Rapidly Mixing Markov Chains and their Applications

Salil Vadhan
Churchill College
Cambridge University

May 21, 1996

I declare that this essay is my own unaided work and does not contain material that I have already used to any substantional extent for any comparable purpose.

Signed Salil Vadhan

500 Ross Place
Oceanside, NY 11572–4520
USA

Contents

1 Introduction

Although the old adage "Better late than never" may be an acceptable guideline for human behavior, we rightfully expect more of our computing devices. An algorithm that takes exponentially long to produce a result is effectively useless in practice, even if we can guarantee its eventual success. This principle is foundational to the modern theory of algorithms, in which tight upper and lower bounds are sought on running times. In this essay, we apply this principle to the theory of Markov chains, as implicit in every Markov chain is an algorithm which can only be of use if bounds on its running time are known.

Informally, a Markov chain on a set X is a random walk on the elements of X. Under certain conditions, we can guarantee that, regardless of the starting point, the position of the random walk at time t approaches some probability distribution π as t tends to infinity. This makes Markov chains a natural solution to random generation problems: If we wish to generate elements from a set X with probability distribution π, we find a Markov chain on X which converges to π, arbitrarily pick a starting point in X, simulate the chain for a large number of steps, and output the final position. The problem with this approach is that, although it is often easy to construct Markov chains with a given limiting distribution, it is much harder to guarantee that the chain is near this distribution after a reasonable number of steps. In fact, there are many chains which take exponentially long to approach their limiting distribution.

The importance of Markov chains stems from the importance of random generation problems in areas such as statistics, combinatorics, and statistical physics, along with theoretical computer science itself. In statistical physics, for example, it is often desirable to be able to randomly sample states of a physical system in order to approximate the values of some physical quantity, such as mean energy, whose exact computation would require summing over an exponentially large set of states. For many years, Markov chains were used to do computations in statistical physics without any bounds on the rates of convergence. Presumably, this resulted in a number of fallacious calculations.

Counting, one of the classical goals of combinatorics, also has close ties with random generation. As first demonstrated by Valiant, many important counting problems are computationally intractable. The natural alternative is approximate counting. It turns out that the problem of approximately counting a class of combinatorial structures, such as the trees in a graph, is often equivalent to the problem of generating those types of structures almost uniformly at random.

Until recently, no one knew how to obtain nontrivial bounds on the rates of convergence for all but the simplest of Markov chains. The first major breakthrough was made less than a decade ago by Jerrum and Sinclair; they related the rate of convergence of a Markov chain to a simple combinatorial quantity known as conductance and devised a novel way of estimating it. This yielded the first provably polynomial-time algorithms for a number of important computational problems. Since Jerrum and Sinclair's work, their technique has been used to solve a host of other interesting problems and their original bounds have been improved.

The goal of this essay is to develop the theory of the rate of convergence of Markov chains and present a couple of its most important applications, using the latest improvments in Jerrum and Sinclair's techniques. In contrast to the papers where these results were first published, we develop the theory from first principles, appealing only to very basic and well-known results. Developing a theory with a specific goal in mind — rapid mixing — enables the mathematics to be streamlined to some degree. Because of this, nearly all the proofs in the first half of this essay (until Section 3.3) have either been devised expressly for this essay or have been simplified from their original form.

In the end, we will be using the Markov chain theory that we develop to construct efficient algorithms. So that our goal in analyzing the Markov chains can be made precise, we begin by discussing some notions from computation theory.

2 Background

2.1 Some Basic Computational Notions

In this section, we discuss the computational notions necessary to analyze the algorithms in which we shall be using our Markov chains. However, we do not wish to obscure the basic ideas with a detailed discussion of the computational model. Thus, we shall assume that what we mean by an **algorithm** and its **running time** are clear to the reader. Most of our algorithms will be **probabilistic**. That is, they will be permitted to branch on independent probabilities that are the ratio of two previously computed integers. We may occasionally speak of **deterministic** algorithms, whose computation is completely determined by its input, and thus cannot branch probabilistically. These models can be made mathematically precise by a discussion of deterministic and probabilistic **Turing machines** (cf. [JVV86]), but this essay is not the place for such a discussion. The definitions of this section generally follow those of [JVV86].

We must introduce a bit of notation before proceeding. Throughout this essay, Σ will represent a fixed finite alphabet and Σ^* will denote the set of strings over Σ. For $x \in \Sigma$, $|x|$ is the **length** of x, *i.e.* the number of symbols from Σ in x. The **empty string** e is the unique string of length 0 in Σ^*.

From the perspective of theoretical computer science, the most important applications of rapidly mixing Markov chains are the uniform generation and approximate counting of combinatorial structures. For reasons that will become clear later, we will restrict our attention to combinatorial structures that are associated with ***p*-relations**.

Definition 2.1 *A relation $R \subset \Sigma^* \times \Sigma^*$ is said to be a* ***p*-relation** *iff the following conditions hold:*
1. it is **polynomially-balanced**, *i.e. there exists a polynomial p such that $\langle x, y\rangle \in R \Rightarrow |y| \leq p(|x|)$*
2. it can be 'checked quickly,' i.e. there is a deterministic algorithm which, on input $\langle x, y\rangle \in \Sigma^* \times \Sigma^*$, can decide whether $\langle x, y\rangle \in R$ in time bounded by a

polynomial in $|x|$. We also write $R(x) = \{y \in \Sigma^* : \langle x, y\rangle \in R\}$. We call x an **instance** *and $R(x)$ the set of* **solutions** *for x.*

An simple example of a p-relation is the relation R that relates graphs to their subgraphs. Precisely, we say $\langle x, y\rangle \in R$ if x is an encoding of a graph $G = (V, E)$ and y is an encoding of a **spanning subgraph** $H = (V, F)$ of G, *i.e.* $F \subset E$. [1] Let n be the number of vertices in G. In any reasonable encoding, we will certainly have $|x| \in \Omega(n^2)$, since we must specify the presence or absence of $\binom{n}{2}$ edges in G. We must also have $|y| \in O(n^2)$, as $O(n^2)$ bits is sufficient to represent a subset of E. So condition 1 is satisfied. Since we can certainly check in time $O(n^2)$ whether H is a subgraph of G, condition 2 is also satisfied.

In terms of p-relations, the counting and generation problems can easily be expressed: Given input x, output $|R(x)|$ (counting); Given input x, output an element of $R(x)$ uniformly at random (generation). In our example, these correspond to counting the number of subgraphs of G and producing a subgraph of G uniformly at random. Now, the reason that we require R to be a p-relation should be clear. If we cannot even efficiently test whether a hypothetical solution y lies in $R(x)$, it is unlikely that we would be able to efficiently count the elements of $R(x)$ or produce an element of $R(x)$ uniformly at random. It is possible to produce examples of relations which are not p-relations whose counting problem is trivial.[2] But these examples are unnatural, and most problems of interest do in fact arise from p-relations.

Armed with the notion of a p-relation, we are in a position to precisely define the generation and counting problems associated with R. Although uniform generation is an interesting problem, in practice a slight bias from uniformity can be tolerated. Since our Markov chains will only be approaching their limiting distributions and cannot be expected to actually achieve them in a finite number of steps, we need the notion of an *almost* uniform generator to describe their behavior.

Definition 2.2 *Let R be a p-relation on Σ^*. An* **almost uniform generator** *for R is a probabilistic algorithm M which takes inputs $\langle x, \epsilon\rangle \in \Sigma^* \times \mathbb{R}^+$ and whose output is a random variable $M(x, \epsilon)$, such that for all $y \in \Sigma^*$:*

1. $M(x, \epsilon) \in R(x) \cup \{N\}$, *where* $N \notin \Sigma^*$
2. $1/|R(x)|(1 + \epsilon) \leq Pr[M(x, \epsilon) = y|\ M(x, \epsilon) \neq N] \leq (1 + \epsilon)/|R(x)|$ *if* $y \in R(x)$
3. $Pr[M(x, \epsilon) \in R(x)] \geq \frac{1}{2}$ *if* $R(x) \neq \emptyset$

Such a generator is said to be a **fully-polynomial almost uniform generator (fpaug)** *if its running time is bounded by a polynomial in $|x|$ and $\log(1/\epsilon)$.*

[1] All our inputs and outputs to algorithms must be somehow encoded as strings. We won't make a fuss about specific choices of encoding, because we will usually be able to specify the running times of our algorithms in terms of natural quantities, such as the number of vertices in a graph. Also, most reasonable encodings will differ by only a polynomial factor in length, and this degree of granularity will be sufficient for any abstract complexity discussion we have.

[2] Assuming $\mathcal{P} \neq \mathcal{NP}$.

In the above definition, the parameter ϵ indicates how much the distribution of outputs is allowed to deviate from uniform. An output of N indicates failure, and this is only allowed to occur with probability smaller than $1/2$. Observe that we can reduce the probability of failure to any $\delta > 0$ by $O(\log \delta^{-1})$ independent runs of the algorithm.

Counting is the other combinatorial problem for which Markov chains will prove useful. Ideally, we would like efficient algorithms which could exactly count the number of solutions for an instance of a p-relation. However, it has been long known that exact counting is intractable for many p-relations. Since Valiant's introduction of the class $\#\mathcal{P}$ in [Val79], he and others have compiled a long list of simple counting problems for which there are no polynomial-time algorithms.[3] [Val79, PB83, DF89, Rot95, Gol93, Vad95] Thus we turn to approximate counting, for which the outlook is much brighter. Later, we will be interested in approximating more general functions from strings to numbers, so we make our definition in this broader context.

Definition 2.3 *Given three nonnegative real numbers α, β, and r, we say that* **α approximates β within ratio r** *iff $\alpha/r \leq \beta \leq \alpha r$. (Clearly $r \geq 1$ is necessary for this to hold, and the smaller $r - 1$ is, the better the estimate.)*

Definition 2.4 *Let $f : \Sigma^* \to \mathbb{R}$ be a non-negative real-valued function on strings. A* **randomized approximation scheme** *for f is a probabilistic algorithm M, which takes input $\langle x, \epsilon \rangle$ and whose output is a random variable $M(x, \epsilon) \in \mathbb{R}$ such that*

$$Pr[M(x, \epsilon) \text{ approximates } f(x) \text{ within ratio } 1 + \epsilon] \geq \frac{3}{4}$$

We call M a **fully-polynomial randomized approximation scheme (fpras)** *its running time is polynomial in $|x|$ and $1/\epsilon$.*

In order to apply this definition to counting the number of solutions for a p-relation R, let f be the function $\#R$, where $\#R(x) = |R(x)|$. Note that, although we allow randomized approximation schemes to fail up to $1/4$ of the time, we can reduce the probability of failure to any $\delta > 0$ by taking the median of $k \in O(\log \delta^{-1})$ independent runs of the algorithm. The probability that the median is not an approximation within ratio $1 + \epsilon$ is bounded above by the probability that the majority of outputs are not approximations within ratio $1 + \epsilon$. It is then easy to show, using the Chernoff bounds, that this happens with probability at most $e^{-k/12}$.

One would expect that approximate counting and uniform generation are closely related; anyone who has spent some time working on generation problems would agree that uniform generation can often be reduced to counting. However, it is somewhat surprising that approximate counting and almost uniform generation are actually *equivalent* for a large class of p-relations. Precisely, it is shown in [JVV86] that if R is a **self-reducible** relation, the existence of a fully-polynomial almost uniform generator for R implies the existence of

[3] Assuming $\mathcal{P} \neq \mathcal{NP}$.

a fully-polynomial randomized approximation scheme for $\#R$ and conversely. The self-reducibility of R essentially means that the solutions for an instance x can be represented recursively in terms of the solutions for smaller instances. Their reduction from almost uniform generation to approximate counting is simply the standard reduction from generation to counting for recursive structures, boosted to cope with slightly inaccurate counts.

Their reduction in the other direction is a bit more interesting. We sketch their technique here: Say the solutions for x can represented in terms of the solutions for smaller instances x_1 and x_2. By repeatedly generating solutions for x, we can estimate which fraction of the solutions for x correspond to solutions for x_1 and which correspond to solutions for x_2. We then recursively obtain an approximate count to the number of solutions for one of the x_i's (using the one yielding a larger fraction of the solutions for x to maximize accuracy). Dividing this by the fraction of x's solutions which come from x_i as we previously estimated, we obtain an approximate count for the number of solutions for x.

From this reduction, it should be clear how random generation algorithms can be used to solve combinatorial counting problems. A problem which demonstrates this is the computation of the permanent, which counts the number of perfect matchings in a bipartite graph. We shall present this example in greater detail in Section 4.1. Another use of random generation algorithms is to compute functions that can be written as expectations of certain operators on the probability space we are generating from. An example of this, computing the partition function in the Ising model of statistical physics, is given in Section 4.2.

The only provably effective generation algorithms known for these problems and others have come from tremendous advances in the study of Markov chains and their rates of convergence. We begin our study of Markov chains in the next section.

2.2 Basic Theory of Markov Chains on a Finite State Space

Although Markov chains can be defined and studied in a very general context, we will only develop the aspects of the theory relevant to this essay. In particular, we shall only discuss homogeneous Markov chains on a finite state space. The definitions of this section follow [GS82].

Definition 2.5 *Let S be a finite set. A* **(time-)homogeneous Markov chain** *X on S is a sequence of random variables $X_0, X_1, \ldots$ taking values in S such that*
1. $Pr(X_{n+1}|X_0, \ldots, X_n) = Pr(X_{n+1}|X_n)$
2. $Pr(X_{n+1} = y|X_n = x) = Pr(X_1 = y|X_0 = x)$ *for all* $x, y \in S$
The **transition matrix** *P is given by* $P(x, y) = Pr(X_{n+1} = y|X_n = x)$.

We will maintain this notation throughout the section, with X denoting an arbitrary homogeneous Markov chain on finite state set S and P denoting its transition matrix. It is easy to verify the following facts about Markov chains:

Proposition 2.6 *Define the matrix* P_n *by* $P_n(x,y) = Pr(X_{m+n} = y | X_m = x)$ *and let* $\mu^{(n)}(x) = Pr(X_n = x)$ *be a row vector. Then,*
1. $P_n = P^n$
2. $\mu^{(n)} = \mu^{(0)} P_n$

Thus, if we are interested in using a Markov chain to randomly sample from its state space, we must study the evolution of $\mu^{(n)}$, which gives the distribution at time n. We want $\mu^{(n)}$ to approach a limiting distribution; it is not surprising that this limiting distribution must be fixed under steps of the chain:

Definition 2.7 *A (row) vector* $\pi \in \mathbb{R}^S$ *is called* **stationary distribution** *for* X *iff the following conditions hold:*
1. $\pi(x) \geq 0$ *for all* $x \in S$
2. $\sum_x \pi(x) = 1$
3. $\pi = \pi P$

In order to simplify analysis and to guarantee the existence of a unique stationary distribution, we impose a few more conditions on our chains.

Definition 2.8 X *is said to be* **irreducible** *iff for all* $x, y \in S$*, there is an* m *such that* $P^m(x,y) > 0$*, i.e.* y *is eventually reachable from* x*.*

Irreducibility guarantees that our chain is connected, so that if we start our chain in one state and simulate, it is possible to reach all the other states. This is clearly desirable when using the chain to sample from the set S. The following proposition verifies the intuitively obvious fact that this condition implies that any stationary distribution must charge each point.

Proposition 2.9 *Let* π *be a stationary distribution for the irreducible Markov chain* X*. Then* $\pi(x) > 0$ *for all* $x \in S$*.*

Proof There must be at least one $y \in S$ such that $\pi(y) > 0$. Now, for any $x \in S$, choose $n \geq 0$ such that $P^n(x,y) > 0$. Then $\pi(x) = (\pi P^n)(x) \geq \pi(y) P^n(x,y) > 0$ $\square$

Definition 2.10 X *is said to be* **aperiodic** *iff for all* $x \in S$*,*

$$\gcd\{m : P^m(x,x) > 0\} = 1$$

Aperiodicity is a condition we impose merely to simplify analysis. This does not cause any loss of generality, because we can turn any periodic chain into an aperiodic one by adding a self-loop probability of 1/2 to each state without affecting the stationary distribution (which will be our limiting distribution). This is done by replacing P with $(I+P)/2$.

Definition 2.11 *Suppose* X *has a stationary distribution* π*.* X *is said to be* **time-reversible** *with respect to* π *iff*

$$\pi(x) P(x,y) = \pi(y) P(y,x) \text{ for all } x, y \in S \tag{1}$$

The conditions of (1) are known as the **detailed balance equations**. Requiring time-reversibility does cause some loss of generality, but, as we shall see in the next section, this sacrifice is well worth the ease of analysis gained. Moreover, it is often easy to check both the existence of a stationary distribution and time-reversibility by means of the following lemma.

Lemma 2.12 *Let X be a Markov chain on state set S. If there exists a vector $\pi \in \mathbb{R}^S$ such that*
1. $\pi(x) \geq 0$ for all $x \in S$
2. $\sum_x \pi(x) = 1$
3. π satisfies the detailed balance equations (1).
Then π is a stationary distribution of X and X is time-reversible with respect to π.

Proof
$(\pi P)(y) = \sum_x \pi(x)P(x,y) = \sum_x \pi(y)P(y,x) = \pi(y)\sum_x P(y,x) = \pi(y)$ $\square$

The conditions of this lemma, coupled with irreducibility, are actually sufficient to guarantee that π is a unique stationary distribution; this will come out in the next section, where we study the rate of convergence of Markov chains. Before proceeding, we state a simple fact about time-reversible chains which will need later.

Proposition 2.13 *If a Markov chain X is time-reversible with respect to π, then*

$$\pi(x)P^k(x,y) = \pi(y)P^k(y,x) \text{ for all } x, y \in S \text{ and } k \geq 1$$

Proof Induction on k. $k = 1$ is by definition.

$$\begin{aligned}
\pi(x)P^{k+1}(x,y) &= \pi(x)(PP^k)(x,y) = \sum_{z \in S} \pi(x)P(x,z)P^k(z,y) = \sum_{z \in S} \pi(z)P(z,x)P^k(z,y) \\
&= \sum_{z \in S} \pi(y)P(z,x)P^k(y,z) = \pi(y)P^{k+1}(y,x)
\end{aligned}$$

$\square$

3 Rates of Convergence

3.1 Spectral Theory of Time-Reversible Markov Chains

A stationary distribution of a Markov chain is, by definition, a left eigenvector of the transition matrix with eigenvalue 1. We shall see that the spectral properties of the transition matrix determine the rate of convergence to a stationary distribution. Fortunately, the detailed balance equations (1) permit us to view P as a self-adjoint operator on an inner-product space, and the spectral theory of self-adjoint operators is fairly simple. The approach taken in this section and the next is a modification of that taken in [DS91]. The main improvement is that the approach illustrates more directly the ties between the spectrum of P and properties of the chain.

Let X be an irreducible homogeneous Markov chain on finite state space S which is reversible with respect to a stationary distribution π. Consider the space $L^2(\pi^{-1})$ of real-valued functions on S, with the following inner product:

$$\langle \phi, \psi \rangle = \sum_{x \in S} \frac{\phi(x)\psi(x)}{\pi(x)}$$

By Proposition 2.9, this is well-defined. Writing elements of $L^2(\pi^{-1})$ as row vectors, the transition matrix P defines an operator on $L^2(\pi^{-1})$ by right-multiplication:

$$(\phi P)(x) = \sum_{y \in S} \phi(y) P(y, x)$$

Note that this action is identical to the way P acts on probability distributions in the time-evolution of the Markov chain. This inner product space simplifies the study of the spectral properties of the transition matrix P, as P is self-adjoint with respect to $\langle \cdot, \cdot \rangle$.

Proposition 3.1 *P is a self-adjoint operator on $L^2(\pi^{-1})$.*

Proof

$$\langle \phi P, \psi \rangle = \sum_{x,y \in S} \frac{\phi(y)P(y,x)\psi(x)}{\pi(x)} = \sum_{x,y \in S} \frac{\phi(y)P(x,y)\psi(x)}{\pi(y)} = \langle \phi, \psi P \rangle$$

where the second equality follows from the detailed balance equations (1). □

By standard linear algebra, this implies that P has an orthonormal basis of eigenvectors. π is an eigenvector of P with eigenvalue 1. We shall show that π is the only eigenvector with eigenvalue 1 and the rest have eigenvalues between -1 and 1.

Proposition 3.2 *Let P be the operator on $L^2(\pi^{-1})$ defined above. Then*
1. Every eigenvalue λ of P satisfies $-1 \le \lambda \le 1$.
2. The only eigenvectors with eigenvalue 1 are multiples of π.

Proof Let ϕ be a non-zero eigenvector of P with eigenvalue λ. Let m be the maximum absolute value ϕ/π achieves on S and let $x \in S$ be a point at which this maximum is achieved. Then we have

$$\begin{aligned} m|\lambda| &\le \left| \frac{\lambda\phi(x)}{\pi(x)} \right| = \left| \frac{(\phi P)(x)}{\pi(x)} \right| \\ &\le \sum_{y \in S} \frac{|\phi(y)|P(y,x)}{\pi(x)} = \sum_{y \in S} \frac{P(x,y)|\phi(y)|}{\pi(y)} \\ &\le m \sum_{y \in S} P(y, z) = m \end{aligned}$$

This proves (1). For (2), let ϕ be an eigenvector with eigenvalue 1. Let m and x be as above. Without loss of generality, we may assume that $\phi(x) > 0$ (or else replace ϕ with $-\phi$.) Now, by irreducibility, for any $y \in S$, there is some $k > 0$ such that $P^k(y,x) > 0$. So we have

$$\begin{aligned}
\frac{\phi(x)}{\pi(x)} &= \frac{(\phi P^k)(x)}{\pi(x)} = \sum_{z \in S} \frac{\phi(z) P^k(z,x)}{\pi(x)} = \sum_{z \in S} \frac{\phi(z) P^k(x,z)}{\pi(z)} \\
&\leq \frac{\phi(y)}{\pi(y)} P^k(x,y) + \frac{\phi(x)}{\pi(x)} \sum_{z \neq y} P^k(x,z) \\
&= \frac{\phi(y)}{\pi(y)} P^k(x,y) + \frac{\phi(x)}{\pi(x)} (1 - P^k(x,y))
\end{aligned}$$

which implies that $\phi(x)/\pi(x) \leq \phi(y)/\pi(y)$ and thus ϕ/π is constant. □

A corollary of this is that if X has a stationary distribution, it is unique.

With this spectral characterization of the transition matrix, it is easy to describe the evolution of our Markov chain with time. Let $1 = \beta_0 \geq \beta_1 \geq \cdots \geq \beta_{m-1} \geq -1$ be the eigenvalues of P, where $m = |S|$, and let $\pi = v_0, v_1, \ldots, v_m$ be an orthonormal basis of eigenvectors. (Note that π does in fact have norm 1 with respect to $\langle \cdot, \cdot \rangle$.) Then, given our initial distribution $\mu^{(0)}$, we can write $\mu^{(0)} = c_0 \pi + c_1 v_1 + \cdots + c_{m-1} v_{m-1}$. The coefficients c_i are given by $\langle \mu^{(0)}, v_i \rangle$. In particular, $c_0 = \sum_{x \in S} \frac{\mu^{(0)}(x)\pi(x)}{\pi(x)} = 1$. The distribution after k steps is now given by

$$\mu^{(k)} = \mu^{(0)} P^k = \pi + c_1 \beta_1^k v_1 + \cdots + c_{m-1} \beta_{m-1}^k v_{m-1} \tag{2}$$

This is where the slight advantage of our approach over that of Diaconis and Strook becomes clear. They consider the action of the adjoint of P on $L^2(\pi)$. Although the spectral properties of P and its adjoint are the same, the action of the adjoint does not relate as directly to the time-evolution of the Markov chain. From the description (2), it is easy to see why the rate of convergence is intimately related to the spectrum of P. As k tends to infinity, all the terms corresponding to eigenvalues of absolute value less than 1 will become insignificant. And if all the eigenvalues other than β_0 are of absolute value strictly less than 1, this means that the chain will approach the stationary distribution. The rate of convergence will clearly depend on how much smaller than 1 the magnitude of the second largest eigenvalue is. A potential hindrance to convergence are eigenvalues of -1. They seem to produce an alternating effect, corresponding to a periodicity of 2. (Time-reversible chains can only have periods of 1 or 2.) The following proposition verifies this intuition.

Proposition 3.3 *If X is aperiodic, then P has no eigenvalues of -1.*

We first need a lemma.

Lemma 3.4 *If X is aperiodic, then there exists an n_0 such that for all $n \geq n_0$, $P^n(x,y) > 0$ for all $x, y \in S$.*

Proof (of Lemma 3.4) For $y \in S$, let $N_y = \{n : P^n(x,y) > 0\}$. Now, given $x, y \in S$, there is a positive integer m such that $P^m(x,y) > 0$, by irreducibility. We also know that $\gcd(N_y) = 1$. Thus there are integers $c_1, \ldots, c_k \in \mathbb{Z}$ and $n_1, \ldots, n_k \in N_y$ such that $c_1 n_1 + \cdots + c_k n_k = 1$. Let $n_0 = m + (|c_1|n_1 + \cdots + |c_k|n_k)^2$. Then for $n > n_0$, we can write

$$\begin{aligned} n - m &= q(|c_1|n_1 + \cdots + |c_k|n_k) + r, \text{ where} \\ r &< |c_1|n_1 + \cdots + |c_k|n_k \text{ and} \\ q &\geq |c_1|n_1 + \cdots + |c_k|n_k \end{aligned}$$

So we can write

$$n = m + (q|c_1| + rc_1)n_1 + \cdots + (q|c_k| + rc_k)n_k$$

Since $q|c_i| + rc_i \geq 0$ for each i, we have

$$P^n(x,y) \geq P^m(x,y)(P^{n_1}(y,y))^{(q|c_1|+rc_1)} \cdots (P^{n_k}(y,y))^{(q|c_k|+rc_k)} > 0$$

Thus we have shown that for any specific pair x, y, $P^n(x,y)$ is positive for all sufficiently large n. It follows from finiteness of S that $P^n(x,y)$ is positive *simultaneously* for all x, y for sufficiently large n. $\square$

Proof (of Proposition 3.3) Suppose ϕ is an eigenvector of P of eigenvalue -1. Let n_0 be as in the lemma. Suppose ϕ/π achieves its maximum absolute value at $x \in S$. Then we have

$$\begin{aligned} \left|\frac{\phi(x)}{\pi(x)}\right| &= \left|\frac{(\phi P^{n_0})(x)}{\pi(x)}\right| = \left|\sum_{y \in S} \frac{P^{n_0}(y,z)\phi(y)}{\pi(x)}\right| \leq \sum_{y \in S} \frac{P^{n_0}(y,x)|\phi(y)|}{\pi(x)} \\ &= \sum_{y \in S} \frac{P^{n_0}(x,y)|\phi(y)|}{\pi(y)} \leq \left|\frac{\phi(x)}{\pi(x)}\right| \sum_{y \in S} P^{n_0}(x,y) = \left|\frac{\phi(x)}{\pi(x)}\right| \end{aligned}$$

In order for equality to hold in the above two inequalities, $\phi(y)$ must be the same sign for all $y \in S$ and $|\phi(y)/\pi(y)| = |\phi(x)/\pi(x)|$ for all $y \in S$. Thus, ϕ is a multiple of π and has eigenvalue 1, which is a contradiction. $\square$

Thus, if we impose a condition of aperiodicity on our chain, convergence to the stationary distribution is guaranteed. We have yet to specify the metric by which we measure distance from the stationary distribution. Although a stronger measure, known as the **relative pointwise distance**, can be used (as in [JS89]), we follow [DS91] and [Sin91] in using the following measure.

Definition 3.5 *Let μ_1 and μ_2 be two probability measures on a finite set S. The* **variation distance** *between μ_1 and μ_2 is defined as*

$$||\mu_1 - \mu_2||_v = \frac{1}{2} \sum_{x \in S} |\mu_1(x) - \mu_2(x)|$$

Using this measure, we can bound the rate of convergence in terms of the eigenvalues of P. For the sake of clarity, we restate all the conditions on our Markov chain in this theorem.

Theorem 3.6 *Let X be an irreducible, aperiodic homogeneous Markov chain on finite state set S. Suppose that X is time-reversible with respect to its (unique) stationary distribution π. Then,*

$$\|\mu^{(n)} - \pi\|_v \leq \beta_*^n \frac{\sqrt{\langle \mu^{(0)}, \mu^{(0)} \rangle - 1}}{2} \quad \textit{where } \beta_* = \max\{\beta_1, |\beta_{m-1}|\}$$

Proof By Equation 2, we have

$$\begin{aligned}
4\|\mu^{(n)} - \pi\|_v^2 &= \left(\sum_{x \in S} \left| \sum_{i=1}^{m-1} \beta_i^n c_i v_i(x) \right| \right)^2 \\
&\leq \left(\sum_{x \in S, 1 \leq i \leq m-1} |\beta_i^n c_i v_i(x)| \right)^2 \\
&\leq \beta_*^{2n} \left(\sum_{x \in S, 1 \leq i \leq m-1} \sqrt{\frac{\pi(x)}{\pi(x)}} |c_i v_i(x)| \right)^2 \\
&\leq \beta_*^{2n} \left(\sum_{x \in S, 1 \leq i \leq m-1} \frac{c_i^2 (v_i(x))^2}{\pi(x)} \right) \left(\sum \pi(x) \right) \\
&= \beta_*^{2n} \sum_{1 \leq i \leq m-1} c_i^2 \langle v_i, v_i \rangle = \beta_*^{2n} (\langle \mu^{(0)}, \mu^{(0)} \rangle - 1)
\end{aligned}$$

where the last inequality is Cauchy-Schwarz and the last equality follows from the orthonormality of the v_i and Parseval's Equation. □

Theorem 3.6 states that, as long as the **spectral gap** $\lambda_* = 1 - \beta_*$ is not too small, the convergence to stationary is quite rapid. In simulations, we will have a fixed initial state $s \in S$. In this case, the initial distribution is $\mu^{(0)}(y) = \delta_{sy}$ so $\langle \mu^{(0)}, \mu^{(0)} \rangle = \pi(s)^{-1}$. Theorem 3.6 also has something like a converse; if the spectral gap is small, then there are initial distributions (of the form $\delta_{sy}(y)$) such that the convergence to stationary is slow. We don't prove the converse here, as we are interested in positive results, but the interested reader is referred to [Sin91].

Although we've determined that lower bounds on the spectral gap yield higher rates of convergence, in applications we need to know how many steps of the chain are needed to be within a certain distance of stationary. As noted above, our initial distribution will be of the form $\mu^{(0)}(y) = \delta_{sy}(y)$ for some initial state s. For $\epsilon > 0$, let $\tau_s(\epsilon)$ be the number of steps of the chain needed to guarantee that the variation distance from stationary is less than ϵ, *i.e.*

$||\mu^{(n)} - \pi||_v < \epsilon$ for all $n \geq \tau_s(\epsilon)$. Suppose we set

$$\begin{aligned}
\tau_s(\epsilon) &= \left\lceil \frac{\log \epsilon^{-1} + \frac{1}{2}\log \pi(s)^{-1} + \log 2}{\lambda_*} \right\rceil &(3)\\
&\geq \left\lceil \frac{\log \epsilon^{-1} + \frac{1}{2}\log \pi(s)^{-1} + \log 2}{\log \beta_*^{-1}} \right\rceil &(4)
\end{aligned}$$

where the inequality comes from the relationship $\lambda_* \leq \log(1-\lambda_*)^{-1} = \log \beta_*^{-1}$. Then, for all $n \geq \tau_s(\epsilon)$, Theorem 3.6 gives us

$$\begin{aligned}
||\mu^{(n)} - \pi||_v &\leq \exp\left[(\log \beta_*)\left(\frac{\log \epsilon^{-1} + \frac{1}{2}\log \pi(s)^{-1} + \log 2}{\log \beta_*^{-1}}\right)\right] \frac{\sqrt{\pi(s)^{-1} - 1}}{2} \\
&= \frac{2\epsilon\sqrt{\pi(s)^{-1} - 1}}{2\sqrt{\pi(s)^{-1}}} < \epsilon
\end{aligned}$$

Let us now use this to determine what sort of lower bounds are needed on λ_* to meet the requirements of our computational model. Specifically, suppose we want to construct a fully-polynomial almost uniform generator (Definition 2.2) M for a p-relation R using Markov chain techniques. The first task is to construct, for each input string x, a Markov chain on state space $R(x)$ whose stationary distribution is uniform. Then, on input $\langle x, \epsilon \rangle$, our algorithm proceeds as follows:

1. Find an element $s = s(x) \in R(x)$ to use as initial state.
2. Simulate the Markov chain for $N = N(x, \epsilon)$ steps.
3. Output the current state of the Markov chain.

The first and third requirements of Definition 2.2 are usually not much of a concern. To satisfy these, we merely need to guarantee that failure occurs with probability less than 1/2. The potential source of failure in our algorithm is the random branching process used to simulate the Markov chain. This is a technical detail that arises because the formal definition of a probabilistic Turing machine (*i.e.* algorithm) only allows branching on the flip of a fair coin. In order to simulate branching on any previously computed rational probabilities in this more restrictive model, we must allow for a small failure probability. In specific applications, we may also need to use a probabilistic algorithm to find a starting state; this could be another source of failure.

The second requirement is the one of most interest to us. It says

$$\pi/(1+\epsilon) \leq \mu^{(N)}(y) \leq \pi(1+\epsilon) \text{ for all } y \in S$$

where $\pi = 1/|R(x)| = \pi(y)$ for all y. Suppose we choose $N(x, \epsilon) = \tau_{s(x)}(\epsilon\pi/4)$ so that the variation distance $||\mu^{(N))} - \pi||_v \leq \epsilon\pi/4$. Then

$$|\mu^{(N)}(y) - \pi| \leq 2||\mu^{(N)} - \pi||_v \leq \epsilon\pi/2 \leq \frac{\epsilon\pi}{(1+\epsilon)}$$

(Assuming, as we may, that $\epsilon < 1$.) So

$$\frac{\pi}{1+\epsilon} = \pi - \frac{\epsilon\pi}{(1+\epsilon)} \leq \mu^{(N)} \leq \pi + \frac{\epsilon\pi}{(1+\epsilon)} \leq \pi(1+\epsilon)$$

as desired. To summarize, the following choice of $N(x, \epsilon)$ ensures that our algorithm is an almost uniform generator for R:

$$N(\epsilon) = \left\lceil \frac{\log \epsilon^{-1} + \frac{3}{2}\log \pi^{-1} + \log 8}{\lambda_*} \right\rceil \tag{5}$$

Now, for our generator to be fully-polynomial, this expression must be bounded by a polynomial in $|x|$ and $\log \epsilon^{-1}$. By the definition of a p-relation (Definition 2.1), $\pi^{-1} = |R(x)|$ will be bounded by $\exp(q(|x|))$ for some polynomial q, so that term does not present any problem. Thus, in order to have a fully-polynomial generation algorithm, all we must do is bound the spectral gap λ_* below by an inverse polynomial in $|x|$. This property is informally called **rapid mixing**, *i.e.* a family of Markov chains used by some algorithm is called rapidly mixing if the spectral gaps are bounded below by an inverse polynomial in the instance size.

Proving that a family of chains is rapidly mixing involves bounding both $1 - \beta_1$ and $1 - |\beta_{m-1}|$ below. The technique mentioned earlier for getting rid of periodicity also eliminates the problem of bounding β_{m-1}. If we replace the transition matrix P by $(I+P)/2$, then the stationary distribution remains the same and all eigenvalues become non-negative. The new spectral gap λ'_* is simply $(1-\beta_1)/2$. Thus, we may focus our attention solely on bounding $1 - \beta_1$.

Fortunately, there is a useful characterization of the second largest eigenvalue of a self-adjoint operator which provides us with a starting point. We explore this in the next section.

3.2 The Variational Characterization

For a general matrix, the eigenvalues are somewhat difficult to get a handle on. But, for self-adjoint matrices, there is a variational characterization of eigenvalues, known as the Rayleigh-Ritz theorem, that will facilitate further analysis. We present only the case which will be of use to us.[4]

Theorem 3.7 (Rayleigh-Ritz) *Let L be a self-adjoint operator on a finite-dimensional inner product space with inner product $(\cdot, \cdot)$. Suppose the eigenvalues of L are $\lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_n$, and v_0 is an eigenvector of eigenvalue λ_0. Then*

$$\lambda_1 = \inf_{x \perp v_0} \frac{(x, L(x))}{(x, x)} \tag{6}$$

[4] The general version of the Rayleigh-Ritz theorem can be found in [HJ85].

Proof Let $v_0, \ldots, v_n$ be an orthonormal basis of eigenvectors corresponding to the eigenvalues $\lambda_0, \ldots, \lambda_n$, respectively. Then, given $x \perp v_0$, we can write $x = c_1 v_1 + \cdots + c_n v_n$. Now we have

$$(x, L(x)) = \sum_{i=1}^{n} \lambda_i c_i^2 \leq \lambda_1 \sum_{i=1}^{n} c_i^2 = \lambda_1 (x, x)$$

Equality is achieved when $x = v_1$. $\square$

In order to apply this result to our situation, we must introduce some new notation. Let $L = I - P$ be the Laplacian of P, which is also a self-adjoint operator on $L^2(\pi^{-1})$. The eigenvalues of L are, in increasing order, $\lambda_i = 1 - \beta_i$, for $i = 0, \ldots, m-1$, and π is an eigenvector of eigenvalue $\lambda_0 = 0$. Let $Q(x, y) = \pi(x)P(x, y)$. Q is symmetric by the detailed balance equations (1). Define a quadratic form $\mathcal{E}$ on $L^2(\pi^{-1})$ by

$$\mathcal{E}(\phi, \phi) = \frac{1}{2} \sum_{x,y \in S} (\phi(x) - \phi(y)^2) Q(x, y)$$

Finally, for $\phi \in L^2(\pi^{-1})$, let $E(\phi) = \langle \pi\phi, \pi \rangle$ be the expectation of ϕ with respect to π and let $\mathrm{Var}(\phi) = E(\phi^2) - E(\phi)^2$ be the variance of ϕ. Then Theorem 3.7 gives us the following variational characterization of λ_1:

Corollary 3.8 *With the notation as above,*

$$\lambda_1 = \inf \left\{ \frac{\mathcal{E}(\phi, \phi)}{\mathrm{Var}(\phi)} : \phi \textit{ nonconstant} \right\} \tag{7}$$

Proof First, observe that, by the symmetry of Q,

$$\begin{aligned}
\mathcal{E}(\psi, \psi) &= \sum_{x,y \in S} \psi(x)^2 Q(x, y) + \sum_{x,y \in S} \psi(x)\psi(y) Q(x, y) \\
&= \sum_{x \in S} \frac{(\psi(x)\pi(x))^2}{\pi(x)} \sum_{y \in S} P(x, y) + \sum_{x \in S} \frac{\psi(x)\pi(x) \sum_{y \in S} \psi(y)\pi(y) P(y, x)}{\pi(x)} \\
&= \langle \pi\psi, \pi\psi \rangle - \langle \pi\psi, \pi\psi P \rangle = \langle \pi\psi, \pi\psi L \rangle
\end{aligned}$$

Now, since π charges every point, we can make the substitution $x = \pi\psi$ in (6) and take the infimum over all ψ such that $\pi\psi \perp \pi$, *i.e.* $E(\psi)$=0. For ψ such that $E(\psi) = 0$, we have $\mathrm{Var}(\psi) = E(\psi^2) = \langle \pi\psi^2, \pi \rangle = \langle \pi\psi, \pi\psi \rangle$. Together, this gives us

$$\lambda_1 = \inf \left\{ \frac{\mathcal{E}(\psi, \psi)}{\mathrm{Var}(\psi)} : E(\psi) = 0 \right\} \tag{8}$$

Now, suppose we're given any nonconstant ϕ. If we let $\psi = \phi - E(\phi)$, we have $E(\psi) = 0$, $\mathrm{Var}(\psi) = \mathrm{Var}(\phi)$, and

$$\begin{aligned}
\mathcal{E}(\phi, \phi) &= \langle \pi\phi, \pi\phi L \rangle \\
&= \langle \pi\psi, \pi\psi L \rangle + E(\phi)\langle \pi\psi, \pi L \rangle + E(\phi)\langle \pi, \pi\psi L \rangle + E(\phi)^2 \langle \pi, \pi L \rangle \\
&= \mathcal{E}(\psi, \psi)
\end{aligned}$$

All the terms vanished in the above calculation because $\pi L = 0$ and L is self-adjoint. We have shown that ratio $\mathcal{E}(\cdot, \cdot)/\mathrm{Var}(\cdot)$ is equal for ϕ and ψ, so extending the range over all nonconstant ϕ will not change the infimum in (8). $\square$

The variational characterization (7) at first glance seems a bit unwieldy, and bounding it as formidable a task as we began with. However, it turns out that fairly natural geometrical arguments about a Markov chain can yield lower bounds on λ_1 via this characterization.

3.3 Canonical Paths and Bounding the Spectral Gap

In this section, we introduce the notion of canonical paths, the only known tool for finding nontrivial lower bounds on the spectral gap of complex Markov chains. Informally, a set of canonical paths for a Markov chain is a set of paths which follow transitions of the chain, with one path between each pair of states in the chain. If it is possible to construct such a set so that no transition of the chain is used by 'too many' paths, then, intuitively, this means that the chain has no severe bottlenecks which could impede mixing. In 1988, Sinclair and Jerrum [SJ89] verified this intuition, discovering that a good set of canonical paths leads to good lower bounds on the spectral gap. Their proof did not use the variational characterization of the previous section. Instead, they showed that a good set of paths leads to lower bounds on a geometrical quantity known as **conductance** and then related the conductance to the spectral gap. Diaconis and Strook [DS91] later noticed that a good set of paths can lead directly to bounds on the spectral gap via the variational characterization, so the intermediate use of conductance is unnecessary. Naturally, the more direct bounds of Diaconis and Strook tend to be tighter than those obtained with conductance. Lastly, with a slight variation on Diaconis and Strook's techniques, Sinclair [Sin91] has improved the bounds for several chains of interest even more. We follow [Sin91], as these methods provide the tightest bounds for the examples we will give later.

As usual, we must begin by introducing some notation. Let X be a reversible Markov chain on a finite state set S. Then we can define a weighted directed graph $G(X) = (S, E(X))$ whose vertices are the states in S and whose edges are the ordered pairs $(x, y) \in S$ such that $P(x, y) > 0$. We assign edge $e = (x, y)$ the weight $w(e) = \pi(x)P(x, y) = Q(x, y)$, where Q is as defined in the previous section. This graph is known as the **underlying graph** of X.

Definition 3.9 *A set of* **canonical paths** *for X is a set Γ of simple (i.e. not self-intersecting) paths in the underlying graph $G(X)$, one between each ordered pair of distinct vertices. We write $\Gamma = \{\gamma_{xy}\}$ where γ_{xy} is the path between x and y.*

A good set of canonical paths should not overload any transition of the Markov chain. We need a measure of this property. Sinclair suggests the following one:

Definition 3.10 *Let $\Gamma = \{\gamma_{xy}\}$ be a set of canonical paths for X. The* **maximum overloading** $\overline{\rho} = \overline{\rho}(\Gamma)$ *is defined as*

$$\overline{\rho} = \max_{e \in E(X)} \frac{1}{Q(e)} \sum_{\gamma_{xy} \ni e} \pi(x)\pi(y)|\gamma_{xy}|$$

It is easy to see that $\overline{\rho}$ measures the degree of mixing that occurs in the chain at the stationary distribution. If we randomly pick two states according to the stationary distribution and consider the canonical path between them, then the expression being maximized is the expected path length for paths that use e relative to the capacity $Q(e)$. This capacity is the probability that the Markov chain in the stationary distribution will use the transition e. The intuition behind considering path length is that if we can construct short paths which do not overload any given edge, then the chain mixes up very rapidly. Of course, this description is all relative to the stationary distribution and we are concerned with the behavior of the chain before it reaches stationary. It is not surprising, however, that the mixing properties of the chain at stationary reflect its general mixing properties. This relationship is made clear by the following theorem.

Theorem 3.11 ([Sin91]) *For any reversible, homogeneous, aperiodic, irreducible Markov chain X and any set of canonical paths Γ, $\lambda_1 = 1 - \beta_1$ is bounded below by $1/\overline{\rho}$, where β_1 is the second largest eigenvalue of the transition matrix P.*

Proof We use the variational characterization of Corollary 3.8. Let ϕ be any nonconstant function in $\mathbb{R}^S$. Then

$$\begin{aligned}
2\mathrm{Var}(\phi) &= 2E(\phi^2) - 2E(\phi)^2 \\
&= \sum_x \phi(x)^2\pi(x) \sum_y \pi(y) + \sum_y \phi(y)^2\pi(y) \sum_x \pi(x) - 2\sum_x \phi(x)\pi(x) \sum_y \phi(y)\pi(y) \\
&= \sum_{x,y} (\phi(x) - \phi(y))^2 \pi(x)\pi(y) \\
&= \sum_{x,y} \pi(x)\pi(y) \left(\sum_{(u,v)\in\gamma_{xy}} (\phi(u) - \phi(v)) \right)^2 \\
&\leq \sum_{x,y} \pi(x)\pi(y)|\gamma_{xy}| \sum_{(u,v)\in\gamma_{xy}} (\phi(u) - \phi(v))^2 \\
&= \sum_e (\phi(u) - \phi(v))^2 \sum_{\gamma_{xy}\ni e} \pi(x)\pi(y)|\gamma_{xy}| \\
&\leq \sum_{(u,v)\in E(X)} (\phi(u) - \phi(v))^2\, Q(u,v)\overline{\rho}(\Gamma) \\
&= 2\mathcal{E}(\phi,\phi)\overline{\rho}(\Gamma)
\end{aligned}$$

where the first inequality is Cauchy-Schwarz. The result now follows from Corollary 3.8. $\square$

As noted in [DS91], Theorem 3.11 is a discrete version of the Poincaré inequality, which is used to estimate the spectral gap of the Laplacian on a continuous domain. Diaconis and Strook give a similar inequality, which uses a slightly different measure than $\overline{\rho}$. For the examples we are interested in, $\overline{\rho}$ yields slightly better bounds.

With Theorem 3.11, our task is reduced to coming up with a 'good' set of canonical paths. How 'good' must these paths be? Consider the case, where π is the uniform distribution, *i.e.* $\pi(x) = 1/|S|$ for all $x \in S$. Then

$$\overline{\rho} \leq \frac{\left(\max_{e=(u,v)} \#\{\gamma_{xy} : e \in \gamma_{xy}\}\right) L}{P(u,v)|S|} \tag{9}$$

where L is the length of the longest path in Γ. In order to have a fully-polynomial almost-uniform generator in the sense of Definition 2.2, it suffices to bound this quantity above by a polynomial in the instance size. It will usually be obvious in the examples we consider that all the transition probabilities $P(u,v)$ are bounded below by a polynomial in the instance size. So the main task is to show that the number of paths using any transition is bounded above by $|S|$ times a polynomial in the instance size. Jerrum and Sinclair introduced a clever technique for doing this: for each transition e, construct an *injective* mapping from the set of paths using transition e to the set of states, possibly crossed with a bit of extra information. Often, the canonical path from x to y can be constructed as a sequence of local changes in structure from x to y in a specified order. Then, for any transition e on this path, we attempt to construct a complementary state $\sigma_e(x,y)$, which agrees with y in the places where the changes have already occurred and agrees with x in the places where the changes have not yet occurred. If we can do this, then, intuitively, we should be able to reconstruct x and y from e and $\sigma_e(x,y)$, completing the proof of rapid mixing. In the next section, we shall see how this technique works in a couple of complex examples of interest.

4 Applications

In this section, we present two of the most significant and interesting applications of the theory developed in this essay — the approximation of the permanent and the estimation of the partition function in the Ising model of statistical physics. In both of these cases, constructing a rapidly mixing Markov chain and proving that it is in fact rapidly mixing is only part of the picture; a significant amount of extra work and ingenuity is required to use the chains to solve the relevant problems. However, as this is an essay on rapid mixing, we shall concentrate on the convergence rates of the chains in question. We will also sketch how the chain is used to solve the problem at hand, but omit detailed proofs.

4.1 Approximating the Permanent

The **permanent** of an $n \times n$ matrix $A = (a_{ij})$ is defined as

$$\text{per}(A) = \sum_{\sigma \in S_n} \prod_{i=1}^{n} a_{i,\sigma(i)}$$

Like its relative, the determinant, the permanent arises naturally in a number of areas of mathematics and its applications. However, unlike the determinent, which can be computed efficiently by Gaussian elimination, efficient algorithms to evaluate the permanent have been elusive. In 1979, Valiant explained this apparent intractability by proving that evaluating the permanent of a $0-1$ matrix is in fact $\mathcal{NP}$-hard. With an algorithm for exact computation out of the question, it is natural to seek approximation algorithms.

The first sign that Markov chain techniques might be useful in computing the permanent comes from the observation that, for $0-1$ matrices, the permanent counts a natural combinatorial structure. Precisely, if we consider A to be the adjacency matrix of a $2n$-vertex bipartite graph $G = (U \cup V, E)$, [5] then per(A) counts the number of **perfect matchings** in G, that is, subsets F of E such that every vertex in $U \cup V$ is an endpoint of exactly one edge in F. Note that any perfect matching must contain exactly n edges. As it is not hard to see that perfect matchings are self-reducible, the result of [JVV86] mentioned in Section 2.1 tells us that efficiently approximating the permanent is computationally equivalent to efficiently generating perfect matchings in bipartite graphs almost uniformly at random.

In 1986, Broder suggested a Markov chain for uniformly generating perfect matchings and near-perfect matchings, where a **near-perfect matching** in a $2n$-vertex bipartite graph $G = (U \cup V, E)$ is a set of edges $F \subset E$ such that no two edges in F share an endpoint and $|F| = n-1$. Broder claimed that his chain is rapidly mixing whenever G is **dense**, that is, every vertex in G has degree $\geq n/2$. Broder's proof of rapid mixing used a complex coupling argument, which was later found to be fallacious. [Mih89, Bro88] In 1989, however, Jerrum and Sinclair used their new techniques to demonstrate that Broder's chain is in fact rapidly mixing. [JS89] The simplicity and success of their proof can be taken as evidence that these techniques have tremendous advantages over other approaches to estimating mixing rates, particularly in complex combinatorial problems.

As we shall see, the proof that Broder's chain is rapidly mixing works only for dense graphs. Unfortunately, perfect matchings appear to lose their self-reducibility in this restricted case, so we cannot simply appeal to the result of [JVV86] to get a fully-polynomial approximation scheme for the dense permanent from our fully-polynomial generator. However, it turns out that approximate counting still reduces to almost uniform generation in this case, as was shown by Broder. Writing $M_n(G)$ for the set of perfect matchings in a graph

[5] When we speak of a $2n$-vertex bipartite graph, we will always mean one in which both sides have n vertices, *i.e.* $G = (U \cup V, E)$ where $|U| = |V| = n$ and $E \subset U \times V$.

G and $M_{n-1}(G)$ for the set of near-perfect matchings, Broder's result can be stated as follows:

Theorem 4.1 (Broder [Bro86]) *Suppose that there is a fully-polynomial almost uniform generator for $M_n(G) \cup M_{n-1}(G)$ for all dense bipartite graphs G. Then there is a fpras for $|M_n(G)|$ for all such graphs.*

Thus, to obtain an fpras for the dense permanent, it suffices to construct a Markov chain $X(G)$ with uniform stationary distribution on state set $S(G) = M_n(G) \cup M_{n-1}(G)$ for each dense bipartite graph $G = (U \cup V, E)$ and to show that this family is rapidly mixing. We now describe Broder's chain, which achieves this goal: While at state $M \in S(G)$, choose $e = (u, v) \in E$ uniformly at random and do the following:
1. If $M \in M_n(G)$ and $e \in M$, move to state $M - e$. (*Type 1 transition*)
2. If $M \in M_{n-1}(G)$ and both u and v are unmatched in M, move to state $M + e$. (*Type 2 transition*)
3. If $M \in M_{n-1}(G)$, u is unmatched in M and v is matched to $w \in M$, then move to state $M + e - (w, v)$. Similarly, if $M \in M_{n-1}(G)$, v is unmatched in M, and u is matched to $w \in V$, then move to state $M + e - (u, w)$. (*Type 0 transition*)
4. If none of the above apply, remain in state M.

It is clear that $P(M_1, M_2) = P(M_2, M_1)$ for all $M_1, M_2 \in S(G)$. Thus, by Lemma 2.12, this chain has a uniform stationary distribution and is time-reversible with respect to this distribution. Since there is always a positive probability of remaining in the same state, the chain is also aperiodic. So, as long as we have irreducibility, we may uses the techniques of the preceeding sections to demonstrate rapid mixing.

Theorem 4.2 ([JS89, DS91]) *For a dense, bipartite graph G on $2n$ vertices, Broder's Markov chain $X(G)$ is irreducible and has $\lambda_1 = 1 - \beta_1 \geq 1/6n^7$.*

Corollary 4.3 *There is a fpras for $|M_n(G)|$ for all dense bipartite graphs G.*

Proof (of Corollary 4.3) It is obvious that, given a starting state, Broder's chain can be simulated efficiently with a probabilistic algorithm. Determining whether there is a near-perfect matching and, if there is one, finding it can both be done in polynomial-time. There are several well-known algorithms to solve these problems. [AHU83] Thus, by the discussion at the end of Section 3.2, there is a fully-polynomial almost uniform generator for perfect and near-perfect matchings. Broder's reduction of Theorem 4.1 yields the desired fpras. □

Proof (of Theorem 4.2) We construct a set of canonical paths between the states in $S(G)$ so that the maximum overloading is not too large. Naturally, the existence of such a set of paths implies irreducibility. We will initially construct two types of canonical paths: one between any state $M \in S(G)$ and its 'nearest' perfect matching $\overline{M}$, and one between any pair of perfect matchings.

The canonical path between any pair $I, F \in S(G)$ will then consist of three segments: the canonical path between I and $\overline{I}$, followed by the canonical path between $\overline{I}$ and $\overline{F}$, followed by the canonical path between $\overline{F}$ and F.

We first describe the canonical path between M and $\overline{M}$.
Case 1: M is a perfect matching. Then $\overline{M} = M$, and the path has 0 transitions.
Case 2: M is a near-perfect matching with vertices $u \in U, v \in V$ unmatched and $e = (u, v) \in E$: Then $\overline{M} = M + e$ and the path from M to $\overline{M}$ consists of a single type 2 transition.
Case 3: M is a near-perfect matching with vertices $u \in U, v \in V$ unmatched and $e = (u, v) \notin E$: Then, since both u and v have degree $\geq n/2$, there must be a neighbor y of u which is matched to a neighbor x of v by M. Choosing any such x and y, $\overline{M} = (M + (u, y) - (x, y)) + (v, x)$, and the canonical path from M to $\overline{M}$ consists of a type 0 transition followed by a type 2 transition as indicated. The canonical path from $\overline{M}$ to M is the same, following the transitions in reverse.

Now we describe the canonical path between any two perfect matchings I and F. Consider the symmetric difference $I \oplus F$. Every vertex v must have degree either 0 or 2 in $I \oplus F$, as both I and F have exactly 1 edge incident to v. This means that $I \oplus F$ is a union of vertex-disjoint cycles, whose edges alternate between I and F. In order to describe the canonical path between I and F, we assume that there's some fixed (but arbitrary) ordering on the set of all cycles in G and that in each cycle a start vertex is specified. So, we can write $I \oplus F = C_1 \oplus C_2 \oplus \cdots C_k$, where $C_1, \ldots, C_k$ are disjoint cycles, in order. The canonical path from I to F consists of unwinding these cycles in order. The unwinding of a cycle C proceeds as follows: Let the vertices in C be (in order) $u_0, v_0, u_1, v_1, \ldots, u_m, v_m$, where u_0 is the start vertex of C. We may assume that $(u_i, v_i) \in I$ for all i, with F containing the rest of the edges of C, by reversing the orientation of the cycle if necessary. The unwinding of C starts with a type 1 transition to remove (u_0, v_0), continues with a sequence of m type 0 transitions to add (u_{i+1}, v_i) and remove (u_{i+1}, v_{i+1}), and ends with a type 2 transition to add (u_0, v_m).

We now bound the number of canonical paths which use a given transition of the Markov chain. We begin by counting the paths which use a specified transition in either the initial or final segment. Note that any one of the transitions used in the canonical path from M to $\overline{M}$ completely determines $\overline{M}$: A type 0 transition (M_1, M_2) is used only if $\overline{M} = M_2$ and a type 2 transition (M_1, M_2) is used only if $\overline{M} = M_2 + (u, v)$, where u and v are the unmatched vertices in M_2. Also, $\overline{M}$ shares at least $n - 2$ edges with M. For any perfect matching N, let us count how many matchings $M \in S(G)$ have $\overline{M} = N$: there is exactly 1 which shares n edges with N; there are exactly n matchings which share $n - 1$ edges with N; and there are strictly less than $2\binom{n}{2}$ which share $n - 2$ edges with N, unless G is the complete bipartite graph on $2n$ vertices. In case G is the complete graph, there are exactly n near-perfect matchings whose nearest perfect matching is N. This gives a total of at most n^2 matchings whose nearest perfect matching is N. Thus, at most $n^2|S(G)|$ canonical paths between arbitrary matchings $I, F \in S(G)$ use any given transition t in the initial segment. (t determines $\overline{I}$, leaving a most n^2 possibilities for I and $|S(G)|$ possibilities for

F). Similarly, at most $n^2|S(G)|$ paths use any given transition in their final segment.

To estimate the number of paths which use a given transition in the central segment, we make use of Jerrum and Sinclair's innovative injective mapping technique. Given a transition t of the Markov chain, we will define a mapping σ_t from the set of pairs of perfect matchings $\langle I, F\rangle$ whose canonical path uses transition t to $S(G)$. Intuitively, $\sigma_t(I, F)$ will be a matching which agrees with I on all the edges which have already been handled and agrees with F on all the edges which haven't. Before defining σ_t precisely, let us make a few observations. Say t is a transition from M to M'. Then, M and M' generally agree with I on the edges already handled and F on the edges not yet handled. So $S = I \oplus F \oplus (M \cup M')$ should be essentially the complementary matching we are seeking.

We need to check that S is in fact a perfect or near-perfect matching. Note that, as the canonical paths do not touch any edges outside of $I \oplus F$, we have $I \cap F \subset M \cup M' \subset I \cup F$. In addition, $|M \cup M'| = n$ by the form of the transitions. Thus,

$$\begin{aligned}
|S| &= |I \cap F| + |I \setminus (M \cup M')| + |F \setminus (M \cup M')| \\
&= |I \cap F \cap (M \cup M')| + |I| - |I \cap (M \cup M')| + |F| - |F \cap (M \cup M')| \\
&= 2n - |(I \cup F) \cap (M \cup M')| \\
&= 2n - |M \cup M'| = n
\end{aligned}$$

Now suppose some vertex u is adjacent to two edges in S. Both of these edges must be in $I \cup F$, which implies that exactly one is in I and exactly one is in F, as I and F are perfect matchings. So neither edge is in $M \cup M'$. By the description of our canonical paths, the only way this can happen is if t is a type 0 transition and u is the start vertex of the cycle currently being unwound. Thus, if we denote by $e_{I,t}$ the edge of I incident to u, the following definition ensures that $\sigma_t(I, F)$ is a perfect or near-perfect matching:

$$\sigma_t(I, F) = \begin{cases} (I \oplus F \oplus (M \cup M')) - e_{I,t} & \text{if } t \text{ is of type 0} \\ (I \oplus F \oplus (M \cup M')) & \text{otherwise} \end{cases}$$

Now, we shall demonstrate that σ_t is in fact injective. First note that $I \oplus F$ can be recovered from $\sigma_t(I, F)$ and t, by taking $\sigma_t(I, F) \oplus (M \cup M')$, and, if t is of type 0, adding the unique edge needed to make this a disjoint union of cycles. Which cycles in $I \oplus F$ have already been unwound is determined by the fixed ordering on all cycles of G and the current transition. As M agrees with I on all the cycles already unwound and with F on all the cycles not yet unwound, we can determine which edges of any cycle in $I \oplus F$, other than the one currently being unwound, are in I and which are in F. Which edges of the current cycle belong to I can be determined by the current transition: Unless t is type 2, $M - M'$ consists of a single edge in the current cycle belonging to I; and if t is of type 2, then $M' - M$ consists of a single edge in the current cycle belonging to F. The remaining edges of I and F are simply those in $I \cap F$, which can be obtained by taking the edges in M not part of any of the cycles in $I \oplus F$. We have shown that both I and F can be recovered from $\sigma_t(I, F)$, so σ_t is injective. So, for any transition t, there are at most $|S(G)|$ pairs of perfect matchings I, F such that the canonical path between them uses t. Now, since there are at most n^2 matchings $A \in S(G)$ such that $\overline{A} = I$ and at most n^2 matchings $B \in S(G)$ such that $\overline{B} = F$, there are at most $n^4|S(G)|$ matchings $A, B \in S(G)$ such that the canonical path between A and B uses the transition t in its middle segment.

Combining our bounds for the three segments, we have that at most $(2n^2 + n^4)|S(G)| \leq 3n^4|S(G)|$ canonical paths use any given transition of the Markov chain. In addition, the longest canonical path is of length at most $n + 3 \leq 2n$ and all the non-zero transition probabilities of the chain are $1/|E| \geq 1/n^2$. Applying the bound (9), we have $\overline{\rho} \leq n^2(2n)(3n^4) = 6n^7$. The result now follows from Theorem 3.11. $\square$

This proof demonstrates how powerful the techniques of Jerrum and Sinclair are. Their relatively simple and intuitive idea of finding a set of canonical paths that do not use any transition too often, combined with their injective mapping technique, allows us to obtain strong bounds on the eigenvalues of a chain which we have very little information about. (We don't even know the number of states!) It is not surprising that these techniques have found a wide variety of other applications, one of which we will see in the next section.

4.2 Estimating the Partition Function of the Ising Model

The Ising model of statistical physics is one of the places that the techniques of this essay has had a dramatic impact. It turns out that most of the important physical information in this model can be obtained from knowledge of a few easily described combinatorial quantities, including, most significantly, the values of the partition function at various temperatures. Efficient algorithms to compute these quantities have been elusive. Indeed, exactly computing the partition function is $\mathcal{NP}$-hard, as shown in [JS93]. In 1993, Jerrum and Sinclair, again using rapidly mixing Markov chains, achieved the breakthrough of finding polynomial-time algorithms to estimate the partition function of the Ising model, along with some other important quantities.

We do not discuss any of the physics related to the Ising model in this essay but jump immediately to the relevant combinatorial problems and their solution. The Ising model can be described as follows: there are a set of **sites** numbered $1, \ldots, n$ and, between each pair of sites i, j, an associated **interaction energy** V_{ij}. We let $G = ([n], E)$ be the undirected graph corresponding to nonzero interaction energies, *i.e.* $E = \{\{i, j\} : V_{ij} \neq 0\}$. A **configuration** is a function σ which assigns each site i a spin $\sigma_i \in \{\pm 1\}$. The **energy** of a configuration is given by

$$H(\sigma) = -\sum_{\{i,j\} \in E} V_{ij}\sigma_i\sigma_j - B\sum_{k \in [n]} \sigma_k$$

The **partition function** Z is defined as

$$Z = Z(V_{ij}, B, \beta) = \sum_{\sigma \in \{\pm 1\}^n} \exp(-\beta H(\sigma))$$

We shall present an efficient algorithm to compute Z in the **ferromagnetic** case, *i.e.* when all the interaction energies V_{ij} are nonnegative. The precise computational problem is "Given a real symmetric matrix V_{ij} of nonnegative interaction energies, a positive real number β and a real number B, output (an approximation to) the partition function $Z(V_{ij}, B, \beta)$." There is an issue of represention of real numbers here. Naturally, real numbers cannot all be expressed exactly, nor can real arithmetic be carried out exactly. As in [JS93], we shall ignore these issues here. As we are only looking for an approximate value for Z, a certain degree of inaccuracy in real arithmetic can be tolerated, and it turns out that the computations can be carried out to the required accuracy without a significant impact on efficiency.

The obvious approach to computing Z via Markov chain techniques is to construct a random walk on the set of configurations so that the stationary distribution has $\pi(\sigma)$ proportional to $\exp(-\beta H(\sigma))$. This can be easily achieved by a Markov chain whose transitions involve changing the sign of only a single σ_i. Unfortunately, this chain is not rapidly mixing.

The first important step in solving this problem is to transform the partition function, which involves a sum over spins, into a quantity which involves a sum over spanning subgraphs of G. A **(spanning) subgraph** of $G = ([n], E)$ is a graph $H = ([n], X)$ such that $X \subset E$. We shall identify spanning subgraphs by their sets of edges. Writing $\lambda_{ij} = \tanh \beta V_{ij}$ and $\mu = \tanh \beta B$, we assign each configuration X a **weight**

$$w(X) = \mu^{|\mathrm{odd}(X)|} \prod_{\{i,j\} \in X} \lambda_{ij}$$

where $\mathrm{odd}(X)$ denotes the set of vertices of odd degree in X. Using these weights, we can define a new partition function

$$Z' = \sum_{X \subset E} w(X)$$

A classical result relates this new partition function to the old one.

Theorem 4.4 *$Z = AZ'$, where $A = (2\cosh \beta B)^n \prod_{\{i,j\} \in E} \cosh \beta V_{ij}$.*

As the quantity A can clearly be computed in time polynomial in n, Theorem 4.4 reduces the problem of estimating Z to estimating Z'. The fact that Z' is a sum of easily-computed weights over the set of subgraphs of G suggests that perhaps a random walk on the set of subgraphs of G with stationary distribution $\pi(X) = w(X)/Z'$ may yield an fpras for Z'. Such a chain is easily constructed. We first describe the chain and prove that it is rapidly mixing, and later sketch how it can be used to estimate Z'.

We describe the chain by its transition probabilities. The state set of our Markov chain will be S — the set of spanning subgraphs of G — or, equivalently, the set of subsets of E. For any two two distinct subgraphs $X, X' \in S$

$$P(X,X') = \begin{cases} 1/2m & \text{if } |X \oplus X'| = 1 \text{ and } w(X') \geq w(X) \\ w(X')/2mw(X) & \text{if } |X \oplus X'| = 1 \text{ and } w(X') < w(X) \\ 0 & \text{if } |X \oplus X'| > 1 \end{cases}$$

where $m = |E|$ is the number of nonzero interactions.

The probability $P(X, X)$ of staying at the same state is implicitly given by

$$P(X,X) = 1 - \sum_{X' \neq X} P(X,X') \geq 1 - \sum_{|X' \oplus X| = 1} \frac{1}{2m} \geq \frac{1}{2}$$

It is clear that this Markov chain can be simulated efficiently and that a starting state ($\emptyset$) can be found. The chain is irreducible, as any two subgraphs X and Y are connected by a path of $X \oplus Y$ transitions. In addition, since there is a nonzero probability of staying in the same place, the chain is aperiodic.

If we write $\pi(X) = w(X)/Z'$ and $Q(X, X') = \pi(X)P(X, X')$, it is easy to see that $Q(X, X') = \min\{\pi(X), \pi(X')\}/2m$ for all pairs X, X' of distinct states which have non-zero transition probabilities. As $Q(X, X')$ is symmetric in X and X', Lemma 2.12, tells us that π is a stationary distribution of the chain and the chain is time-reversible with respect to π. Thus, we may use the techniques of this essay to show that the chain is rapidly mixing. We will obtain the following theorem:

Theorem 4.5 ([JS93, Sin91]) *For the Markov chain described above, $\lambda_1 = 1 - \beta_1$ is bounded below by $\mu^6/2m^2$, where β_1 is the second-largest eigenvalue of the transition matrix P and $m = |E|$ is the number of nonzero interactions.* [6]

Corollary 4.6 *There is a probabilistic algorithm M, which takes inputs $0 \leq \lambda_{ij} \leq 1, 0 < \mu \leq 1$, and $\epsilon > 0$, and outputs a random variable $M = M(\lambda_{ij}, \mu, \epsilon) \in S \cup \{N\}$ such that*
1.$Pr[M = N] < 1/2$, and
2.$\|(M|M \neq N) - \pi\|_v < \epsilon$, where $(M|M \neq N)$ is the probability distribution on S obtained by conditioning M on the event $M \neq N$
Moreover the runtime of M is bounded by a polynomial in $m, \mu^{-1}, \log \epsilon^{-1}$, where m is the number of nonzero λ_{ij}.

Proof(of Corollary 4.6) This follows immediately from Theorems 4.5 and the expression (3), along our earlier observation that the chain can be efficiently simulated. Again, allowing a probability of failure is simply a technicality of the computational model. □

Proof(of Theorem 4.5) By Theorem 3.11, it suffices to come up with a

[6] Jerrum and Sinclair actually prove the slightly stronger result $\lambda_1 \geq \mu^4/2m^2$, which can be obtained by being slightly more careful in the same proof.

good set of canonical paths as described in Section 3.3. Since we are no longer dealing with a uniform stationary distribution, it does not suffice to merely construct a set of paths which do not use any edge too often; the weights must be considered, too. We need $\sum_{\gamma_{xy} \ni t} \pi(x)\pi(y)|\gamma_{xy}|$ to be bounded relative to $Q(t)$ for each transition t. We will do this by a modification of the injective mapping technique used in the previous section. For any transition $t = (T, T')$, we will still construct an injective mapping σ_t from the pairs $I, F \in S$ whose canonical paths use t to S. Our injective mapping will satisfy the following condition:

$$\pi(\sigma_t(I, F))Q(T, T') \geq \frac{\mu^6}{2m}\pi(I)\pi(F) \tag{10}$$

Once we have done this, we will have

$$\begin{aligned} \sum_{\gamma_{xy} \ni t} \pi(x)\pi(y)|\gamma_{xy}| &\leq \frac{2mLQ(T, T')}{\mu^6} \sum_{\gamma_{xy} \ni t} \pi(\sigma_t(x, y)) \\ &\leq \frac{2mLQ(T, T')}{\mu^6} \end{aligned}$$

where L is the length of the longest path. The last inequality above follows from injectivity of σ_t. This gives us $\overline{\rho} \leq 2mL/\mu^6$.

Now, let's construct these paths. Consider two subgraphs $I, F \subset E$. The symmetric difference $\Delta = I \oplus F$ is some arbitrary subgraph of G. As all graphs have an even number of vertices of odd degree, we may write the number of vertices of odd degree in Δ as $2k$. Suppose $k > 0$. Consider any vertex u of odd degree in Δ. The connected component of Δ containing u must have an even number of vertices of odd degree, so there must be another vertex v of odd degree in Δ. We may connect u and v by an open **trail** T_1, which is a path in Δ with no repeated edges. If $k > 1$, we can repeat the same argument with $\Delta_1 = \Delta - T_1$, which has $2k - 2$ vertices of odd degree, finding a trail T_2 which connect two vertices of odd degree and obtaining a graph $\Delta_2 = \Delta - T_1 - T_2$ with $2k - 4$ vertices of odd degree. If we continue inductively, we obtain k edge-disjoint trails $T_1, \ldots, T_k$ in Δ, whose removal produces a graph Δ_k, in which all vertices have even degree. A similar inductive procedure enables us to cover Δ_k with some number of disjoint **circuits** $C_1, \ldots, C_r$, which are simply trails which have the same start and end vertex: Remove any edge e of Δ_k to obtain a graph with exactly two vertices of odd degree. These must be in the same connected component, so we may connect them by a trail. Adding e to this trail produces a circuit C_1, and $\Delta_k - C_1$ also has all vertices of even degree, so we may proceed inductively. [7]

Thus we have a covering of Δ by disjoint trails $T_1, \ldots, T_k$ and circuits $C_1, \ldots, C_r$, where $2k = \text{odd}(\Delta)$. We assume that for each subgraph Δ of G, a

[7] Actually, a standard theorem of graph theory tells us that Δ_k may be covered by c disjoint circuits, where c is the number of connected components of Δ_k. As the number of circuits used does not matter to us, we do not appeal to this result.

fixed choice is made for the covering of Δ by trails and circuits in this manner and the ordering of these trails and circuits is also fixed. In addition, in each trail and circuit, a start vertex (which must be an endpoint in the case of a trail) and orientation must be fixed. Now, like the canonical paths for matchings in the previous section, the canonical path from I to F will consist of 'unwinding' each of the trails and circuits in succession. The unwinding of a trail or circuit proceeds by considering each edge e in order, starting with the one adjacent to the start vertex in the direction of orientation, removing e if e is in $I - F$, and adding e if e is in $F - I$.

For any transition t of the Markov chain, the injective mapping σ_t is similar to the one used in the matching chain. If t takes subgraph T to subgraph T', then $\sigma_t(I, F) = I \oplus F \oplus (T \cup T')$. As in the matching chain, this is essentially a 'complementary' subgraph which agrees with I on the edges already handled and agrees with F on the edges not yet handled. The injectivity of σ_t is not too difficult to see. From knowledge of t and $\sigma_t(I, F)$, we can recover $I \oplus F = (T \cup T') \oplus \sigma_t(I, F)$. The fixed covering of $I \oplus F$ by trails and circuits is then determined. The transition t determines our current position in the unwinding, from which we can tell which edges of $I \oplus F$ belong to I — for edges $e \in I \oplus F$ that have already been handled, e is in I iff e is not in T; for edges that have not yet been handled, e is in I iff e is in T; and for the edge e that is currently being added or removed, e is in I iff it is being removed. As the edges that are in both or neither of I and F are never touched in the canonical path, the remainder of I and F is simply $T - (I \oplus F)$, which we can also determine. Thus σ_t is injective as desired.

We now wish to show that (10) holds. Let $U = \sigma_t(I, F)$. Since $Q(T, T') = \min\{\pi(T), \pi(T')\}/2m$ and $\pi(X) = w(X)/Z'$, (10) is equivalent to the following two inequalities:

$$w(U)w(T) \geq \mu^6 w(I)w(F)$$
$$w(U)w(T') \geq \mu^6 w(I)w(F)$$

Now, $w(X) = \mu^{\mathrm{odd}(X)}\Lambda(X)$, where $\Lambda(X) = \prod_{\{i,j\}\in X} \lambda_{ij}$. So it suffices to demonstrate that the following two pairs of inequalities hold:

$$\Lambda(U)\Lambda(T) \geq \Lambda(I)\Lambda(F) \text{ and } \Lambda(U)\Lambda(T') \geq \Lambda(I)\Lambda(F) \tag{11}$$

$$\begin{gathered} \mathrm{odd}(U) + \mathrm{odd}(T) \leq \mathrm{odd}(I) + \mathrm{odd}(F) + 6, \text{ and} \\ \mathrm{odd}(U) + \mathrm{odd}(T') \leq \mathrm{odd}(I) + \mathrm{odd}(F) + 6 \end{gathered} \tag{12}$$

The inequalities (11) will certainly hold if we show that the *multisets* $U \sqcup T$ and $U \sqcup T'$ are contained in the multiset $I \sqcup F$, as all the λ_{ij}'s are between 0 and 1, where $\sqcup$ denotes union with multiplicity. We show that $A = U \sqcup (T \cup T')$ is contained in $B = I \sqcup F$. We know that $U \oplus (T \cup T') = I \oplus F$, so the edges that occur with multiplicity one are the same in A and B. We also have that $T \cup T' \subset I \cup F$ and consequently $U \subset I \cup F$, so any edge that appears in A also appears in B. Since every edge appears with multiplicity at most 2 in A, this proves that A is contained in B. Thus, (11) holds for both T and T'.

The inequalities (12) can also be shown with an argument about multisets. It suffices to show that $|A| \leq |B| + 6$, where $A = \mathrm{odd}(U) \sqcup (\mathrm{odd}(T) \cup \mathrm{odd}(T'))$ and $B = \mathrm{odd}(I) \sqcup \mathrm{odd}(F)$. Before we show this, observe that if $G, H \subset S$ are both subgraphs of G and $\mathrm{par}_G(w), \mathrm{par}_H(w) \in \{0, 1\}$ denote the parity of the degree of w in G and H respectively, then $\mathrm{par}_{G \oplus H}(w) = \mathrm{par}_G(w) \oplus \mathrm{par}_H(w)$, where the last $\oplus$ indicates sum modulo 2.

Now, let w be a vertex which does not appear in B. Then w has even degree in both I and F, so it has even degree in $I \oplus F$ and is not the endpoint of a trail. The only times the degree of w can become odd during the canonical path from I to F are when w lies on the trail or circuit being unwound and exactly one of the two edges on this trail or circuit incident to w has been handled and the other hasn't. Clearly, this only can happen if w is one of the two endpoints of the edge currently being changed or if w is the start vertex of a circuit currently being unwound. Now, as T and T' differ in only the edge being changed, w can have odd degree in $T \cup T'$ only in these three cases. Since $U \oplus (T \cup T') = I \oplus F$ and w has even degree in $I \oplus F$, w has odd degree in U if and only if w has odd degree in $(T \cup T')$. Again, this can only happen if w is one of the three vertices mentioned above.

Now let w be a vertex which appears in B with multiplicity 1, *i.e.* w has odd degree in exactly one of I and F. So w has odd degree in $I \oplus F = U \oplus (T \cup T')$ and therefore in exactly one of either U or $T \cup T'$. If w has odd degree in $T \cup T'$, then it has multiplicity at most 1 in A. If w has even degree in $T \cup T'$, then it can only have odd degree in either T or T' if it is an endpoint of the edge being changed. So w has multiplicity 1 in A unless w is an endpoint of the edge being changed.

We have shown that only three vertices can have greater multiplicity in A than in B — the two endpoints of the edge being changed and the start vertex of a circuit currently being unwound. As every vertex in A has multiplicity at most 2, this tells us that $|A| \leq |B| + 6$.

To complete the proof, simply note that the length of the longest path is at most m, so we have $\overline{\rho} \leq 2m^2/\mu^6$ and we can apply Theorem 3.11. □

Now, we have an efficient generator which produces subgraphs of of the interaction graph with probability nearly proportional to their weights. It is not clear how to use this generator to achieve our goal of estimating Z', the sum of the weights. Accomplishing this requires a bit of ingenuity and work. Here, we sketch how Jerrum and Sinclair use the generator to estimate Z'.

Theorem 4.7 ([JS93]) *There is a fpras for Z', and hence also for Z, the partition function of the ferromagnetic Ising model.*

Proof (sketch) The first observation to make is that, given our efficient generator which produces subgraphs with a distribution near π, we should be able to estimate the expectations of functions on the space of subgraphs with respect to π. Specifically, they prove that if f is a nonnegative, real-valued function on the space S of subgraphs, then there is an algorithm which takes inputs $\xi > 0$ and

$\eta > 0$ and, by making a number of calls to the generator, produces an output Y which approximates $E(f)$ within ratio $1+\xi$ with probability greater than $1-\eta$. Moreover, the runtime of the generator is bounded by a polynomial in $\xi^{-1}, \log \eta^{-1}, \max(f)/E(f), m$ and μ^{-1}. A detailed proof of this involves some tedious but elementary statistics.

The second observation required is that we can relate the values of Z' at different values of μ by the expectation of a simple function. As we will be holding the λ_{ij}'s constant and varying μ, we write $Z' = Z'(\mu)$. Specifically, if we let $f(X) = (\mu_1/\mu_0)^{|\mathrm{odd}(X)|}$, then the expectation of f with respect to the distribution on S obtained by using $\mu = \mu_0$ is

$$E_{\mu_0}(f) = \frac{\sum_X \left(\frac{\mu_1}{\mu_0}\right)^{|\mathrm{odd}(X)|} \mu_0^{|\mathrm{odd}(X)|} \Lambda(X)}{Z'(\mu_0)} = \frac{Z'(\mu_1)}{Z'(\mu_0)}$$

Now, noting that we can compute $Z'(1) = \sum_X \prod_{(i,j)\in E} \lambda_{ij} = \prod_{(i,j)\in E}(1+\lambda_{ij})$ easily, the obvious thing to do is to take $\mu_0 = 1$ and $\mu_1 = \tanh \beta B$ and get $Z'(\tanh \beta B) = E_{\mu_0}(f) Z'(1)$. Unfortunately, $\max(f)/E(f)$ can be exponentially large in this case. The solution is to observe that if $\mu_1 < \mu_0 \leq \mu_1 + 1/n$, where n is the number of sites, then $\max(f)/E(f)$ is bounded above by a constant. Thus, by 'stepping down' μ from 1 to $\tanh \beta B$ in steps of size $1/n$, with a last step of size at most $1/n$, we can successively estimate $Z'(\mu)$, finally obtaining $Z'(\tanh \beta B)$, as desired. This completes the proof, except that the algorithm is inefficient when $\tanh \beta B$ is exponentially small, as the runtime of the generator depends on μ^{-1}. Jerrum and Sinclair show how to remove this dependence in their paper. $\square$

This example illustrates how, with a bit of creativity, the efficient generators obtained from rapidly mixing Markov chains can be used to solve a wide variety of computational problems. In their paper, Jerrum and Sinclair show how the subgraphs-generator of Theorem 4.6 can also be used to compute the mean energy and mean magnetic moment of the Ising model efficiently. The approach is similar to that in Theorem 4.7: relate these operators to the expectations of simple functions on the probability space of subgraphs and show that we can estimate the expectations efficiently. Algorithms such as these enable statistical physicists to run computational experiments with a rigorous mathematical foundation. Hopefully, further work like this will eliminate the threat of faulty results due to incorrect guesses about how long a Markov chain needs to be simulated.

5 Conclusion

The examples presented in last two sections demonstrate the usefulness of rapidly mixing Markov chains and the power of the techniques recently developed to analyze them. Although it was not trivial to prove rapid mixing in the two applications, the intuitive geometric ideas of canonical paths and the injective mapping technique remain at the heart of the proofs. It would seem that these techniques are quite generally applicable, and this has been verified by the sudden proliferation of algorithms based on rapidly mixing Markov chains. Other recent applications include approximately computing the partition function in the monomer-dimer model of statistical physics, estimating the volume of an convex set in Euclidean space, approximately counting the number of graphs with a given degree sequence, and couting the number of Hamiltonian cycles in a dense graph. [JS89, Kan94, SJ89]. However, there are still many open approximation problems which may or may not admit Markov chain solutions. These include counting independent sets in various restricted classes of graphs, approximating the permanent for general graphs, and estimating the chance that a probabilistic graph remains connected.[Kan94, Vad95] Thus, the study of rapidly mixing Markov chains is a rich and active field. The techniques presented in this essay have already solved several interesting and important problems; perhaps they can be used to solve many more.

References

[AHU83] Alfred V. Aho, John E. Hopcroft, and Jeffrey D. Ullman. *Data Structures and Algorithms*, chapter 7.5, pages 246–250. Addison-Wesley, 1983.

[Bro86] A.Z. Broder. How hard is it to marry at random? (on the approximation of the permanent). In *Proceedings of the 18th Annual ACM Symposium on the Theory of Computing*, pages 50–58. Association for Computing Machinery, 1986.

[Bro88] A.Z. Broder. How hard is it to marry at random? (on the approximation of the permanent), erratum. In *Proceedings of the 20th Annual ACM Symposium on the Theory of Computing*, page 551. Association for Computing Machinery, 1988.

[DF89] M. E. Dyer and A. M. Frieze. On the complexity of computing the volume of a polyhedron. *Society for Industrial and Applied Mathematics Journal on Computing*, 18:205–226, 1989.

[DS91] Persi Diaconis and Daniel Strook. Geometric bounds for eigenvalues of Markov chains. *The Annals of Applied Probability*, 1(1):36–61, 1991.

[Gol93] Leslie Ann Goldberg. Automating Polya theory: The computational complexity of the cycle index polynomial. *Information and Computation*, 105:268–288, 1993.

[GS82] G.R. Grimmett and D.R. Stirzaker. *Probability and Random Processes*, chapter 6, pages 115–174. Oxford University Press, 1982.

[HJ85] Roger A. Horn and Charles A. Johnson. *Matrix Analysis*, chapter 4.2, pages 176–181. Cambridge University Press, 1985.

[JS89] Mark Jerrum and Alistair Sinclair. Approximating the permanent. *Society for Industrial and Applied Mathematics Journal on Computing*, 18(6):1149–1178, December 1989.

[JS93] Mark Jerrum and Alistair Sinclair. Polynomial-time approximation algorithms for the Ising model. *Society for Industrial and Applied Mathematics Journal on Computing*, 22(5):1087–1116, October 1993.

[JVV86] Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, 1986.

[Kan94] Ravi Kannan. Markov chains and polynomial time algorithms. In *Annual Symposium on the Foundations of Computer Science*, pages 656–669. Institute of Electrical and Electronics Engineers, 1994.

[Mih89] M. Mihail. On coupling and approximation of the permanent. *Information Processing Letters*, 30:91–96, 1989.

[PB83] J. Scott Provan and Michael O. Ball. The complexity of counting cuts and of computing the probability that a graph is connected. *Society for Industrial and Applied Mathematics Journal on Computing*, 12(4):777–788, 1983.

[Rot95] Dan Roth. On the hardness of approximate reasoning. *Artificial Intelligence Journal*, 1995. To appear.

[Sin91] Alistair Sinclair. Improved bounds for mixing rates of Markov chains on combinatorial structures. Technical report, Department of Computer Science, University of Edinburgh, July 1991.

[SJ89] Alistair Sinclair and Mark Jerrum. Approximate counting, uniform generation, and rapidly mixing Markov chains. *Information and Computation*, 82:93–133, 1989.

[Vad95] Salil P. Vadhan. The complexity of counting. Undergraduate Thesis, Harvard University, 1995.

[Val79] Leslie G. Valiant. The complexity of enumeration and reliability problems. *Society for Industrial and Applied Mathematics Journal on Computing*, 8(3):410–421, August 1979.