimensional volume) of such a facet is

$$\mathrm{vol}_{nq-1}(R(c) \cap R(c')) = \frac{\sqrt{2}}{q^{n-1}(q-1)}. \tag{4.5}$$

(See the Appendix for a proof of this claim.) Thus the number of transitions $(c, c') \in (S, \overline{S}\,)$ from a state in S to one in $\overline{S}$ is

$$\mathrm{vol}_{nq-1}(\partial\, R(S) \setminus \partial\, K) \times \frac{q^{n-1}(q-1)}{\sqrt{2}},$$

and, since the $\widetilde{P}(c, c') = (2nq\,|\Omega|)^{-1}$ for any pair of adjacent states c, c',

$$\widetilde{P}(S, \overline{S}\,) = \frac{q^{n-1}(q-1)\,\mathrm{vol}_{nq-1}(\partial\, R(S) \setminus \partial\, K)}{2\sqrt{2}\,nq\,|\Omega|}. \tag{4.6}$$

Furthermore the unit vector u normal to any facet has l_1-norm $\|u\|_1 = \sqrt{2}$. Taking f identically 1 in Theorem 4.3, we have, for $|S| \le \frac{1}{2}|\Omega|$,

$$\frac{\mathrm{vol}_{nq} R(S)}{\sqrt{2}\ \mathrm{vol}_{nq-1}(\partial\, R(S) \setminus \partial\, K)} \le \frac{\operatorname{diam} K}{2},$$

which, in the light of (4.4), is equivalent to

$$\mathrm{vol}_{nq-1}(\partial R(S) \setminus \partial K) \geq \frac{\sqrt{2}\,|S|}{|\Omega|}.$$

Combining this inequality with (4.6) yields

$$\widetilde{P}(S,\overline{S}\,) \geq \frac{(q-1)|S|}{2nq^2|\Omega|},$$

whence, by definition of conductance (4.3),

$$\Phi \geq \frac{q-1}{2nq^2}.$$

Thus, by Theorem 4.2, the mixing time of $\mathfrak{M}'_{\mathrm{col}}(O_n,q)$ is

$$\tau(\varepsilon) \leq 8n^2q^4(q-1)^{-2}(n\ln q + \ln\varepsilon^{-1}).$$

Again, we have demonstrated that $\mathfrak{M}'_{\mathrm{col}}(O_n,q)$ is rapidly mixing, though the bound is worse by a factor of order q^2 than the one we had already obtained using the canonical paths argument.

4.3 Coupling

Suppose $\mathfrak{M}$ is a countable, ergodic (though not necessarily time-reversible) Markov chain with transition probabilities $P(\cdot\,,\cdot)$ and stationary distribution π. As usual, the assumption of countability is for expositional convenience only, and the ideas easily extend to uncountably infinite state spaces. In its basic form, the coupling technique was introduced by Doeblin in the 1930s. The word "coupling" in probability theory is applied to a variety of related notions, and it would be difficult to provide a general definition. In the current context, we mean by coupling a Markov process (X_t, Y_t) on $\Omega\times\Omega$ such that each of the processes (X_t) and (Y_t), considered in isolation, is a faithful copy of $\mathfrak{M}$. More precisely, we require that

$$\Pr(X_{t+1} = x' \mid X_t = x \wedge Y_t = y) = P(x,x') \tag{4.7}$$

and

$$\Pr(Y_{t+1} = y' \mid X_t = x \wedge Y_t = y) = P(y,y'), \tag{4.8}$$

for all $x,y,x',y' \in \Omega$. This condition is consistent with (X_t) and (Y_t) being independent evolutions of $\mathfrak{M}$, but does not imply it. In fact, we shall use the possibility that

$$\Pr\left(X_{t+1} = x' \wedge Y_{t+1} = y' \;\middle|\; X_t = x \wedge Y_t = y\right) \neq P(x,x')\,P(y,y')$$

to encourage (X_t) and (Y_t) to *coalesce* rapidly, so that $X_t = Y_t$ for all sufficiently large t. (Note that it is easy to design the coupling so that, if t is the first time step such that $X_t = Y_t$, then $X_{t'} = Y_{t'}$ for all $t' > t$.)

If it can be arranged that coalescence occurs rapidly—independently of the initial states X_0 and Y_0—we may deduce that $\mathfrak{M}$ is rapidly mixing. The key result we use here is the "Coupling Lemma," which apparently makes its first explicit appearance in the work of Aldous [1, Lemma 3.6] (see also Diaconis [17, Chap. 4, Lemma 5]).

Lemma 4.1. *Suppose that $\mathfrak{M}$ is a countable, ergodic Markov chain with transition probabilities $P(\cdot,\cdot)$, and let $\big((X_t,Y_t):t\in\mathbb{N}\big)$ be a coupling, i.e., a Markov process satisfying (4.7) and (4.8). Suppose further that $t:(0,1]\to\mathbb{N}$ is a function such that $\Pr(X_{t(\varepsilon)}\neq Y_{t(\varepsilon)})\leq\varepsilon$ for all $\varepsilon\in(0,1]$, uniformly over the choice of initial state (X_0,Y_0). Then the mixing time $\tau(\varepsilon)$ of $\mathfrak{M}$ is bounded above by $t(\varepsilon)$.*

Proof. Let $X_0=x\in\Omega$ be arbitrary, and choose Y_0 according to the stationary distribution π. Fix $\varepsilon\in(0,1]$ and for convenience abbreviate $t(\varepsilon)$ to t. Let $A\subseteq\Omega$ be an arbitrary event. Then

$$\begin{aligned}\Pr(X_t\in A)&\geq\Pr(Y_t\in A\wedge X_t=Y_t)\\&\geq 1-\Pr(Y_t\notin A)-\Pr(X_t\neq Y_t)\\&\geq\Pr(Y_t\in A)-\varepsilon\\&=\pi(A)-\varepsilon,\end{aligned}$$

with a similar inequality holding for the complementary event $\Omega\setminus A$. Since A was chosen arbitrarily, $D_{\mathrm{tv}}(P^t(x,\cdot),\pi)\leq\varepsilon$, i.e., the total variation distance between the t-step distribution and the stationary distribution is bounded by ε.

For the toy example, the coupling may be very simple indeed. The transition $(X_t,Y_t)\to(X_{t+1},Y_{t+1})$ in the coupling is defined by the following experiment:

(1) Select a vertex $v\in V$, u.a.r.
(2) Select a colour $c\in Q$ u.a.r., and recolour vertex v in X_t (respectively Y_t) with colour c and let the resulting colouring be X_{t+1} (respectively Y_{t+1}).

Note that (X_t) and (Y_t) are both faithful copies of $\mathfrak{M}$; specifically, (4.7) and (4.8) are satisfied. Nevertheless it is also clear that (X_t) and (Y_t) are "highly coupled" and we can expect rapid coalescence.

As before, regard states (colourings) as functions $V\to Q$. Denote by D_t the random variable

$$D_t=\{v\in V:X_t(v)\neq Y_t(v)\},$$

i.e., the set of vertices on which the two colouring X_t and Y_t disagree. If step (1) of the colouring selects a vertex v in D_t, then $D_{t+1}=D_t\setminus\{v\}$; otherwise $D_{t+1}=D_t$. Since v is selected u.a.r.,

$$\operatorname{Exp}\big(|D_{t+1}|\bigm|D_t\big)=\Big(1-\frac{1}{n}\Big)|D_t|,$$

and hence
$$\operatorname{Exp}\big(|D_t| \bigm| D_0\big) = \Big(1 - \frac{1}{n}\Big)^t |D_0|.$$
Since $|D_t|$ is a non-negative integer r.v., we obtain
$$\begin{aligned}\Pr\big(|D_t| > 0 \bigm| D_0\big) &\le \operatorname{Exp}\big(|D_t| \bigm| D_0\big)\\ &\le n\Big(1 - \frac{1}{n}\Big)^t\\ &\le n\operatorname{Exp}(-t/n),\end{aligned}$$
which is bounded by ε, provided $t \ge n \ln n\varepsilon^{-1}$. Invoking the Coupling Lemma we obtain $\tau_x(\varepsilon) \le n(\ln n + \ln \varepsilon^{-1})$, independent of the starting state x, the correct asymptotic result.

5. Some more challenging applications

We now reprise the three techniques for proving rapid mixing in the context of three more realistic problems. In each case, the chosen solution technique will be "natural" for the application. Indeed, for our first example, we are forced to use the canonical paths method, as it provides the only known solution technique.

5.1 Monomer-dimer coverings via canonical paths

The presentation of this topic is condensed from Jerrum and Sinclair [36], which in turn is an improved version of the original source [34]. See also Sinclair [56].

We shall be concerned with the classical monomer-dimer model from statistical physics. A *monomer-dimer system* is defined by a graph $G = (V, E)$ and a positive real parameter λ. A *configuration* of the system is just a *matching* in G, that is to say, a subset $M \subseteq E$ such that no two edges in M share an endpoint. In physical terms, the pairs of matched vertices are *dimers* and the uncovered vertices *monomers*. Thus a matching of cardinality k, or *k-matching*, corresponds precisely to a monomer-dimer configuration with k dimers and $2(n-k)$ monomers, where $2n = |V|$ is the number of vertices in G. (The assumption that the number of vertices in G is even is inessential and is made for notational convenience.) Typically, G is a regular lattice in some fixed number of dimensions, but we shall make no such assumption what follows. For a detailed account of the history and significance of monomer-dimer systems, the reader is referred to the seminal paper of Heilmann and Lieb [32] and the references given there.

To each matching M, a *weight* $w(M) = \lambda^{|M|}$ is assigned; thus the parameter λ reflects the contribution of a dimer to the energy of the system. The *partition function* of the system is defined as

$$Z = Z(G, \lambda) := \sum_M w(M) = \sum_{k=0}^{n} m_k \lambda^k, \tag{5.1}$$

where $m_k = m_k(G)$ is the number of k-matchings in G. For a physical interpretation of (5.1), see [32]. The partition function may be efficiently approximated (in the FPRAS sense) using the method of Section 2., provided we can efficiently sample matchings from the distribution that assigns probability

$$\pi(M) = \frac{w(M)}{Z} \tag{5.2}$$

to matching M (see [36] for details). We therefore concentrate on the sampling problem.

Following an idea of Broder [8], we construct a Markov chain $\mathfrak{M}_{\mathrm{match}} = \mathfrak{M}_{\mathrm{match}}(G, \lambda)$, parameterised by the underlying graph G and the edge weight λ. The state space, Ω, is the set of all matchings in G, and the transitions are constructed so that the chain is ergodic with stationary distribution π given by (5.2). In other words, the stationary probability of each matching (monomer-dimer configuration) is proportional to its weight in the partition function (5.1). The Markov chain $\mathfrak{M}_{\mathrm{match}}$, if simulated for sufficiently many steps, provides a method of sampling matchings from the distribution π.

It is not hard to construct a Markov chain $\mathfrak{M}_{\mathrm{match}}$ with the right asymptotic properties. Let the state of $\mathfrak{M}_{\mathrm{match}}$ at time t be X_t. The probability distribution of the next state X_{t+1} is defined by the following experiment:

(1) With probability $\frac{1}{2}$ let $X_{t+1} := X_t$ and halt.
(2) Otherwise (with the remaining probability $\frac{1}{2}$), select an edge $e = \{u, v\} \in E$, u.a.r., and set

$$M' := \begin{cases} M - e & \text{if } e \in M; \\ M + e & \text{if both } u \text{ and } v \text{ are unmatched in } M; \\ M + e - e' & \text{if exactly one of } u \text{ and } v \text{ is matched in } M \\ & \qquad \text{and } e' \text{ is the matching edge;} \\ M & \text{otherwise.} \end{cases}$$

(3) With probability $\min\{1, \pi(M')/\pi(M)\}$ let $X_{t+1} := M'$; otherwise (with the complementary probability) let $X_{t+1} := M$.

It is helpful to view this chain as follows. There is an underlying graph defined on the set of matchings Ω in which the neighbours of matching M are all matchings M' that differ from M via one of the following local perturbations: an edge is removed from M (a $\downarrow$-*transition*); an edge is added to M (a $\uparrow$-*transition*); or a new edge is exchanged with an edge in M (a $\leftrightarrow$-*transition)*. Transitions from M are made by first selecting a neighbour M' u.a.r., and then actually making, or *accepting* the transition with probability $\max\{1, \pi(M')/\pi(M)\}$. Note that the ratio appearing in this expression is easy to compute: it is just λ^{-1}, λ or 1 respectively, according to the type of the transition.

As the reader may easily verify, this acceptance probability is constructed so that the transition probabilities $P(M, M')$ of $\mathfrak{M}_{\text{match}}$ satisfy the detailed balance condition (3.1). Furthermore $\mathfrak{M}_{\text{match}}$ is irreducible (i.e., all states communicate via the empty matching) and aperiodic (by step (1), the self-loop probabilities $P(M, M)$ are all non-zero), and hence ergodic. Thus, by Lemma 3.1, the distribution π defined in (5.2) is indeed the stationary distribution of $\mathfrak{M}_{\text{match}}$.[5]

Proposition 5.1. *The mixing time of the Markov chain $\mathfrak{M}_{\text{match}}$ satisfies*

$$\tau(\varepsilon) \le 4\,|E|\,n\bar{\lambda}\big(n(\ln n + \ln\bar{\lambda}) + \ln\varepsilon^{-1}\big),$$

where $\bar{\lambda} = \max\{1, \lambda\}$.

Proof (sketch). Our strategy will be to carefully choose a collection of canonical paths $\Gamma = \{\gamma_{XY} : X, Y \in \Omega\}$ in the Markov chain $\mathfrak{M}_{\text{match}}$ for which the "bottleneck" measure $\bar{\varrho}(\Gamma)$ of Section 4.1 is small. We can then appeal to Theorem 4.1 to bound the mixing time. Specifically, we shall show that our paths satisfy

$$\bar{\varrho}(\Gamma) \le 4\,|E|\,n\bar{\lambda}. \tag{5.3}$$

Since the number of matchings in G is certainly bounded above by $(2n)!$, the stationary probability $\pi(X)$ of any matching X is bounded below by $\pi(X) \ge 1/\bar{\lambda}^n(2n)!$. Using (5.3) and the fact that $\ln n! \le n\ln n$, the bound on the mixing time in Proposition 5.1 can now be read off Theorem 4.1.

It remains for us to find a set of canonical paths Γ satisfying (5.3). For each pair of matchings X, Y in G, we construct a canonical path γ_{XY} from X to Y as indicated in Figure 5.1. (A rigorous description of the canonical paths together with all other details missing from this sketch proof may be found in [36].)

The interpretation of Figure 5.1 is as follows. Consider the symmetric difference $X \oplus Y$. A moment's reflection should convince the reader that this consists of a disjoint collection of paths in G (some of which may be closed cycles), each of which has edges that belong alternately to X and to Y. Now suppose that we have fixed some arbitrary ordering on all simple paths in G, and designated in each of them a so-called "start vertex," which is arbitrary if the path is a closed cycle but must be an endpoint otherwise. This ordering induces a unique ordering $P_1, P_2, \ldots, P_m$ on the paths appearing in $X \oplus Y$. The canonical path from X to Y involves "unwinding" each of the P_i in turn. In Figure 5.1 the path P_i (which happens to be a cycle) is the one currently being unwound; the paths $P_1, \ldots, P_{i-1}$ to the left have already been processed, while the ones $P_{i+1}, \ldots, P_m$ are yet to be dealt with.

[5] The device of performing random walk on a connected graph with acceptance probabilities of this form is well known in computational physics under the name of the "Metropolis process" [51]. Clearly, it can be used to achieve any desired stationary distribution π for which the ratio $\pi(u)/\pi(v)$ for neighbours u, v can be computed easily.

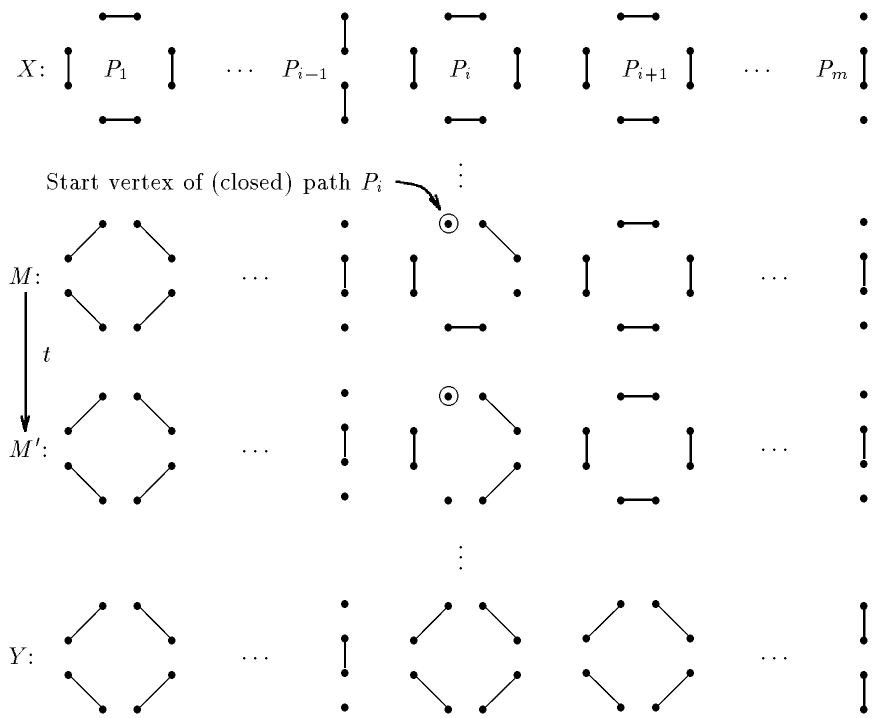

Fig. 5.1. A transition t in the canonical path from X to Y

Unwinding a cycle is done by removing the edge adjacent to the start vertex using a $\downarrow$-transition; then moving round the cycle using $\leftrightarrow$-transitions to swap Y-edges for X-edges; and finally completing the cycle with a single $\uparrow$-transition. A path is processed similarly, working from one end to the other using a sequence of $\leftrightarrow$-transitions to swap Y-edges for X-edges, starting or finishing with the job with single $\uparrow$- or $\downarrow$-transitions as required.

We now proceed to bound the "bottleneck" measure $\bar{\varrho}(\Gamma)$ for these paths, using the injective mapping technology introduced in Section 4.1. Let t be an arbitrary edge in the Markov chain, i.e., a transition from M to $M' \neq M$, and let $\mathrm{cp}(t) = \{(X, Y) : \gamma_{XY} \ni t\}$ denote the set of all canonical paths that use t. Just as in Section 4.1, we shall obtain a bound on the total weight of all paths that pass through t by defining an injective mapping $\eta_t : \mathrm{cp}(t) \to \Omega$. By analogy with the toy example in Section 4.1, what we would like to do is to set $\eta_t(X, Y) = X \oplus Y \oplus (M \cup M')$; the intuition for this is that $\eta_t(X, Y)$ should agree with X on paths that have already been unwound, and with Y on paths that have not yet been unwound (just as $\eta_t(x, y)$ agreed with x on positions $1, \ldots, i-1$ and with y on positions $i+1, \ldots, n-1$). This will not quite do, since the set of edges $\eta_t(X, Y)$ defined in this way may fail to be a matching; however, the problem is a small one, and can be rectified by removing a single

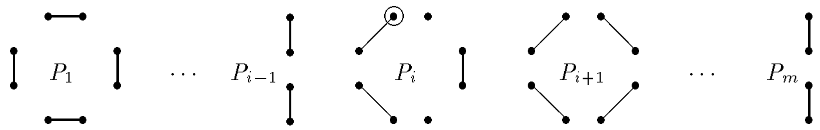

Fig. 5.2. The corresponding encoding $\eta_t(X,Y)$

offending edge. Figure 5.2 illustrates the encoding $\eta_t(X,Y)$ that would result from the transition t on the canonical path sketched in Figure 5.1.

We now have to check that η_t is injective, which amounts to demonstrating that X and Y can be unambiguously reconstructed from a knowledge of $t=(M,M')$ and $\eta_t(X,Y)$. Roughly, the way this is done is to note that, modulo the single offending edge,

$$X \oplus Y = \eta_t(X,Y) \oplus (M \cup M');$$

so that, given $t=(M,M')$ and $\eta_t(X,Y)$, we may compute the path decomposition $P_1,\ldots,P_m$. The path P_i being unwound during the transition t is immediately apparent from an examination of $M \oplus M'$. From there, it is a straightforward matter to apportion edges in $P_1 \cup \cdots \cup P_m$ to X or Y as appropriate. Finally, edges in $\eta_t(X,Y)\cap M$ are the ones which are common to X and Y.

We are almost done. However, the fact that η_t is injective is not sufficient in this case because, in contrast to the toy example, the stationary distribution π is highly non-uniform. What we require in addition is that η_t be "weight-preserving," in the sense that $\widetilde{P}(t)\pi(\eta_t(X,Y))$ is reasonably close to $\pi(X)\pi(Y)$. Roughly speaking, this occurs because each edge $e\in E$ (with a couple of exceptions) contributes an equal factor—1, λ or λ^2—to the two terms $\pi(M)\pi(\eta_t(X,Y))$ and $\pi(X)\pi(Y)$. Specifically it can be shown that

$$\pi(X)\pi(Y) \le 2\,|E|\,\bar\lambda\,\widetilde{P}(t)\pi(\eta_t(X,Y)). \tag{5.4}$$

It is not too difficult to achieve a looser variant of (5.4) with $\bar\lambda^2$ replacing $\bar\lambda$ on the right hand side, but the inequality as given requires a little care. The full calculation can be found in [36].

A bound on $\bar\varrho$ follows easily from (5.4). We have

$$\begin{aligned}
\bar\varrho(\varGamma) &:= \frac{1}{\widetilde{P}(t)} \sum_{\gamma_{XY}\ni t} \pi(X)\pi(Y)\,|\gamma_{XY}| && (5.5)\\
&\le 2\,|E|\,\bar\lambda \sum_{\gamma_{XY}\ni t} \pi(\eta_t(X,Y))\,|\gamma_{XY}| \\
&\le 4\,|E|\,n\bar\lambda \sum_{\gamma_{XY}\ni t} \pi(\eta_t(X,Y)) \\
&\le 4\,|E|\,n\bar\lambda, && (5.6)
\end{aligned}$$

where the second inequality follows from the fact that the length of any canonical path is bounded by $2n$, and the last inequality from the facts that η_t is injective and π is a probability distribution. The claimed bound on mixing time follows quickly from (5.6) and Theorem 4.1, as described at the outset.

Aside from the monomer-dimer example presented in this section, applications of the canonical path method include: counting dimer coverings (perfect matchings) of lattice graphs (Kenyon, Randall and Sinclair [47]), evaluating the partition function of the ferromagnetic Ising model (Jerrum and Sinclair [35]) and counting configurations in the "six-point ice model" (Mihail and Winkler [52]). All these applications share similarities with the monomer-dimer one.

An application which is further removed from the monomer-dimer example is to the "basis-exchange" random walk for graphic matroids. The state space here is the set of spanning trees of a graph, and a transition from tree T to T' is possible iff the symmetric difference of T and T' consists of just two edges. The canonical paths argument for spanning trees has not, as far as I am aware, appeared explicitly in the literature, but Cordovil and Moreira have presented a construction (see [16, Thm 1.6]) for paths between pairs of spanning trees that is ideally suited to this purpose. However, there are many other approaches to proving rapid mixing in this instance (see Aldous [3], Dyer and Frieze [21] and Feder and Mihail [27]). Refer to Section 8. for a related open problem.

5.2 Linear extensions of a partial order via geometry

In this example, we essentially follow Karzanov and Khachiyan [44], though we achieve a sharper bound by invoking an enhanced isoperimetric inequality due to Dyer and Frieze [20].

We are given a partially ordered set $(V, \prec)$, where $V = [n]$. Denote by $\operatorname{Sym} V$ the symmetric group on V. We are interested in sampling, u.a.r., a member of the set

$$\Omega = \bigl\{g \in \operatorname{Sym} V : g(i) \prec g(j) \Rightarrow i \le j, \text{ for all } i, j \in V\bigr\}$$

of linear extensions of $\prec$. In forming a mental picture of the state space Ω, the following characterisation may be helpful: $g \in \Omega$ iff the linear order

$$g(0) \sqsubset g(1) \sqsubset \cdots \sqsubset g(n-1) \tag{5.7}$$

extends, or is consistent with, the partial order $\prec$.

As usual, we propose to sample from Ω by constructing an ergodic Markov chain on state space Ω, whose stationary distribution is uniform. Transitions from a linear extension $g \in \Omega$ are generated by composing g with a random transition $(p, p{+}1)$;[6] equivalently, by swapping adjacent elements in the linear order (5.7). Formally, transition probabilities from state $X_t \in \Omega$ are defined by the following experiment:

(1) Select $p \in [n-1]$ and $r \in \{0,1\}$, u.a.r.
(2) If $r = 1$ and $X_t \circ (p, p+1) \in \Omega$ then $X_{t+1} := X_t \circ (p, p+1)$; otherwise $X_{t+1} := X_t$.

Here, the operator $\circ$ denotes function composition (read right to left). Let us refer to this Markov chain as $\mathfrak{M}_{\mathrm{le}}$. As in Section 4.2, the loop probabilities are artificially raised to permit convenient application of Theorem 4.2.

Proposition 5.2. *The mixing time of the Markov chain $\mathfrak{M}_{\mathrm{le}}$ satisfies*

$$\tau(\varepsilon) \le 8n^2(n-1)^2(\ln \Omega^{-1} + \ln \varepsilon^{-1}) = O\big(n^4(n \ln n + \ln \varepsilon^{-1})\big).$$

We shall see in Section 6. that this bound can be tightened considerably.

Proof. We adopt the notation introduced in Section 4.2. To each permutation $g \in \mathrm{Sym}\, V$, associate the simplex

$$R(g) = \big\{x = (x_i) \in \mathbb{R}^n : 0 \le x_{g(0)} \le x_{g(1)} \le \cdots \le x_{g(n-1)} \le 1\big\}.$$

For any $S \subseteq \mathrm{Sym}\, V$, let $R(S) = \bigcup_{g \in S} R(g)$, and observe that $R(\mathrm{Sym}\, V) = \frac{1}{2}B_\infty$, where $\frac{1}{2}B_\infty$ denotes the l_∞-ball of radius $\frac{1}{2}$, or unit cube. Define $K := R(\Omega)$, and observe that K is a convex set. (Take any two points in K and join them by a straight line segment. It is routine to check that every intermediate point is contained in a simplex $R(g)$, where g is a linear extension of $\prec$.) Clearly, $\mathrm{diam}\, K \le \mathrm{diam}(R(\mathrm{Sym}\, V)) \le 1$, where diameter is measured with respect to l_∞-norm. Note that, by symmetry, $\mathrm{vol}_n\, R(g) = |\,\mathrm{Sym}\, V|^{-1} = 1/n!$ for any $g \in \Omega$, and hence

$$\mathrm{vol}_n\, R(S) = \frac{|S|}{n!}. \tag{5.8}$$

A transition is available between linear extensions g and g' (we say that g and g' are *adjacent*) if they differ in an adjacent transposition; equivalently, if $R(g)$ and $R(g')$ share a common $(n-1)$-dimensional face. By an argument very similar to that used in Section 4.2 (see also the Appendix), if g and g' are adjacent,

$$\mathrm{vol}_{n-1}(R(g) \cap R(g')) = \frac{\sqrt{2}}{(n-1)!},$$

so the number of transitions $(g, g') \in (S, \overline{S}\,)$ from a state in S to one in $\overline{S}$ is

$$\mathrm{vol}_{n-1}(\partial\, R(S) \setminus \partial\, K) \times \frac{(n-1)!}{\sqrt{2}},$$

[6] The transposition is to be performed first, followed by the permutation g.

and

$$\widetilde{P}(S,\overline{S}\,) = \frac{(n-1)!\ \mathrm{vol}_{n-1}(\partial\, R(S) \setminus \partial\, K)}{2\sqrt{2}\,(n-1)\,|\Omega|}. \tag{5.9}$$

Furthermore the unit vector u normal to any facet has l_1-norm $\|u\|_1 = \sqrt{2}$. Taking f identically 1 in Theorem 4.3, we have, for $|S| \leq \frac{1}{2}|\Omega|$,

$$\frac{\mathrm{vol}_n\, R(S)}{\sqrt{2}\ \mathrm{vol}_{n-1}(\partial\, R(S) \setminus \partial\, K)} \leq \frac{\mathrm{diam}\, K}{2},$$

which, in the light of (5.8), is equivalent to

$$\mathrm{vol}_{n-1}(\partial\, R(S) \setminus \partial\, K) \geq \frac{\sqrt{2}\,|S|}{n!}.$$

Combining this inequality with (5.9) yields

$$\widetilde{P}(S,\overline{S}\,) \geq \frac{|S|}{2n(n-1)|\Omega|},$$

whence

$$\Phi \geq \frac{1}{2n(n-1)}.$$

The claimed bound on mixing time now follows from Theorem 4.2.

By far the most important application of the techniques deployed here and in Section 4.2 is to the analysis of random walks in convex bodies. The groundbreaking work on this topic was done by Dyer, Frieze and Kannan [22], who showed that a certain natural random walk in a convex body $K \subset \mathbb{R}^n$ is rapidly mixing. As a consequence, they were able to exhibit the first FPRAS for approximating the volume of a convex body. (The significant point here is that the running time of the algorithm is polynomial in the dimension n, whereas all previous approaches were exponential in n.) In this application, the state space comes ready equipped with a geometric interpretation, so the conductance argument is a natural candidate.

The random walk employed in [22] was akin to a traditional unbiased random walk on a (sufficiently fine) n-dimensional lattice, but restricted to the interior of the body. The time complexity of the resulting sampling procedure was a high-degree polynomial in the dimension n. The perceived importance of the volume estimation problem spurred various authors to improve on Dyer et al.'s proposal in various directions: widening the range of applicability refining the algorithmic techniques and sharpening the analytical tools. Applegate and Kannan [5] extended the method to cover integration of log-concave functions; Lovász and Simonovits [41] replaced the grid walk with a kind of discretised Brownian motion known as the "ball walk"; and Dyer and Frieze [20] introduced an improved isoperimetric inequality. Refer to Kannan [40] for an overview of the topic, and Kannan, Lovász and Simonovits [42] to learn the state of the art.

5.3 Colourings of a low-degree graph via coupling

We return to the Markov chain $\mathfrak{M}_{\rm col}(G, q)$ of Section 3., and use the coupling method to analyse its mixing time for graphs G of low degree.

Lemma 5.1. *Let G be a graph of maximum degree Δ on n vertices. Assuming $q \geq 2\Delta + 1$, the mixing time $\tau(\varepsilon)$ of the Markov chain $\mathfrak{M}_{\rm col}(G, q)$ is bounded above by*

$$\tau(\varepsilon) \leq \frac{q - \Delta}{q - 2\Delta}\, n \ln\left(\frac{n}{\varepsilon}\right).$$

In order to define an appropriate coupling in this instance, the following easy technical lemma is useful.

Lemma 5.2. *Let U be a finite set, A, B be subsets of U, and X_A, X_B be random variables, taking values in U, which are uniform and supported on A (respectively B). Then there is a joint sample space for X_A and X_B such that*

$$\Pr(X_A(\omega) = X_B(\omega)) = \frac{|A \cap B|}{\max\{|A|, |B|\}}$$

The proof of Lemma 5.2 is left as an easy exercise.

Proof (of Lemma 5.1). The proof is adapted from [33]; note however that the proof there applies to a Metropolis-style Markov chain rather than the heat-bath dynamics version considered here.

We construct a coupling, as in section 4.3, but now taking account of the constraints imposed by the edges of G. For all $v \in V$ denote by $\Gamma(v) \subseteq V$ the set of all neighbours of v in G, and by $X_t(v)$ (respectively, $Y_t(v)$) the colour of vertex v in colouring X_t (respectively, Y_t). Further, for all $U \subseteq V$, let $X_t(U) = \{X_t(u) : u \in U\}$. The transition $(X_t, Y_t) \to (X_{t+1}, Y_{t+1})$ in the coupling is defined by the following experiment:

(1) Select a vertex $v \in V$, u.a.r.
(2) Choose a colour $c_X \in Q \setminus X_t(\Gamma(v))$ and a colour $c_Y \in Q \setminus Y_t(\Gamma(v))$ u.a.r., using the joint sample space of Lemma 5.2.
(3) In the colouring X_t (respectively Y_t), recolour vertex v with colour c_X (respectively c_Y) to obtain a new colouring X_{t+1} (respectively Y_{t+1}).

Let $A = A_t \subseteq V$ be the set of vertices on which the colourings X_t and Y_t agree, and $D = D_t \subseteq V$ be the set on which they disagree. Let $d'(v)$ denote the number of edges incident at vertex v that have one endpoint in A and one in D. Observe that

$$\sum_{v \in A} d'(v) = \sum_{v \in D} d'(v) = m', \tag{5.10}$$

where m' is the number of edges of G that span A and D.

It is clear that $|D_{t+1}| - |D_t| \in \{-1, 0, 1\}$. Consider first the probability that $|D_{t+1}| = |D_t| + 1$. For this event to occur, the vertex v selected in step (1) must lie in A, and the new colours c_X and c_Y selected in step (2) must be unequal. Fix a vertex $v \in A$, and denote by $\xi = |Q \setminus X_t(\Gamma(v))|$ (respectively, $\eta = |Q \setminus Y_t(\Gamma(v))|$) the number of possible values for c_X (respectively, c_Y), and by $\zeta = |Q \setminus (X_t(\Gamma(v)) \cup Y_t(\Gamma(v)))|$ the number of possible common values. By Lemma 5.2, conditional on vertex v being selected in step (1), the probability that the same colour is selected for vertex v in both X_{t+1} and Y_{t+1} is

$$\Pr(c_X = c_Y) = \frac{\zeta}{\max\{\xi, \eta\}}. \tag{5.11}$$

A moment's reflection reveals that the quantities ξ, η and ζ satisfy the following linear inequalities:

$$\xi - \zeta \leq d'(v), \tag{5.12}$$

$$\eta - \zeta \leq d'(v) \tag{5.13}$$

and

$$\zeta \geq q - \Delta - d'(v). \tag{5.14}$$

Thus, starting from (5.11),

$$\Pr(c_X = c_Y) \geq \frac{\zeta}{d'(v) + \zeta} \geq 1 - \frac{d'(v)}{q - \Delta}, \tag{5.15}$$

where the first inequality is from (5.12) and (5.13), and the second from (5.14). Hence,

$$\begin{aligned} \Pr\big(|D_{t+1}| = |D_t| + 1\big) &\leq \frac{1}{n} \sum_{v \in A} \frac{d'(v)}{q - \Delta} \\ &= \frac{m'}{(q - \Delta)n}, \end{aligned} \tag{5.16}$$

where the equality is by equation (5.10).

Now consider the probability that $|D_{t+1}| = |D_t| - 1$. For this event to occur, the vertex v selected in line (1) must lie in D, and the new colours c_X and c_Y selected in step (2) must be equal. Equation (5.11) continues to hold, with ξ, η and ζ defined as before. The analogues of inequalities (5.12)–(5.14) for the case $v \in D$ are

$$\xi - \zeta \leq \Delta - d'(v),$$

$$\eta - \zeta \leq \Delta - d'(v)$$

and

$$\zeta \geq q - 2\Delta + d'(v).$$

By reasoning similar to that leading to (5.15),

$$\Pr(c_X = c_Y) \geq \frac{\zeta}{\Delta - d'(v) + \zeta} \geq \frac{q - 2\Delta}{q - \Delta} + \frac{d'(v)}{q - \Delta},$$

conditional on v being selected in step (1). Hence

$$\begin{aligned}\Pr\bigl(|D_{t+1}| = |D_t| - 1\bigr) &\geq \frac{1}{n}\sum_{v\in D}\left(\frac{q-2\Delta}{q-\Delta} + \frac{d'(v)}{q-\Delta}\right)\\ &= \frac{q-2\Delta}{(q-\Delta)n} \times |D| + \frac{m'}{(q-\Delta)n}. \qquad (5.17)\end{aligned}$$

Define

$$a = \frac{q-2\Delta}{(q-\Delta)n} \quad \text{and} \quad b = b(m') = \frac{m'}{(q-\Delta)n},$$

so that $\Pr\bigl(|D_{t+1}| = |D_t| + 1\bigr) \leq b$ and $\Pr\bigl(|D_{t+1}| = |D_t| - 1\bigr) \geq a\,|D_t| + b$. Provided $a > 0$, i.e., $q > 2\Delta$, the size of the set D_t tends to decrease with t, and hence, intuitively at least, the event $D_t = \emptyset$ should occur with high probability for some $t \leq T$ with T not too large. Since $D_t = \emptyset$ is precisely the event that coalescence has occurred, it only remains to confirm this intuition, and quantify the rate at which D_t converges to the empty set. From equations (5.16) and (5.17),

$$\begin{aligned}\operatorname{Exp}\bigl(|D_{t+1}| \bigm| D_t\bigr) &\leq b(|D_t|+1) + (a|D_t|+b)(|D_t|-1)\\ &\qquad + (1 - a|D_t| - 2b)|D_t|\\ &= (1-a)|D_t|.\end{aligned}$$

Thus $\operatorname{Exp}|D_t| \leq (1-a)^t|D_0| \leq n(1-a)^t$, and, because $|D_t|$ is an non-negative integer random variable, $\Pr(|D_t| \neq 0) \leq n(1-a)^t \leq n\mathrm{e}^{-at}$. Note that $\Pr(D_t \neq \emptyset) \leq \varepsilon$, provided $t \geq a^{-1}\ln(n\varepsilon^{-1})$, establishing the result.

Observe that this result, combined with Proposition 2.1, implies the existence of an FRPAS for q-colourings in graphs of maximum degree Δ, provided $q \geq 2\Delta + 1$. With a little care, the argument can be pushed to $q \geq 2\Delta$, though the bound on mixing time worsens by a factor of about n^2.

The (direct) coupling technique described here has been used in a number of other applications, such as approximately counting independent sets in a low-degree graph (Luby and Vigoda [50]), and estimating the volume of a convex body (Bubley, Dyer and Jerrum [14]).[7] In practice, the versatility of the approach is limited by our ability to design couplings that work well in situations of algorithmic interest. The next section reports on a new technique that promises to extend the effective range of the coupling argument by providing us with a powerful design tool.

[7] The latter application draws inspiration from Lindvall and Rodgers's [49] idea of coupling diffusions by reflection.

6. A new technique: path coupling

The coupling technique described and illustrated in Sections 4.3 and 5.3 is conceptually very simple and appealing. Unfortunately, it may be very difficult or indeed virtually impossible to design couplings appropriate to specific situations of practical interest. The problem, which began to surface even in Section 5.3, is one of engineering: how do we encourage (X_t) and (Y_t) to coalesce, while satisfying the demanding constraints (4.7) and (4.8)? Path coupling is an engineering solution to this problem, invented by Bubley and Dyer [10, 11]. Their idea is to define the coupling only on pairs of "adjacent" states, for which the task of satisfying (4.7) and (4.8) is relatively easy, and then to extend the coupling to arbitrary pairs of states by composition of adjacent couplings along a path. The approach is not entirely distinct from classical coupling, and the Coupling Lemma (Lemma 4.1) still plays a vital role.

We illustrate path coupling in the context of the Markov chain $\mathfrak{M}_{\mathrm{le}}$, of Section 5.2, on linear extensions of a partial order. Our treatment will closely follow that of Bubley and Dyer [12]. For convenience, we work with a slightly modified version of $\mathfrak{M}_{\mathrm{le}}$. The transitions from one linear extension to another are still obtained by pre-composing with a random transition $(p, p+1)$; however, instead of selecting $p \in [n-1]$ uniformly, we select p from a probability distribution f on $[n-1]$ that gives greater weight to values near the centre of the range. It is possible that this refinement actually reduces the mixing time; in any case, it leads to a simplification of the proof. Formally, transition probabilities from state X_t are defined by the following experiment:

(1) Select $p \in [n-1]$ according to the distribution f, and $r \in \{0,1\}$ u.a.r.
(2) If $r = 1$ and $X_t \circ (p, p+1) \in \Omega$, then $X_{t+1} := X_t \circ (p, p+1)$; otherwise, $X_{t+1} := X_t$.

Let us refer to this Markov chain as $\mathfrak{M}_{\mathrm{le}}^f$. Provided the probability distribution f is supported on the whole interval $[n-1]$, the Markov chain $\mathfrak{M}_{\mathrm{le}}^f$ is irreducible and aperiodic. It is easy to verify, for example using Lemma 3.1, that the stationary distribution of $\mathfrak{M}_{\mathrm{le}}^f$ is uniform. As in Section 5.2, the explicit loop probability of $\frac{1}{2}$ is introduced mainly for convenience in the proof. However, some such mechanism for destroying periodicity is necessary in any case if we wish to treat the empty partial order consistently.

To apply path coupling, we need first to decide on an adjacency structure for the state space Ω. In this instance we decree that two states g and g' (linear extensions of $\prec$) are adjacent iff $g' = g \circ (i, j)$ for some transposition (i, j) with $0 \le i < j \le n-1$; in this case, the distance $d(g, g')$ from g to g' is defined to be $j - i$. Note that the notions of adjacency and distance are symmetric with respect to interchanging g and g', so we can regard this imposed adjacency structure as a weighted, undirected graph on vertex set Ω; let us refer to this structure as the adjacency graph. It is easily verified that the shortest path in the adjacency graph between two adjacent states is the direct one using a single edge. Thus d may be extended to a metric on Ω by defining $d(g, h)$, for arbitrary states g and h, to be the length of a shortest path from g to h in the adjacency graph.

Next we define the coupling. We need to do this just for adjacent states, as the extension of the coupling via shortest paths to arbitrary pairs of states will be automatic. Suppose the current pair of states is (X_t, Y_t) and that $Y_t = X_t \circ (i, j)$ for some transposition (i, j) with $0 \le i < j \le n-1$; then the transition to (X_{t+1}, Y_{t+1}) is defined by the following experiment:

(1) Select $p \in [n-1]$ according to the distribution f, and $r_X \in \{0,1\}$ u.a.r. If $j - i = 1$ and $p = i$, set $r_Y := 1 - r_X$; otherwise, set $r_Y := r_X$.
(2) If $r_X = 1$ and $X_t \circ (p, p+1) \in \Omega$ then set $X_{t+1} := X_t \circ (p, p+1)$; otherwise, set $X_{t+1} := X_t$.
(3) If $r_Y = 1$ and $Y_t \circ (p, p+1) \in \Omega$ then set $Y_{t+1} := Y_t \circ (p, p+1)$; otherwise, set $Y_{t+1} := Y_t$.

We need to show:

Lemma 6.1. *For adjacent states X_t and Y_t,*

$$\operatorname{Exp}\left(d(X_{t+1}, Y_{t+1}) \mid X_t, Y_t\right) \le \varrho\, d(X_t, Y_t), \tag{6.1}$$

where $\varrho < 1$ is a constant depending on f. For a suitable choice for f, one has $\varrho = 1 - \alpha$, where $\alpha = 6/(n^3 - n)$.

Before proceeding with the proof of Lemma 6.1, let us pause to consider why it is sufficient to establish (6.1) just for adjacent states.

Lemma 6.2. *Suppose a coupling (X_t, Y_t) has been defined for $\mathfrak{M}_{\mathrm{le}}^f$ on adjacent pairs of states, and suppose that the coupling satisfies the contraction condition (6.1) on adjacent pairs. Then the coupling can be extended to all pairs of states in such a way that (6.1) holds unconditionally.*

Proof (sketch). For notational convenience set $X := X_t$ and $Y := Y_t$, where $X_t, Y_t \in \Omega$ are now arbitrary. Denote by $P(\cdot, \cdot)$ the transition probabilities of $\mathfrak{M}_{\mathrm{le}}^f$. Let $X = Z_0, Z_1, \dots, Z_l = Y$ be a shortest path from X to Y in the adjacency graph. (Assume a deterministic choice rule for resolving ties.) First select $X' = Z_0' \in \Omega$ according to the probability distribution $P(X, \cdot)$. Now select Z_1' according to the distribution induced by the pairwise coupling of the adjacent states Z_0 and Z_1, conditioned on the choice of Z_1'; then select Z_2' using the pairwise coupling of Z_1 and Z_2, and so on, ending with $Z_l' = Y'$. Let $X_{t+1} := X'$ and $Y_{t+1} := Y'$. It is routine to verify, by induction on path length l, that Y_{t+1} has been selected according to the (correct) distribution $P(Y_t, \cdot)$. Moreover, by linearity of expectation and (6.1),

$$\operatorname{Exp}\left(d(X_{t+1}, Y_{t+1}) \mid X_t, Y_t\right) \le \sum_{i=0}^{l-1} \operatorname{Exp}\left(d(Z_i', Z_{i+1}') \mid Z_i, Z_{i+1}\right)$$

$$\begin{aligned} &\leq \varrho \sum_{i=0}^{l-1} d(Z_i, Z_{i+1}) \\ &= \varrho\, d(X_t, Y_t). \end{aligned}$$

Proof (of Lemma 6.1). If $p \notin \{i-1, i, j-1, j\}$ then the tests made in steps (2) and (3) either both succeed or both fail. Thus $Y_{t+1} = X_{t+1} \circ (i,j)$ and $d(X_{t+1}, Y_{t+1}) = j - i = d(X_t, Y_t)$. Summarising:

$$d(X_{t+1}, Y_{t+1}) = d(X_t, Y_t), \quad \text{if } p \notin \{i-1, i, j-1, j\}. \tag{6.2}$$

Next suppose $p = i-1$ or $p = j$. These cases are symmetrical, so we consider only the former. With probability at least $\frac{1}{2}$, the tests made in steps (2) and (3) both fail, since $\Pr(r_X = r_Y = 0) = \frac{1}{2}$. If this happens, clearly, $d(X_{t+1}, Y_{t+1}) = j - i = d(X_t, Y_t)$. Otherwise, with probability at most $\frac{1}{2}$, one or other test succeeds. If they both succeed, then

$$\begin{aligned} Y_{t+1} &= Y_t \circ (i-1, i) \\ &= X_t \circ (i,j) \circ (i-1, i) \\ &= X_{t+1} \circ (i-1, i) \circ (i,j) \circ (i-1,i) \\ &= X_{t+1} \circ (i-1, j), \end{aligned}$$

and $d(X_{t+1}, Y_{t+1}) = j - i + 1 = d(X_t, Y_t) + 1$; if only one (say the one in step 2) succeeds, then $Y_{t+1} = Y_t = X_t \circ (i,j) = X_{t+1} \circ (i-1, i) \circ (i,j)$, and $d(X_{t+1}, Y_{t+1}) \leq j - i + 1 = d(X_t, Y_t) + 1$. Summarising:

$$\operatorname{Exp}\big(d(X_{t+1}, Y_{t+1}) \bigm| X_t, Y_t, p = i-1 \ \vee p = j\big) \leq d(X_t, Y_t) + \frac{1}{2}. \tag{6.3}$$

Finally suppose $p = i$ or $p = j-1$. Again, by symmetry, we need only consider the former. There are two subcases, depending on the value of $j - i$. The easier subcase is $j - i = 1$. If $r_X = 1$ then $r_Y = 0$ and

$$X_{t+1} = X_t \circ (i, i+1) = Y_t \circ (i, i+1) \circ (i, i+1) = Y_t = Y_{t+1},$$

with a similar conclusion when $r_X = 0$. Thus $d(X_{t+1}, Y_{t+1}) = 0 = d(X_t, Y_t) - 1$. The slightly harder subcase is the complementary $j - i \geq 2$. The crucial observation is that $X_t \circ (i, i+1), Y_t \circ (i, i+1) \in \Omega$ and hence the tests in steps (2) and (3) either both succeed or both fail, depending only on the value of $r_X = r_Y$. To see this, observe that

$$X_t(i) \not\prec X_t(i+1) = Y_t(i+1) \not\prec Y_t(j) = X_t(i),$$

from which we may read off the fact that $X_t(i)$ and $X_t(i+1)$ are incomparable in $\prec$. The same argument applies equally to $Y_t(i)$ and $Y_t(i+1)$. If $r_X = 0$ there is no change in state; otherwise, if $r_X = 1$,

$$\begin{aligned} X_{t+1} &= X_t \circ (i, i+1) \\ &= Y_t \circ (i, j) \circ (i, i+1) \\ &= Y_{t+1} \circ (i, i+1) \circ (i, j) \circ (i, i+1) \\ &= Y_{t+1} \circ (i+1, j), \end{aligned}$$

and $d(X_{t+1}, Y_{t+1}) = j - i - 1 = d(X_t, Y_t) - 1$. Summarising both the $j - i = 1$ and $j - i \geq 2$ subcases:

$$\operatorname{Exp}\big(d(X_{t+1}, Y_{t+1}) \bigm| X_t, Y_t, p = i \vee p = j-1\big) \leq e(X_t, Y_t), \tag{6.4}$$

where

$$e(X_t, Y_t) = \begin{cases} 0, & \text{if } d(X_t, Y_t) = 1; \\ d(X_t, Y_t) - \frac{1}{2}, & \text{otherwise.} \end{cases}$$

Note that, in the case $j - i = 1$, inequality (6.4) covers just one value of p, namely $p = i = j - 1$, instead of two; however, this effect is exactly counterbalanced by an expected reduction in distance of 1 instead of just $\frac{1}{2}$. Combining (6.2)–(6.4) we obtain

$$\begin{aligned} &\operatorname{Exp}\big(d(X_{t+1}, Y_{t+1}) \bigm| X_t, Y_t\big) \\ &\qquad \leq d(X_t, Y_t) - \frac{-f(i-1) + f(i) + f(j-1) - f(j)}{2}. \end{aligned}$$

Specialising the probability distribution $f(\cdot)$ to be $f(i) := \alpha(i+1)(n-i-1)$—where $\alpha := 6/(n^3 - n)$ is the appropriate normalising constant—we have, by direct calculation, $-f(i-1) + f(i) + f(j-1) - f(j) = 2\alpha(j - i)$. Since $d(X_t, Y_t) = j - i$, we obtain (6.1) with $\varrho = 1 - \alpha$.

From Lemmas 6.1 and 6.2 it is now a short step to:

Proposition 6.1. *The mixing time of the Markov chain* $\mathfrak{M}_{\mathrm{le}}^f$ *is bounded by*

$$\tau(\varepsilon) \leq (n^3 - n)(2 \ln n + \ln \varepsilon^{-1})/6.$$

Proof. By iteration, $\operatorname{Exp}\big(d(X_t, Y_t) \bigm| X_0, Y_0\big) \leq \varrho^t d(X_0, Y_0)$. For any pair of linear extensions g and h, there is a path in the adjacency graph using only *adjacent* transpositions (i.e., length one edges) that swaps each incomparable pair at most once. Thus $d(X_0, Y_0) \leq \binom{n}{2} \leq n^2$, and

$$\Pr(X_t \neq Y_t) \leq \operatorname{Exp}(d(X_t, Y_t)) \leq (1 - \alpha)^t n^2.$$

The latter quantity is less than ε, provided $t \geq (n^3 - n)(2 \ln n + \ln \varepsilon^{-1})/6$. The result follows directly from Lemma 4.1.

David Wilson has recently derived a similar $O(n^3 \log n)$ bound on mixing time when f is uniform, i.e, when the transposition $(p, p+1)$ is selected u.a.r.

New applications of path coupling are regularly being discovered. Bubley, Dyer and Greenhill [13] have presented an FPRAS for q-colourings of a low degree graph that extends the range of applicability of the one described earlier. They were able, for example, to approximate in polynomial time the number of 5-colourings of a graph of maximum degree 3, thus "beating the 2Δ bound" that appeared to exist following the result described in Section 5.3. It is fair to say that this improvement would not have been possible without the aid of path coupling. Dyer and Greenhill have also considered independent sets in a low degree graph [23], and obtained a result similar to, but apparently incomparable with, that of Luby and Vigoda [50]. One further example must suffice: Cooper and Frieze [15] have applied path coupling to analyse the "Swendsen-Wang process," which is commonly used to sample configurations of the "random cluster" or ferromagnetic Potts model in statistical physics.

7. Exact sampling by coupling from the past (CFTP)

The previous section perhaps struck an overly optimistic note. In the majority of cases, we do not have good a priori bounds, using any of the techniques in the previous sections, on the mixing time of the Markov chains used in actual MCMC applications. When analytical bounds are weak or non-existent, we can sometimes use coupling as an *algorithmic* (as opposed to *proof*) technique. Propp and Wilson's remarkable contribution is to demonstrate that in certain circumstances, "algorithmic coupling" may be used to obtain samples from the *exact* stationary distribution, rather than just a t-step approximation. This section is based on Propp and Wilson's seminal article on exact sampling [53], and a paper of Kendall's that describes an extension to their technique [45].

Suppose $\mathfrak{M}$ is an ergodic (irreducible, aperiodic) Markov chain on finite state space Ω and with transition probabilities $P : \Omega \times \Omega \to [0,1]$. (The finiteness assumption is for ease of presentation only, and plays no crucial role in what follows.) Suppose $\mathcal{F}$ is a probability distribution on functions $f : \Omega \to \Omega$ that is consistent with P is the sense that

$$\Pr_{\mathcal{F}}(f(x) = y) = P(x,y), \quad \text{for all } x, y \in \Omega. \tag{7.1}$$

A special example of this situation arises when $\mathcal{F}$ is constructed as a product distribution from P. Thus, to sample $f \in \mathcal{F}$: (i) sample, independently for each $x \in \Omega$, a state y_x from the distribution $P(x,\cdot)$, and then (ii) let $f : \Omega \to \Omega$ be the function mapping x to y_x for all $x \in \Omega$. But just as with the vanilla coupling in Section 4.3, we are in practice interested in distributions $\mathcal{F}$ that strongly couple evaluations of f at different states (elements in the domain).

If $s < t$, and $f_s, \dots, f_{t-1} : \Omega \to \Omega$ is a indexed sequence of functions (usually the f_i will be sampled independently from $\mathcal{F}$), we denote by $F_s^t : \Omega \to \Omega$ the iterated function composition

$$F_s^t = f_{t-1} \circ f_{t-2} \circ \cdots \circ f_{s+1} \circ f_s. \tag{7.2}$$

We may perform a rather perverse t-step simulation of $\mathfrak{M}$ from some initial state $x_0 \in \Omega$ by the following procedure: (i) select $f_0, \dots, f_{t-1}$ independently from distribution $\mathcal{F}$, (ii) compute the composition $F_0^t = f_{t-1} \circ f_{t-2} \circ \cdots \circ f_1 \circ f_0$ as in (7.2), and (iii) return $F_0^t(x_0)$ as the required sample from the t-step distribution. Of course, this would be a very inefficient way of simulating $\mathfrak{M}$, requiring about $|\Omega|$ times the work of a direct simulation of a single trajectory. However, this view of proceedings will be convenient to bear in mind in what follows.

As hinted at earlier, for fixed transition probabilities $P(\cdot\,,\cdot)$, there is considerable flexibility in the choice of the distribution $\mathcal{F}$, allowing us to encode uniform couplings over the entire state space. The Coupling Lemma—at least an important special case of it—can be stated in this setting. Suppose $f_0, \dots, f_{t-1}$ are sampled independently from $\mathcal{F}$, and let F_0^t be as before. If there exists a function $t : (0,1] \to \mathbb{N}$ such that

$$\Pr\left(F_0^{t(\varepsilon)}(\cdot) \text{ is not a constant function}\right) \le \varepsilon,$$

then the mixing time $\tau(\varepsilon)$ of $\mathfrak{M}$ is bounded by $t(\varepsilon)$. In principle, this observation permits us to estimate the mixing time of $\mathfrak{M}$ empirically, by observing the coalescence time of the coupling defined by $\mathcal{F}$. We could then obtain samples from an *approximation* to the stationary distribution of $\mathfrak{M}$ by simulating $\mathfrak{M}$ for a number of steps comparable with the empirically observed mixing time. In practice, as we have already observed, the explicit evaluation of F_0^t would be computationally infeasible.

The first of the two ideas that underlie Propp and Wilson's proposal is completely original and surprising: by working with F_{-t}^0 in place of F_0^t, i.e., by "coupling from the past," (CFTP) it is possible to obtain samples from the *exact* stationary distribution.

Theorem 7.1. *Suppose that $f_{-1}, f_{-2}, \dots$ is a sequence of independent samples from $\mathcal{F}$. Let the stopping time T be defined as the smallest number t for which $F_{-t}^0(\cdot)$ is a constant function, and assume that* $\operatorname{Exp} T < \infty$. *Denote by $\widehat{F}_{-\infty}^0$ the unique value of F_{-T}^0 (which is defined with probability 1). Then $\widehat{F}_{-\infty}^0$ is distributed according to the stationary distribution of $\mathfrak{M}$.*

Note that the constant function F_{-t}^0 is the same constant function for all sufficiently large t, specifically for all $t \ge T$. Thus, coupling from time $-T$ is equivalent to "coupling from time $-\infty$," which is the rationale behind both the choice of notation $\widehat{F}_{-\infty}^0$ and the CFTP method itself.

Proof (of Theorem 7.1). Let π_0 be the distribution of the random variable $\widehat{F}_{-\infty}^0$. Take one further independent sample f_0 from $\mathcal{F}$, and let $T' \le T$ be the smallest number such that $F_{-T'}^1$ is a constant function. Let $\widehat{F}_{-\infty}^1$ denote the unique value of $F_{-T'}^1$, and let π_1 denote the distribution of the random variable $\widehat{F}_{-\infty}^1$. By translational symmetry, $\pi_0 = \pi_1$. But $\widehat{F}_{-\infty}^1 = f_0(\widehat{F}_{-\infty}^0)$, which implies that $\pi_0 = \pi_1$ is a stationary distribution for $\mathfrak{M}$. ($\widehat{F}^1_{-\infty}$ is obtained from $\widehat{F}^0_{-\infty}$ by effecting a single transition of $\mathfrak{M}$.) But $\mathfrak{M}$ is ergodic.

Note that we did not really need to assume that $\mathfrak{M}$ is ergodic, since the condition $\operatorname{Exp} T < \infty$ implies the existence of a stationary distribution π_0—we *constructed* it!—and it is easily verified that this stationary distribution must be unique.

The second idea underlying Propp and Wilson's proposal—independently discovered by others, e.g., Johnson [39]—is that in certain circumstances, specifically when the coupling $\mathcal{F}$ is "monotone," it is possible to evaluate F^0_{-T} without explicitly computing the function composition $f_1 \circ f_2 \circ \cdots \circ f_{-T+1} \circ f_{-T}$. Suppose that the state space Ω is partially ordered by $\preceq$, with a unique maximal element $\top$ and a unique minimum element $\perp$. We say that the coupling $\mathcal{F}$ is *monotone* if, for every $x, y \in \Omega$ and $f : \Omega \to \Omega$ in the support of $\mathcal{F}$, the condition $x \preceq y$ entails $f(x) \preceq f(y)$. When $\mathcal{F}$ is monotone, the test for F^0_{-t} being a constant function is equivalent to the test $F^0_{-t}(\perp) = F^0_{-t}(\top)$. Moreover, if equality holds between $F^0_{-t}(\perp)$ and $F^0_{-t}(\top)$ then their common value is just $\widehat{F}^0_{-\infty}$. Roughly speaking, rather than tracking $|\Omega|$ trajectories of $\mathfrak{M}$, in the monotone case we just need to track two, namely the ones starting at $\perp$ and $\top$.

```
T ← 1;
repeat
    lower ← ⊥;
    upper ← ⊤;
    for t ← −T to −1:
        lower ← f_t(lower);
        upper ← f_t(upper);
    T ← 2T
until lower = upper;
return lower
```

Fig. 7.1. Coupling from the past: the monotone case

Note that to compute $\widehat{F}^0_{-\infty}$ it is not necessary to know T exactly, only an upper bound. Rather than iteratively computing F^0_{-t} for $t = 0, 1, 2, 3, 4, \ldots$, until convergence, it is much more efficient to iterate according to the doubling scheme $t = 1, 2, 4, 8, 16, \ldots$. A general procedure for (monotone) CFTP, incorporating this algorithmic refinement, is presented as Figure 7.1.

7.1 A monotone example: the random cluster model

The random cluster model arises in statistical physics as a dual (in some sense) of the ferromagnetic Potts model. An instance of the random cluster model is defined by an undirected graph $G = (V, E)$, and real numbers $0 \leq$

$p \leq 1$ and $q \geq 0$. A configuration (state) of the model is a subset $X \subseteq E$; denote by $\Omega = 2^E$ the set of all configurations. Each configuration X is assigned a weight $w(X) := p^{|X|}(1-p)^{m-|X|}q^{c(X)}$, where $m = |E|$ and $c(X)$ is the number of connected components of the graph $H = (V, X)$. Let $Z := \sum_{X \subseteq E} w(X)$. Then the random cluster model specifies a probability distribution (Gibbs distribution) $\pi : \Omega \to [0,1]$ on the set of configurations, where

$$\pi(X) := w(X)/Z, \tag{7.3}$$

for all $X \subseteq E$. In the special case $q = 1$ and $G = K_n$ (the complete graph on n vertices), the random cluster model reduces to the standard random graph model $\mathcal{G}_{n,p}$ [7]. When q is a positive integer, the random cluster model is equivalent (in a strong sense) to the ferromagnetic q-state Potts model, as was first observed by Fortuin and Kasteleyn [29]. For more on this, see, e.g., Edwards and Sokal [26].

Suppose we wish to obtain random samples from the Gibbs distribution with the aim, for example, of estimating the average size of a "cluster" (connected component of the graph (V, X)). We construct a Markov chain $\mathfrak{M}_{\mathrm{rc}} = \mathfrak{M}_{\mathrm{rc}}(G, p, q)$ on the set of configurations Ω by defining transition probabilities according to the following trial.

(1) Suppose the current state is $X \subseteq E$. Select $e \in E$, u.a.r., and let

$$\theta_{X,e} := \frac{w(X+e)}{w(X+e) + w(X-e)}.$$

(2) Select $\alpha \in [0,1)$, u.a.r. If $\alpha < \theta_{X,e}$, set $X' := X + e$; otherwise, set $X' := X - e$. The next state is X'.

It is easily to verify that $\mathfrak{M}_{\mathrm{rc}}$ is ergodic and, using Lemma 3.1, that its stationary distribution is the Gibbs distribution (7.3).

The threshold $\theta_{X,e}$ can be interpreted as the probability, in the Gibbs distribution, that edge e is present in a random configuration X', conditioned on the event $X' - e = X - e$, i.e., that X' and X agree except perhaps on e. The transition probabilities defined above are another example application of the heat-bath dynamics. Note that $\theta_{X,e}$ is easy to compute from the explicit expression

$$\theta_{X,e} = \begin{cases} p, & \text{if } c(X+e) = c(X-e); \\ p/(p+(1-p)q), & \text{otherwise.} \end{cases} \tag{7.4}$$

The trial just described is easily extended to a (uniform) coupling, simply by ruling that the same choice of random edge e and number α are used independently of X. Specifically, the probability distribution $\mathcal{F}$ is defined by the following trial:

(1) Select $e \in E$ and $\alpha \in [0,1)$ u.a.r.

(2) Define the function $f : \Omega \to \Omega$ by

$$f(X) = \begin{cases} X + e, & \text{if } \alpha < \theta_{X,e}; \\ X - e, & \text{otherwise.} \end{cases}$$

The function f is a random sample from $\mathcal{F}$.

This coupling is monotone with respect to the inclusion ordering on configurations (states), provided $q \geq 1$; i.e., for any two states $X, Y \in \Omega$ with $X \subseteq Y$, and any function f in the support of $\mathcal{F}$, it is the case that $f(X) \subseteq f(Y)$. To see this, simply observe that for any such pair of states, $\theta_{X,e} \leq \theta_{Y,e}$, for all $e \in E$.

For any integer $q \geq 3$, Gore and Jerrum [30] have shown that the mixing time of $\mathfrak{M}_{\mathrm{rc}}(G, p, q)$ may be exponential in n, the number of vertices in the graph G. The important special case $q = 2$, equivalent to the celebrated (ferromagnetic) Ising model in statistical physics, is completely open: it may be the case that the mixing time of $\mathfrak{M}_{\mathrm{rc}}(G, p, q)$ is bounded by $\mathrm{poly}(n, \varepsilon^{-1})$ uniformly over G, but there is little evidence either way. Nevertheless, the point about coupling from the past is exactly that we don't need a priori analytical bounds on the mixing time: we can just implement the coupling suggested above and proceed empirically.

Figure 7.2 illustrates the result of one such experiment. Here we see Propp-Wilson CFTP applied to the random cluster model on a 10×10 square grid, at $q = 2$ and $p = \sqrt{2}/(1+\sqrt{2}\,)$. (The chosen values for p and q correspond to the Ising model at the critical temperature for the infinite 2-dimensional square lattice.) To save space, not all the doubling steps demanded by the procedure of Figure 7.1 are illustrated. Salient features to note are that $F^0_{-t}(\bot)$ (respectively, $F^0_{-t}(\top)$) is monotonically increasing (respectively decreasing) with t, and that $F^0_{-t}(\bot) \leq \widehat{F}^0_{-\infty} \leq F^0_{-t}(\top)$ for all $t \geq 0$. As t increases, we learn more about the identity of $\widehat{F}^0_{-\infty}$. Convergence in this case is surprisingly rapid when one considers that the expected number of steps before all 180 edges in the grid have been selected is about 1039 (c.f. the "coupon collector" problem). Note that after 1024 steps the lower and upper bounds differ in just five edges, and that convergence proper occurs in at most twice that many steps.

7.2 A non-monotone example: random forests

When $q < 1$, the coupling just devised for the random cluster model ceases to be monotone; worse still, no monotone coupling exists. (The existence of a monotone coupling when $q \geq 1$ is connected to the "FKG inequality," which fails when $q < 1$.) Fortunately, Kendall [45] has shown how to extend the Propp-Wilson framework to encompass many non-monotone situations. In the original Propp-Wilson proposal, the two extreme trajectories of a Markov chain $\mathfrak{M}$—starting from the extreme states $\bot$ and $\top$—are known to bound all

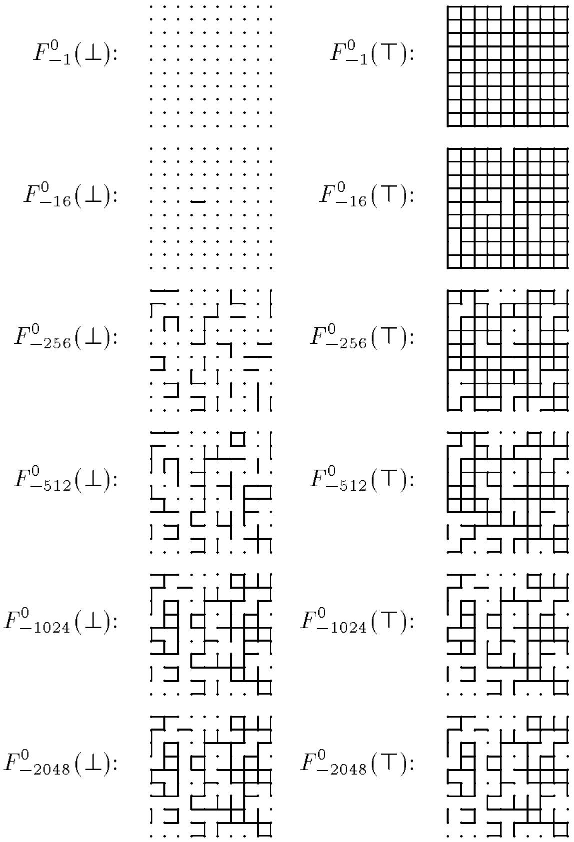

Fig. 7.2. A sample run of coupling from the past

the others, so we can be certain that once those two extreme trajectories have converged then so have all the others. Kendall's idea is that the two bounding trajectories do not have to be honest simulations of $\mathfrak{M}$; it is enough that the upper one remains above all of the actual trajectories (in the specified partial order), while the lower one remains below.

In general, the situation is as follows. Recall that Ω is endowed with a partial order $\preceq$. An *interval* I of Ω is defined by two *endpoints* $l, u \in \Omega$ with $l \preceq u$, and consists of all points lying between l and u, thus: $I = \{x \in \Omega : l \preceq x \preceq u\}$. Denote by $\mathcal{I} = \mathcal{I}(\Omega)$ the set of all intervals of Ω. The probability distribution $\mathcal{F}$ is extended to a distribution $\mathcal{F}'$ on pairs (f, g), where $f : \Omega \to \Omega$ and $g : \mathcal{I} \to \mathcal{I}$. As before, we stipulate that the component f satisfies (7.1), which roughly says that the coupling defined by $\mathcal{F}'$ has the correct marginals. The condition that replaces monotonicity is

$$x \in I \text{ entails } f(x) \in g(I), \text{ for all } I \in \mathcal{I} \text{ and } (f, g) \in \operatorname{supp} \mathcal{F}'. \tag{7.5}$$

By analogy with (7.2), define

$$G_s^t = g_{t-1} \circ g_{t-2} \circ \cdots \circ g_{s+1} \circ g_s, \tag{7.6}$$

where $(f_s, g_s), \ldots, (f_{t-1}, g_{t-1})$ are random samples from $\mathcal{F}'$. It follows from condition (7.5) that $G_{-t}^0(\bot, \top) = (y_0, y_0)$ implies that $F_{-t}^0(\,\cdot\,)$ is the constant function y_0, which in turn implies $\widehat{F}_{-\infty}^0 = y_0$. So we have the following extension to Theorem 7.1:

Theorem 7.2. *Suppose that $(f_{-1}, g_{-1}), (f_{-2}, g_{-2}), \ldots$ is a sequence of independent samples from $\mathcal{F}'$. Let the stopping time T be defined as the smallest number t for which $G_{-t}^0(\bot, \top) = (y_0, y_0)$, for some $y_0 \in \Omega$, and assume that* $\operatorname{Exp} T < \infty$. *Then y_0 (which is defined with probability 1), is distributed according to the stationary distribution of $\mathfrak{M}$.*

Note that the samples $f_{-1}, f_{-2}, \ldots$, are a conceptual convenience only, having no algorithmic significance. The algorithm for the Kendall variant of CFTP is a simple modification of the monotone one presented in Figure 7.1: simply replace the lines

$lower \leftarrow f_t(lower)$;
$upper \leftarrow f_t(upper)$;

by

$(lower, upper) \leftarrow g_t(lower, upper)$;

As an illustrative example, let us consider how CFTP might be applied to the random cluster model with $0 \leq q < 1$. The probability distribution $\mathcal{F}'$ is specified by the following trial:

(1) Select $e \in E$ and $\alpha \in [0, 1)$ u.a.r.
(2) Define the function $f : \Omega \to \Omega$ by

$$f(X) = \begin{cases} X + e, & \text{if } \alpha < \theta_{X,e}; \\ X - e, & \text{otherwise}; \end{cases}$$

where $\theta_{X,e}$ is defined as in (7.4).
(3) Define the function $g : \mathcal{I} \to \mathcal{I}$ by

$$g(L, U) = \begin{cases} (L + e, U + e), & \text{if } \alpha < \theta_{U,e}; \\ (L - e, U + e), & \text{if } \theta_{U,e} \leq \alpha < \theta_{L,e}; \\ (L - e, U - e), & \text{if } \alpha \geq \theta_{L,e}. \end{cases}$$

(4) The pair (f, g) is a random sample from $\mathcal{F}'$.

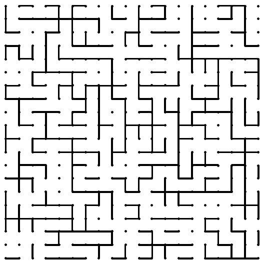

Fig. 7.3. Exact sampling: a random forest in a 20×20 square grid

Informally, the function g updates its first or "lower" argument using the threshold $\theta_{U,e}$ appropriate for its second or "upper" argument, and vice versa. This artifice ensures that g preserves intervals—that is to say, $L \subseteq U$ and $(L', U') = g(L, U)$ entail $L' \subseteq U'$—even though f itself is not monotone. Indeed it is routine to verify that condition (7.5) holds with $\mathcal{F}'$ defined as above.

The picture to have in mind is that the iterates F_0^t of f define coupled sample paths of $\mathfrak{M}_{\mathrm{rc}}$ starting at all possible initial states. When $q \geq 1$ (the monotone case) these paths behave in an orderly fashion, and their joint evolution is summarised by the lower and uppermost sample paths $F_0^t(\bot)$ and $F_0^t(\top)$. When $q < 1$ the sample paths are unruly, crossing and recrossing each other; nevertheless, the iterates $G_0^t(\bot, \top)$ continue to provide conservative lower and upper bounds on their joint evolution.

The set of forests (acyclic, spanning, not necessarily connected subgraphs) of a graph G endowed with the uniform distribution can be regarded as the set of configurations of the limit of the random cluster model as $p, q \to 0^+$ with $p/q = 1$. Explicitly, the threshold $\theta_{X,e}$ in this limit is

$$\theta_{X,e} = \begin{cases} 0, & \text{if } c(X+e) = c(X-e); \\ \frac{1}{2}, & \text{otherwise.} \end{cases}$$

Plugging this threshold into the non-monotone coupling for $\mathfrak{M}_{\mathrm{rc}}(G, p, q)$ with $q < 1$, we obtain (in principle) an exact sampler for forests in a graph G. As an experiment, ten runs of this sampler were conducted, with G being the 20×20 square grid. Figure 7.3 illustrates the end result of a typical run. All ten runs terminated within 2^{20} steps (about 12 minutes on a Sun Ultra-SPARC E150), with an average run time of about 7 minutes. This seems to be the limit of the method; the run time degrades rapidly beyond the 20×20 grid, and the 30×30 grid appears to be inaccessible. Nevertheless, it is perhaps surprising that the apparently very conservative lower and upper bounds provided by $G_{-t}^0(\bot, \top)$ should converge in any realistic time bound. It certainly seems worth experimenting further with this approach. See Häggström and Nelander [31] for some more extensive experiments with non-monotone CFTP.

7.3 Further applications

Exact sampling by CFTP and other methods is a thriving research topic, and only a small sample of the burgeoning literature will be mentioned here. Refer to Wilson's online bibliography [59] for a much wider selection. Sampling from Markov random fields was covered (in the monotone case) in Propp and Wilson's original article [53], and (more generally) by Häggström and Nelander [31]. A further twist was introduced by Kendall [45] in applying CFTP to a situation—area interaction point processes—where there is no natural "top state" $\top$.

In statistical physics, one is concerned with infinite Markov random fields, the Ising model on the infinite 2-dimensional square lattice being a prime example. In a remarkable development, van den Berg and Steif [6] point out that is possible in some cases to sample exactly from infinite random fields, even though the configurations are unbounded in extent. The sense in which infinite configurations may be "sampled" is the following: given an positive integer N, the sampler produces, with probability 1, a configuration on the $[-N,N]\times[-N,N]$ grid which is a $(2N+1)\times(2N+1)$ "window" onto a perfectly sampled infinite configuration. The rough idea is that, with probability 1, the spin (= state = colour) at a given lattice site (= vertex) at time 0 can be computed by coupling from a point in time only finitely many steps before and within a region of the lattice stretching only finitely far from the site in question. To get a picture of this, think of "light cones" of relativistic physics, which if bounded temporally must be bounded spatially too. See also Kendall [46].

CFTP à la Propp and Wilson requires a simultaneous coupling on all states Ω—encapsulated in the probability distribution $\mathcal{F}$—rather than the more familiar and less demanding pairwise coupling. Fill's version of exact sampling [28] requires only pairwise coupling, and deals with the (at least philosophically significant) problem of bias induced by "user impatience." Since the running time of the Propp-Wilson sampler is unbounded, there is a danger that an impatient user will abort a run, leading to a biased sample. Fill's proposal has the property that if the user decides to abort a run after some number of steps have elapsed, the samples obtained are not biased.

8. Key open problems

There are many unresolved questions in the area of rapid mixing and approximate counting. A few of the most pressing are collected together in this section.

8.1 Matroid bases

Perhaps the major open problem in this area, and one that would be very rich in terms of consequences, is to determine useful bounds on the mixing time of the *basis-exchange* Markov chain for a general matroid. (A matroid is an algebraic structure that provides an abstract treatment of the concept of linear independence.) The states of this Markov chain are the bases (maximum independent sets) of a given matroid, and a transition is available from base B to base B' if the symmetric difference of B and B' consists of precisely two elements of the ground set. All transition probabilities are equal, so the chain is ergodic and reversible with uniform stationary distribution.

A concrete example is provided by the *graphic matroid* associated with an undirected graph G. In this case, the bases are spanning trees of G, and a transition from a given tree T is effected by adding a single edge (selected u.a.r.) to T, thus creating a cycle, and then breaking the cycle by deleting one of its edges (selected u.a.r.). The basis-exchange Markov chain is known to be rapidly mixing for graphic matroids, and, somewhat more generally, for matroids satisfying a certain "balance condition" (see Feder and Mihail [27]). A proof of rapid mixing in the general case would imply the existence of an FPRAS for a number of important problems in combinatorial enumeration, all of which are #P-complete, including counting connected spanning subgraphs of a graph (network reliability), forests of given size in a graph, and independent subsets of vectors in a set of n-vectors over GF(2).

8.2 Permanent of a 0,1 matrix

Is there an FPRAS for the permanent of a general 0,1 matrix? Equivalently, is there an FPRAS for the number of perfect matchings in a bipartite graph? Note that this problem is not phrased as a question about the mixing time of a specific Markov chain, and certainly the chain $\mathfrak{M}_{\mathrm{match}}$ described in Section 5.1 is not directly applicable. To have a good chance of observing *perfect* matchings (or "dimer covers") the parameter λ must be of order m_{n-1}/m_n; however, it is possible to construct graphs where this ratio is exponential in n. Nevertheless, the Markov chain Monte Carlo method seems to offer the best hope for a positive resolution of this question. Essentially, the issue is whether the Markov chain $\mathfrak{M}_{\mathrm{match}}$ can be suitably adapted to provide a general solution, or perhaps used as a "black box" following some ingenious preprocessing of the input matrix. (This latter idea has been used in a weaker way by Jerrum and Vazirani [38] to obtain a randomised approximation scheme for the general 0,1 permanent whose running time, while still not polynomial, is asymptotically significantly faster than that of more naive methods.)

8.3 Contingency tables

Consider the following task: given $m+n$ positive integers $r_1, \ldots, r_m$ and $c_1, \ldots, c_n$, sample, u.a.r., from the set of $m \times n$ non-negative integer matrices ("contingency tables") with row-sums $r_1, \ldots, r_m$ and column-sums $c_1, \ldots, c_n$. This problem arises in the interpretation of the results of certain kinds of statistical experiment; see, for example, Diaconis and Efron [18].

An elegant direct approach to sampling contingency tables has been proposed by Diaconis. Consider the Markov chain $\mathfrak{M}_{\mathrm{ct}}$, whose state space is the set of all matrices with specified row and column sums, and whose transition probabilities are defined as follows. Let the current state (matrix) be $A = (a_{ij})$. Select a pair of rows (i, i') with $i \neq i'$, and a pair of columns (j, j') with $j \neq j'$, both u.a.r. Form a new matrix A' from A by incrementing by one the array elements $a_{ij}, a_{i'j'}$, and decrementing by one the elements $a_{ij'}, a_{i'j}$. Note that A' has the same row- and column-sums as A. If A' is non-negative then we accept it as the next state; otherwise the chain remains at state A. It is easy to verify that $\mathfrak{M}_{\mathrm{ct}}$ is ergodic and reversible with uniform stationary distribution. Moreover, it appears to work well in practice as a uniform sampling procedure for contingency tables. However, its mixing time is not known to be bounded by any polynomial in the size of the input. (We assume that the row- and column-sums are expressed in unary notation when defining the input size, otherwise even the direct path between two states may be exponentially long.) Dyer, Kannan and Mount [25] have a partial result.

To deal with tables with large entries, a natural idea is to use a kind of heat-bath dynamics. As before, select a pair of rows (i, i') with $i \neq i'$, and a pair of columns (j, j') with $j \neq j'$. Now choose the new matrix A' u.a.r. from those which agree with A except at the four entries a_{ij}, $a_{i'j'}$, $a_{ij'}$, and $a_{i'j}$ (and have the correct row and column sums). Again, little is known about the mixing time in general, but see Dyer and Greenhill [24] for a special case.

Appendix

Proof (of Proposition 2.1). The techniques we employ are standard in the area [36]. Recall from Section 2. (refer to equation (2.2)) that we have expressed the number of q-colourings of G as a product

$$|\Omega(G)| = q^n \varrho_1 \ldots \varrho_m, \tag{8.1}$$

where

$$\varrho_i = \frac{|\Omega(G_i)|}{|\Omega(G_{i-1})|}.$$

Suppose that the graphs G_i and G_{i-1} differ in the edge $\{u, v\}$, which is present in G_i but absent from G_{i-1}. Clearly, $\Omega(G_i) \subseteq \Omega(G_{i-1})$. Any colouring in $\Omega(G_{i-1}) \setminus \Omega(G_i)$ assigns the same colour to u and v, and may be perturbed to a colouring in $\Omega(G_i)$ by recolouring vertex u with one of at least $q - \Delta \geq 1$ colours. (To resolve ambiguity, let u be the smaller of the two vertices.) On the other hand, each colouring in $\Omega(G_i)$ can be obtained in at most one way as the result of such a perturbation; hence $|\Omega(G_{i-1}) \setminus \Omega(G_i)| \leq |\Omega(G_i)|$ and

$$\frac{1}{2} \leq \varrho_i \leq 1. \tag{8.2}$$

To avoid trivialities, assume $0 < \varepsilon \leq 1$ and $m \geq 1$. Let $Z_i \in \{0,1\}$ denote the random variable which results from running the postulated almost uniform sampler on the graph G_{i-1} and returning one if the resulting q-colouring is also a colouring of G_i and zero otherwise. Denote by $\mu_i = \operatorname{Exp} Z_i$ the expectation of Z_i. By setting $\delta = \varepsilon/6m$, we may ensure

$$\varrho_i - \frac{\varepsilon}{6m} \leq \mu_i \leq \varrho_i + \frac{\varepsilon}{6m}, \tag{8.3}$$

or, noting inequality (8.2),

$$\left(1 - \frac{\varepsilon}{3m}\right) \varrho_i \leq \mu_i \leq \left(1 + \frac{\varepsilon}{3m}\right) \varrho_i; \tag{8.4}$$

so the mean of a sufficiently large (but still polynomial) number of independent copies of Z_i will provide a good estimate for ϱ_i. Note that, by inequalities (8.2) and (8.3), $\mu_i \geq \frac{1}{3}$.

So let $Z_i^{(1)}, \ldots, Z_i^{(s)}$ be a sequence of $s = \lceil 74\varepsilon^{-2} m \rceil \leq 75\varepsilon^{-2} m$ independent copies of the random variable Z_i, obtained from independent trials using the postulated almost uniform sampler, and let $\overline{Z}_i = s^{-1} \sum_{j=1}^{s} Z_i^{(j)}$ be their mean. Since Z_i is a random variable taking values from $\{0,1\}$, it follows easily that $\mu_i^{-2} \operatorname{Var} Z_i = \mu_i^{-1} - 1 \leq 2$, and hence $\mu_i^{-2} \operatorname{Var} \overline{Z}_i \leq 2s^{-1}$. As our estimator for $|\Omega(G)|$, we use the random variable $Y = q^n\, \overline{Z}_1 \overline{Z}_2 \ldots \overline{Z}_m$. Note that $\operatorname{Exp} Y = q^n \mu_1 \mu_2 \ldots \mu_m$.

The performance of this estimator is characterised by its variance, which may be bounded as follows:

$$\begin{aligned}
\frac{\operatorname{Var} \overline{Z}_1 \overline{Z}_2 \ldots \overline{Z}_m}{(\mu_1 \mu_2 \ldots \mu_m)^2} &= \prod_{i=1}^{m} \left(1 + \frac{\operatorname{Var} \overline{Z}_i}{\mu_i^2}\right) - 1 \\
&\leq \left(1 + \frac{2}{s}\right)^m - 1 \\
&\leq \operatorname{Exp}\left(\frac{\varepsilon^2}{37}\right) - 1 \\
&\leq \frac{\varepsilon^2}{36},
\end{aligned}$$

since $\mathrm{e}^{x/37} \leq 1 + x/36$ provided $0 \leq x \leq 1$. Thus, by Chebychev's inequality,

$$\left(1-\frac{\varepsilon}{3}\right)\mu_1\mu_2\dots\mu_m \le q^{-n}Y \le \left(1+\frac{\varepsilon}{3}\right)\mu_1\mu_2\dots\mu_m$$

with probability at least $\frac{3}{4}$. But from inequality (8.4), we have

$$\left(1-\frac{\varepsilon}{2}\right)\varrho_1\varrho_2\dots\varrho_m \le \mu_1\mu_2\dots\mu_m \le \left(1+\frac{\varepsilon}{2}\right)\varrho_1\varrho_2\dots\varrho_m,$$

which, combined with the previous inequality and (8.1), implies that the estimator Y satisfies the requirements of a randomised approximation scheme for the number of colourings $|\Omega(G)|$.

To estimate each ratio ϱ_i we need $75\varepsilon^{-2}m$ samples from the almost uniform sampler, and there are m such ratios in all to estimate. The claimed time complexity for approximate counting follows.

Proof (of equation (4.5)). Consider a facet $R(c) \cap R(c')$, where c and c' are adjacent states (colourings). Up to symmetry, such a facet is a $(nq-1)$-dimensional polytope defined by inequalities

$$1 \ge x_{0,0} = x_{0,1} \ge x_{0,2}, x_{0,3}, \dots, x_{0,q-1} \ge 0 \tag{8.5}$$

$$1 \ge x_{1,0} \ge x_{1,1}, x_{1,2}, \dots, x_{1,q-1} \ge 0 \tag{8.6}$$

$$\vdots$$

$$1 \ge x_{n-1,0} \ge x_{n-1,1}, x_{n-1,2}, \dots, x_{n-1,q-1} \ge 0. \tag{8.7}$$

This particular facet corresponds to the boundary between the all-0 state and the and adjacent state in which vertex 0 acquires colour 1; the facet clearly lies in the plane defined by $x_{0,0} = x_{0,1}$.

We wish to compute $\mathrm{vol}_{nq-1}(R(c) \cap R(c'))$, the area (i.e., $(nq-1)$-dimensional volume) of the facet $R(c) \cap R(c')$. Each line of the above display relates a different set of q variables, so the required volume is the product of the volumes of the polytopes defined by each line. The polytope defined by (8.5) is of dimension $q-1$, and all the others, namely (8.6)–(8.7), are of dimension q. The q-dimensional volume of the polytope defined by any of (8.6)–(8.7) is simply

$$\int_0^1 x^{q-1}\,dx = \left[\frac{x^q}{q}\right]_0^1 = \frac{1}{q}. \tag{8.8}$$

To calculate the volume of the polytope defined by (8.5), project it onto the plane $x_{0,0} = 0$ to obtain the polytope

$$1 \ge x_{0,1} \ge x_{0,2}, x_{0,3}, \dots, x_{0,q-1} \ge 0,$$

which, by comparison with (8.8), has $(q-1)$-dimensional volume $(q-1)^{-1}$. Projecting from the plane $x_{0,0} = x_{0,1}$ to the plane $x_{0,0} = 0$ contracts volume by a factor $\sqrt{2}$ (the scalar product of the normals to the two planes) so the actual volume before projection is $\sqrt{2}(q-1)^{-1}$.

Multiplying the n factors just computed together, we obtain

$$\mathrm{vol}_{nq-1}(R(c) \cap R(c')) = \frac{\sqrt{2}}{q^{n-1}(q-1)},$$

as claimed.

Bibliography

1. David ALDOUS, Random walks on finite groups and rapidly mixing Markov chains, *Séminaire de Probabilités XVII 1981/82* (A. Dold and B. Eckmann, eds), Springer Lecture Notes in Mathematics **986**, 1986, 243–297.
2. David ALDOUS, On the Markov chain simulation method for uniform combinatorial distributions and simulated annealing, *Probability in the Engineering and Informational Sciences* **1** (1987), 33–46.
3. David ALDOUS, The random walk construction of uniform spanning trees and uniform labelled trees, *SIAM Journal of Discrete Mathematics* **3** (1990), 450–465.
4. N. ALON, Eigenvalues and expanders, *Combinatorica* **6** (1986), 83–96.
5. David APPLEGATE and RAVI Kannan, Sampling and integration of near log-concave functions, *Proceedings of the 23rd Annual ACM Symposium on Theory of Computing* (STOC), ACM Press, 1991, 156–163.
6. J. VAN DEN BERG and J. E. STEIF, *On the existence and non-existence of finitary codings for a class of random fields*, Preprint, 1998.
7. Béla BOLLOBÁS, *Extremal Graph Theory*, Academic Press, 1978.
8. Andrei Z. BRODER, How hard is it to marry at random? (On the approximation of the permanent), *Proceedings of the 18th Annual ACM Symposium on Theory of Computing* (STOC), ACM Press, 1986, 50–58. Erratum in *Proceedings of the 20th Annual ACM Symposium on Theory of Computing*, 1988, p. 551.
9. R. L. BROOKS, On colouring the nodes of a network, *Proceedings of the Cambridge Philosophical Society* **37** (1941), 194–197.
10. Russ BUBLEY and Martin DYER, Path coupling, Dobrushin uniqueness, and approximate counting, *Report 97.04*, School of Computer Studies, University of Leeds, January 1997.
11. Russ BUBLEY and Martin DYER, Path coupling: a technique for proving rapid mixing in Markov chains, *Proceedings of the 38th IEEE Symposium on Foundations of Computer Science* (FOCS), IEEE Computer Society Press 1997, 223–231.
12. Russ BUBLEY and Martin DYER, Faster random generation of linear extensions, *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms* (SODA), ACM/SIAM, 1998, 350–354.
13. Russ BUBLEY, Martin DYER and Catherine GREENHILL, Beating the 2Δ bound for approximately counting colourings: a computer-assisted proof of rapid mixing, *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms* (SODA), ACM/SIAM, 1998, 355–363.
14. Russ BUBLEY, Martin DYER and Mark JERRUM, An elementary analysis of a procedure for sampling points in a convex body. *Random Structures and Algorithms* **12** (1998), 213–235.
15. Colin COOPER and Alan M. FRIEZE, *Mixing Properties of the Swendsen-Wang Process on Classes of Graphs*, Preprint, 1998.
16. R. CORDOVIL and M. L. MOREIRA, Bases-cobases graphs and polytopes of matroids, *Combinatorica* **13** (1993), 157–165.

17. Persi DIACONIS, *Group Representations in Probability and Statistics*, Institute of Mathematical Statistics, Hayward CA, 1988.
18. Persi DIACONIS and Bradley EFRON, Testing for independence in a two-way table: new interpretations of the chi-squared statistic, *Annals of Statistics* **13**, 1985, 845–913.
19. Persi DIACONIS and Daniel STROOCK, Geometric bounds for eigenvalues of Markov chains, *Annals of Applied Probability* **1** (1991), 36–61.
20. Martin DYER and Alan FRIEZE, Computing the volume of convex bodies: a case where randomness provably helps. In Probabilistic Combinatorics and its Applications, *Proceedings of AMS Symposia in Applied Mathematics* **44** (1991), 123–170.
21. Martin DYER and Alan FRIEZE, Random walks, totally unimodular matrices and a randomized dual simplex method, *Mathematical Programming* **64** (1994), 1–16.
22. Martin DYER, Alan FRIEZE and Ravi KANNAN, A random polynomial time algorithm for approximating the volume of convex bodies, *Journal of the ACM* **38** (1991), 1–17.
23. Martin DYER and Catherine GREENHILL, *On Markov chains for independent sets.* Preprint, 1997. (Visit `http://www.scs.leeds.ac.uk/rand/acg.html`)
24. Martin DYER and Catherine GREENHILL, A genuinely polynomial-time algorithm for sampling two-rowed contingency tables, *25th EATCS International Colloquium on Automata, Languages and Programming*, Aalborg, Denmark, Springer-Verlag LNCS Series, 1998, to appear.
25. Martin DYER, Ravi KANNAN and John MOUNT, Sampling contingency tables, *Random Structures and Algorithms* **10** (1997), 487–506.
26. R. G. EDWARDS and A. D. SOKAL, Generalizations of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm, *Physical Review D* **38** (1988), 2009–2012.
27. Tomás FEDER and Milena MIHAIL, Balanced matroids, *Proceedings of the 24th Annual ACM Symposium on Theory of Computing* (STOC), ACM Press, 1992, 26–38.
28. James Allen FILL, An interruptible algorithm for perfect sampling via Markov chains, *Proceedings of the 29th Annual ACM Symposium on Theory of Computing* (STOC), ACM Press, 1997, 688–695.
29. C. M. FORTUIN and P. W. KASTELEYN, On the random cluster model I: introduction and relation to other models, *Physica* **57** (1972) 536–564.
30. Vivek GORE and Mark JERRUM, The Swendsen-Wang process does not always mix rapidly, *Proceedings of the 29th Annual ACM Symposium on Theory of Computing* (STOC), ACM Press, 1997, 674–681.
31. Olle HÄGGSTRÖM and Karin NELANDER, Exact sampling from anti-monotone systems, Preprint, February 1997. To appear in *Statistica Neerlandica.*
32. O. J. HEILMANN and E. H. LIEB, Theory of monomer-dimer systems, *Communications in Mathematical Physics* **25** (1972), 190–232.
33. Mark JERRUM, A very simple algorithm for estimating the number of k-colourings of a low-degree graph, *Random Structures and Algorithms* **7** (1995), 157–165.
34. Mark JERRUM and Alistair SINCLAIR, Approximating the permanent, *SIAM Journal on Computing* **18** (1989), 1149–1178.
35. Mark JERRUM and Alistair SINCLAIR, Polynomial-time approximation algorithms for the Ising model, *SIAM Journal on Computing* **22** (1993), 1087–1116.
36. Mark JERRUM and Alistair SINCLAIR, The Markov chain Monte Carlo method: an approach to approximate counting and integration. In *Approximation Algorithms for NP-hard Problems* (Dorit Hochbaum, ed.), PWS, 1996, 482–520.

37. Mark R. JERRUM, Leslie G. VALIANT, and Vijay V. VAZIRANI, Random generation of combinatorial structures from a uniform distribution, *Theoretical Computer Science* **43** (1986), 169–188.
38. Mark JERRUM and Umesh VAZIRANI, A mildly exponential approximation algorithm for the permanent, *Algorithmica* **16** (1996), 392–401.
39. Valen E. JOHNSON, Studying convergence of Markov chain Monte Carlo algorithms using coupled sample paths, *Journal of the American Statistical Association* **91** (1996), 154–166.
40. Ravi KANNAN, Markov chains and polynomial time algorithms, *Proceedings of the 35th Annual IEEE Symposium on Foundations of Computer Science* (FOCS), Computer Society Press, 1994, 656–671.
41. L. LOVÁSZ and M. SIMONOVITS, Random walks in a convex body and an improved volume algorithm, it Random Structures and Algorithms **4** (1993), 359–412.
42. Ravi KANNAN, László LOVÁSZ and Miklós SIMONOVITS, *Random Walks and an $O^*(n^5)$ Volume Algorithm for Convex Bodies.* Preprint, January 1996.
43. R. M. KARP and M. LUBY, Monte-Carlo algorithms for enumeration and reliability problems, *Proceedings of the 24th Annual IEEE Symposium on Foundations of Computer Science*, Computer Society Press, 1983, 56–64.
44. Alexander KARZANOV and Leonid KHACHIYAN, *On the conductance of order Markov chains*, Technical Report DCS 268, Rutgers University, June 1990.
45. W. S. KENDALL, Perfect simulation for the area-interaction point process, University of Warwick, Department of Statistics Research Report **292**, April 1996. To appear in *Probability Perspective* (C. C. Heyde and L. Accardi, editors), World Scientific Press, Singapore.
46. W. S. KENDALL, Perfect simulation for spatial point processes, University of Warwick, Department of Statistics Research Report **308**, 1997. To appear in *Proceedings of ISI 51st session*, Istanbul, August 1997.
47. Claire KENYON, Dana RANDALL and Alistair SINCLAIR, Matchings in lattice graphs, *Proceedings of the 25th Annual ACM Symposium on Theory of Computing* (STOC), ACM Press, 1993, 738–746.
48. Donald E. KNUTH, Estimating the efficiency of backtrack programs, *Mathematics of Computation* **29** (1975), 121–136.
49. Torgny LINDVALL and L. C. G. ROGERS, Coupling of Multidimensional Diffusions by Reflection, *Annals of Probability* **14** (1986), 860–872.
50. Michael LUBY and Eric VIGODA, Approximately counting up to four, *Proceedings of the 29th Annual ACM Symposium on Theory of Computation* (STOC), ACM Press, 1997, 682–687.
51. N. METROPOLIS, A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER and E. TELLER, Equation of state calculation by fast computing machines, *Journal of Chemical Physics* **21** (1953), 1087–1092.
52. Milena MIHAIL and Peter WINKLER, On the number of Eulerian orientations of a graph, *Proceedings of the 3rd Annual ACM-SIAM Symposium on Discrete Algorithms* (SODA), ACM Press, 1992, 138–145.
53. James G. PROPP and David B. WILSON, Exact sampling with coupled Markov chains and applications to statistical mechanics, *Random Structures and Algorithms* **9** (1996), 223–252.
54. Lars Eilstrup RASMUSSEN, Approximating the permanent: a simple approach, *Random Structures and Algorithms* **5** (1994), 349–361.
55. Alistair SINCLAIR, Improved bounds for mixing rates of Markov chains and multicommodity flow, *Combinatorics, Probability and Computing* **1** (1992), 351–370.

56. Alistair SINCLAIR, *Randomised algorithms for counting and generating combinatorial structures*, Advances in Theoretical Computer Science, Birkhäuser, Boston, 1993.
57. Alistair SINCLAIR and Mark JERRUM, Approximate counting, uniform generation and rapidly mixing Markov chains, *Information and Computation* **82** (1989), 93–133.
58. Dominic WELSH, Approximate Counting. In *Surveys in Combinatorics*, London Mathematical Society Lecture Note **241**, Cambridge University Press, 1997, 287–323.
59. David WILSON, *Annotated Bibliography of Perfectly Random Sampling with Markov Chains*, `http://dimacs.rutgers.edu/~dbwilson/exact.html/`