CONVERGENCE RATES FOR MONTE CARLO EXPERIMENTS

ALISTAIR SINCLAIR*

Abstract. This paper gives a brief overview of techniques developed recently for analyzing the rate of convergence to equilibrium in Markov chain Monte Carlo experiments. A number of applications in statistical physics are mentioned, and extensive references provided.

Key words. Statistical physics, Monte Carlo simulation, Markov chains, Metropolis rule, mixing rates, coupling, multicommodity flow.

AMS(MOS) subject classifications. 05C85, 60J10, 60J20, 60K35, 68Q20, 68Q25, 82B20, 82B31, 82B80.

1. Introduction.

This short paper is a summary of a talk given at the workshop on “Numerical Methods for Polymeric Systems” at the IMA in May 1996. The purpose of the talk, and of this paper, is to bring to the attention of the computational physics and chemistry communities some techniques developed recently in computer science and discrete probability for the analysis of convergence rates of Markov chains. When applied to Markov chains arising in Monte Carlo experiments on physical systems, these techniques can potentially yield rigorous bounds on the time to reach equilibrium; this in turn leads to precise performance guarantees for the experiments, in contrast to the heuristic error bars that are conventionally quoted.

Since the techniques are well documented in survey articles and specific applications elsewhere (see, e.g., [9,19,21,36] and the references given there), this paper will aim only to summarize the basic ideas from the perspective of statistical mechanics applications. Pointers to the literature are provided for those wishing to dig deeper. My hope is that practitioners in the Monte Carlo world will perceive the value of these analytical tools, and apply them (probably with refinements) to their own experiments. There is by now a sufficient body of examples to suggest that this line of enquiry should be quite fruitful.

1.1. The framework.

We begin by introducing a general framework that captures the essence of Markov chain Monte Carlo experiments. Consider a statistical mechanical system that has a finite set Ω of possible configurations. Let $w : \Omega \to \mathbb{R}^+$ be a positive function defined on Ω; we shall refer to $w(x)$ as the *weight* of configuration x. Typically, w will take the form $w(x) = \exp\big(-\beta H(x)\big)$, where $H(x)$ is the energy of x and the

* Computer Science Division, University of California, Berkeley, CA 94720-1776. Email: sinclair@cs.berkeley.edu. Supported in part by NSF Grant CCR-9505448.

constant β depends inversely on temperature. The goal of a Monte Carlo experiment can then be simply stated as follows:

> *Sample configurations at random from the probability distribution*
>
> $$\pi(x) = w(x)/Z \qquad \forall x \in \Omega,$$
>
> *where $Z = \sum_{x\in\Omega} w(x)$ is a normalizing constant, known as the* partition function *of the system. (When w has the exponential form stated above, π is of course the Gibbs distribution.)*

As a concrete example, consider the ferromagnetic Ising model on a finite graph $G = (V, E)$ (which might typically be a region of the 3-dimensional rectangular lattice). The vertex set V, which we shall take to be $[n] = \{1, 2, \dots, n\}$, represents the sites, and the edge set E the pairs of adjacent (or neighboring) sites. A configuration x of the system is an assignment of ±1 spins to each site, i.e., $\Omega = \{+1, -1\}^{[n]}$; we shall write x_i for the spin at site i. The energy of configuration x is

$$H(x) = -J \sum_{\{i,j\}\in E} x_i x_j,$$

where $J > 0$ is the interaction energy. (Since the system is ferromagnetic, configurations with larger numbers of aligned neighbors have lower energy.) With $\beta = 1/kT$, where k is Boltzmann's constant and T is temperature, the Gibbs distribution is then

$$\pi(x) = \exp\bigl(-\beta H(x)\bigr)/Z,$$

and the weight function is $w(x) = \exp\bigl(-\beta H(x)\bigr)$.

Note that in this example $|\Omega| = 2^n$, where n is the volume (number of sites). Thus the configuration space is exponentially large as a function of the size of the system, making exhaustive enumeration of it infeasible. This is a property shared by all statistical mechanical systems.

The size of the configuration space, and the complexity of the distribution π, motivate the Monte Carlo approach. The idea is to construct a discrete-time ergodic Markov chain $(X_t)_{t=0}^{\infty}$ whose state space is Ω and which converges to the desired equilibrium (or stationary) distribution π as $t \to \infty$, regardless of the initial state X_0. This much is usually a straightforward task.[1] All that is needed is to define a connected *neighborhood structure* on Ω, i.e., a connected graph whose vertices are configurations $x \in \Omega$. This is usually done by introducing edges between configurations which differ by some small local perturbation;[2] for example, in the Ising model the

[1] Though of course it is *not* so straightforward to construct a chain in which the convergence is fast; we will have a lot more to say about this shortly.

[2] Non-local perturbations are also possible, and potentially very powerful: perhaps the most famous example is the Swendsen-Wang algorithm for the Ising and Potts models [37].

neighbors of a configuration x might be all those configurations obtained from x by flipping the spin value at a single site. Generally, we will write $\mathcal{N}(x)$ for the set of neighbors of x.

Given such a neighborhood structure, a Markov chain with the desired properties is immediately obtained using the *Metropolis rule.* If the chain is at state[3] $X_t = x$ at time t, a transition is made to a new state X_{t+1} as follows:

select a neighbor y of x with probability $q(x, y)$

with probability $\min\{\frac{w(y)}{w(x)}, 1\}$, set $X_{t+1} = y$

else set $X_{t+1} = x$

Here $q(x, \cdot)$ is a probability distribution over $\mathcal{N}(x)$ for each $x \in \Omega$, and the function q is symmetric, i.e., $q(x, y) = q(y, x)$. We may in fact allow $\sum_{y \in \mathcal{N}(x)} q(x, y) < 1$, in which case we set $X_{t+1} = x$ with the remaining probability. The simplest choice for q is to set $q(x, y) = \Delta^{-1}$ for all pairs of neighbors x, y, where $\Delta = \max_{x \in \Omega} |\mathcal{N}(x)|$ is the maximum degree of the neighborhood graph. It should be clear that implementing this Markov chain is a simple task, requiring knowledge only of the local neighborhood $\mathcal{N}(x)$ and the weight function w (and not of global quantities such as the partition function Z).

We write $P(x, y)$ for the transition probability $\Pr[X_{t+1} = y \mid X_t = x]$. Note that $P(x, y) > 0$ if and only if x and y are neighbors (or if $x = y$). Moreover, it is easy to check that the Markov chain is *reversible* with respect to the distribution π, i.e., it satisfies the detailed balance conditions

$$\pi(x)P(x, y) = \pi(y)P(y, x) \qquad \forall x, y \in \Omega. \tag{1.1}$$

This immediately implies that the chain converges to π. (Strictly speaking, we also need to ensure that the chain is *aperiodic*; this can be achieved easily by the simple trick of adding an artificial holding probability to every state. See section 1.2 for an example.)

To sample from π, it therefore suffices to simulate the above process, starting in some arbitrary initial configuration, for sufficiently many steps; the final configuration will then be distributed (approximately) according to π. This is the essence of the Markov chain Monte Carlo approach. The central question, however, is the following:

How many steps is "sufficiently many" ?

Since it is in general not possible to determine whether a Markov chain has reached equilibrium simply by observing it, we actually need an *a priori* bound on its rate of convergence.

[3] Henceforth, we shall use the terms "state" (of the Markov chain) and "configuration" (of the physical system) interchangeably. Note that this deviates from some uses of the word "state" in statistical physics.

In practice, this problem is generally sidestepped by non-rigorous methods such as auto-correlation times or appeals to physical intuition. The purpose of this paper is to demonstrate that the machinery exists for answering the above question rigorously.

To phrase the question precisely we need a little notation. The *variation distance* between two probability distributions μ, ν on Ω is defined by

$$\|\mu - \nu\| = \tfrac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)| = \max_{S \subseteq \Omega} |\mu(S) - \nu(S)|.$$

Let $p^t(x, \cdot)$ denote the distribution of the Markov chain at time t, given that it starts in state $X_0 = x$. Following standard practice, we will measure the distance of the chain from stationarity by the quantity

$$\Delta_x(t) = \|p^t(x, \cdot) - \pi\|.$$

It should be clear that $\Delta_x(t)$ is directly related to error bars in statistical estimates of quantities obtained from observations of the chain at time t.

Convergence of the chain means that $\Delta_x(t) \to 0$ as $t \to \infty$, for all x. To measure the *rate* of convergence, we introduce the quantity

$$\tau_x(\epsilon) = \min\{t : \Delta_x(t) \leq \epsilon \text{ for all } t' \geq t\},$$

i.e., the time required to reduce the variation distance to ϵ. We will refer to $\tau_x(\epsilon)$ as the *mixing time* of the chain (from initial state x). Our goal will be to calculate *a priori* upper bounds on the mixing time. These bounds will tell us how long we need to run our Monte Carlo simulation in order to be sure of achieving any specified variation distance ϵ, or equivalently, any desired error bars in our experiment.

In the next two sections, we will describe two very different approaches to this problem: *coupling* and *flows*. Each of these has a powerful intuitive appeal, and each has been successfully applied to the analysis of several interesting Markov chains, both within statistical mechanics and outside. Moreover, there are chains that are amenable to each of these approaches but apparently not to the other.

We shall illustrate both approaches with a single toy example. This has the advantage of keeping the technical difficulties to a minimum, though of course it does not do justice to the full power of the techniques. For more significant examples, the reader is urged to consult the references provided at the end of each section. In keeping with our desire not to obscure the main ideas with technical details, we shall also be content with suboptimal constants in our bounds. The reader should appreciate that these can (and should) be significantly sharpened in any real application.

1.2. A toy example. We close this introductory section by defining the simple Markov chain which we shall use for illustration. The state space (set of configurations) will be $\Omega = \{0,1\}^n$, the set of all 0-1 vectors of length n; we shall write a vector $x \in \{0,1\}^n$ as $(x_1, \dots, x_n)$. The weight function w will be constant, so the distribution π is uniform, i.e., $\pi(x) = 2^{-n}$ for all $x \in \{0,1\}^n$. Configurations x, y are adjacent if and only if they differ in exactly one position. The Metropolis Markov chain in this case is therefore simply nearest-neighbor random walk on the vertices of the n-dimensional unit hypercube. (Another way to view this process is single spin-flip dynamics for the ferromagnetic Ising model with n sites in the infinite temperature limit.)

In order to avoid tiresome technical complications connected with periodicity, we add a holding probability of $\frac{1}{2}$ to every state; that is, at every step of the Markov chain, we either (with probability $\frac{1}{2}$) do nothing, or (with probability $\frac{1}{2}$) make a step as above. Of course, this will slow down the chain (and increase the mixing time) by at most a factor of 2, but makes the results simpler to state; in practice, a much smaller holding probability can be used.

In summary, then, our Markov chain makes transitions as follows from any state $x \in \Omega$:

```
pick a position $i \in \{1, \dots, n\}$ uniformly at random
with probability $\frac{1}{2}$, flip the ith bit of $x$ (i.e., replace $x_i$ by $1 - x_i$)
else do nothing
```

2. Coupling.

2.1. The idea. Coupling is an elementary probabilistic method for bounding the mixing time of a Markov chain $\mathcal{M}$ by relating it to the stopping time of an associated stochastic process. This process consists of a *pair* (X_t, Y_t), evolving in time in such a way that

1. each of the processes (X_t) and (Y_t) is a faithful copy of $\mathcal{M}$, given initial states $X_0 = x$ and $Y_0 = y$ respectively; and
2. if $X_t = Y_t$, then $X_{t+1} = Y_{t+1}$.

We refer to such a process as a *coupling* for $\mathcal{M}$.

The idea here is the following. Although each of $(X_t), (Y_t)$, viewed in isolation, behaves exactly like $\mathcal{M}$, they need not be independent; on the contrary, we will construct a joint distribution for the two processes in such a way that they tend to move closer together. By the second condition above, once they have met they must remain together at all future times.

For fixed initial states $X_0 = x$, $Y_0 = y$, we let $T_{xy} = \min\{t : X_t = Y_t\}$, i.e., the (random) time until the processes meet. The *coupling time* $\mathcal{T}$ of $\mathcal{M}$ is defined as the time that must elapse before the processes have met with some prescribed probability, which we take to be $1 - \mathrm{e}^{-1}$. (There is nothing magical about this constant; it merely affects the base of the logarithm in

Theorem 2.1 below.) In other words, we define

$$\mathcal{T} = \min\{t : \Pr[T_{xy} > t] \le \mathrm{e}^{-1} \text{ for all } x, y\}.$$

Now it is not hard to see that the coupling time provides a bound on the mixing time of $\mathcal{M}$. In fact, we can show:

THEOREM 2.1. *The mixing time of an ergodic Markov chain satisfies*

$$\tau_x(\epsilon) \le \mathcal{T}\lceil \ln \epsilon^{-1} \rceil \qquad \text{for all } x \in \Omega.$$

Thus, to obtain an upper bound on the mixing time, it suffices to find a coupling with a small coupling time.

The proof of Theorem 2.1 is so simple that we sketch it here; a more systematic development can be found in [1]. The key is the following general observation known as the "coupling lemma," whose proof is an easy exercise. Given any two random variables X, Y on Ω with distributions μ, ν respectively, we have

$$\|\mu - \nu\| \le \Pr[X \ne Y].$$

To apply this in the context of our Markov chain, imagine that (Y_t) is the stationary process, i.e., $Y_0 = y$ is distributed according to the stationary distribution π, and hence the same holds for Y_t at all later times t. Then, by the coupling lemma applied to the random variables X_t, Y_t, we have

$$\Delta_x(t) \le \Pr[X_t \ne Y_t] \le \max_{x,y} \Pr[T_{xy} > t]. \tag{2.1}$$

Also, by the definition of $\mathcal{T}$ we have, for any positive integer k and all pairs $x, y \in \Omega$,

$$\Pr[T_{xy} > k\mathcal{T}] \le \mathrm{e}^{-k}; \tag{2.2}$$

to see this, consider a sequence of k epochs each of length $\mathcal{T}$, during each of which coupling fails to occur with probability at most e^{-1}. Putting (2.1) and (2.2) together yields the theorem.

Remarks (i) Theorem 2.1 has a converse which (very loosely stated) says that there always exists a coupling that captures the time taken for the chain to converge. For the details, see [15]. For applications of coupling to more general processes see, e.g., the lectures by Lindvall [25] or the paper by Thorisson [38].

(ii) It is often convenient to work with the *expected coupling time*, which we define as $\overline{\mathcal{T}} = \max_{x,y} \mathrm{E}(T_{xy})$, where $\mathrm{E}(\,\cdot\,)$ denotes expectation. Applying Markov's inequality to Theorem 2.1 shows that $\tau_x(\epsilon) \le \mathrm{e}\overline{\mathcal{T}}\lceil \ln \epsilon^{-1} \rceil$. This is somewhat cruder than Theorem 2.1, but often easier to use in practice when an upper bound on the expectation of T_{xy} is readily available but its distribution is more complicated. We shall use expectations in the example in the next subsection.

2.2. An example. Let us now apply Theorem 2.1 to obtain an upper bound on the mixing time of our hypercube Markov chain from section 1.2. First, we need to define a suitable coupling. How can we define a joint distribution on two copies of this process so as to bring them together quickly? The intuitively obvious idea is to make both processes choose the *same* random bit at every step, thus tending to bring individual bits into agreement.

To make this idea precise, it helps to rephrase the transitions of the Markov chain very slightly. In state $x \in \Omega$, we do the following:

```
pick a position $i \in \{1,\ldots,n\}$ uniformly at random
pick a value $b \in \{0,1\}$ uniformly at random
set $x_i = b$
```

It should be clear that this is entirely equivalent to our original definition at the end of section 1.2.

Now we can define our coupling as follows. If the pair process (X_t, Y_t) is in state $(x,y) \in \Omega \times \Omega$, we do the following:

```
pick a position $i \in \{1,\ldots,n\}$ uniformly at random
pick a value $b \in \{0,1\}$ uniformly at random
set $x_i = b$ and $y_i = b$
```

Thus both copies of the process choose the *same* position i and the *same* new value b for the associated bit.

It should be clear that this *is* a coupling: plainly, each copy viewed in isolation is evolving exactly according to the original chain, so condition 1 is satisfied. And the pair process can never cause agreeing bits to disagree, so condition 2 is also satisfied.

What is the coupling time? To analyze this, we introduce a measure of distance between X_t and Y_t. Let D_t denote the number of bit positions in which X_t and Y_t differ. Thus D_t is a process taking integer values in the interval $[0,n]$, and $D_t = 0$ if and only if $X_t = Y_t$. The quantity T_{xy} is the time required for D_t to reach zero, given that $X_0 = x$ and $Y_0 = y$.

How does D_t change with time? The key observation is that, as soon as a bit position $i \in \{1,2,\ldots,n\}$ has been chosen, the values x_i, y_i agree, and this persists for all times thereafter. This implies that D_t is monotonically decreasing; more precisely, it implies that, if $D_t = d$, then

$$D_{t+1} = \begin{cases} d-1 & \text{with probability } \frac{d}{n}; \\ d & \text{otherwise.} \end{cases} \tag{2.3}$$

Thus, for any initial values x, y, the time T_{xy} is stochastically dominated by the random variable $T_n + T_{n-1} + \cdots + T_1$, where T_d is the time for D_t to decrease from d to $d-1$. But from (2.3) T_d is just the number of tosses of a biased coin with heads probability $\frac{d}{n}$ until the first head appears. Thus $\mathrm{E}\,(T_d) = \frac{n}{d}$, and so $\mathrm{E}\,(T_{xy}) \leq \sum_{d=1}^{n} \mathrm{E}\,(T_d) \sim n(\ln n + \gamma)$ as $n \to \infty$, where

γ is Euler's constant. By Markov's inequality, as in Remark (ii) at the end of the previous subsection, we have $\Pr[T_{xy} > \mathrm{eE}\,(T_{xy})] \le \mathrm{e}^{-1}$, and hence $\mathcal{T} \le \max_{x,y} \mathrm{eE}\,(T_{xy}) \le en(\ln n + \mathrm{O}(1))$.

Appealing to Theorem 2.1, we have therefore established:

THEOREM 2.2. *The mixing time of the hypercube Markov chain is bounded above by*

$$\tau_x(\epsilon) \le en(\ln n + \mathrm{O}(1))\lceil \ln \epsilon^{-1} \rceil.$$

This bound is in fact asymptotically tight, up to a small constant factor; see [1].

Remarks (i) The reader familiar with discrete probability may have noticed that a similar bound on the time for D_t to hit 0 (with a slightly better constant) could have been obtained immediately by analogy with the *coupon collector's problem*: if each cereal box contains one of a set of n different coupons, each equally likely, how many boxes does one need to buy in order to collect at least one copy of every coupon? The distribution of this random variable is well understood (see, e.g., Feller [13]). We have presented the above more hands-on approach because it illustrates the general form such arguments usually take in less tidy examples.

(ii) The following slightly more involved coupling shaves off a factor of 2 from the bound of Theorem 2.2. As before, let (x, y) denote the state of the pair process (X_t, Y_t), and now let $A = \{i_1, \ldots, i_r\}$ be the set of positions in which x, y differ. Now let process X_t choose position $i \in \{1, \ldots, n\}$ uniformly at random. If $i \notin A$, let Y_t choose the same i; if $i = i_j \in A$, let Y_t choose i_{j+1} (where we interpret i_{r+1} as i_1). In either case, let both processes pick the same value b. The reader should check that this is also a valid coupling. Now it should be clear that, under this coupling, (2.3) becomes

$$D_{t+1} = \begin{cases} d-2 & \text{with probability } \frac{d}{n} \text{ if } d \ge 2; \\ 0 & \text{with probability } \frac{d}{n} \text{ if } d = 1; \\ d & \text{otherwise.} \end{cases}$$

Hence the time for D_t to reach zero is stochastically dominated by the random variable $T_{n^*} + T_{n^*-2} + \cdots + T_3 + T_1$, where $n^* = n$ if n is odd and $n^* = n - 1$ if n is even. This leads to a factor of 2 improvement.

2.3. Applications of coupling in statistical physics. Until recently, applications of the coupling approach had been confined to Markov chains that possess a high degree of symmetry, like the hypercube example above. It was felt that coupling was not sophisticated enough to handle the kind of complex chains that occur in Monte Carlo experiments in physics. Recently, however, there have been some physical examples in which coupling has turned out to provide the only known analysis.

The first of these was Jerrum's analysis of a Markov chain for the antiferromagnetic q-state Potts model, where q is sufficiently large (specifically,

q must be at least $2d+1$, where d is the maximum degree of the interaction graph) [16]. The second was Luby, Randall and Sinclair's analysis of Markov chains for several structures on 2-dimensional lattices, including dimer coverings and Eulerian orientations (configurations of the ice model) [27]. Other structures on 2-dimensional grids that have recently been tackled are the 3-state Potts model (Madras and Randall [29]) and independent sets, or configurations of the hard-core gas model (Luby and Vigoda [28]). For a unifying view of some of these examples, and others, see the interesting recent paper of Bubley and Dyer [6].

In all these cases, coupling is used to obtain an upper bound on the mixing time of the form $\mathrm{O}(n^k)$, where n is the volume and k is a small constant and we have absorbed the usual dependence on $\ln \epsilon^{-1}$, as well as constant factors, into the O. Note that such a bound with $k=1$ is the best we could possibly hope for for any Markov chain that makes only "local" moves; and such a polynomial bound for *any* fixed k is quite non-trivial, since the number of configurations in Ω is *exponentially* large as a function of n. Currently, typical values of k are in the range [2..6], often rather too big for Monte Carlo experiments on large systems. It is an area of active research to tune the results so as to make k as small as possible.

The arguments in the above papers are quite straightforward, and only slightly more complicated than that for the toy hypercube example above. The principal complication is usually that the natural distance measure D_t is not Markovian and not monotonically decreasing under the coupling, as it was above. However, it is usually enough to show that the *expected* change in D_t at each time step is negative, and then appeal to a simple martingale argument. These examples give much hope that other Markov chains in statistical physics might be amenable to the coupling approach.

Another recent related development is due to Propp and Wilson [32]. They observe that, if the state space of the Markov chain is equipped with a partial order with unique maximum and minimum elements a, b respectively, and if the coupling preserves this order (in a certain strong sense), then the coupling time is stochastically dominated by T_{ab}, the time for a pair of processes starting in the maximum and minimum states to meet. This observation can dramatically simplify the task of bounding the coupling time analytically. It also allows the coupling time to be estimated rigorously by a simple experiment: namely, simulate the coupling with $X_0 = a$ and $Y_0 = b$ until the processes meet. Propp and Wilson give some examples from statistical mechanics where such a partial order exists. This area seems ripe for further investigation.

We should also mention briefly that coupling has been successfully applied to analyze the mixing time of a number of Markov chains arising in computational applications outside statistical physics. Examples include approximating the volume of convex bodies [7], generating acyclic orientations of certain graphs [5], and protocol testing [30].

3. Flows.

3.1. The idea. The method of "flows" is a more sophisticated approach to bounding the mixing time which has proved successful in handling several rather complex Markov chains in statistical physics, including ones related to monomer-dimer systems, the Ising model and self-avoiding walks.

The intuition we are trying to capture here is the following: if a Markov chain is "globally well connected," in the sense that it contains no bottlenecks, then it should converge rapidly to equilibrium, i.e., the mixing time should be small. The concept of "bottleneck" is most conveniently captured in the language of flow networks. We now proceed to set up the appropriate framework.

We will view the Markov chain as a network whose vertices are the elements of Ω. There is a directed edge $e = (x, y)$ between distinct states x and y if and only if the transition probability $P(x, y) > 0$. This edge has *capacity* $c(e) = \pi(x)P(x, y)$, where as usual π is the stationary distribution. We shall assume that the chain is *reversible*, as defined in equation (1.1); this implies that to every edge $e = (x, y)$ there corresponds a reversed edge $\overline{e} = (y, x)$ with $c(\overline{e}) = c(e)$.

Our task is to route $\pi(x)\pi(y)$ units of flow from x to y along the edges of the network, for all ordered pairs of distinct states (x, y) simultaneously. (We should think of there being a distinct "commodity" for each pair (x, y), so that the flows between different pairs do not interact.) Such a routing is called a *flow*. The quality of the flow is measured as the maximum over edges of the total flow (of all commodities) along the edge divided by the capacity of the edge.

More formally, let $\mathcal{P}_{xy}$ denote the set of all simple paths (i.e., paths that touch each vertex at most once) from x to y in the network, and let $\mathcal{P} = \bigcup_{(x,y)} \mathcal{P}_{xy}$. A *flow* is a function $f : \mathcal{P} \to \mathbb{R}^+$ such that

$$\sum_{p \in \mathcal{P}_{xy}} f(p) = \pi(x)\pi(y) \qquad \forall x, y \in \Omega, x \neq y.$$

We extend f to edges in the obvious way: the flow along edge e is just $f(e) = \sum_{p \ni e} f(p)$. The *cost* of the flow f is then defined as

$$\rho(f) = \max_e \frac{f(e)}{c(e)}.$$

Our earlier informal intuition can now be expressed as follows. If the Markov chain supports a flow of low cost, then it can have no bottlenecks, and hence its mixing time should be small. This intuition is formalized in the following theorem.

THEOREM 3.1. *Let $\mathcal{M}$ be an ergodic reversible Markov chain with holding probabilities $P(x, x) \geq \frac{1}{2}$ at all states x. The mixing time of $\mathcal{M}$*

satisfies

$$\tau_x(\epsilon) \leq \rho(f)\ell(f)\bigl(\ln \pi(x)^{-1} + \ln \epsilon^{-1}\bigr),$$

for any flow f, where $\ell(f)$ is the length (number of edges) of a longest path that carries non-zero flow in f.

Thus, in order to apply Theorem 3.1, we must find a flow f of low cost. (The factor $\ell(f)$ is rarely problematic since usually we route all flow along geodesic paths and the diameter of the Markov chain is relatively small.) Any such flow gives an upper bound on the mixing time.

Remarks (i) This theorem follows by combining Proposition 1 of [35] and Corollary $6'$ of [35]. The proof proceeds via a bound on the second eigenvalue of the transition matrix P, and is an instance of the general technique of obtaining geometric bounds on eigenvalues. For more on this large topic, see, e.g., [2,3,8,10,20,24,36].
(ii) As usual, the requirement that $P(x,x) \geq \frac{1}{2}$ is introduced only to handle periodic behavior in a way that simplifies the statement of the theorem. In practice, a much smaller holding probability can be used.
(iii) An alternative version of Theorem 3.1 has the quantity $8\rho(f)^2$ in place of $\rho(f)\ell(f)$. Usually, however, this bound is inferior to that of Theorem 3.1. For a detailed discussion of Theorem 3.1 and its relatives, see [35] and [10].
(iv) Theorem 3.1 has a suitably stated converse, which says (roughly) that there always exists a flow whose cost is close to the mixing time. See Theorem 8 of [35].

3.2. An example. We now apply Theorem 2.1 to obtain an upper bound on the mixing time of our hypercube Markov chain from section 1.2. To do this, we need to define a suitable flow. How can we route flow between all pairs of vertices of the hypercube in such a way that no edge is overloaded? Let's consider the simplest type of flow, namely one in which all the flow between a given pair of vertices (x,y) travels along a single path, γ_{xy}. A canonical choice for γ_{xy} is the "bit-fixing" path, i.e., the path which flips the bit values from x_i to y_i in the order $i = 1, 2, \ldots, n$.

More formally, this is the path whose ith edge is

$$\bigl((y_1, \ldots, y_{i-1}, x_i, x_{i+1}, \ldots, x_n)\,,\,(y_1, \ldots, y_{i-1}, y_i, x_{i+1}, \ldots, x_n)\bigr).$$

Note that some of these edges (those for which $x_i = y_i$) are self-loops, and hence redundant; we eliminate these from the path. The length of the path is then precisely equal to the number of positions in which x and y differ. Thus it is a *geodesic* (shortest path) between x and y.

Now in our flow f, we route all the (x,y) flow along the path γ_{xy}; i.e., we have $f(\gamma_{xy}) = \pi(x)\pi(y)$ for each $x \neq y$, and $f(p) = 0$ for all other paths p. The intuition for this choice of flow is that, by symmetry of the hypercube, the flow along every edge is the same. Since the total quantity of all commodities flowing in the system is $\sum_{x \neq y} \pi(x)\pi(y) \leq 1$ unit, and since no commodity travels a distance greater than n, the total flow $\sum_e f(e)$ along all edges is at most n. Now the number of edges is nN, where $N = 2^n$ is the number of vertices; so by symmetry we have $f(e) \leq \frac{n}{nN} = \frac{1}{N}$ for every edge e. But since the transition probability along every edge is $\frac{1}{2n}$, the capacity of each edge is $c(e) = \frac{1}{2nN}$, so the cost of the flow is

$$\rho(f) = \max_e \frac{f(e)}{c(e)} \leq \frac{1/N}{1/2nN} = 2n.$$

Finally, since $\ell(f) = n$, we can apply Theorem 3.1 to obtain the following bound on the mixing time:

THEOREM 3.2. *The mixing time of the hypercube Markov chain is bounded above by*

$$\tau_x(\epsilon) \leq 2n^2(n \ln 2 + \ln \epsilon^{-1}).$$

This bound is significantly weaker than that of Theorem 2.2, and this slackness is typical of this heavier-duty method. However, there are examples for which flows provide the only known approach to obtaining good bounds on the mixing rate (see section 3.3 below).

The above analysis of the cost of the flow leaves something to be desired since it relies crucially on the strong symmetry properties of the hypercube, and also on the fact that π is uniform. Obviously, interesting statistical mechanical systems do not possess such a simple structure. We therefore explain now an additional technique for analyzing the cost of a flow which does not appeal to symmetry and which has proved essential in more complex examples. For illustrative purposes we shall again use the above simple flow f on the hypercube.

Recall that our goal is to bound the flow along any edge of the hypercube. So let $e = (z, z')$ be any edge, where $z = (z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_n)$ and $z' = (z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_n)$, i.e., edge e flips the ith bit of z. Let paths(e) denote the set of flow-carrying paths that pass through e, i.e., paths$(e) = \{(x, y) : \gamma_{xy} \ni e\}$. The trick is to *use the configuration space Ω itself* to measure the flow along paths in paths(e). To do this, we set up a mapping $\eta_e :$ paths$(e) \to \Omega$ with the following properties:

1. η_e is an injection; and
2. $\pi(x)\pi(y) = \pi(z)\pi(\eta_e(x, y))$ for all $(x, y) \in$ paths(e).

Property 1 means that each flow-carrying path through e is uniquely encoded by an element of Ω: this places a bound on the total number of such paths. Property 2 means that this encoding scheme is "flow-preserving," in the sense that the flow $\pi(x)\pi(y)$ along each path is proportional to the weight $\pi(\eta_e(x, y))$ of its encoding in the stationary distribution (note that $\pi(z)$ is fixed).

Before we demonstrate the existence of such a mapping η_e for our hypercube example, let's first see that it will immediately give us a bound on $f(e)$. For we have

$$\begin{aligned} f(e) = \sum_{p \ni e} f(p) &= \sum_{\gamma_{xy} \ni e} \pi(x)\pi(y) \\ &= \sum_{\gamma_{xy} \ni e} \pi(z)\pi(\eta_e(x,y)) \\ &\le \pi(z). \end{aligned}$$

The second line here follows from property 2, and the third line from property 1 (since η_e is an injection and π is a probability distribution). Finally, since $c(e) = \pi(z)P(z,z')$, we have $\rho(f) = \max_e \frac{f(e)}{c(e)} \le \frac{1}{P(z,z')} = 2n$, the same bound as we obtained earlier using the symmetry argument.[4]

It remains to specify the mapping η_e with the required properties; here, of course, we will need to use some combinatorial insight about the hypercube. Let (x,y) be any element of paths(e). What can we say about x and y? Well, since γ_{xy} is a bit-fixing path that passes through the edge $e = (z,z')$, and this edge flips the ith bit, it must be the case that the first $i-1$ bits of y are exactly the first $i-1$ bits of z, and the last $n-i+1$ bits of x are exactly the last $n-i+1$ bits of z. So, since z is fixed, to specify x and y uniquely it is enough to record the *first* $i-1$ bits of x and the *last* $n-i+1$ bits of y. Thus we define the encoding

$$\eta_e(x,y) = (x_1, \ldots, x_{i-1}, y_i, y_{i+1}, \ldots, y_n),$$

which is certainly a valid member of Ω.

Now this mapping satisfies property 1, since, as we have seen, we can recover x and y uniquely from $\eta_e(x,y)$: to be precise, if we let $\eta_e(x,y) = (u_1, \ldots, u_n)$ then we can write down the explicit expressions

$$x = (u_1, \ldots, u_{i-1}, z_i, z_{i+1}, \ldots, z_n) \text{ and } y = (z_1, \ldots, z_{i-1}, u_i, u_{i+1}, \ldots, u_n).$$

It also satisfies property 2, trivially, since π is the uniform distribution.[5] This concludes the argument.

We stress that this second argument is much more general than the first, and does not appeal to symmetry. Moreover, since it uses the space of configurations Ω to measure flow *implicitly*, it does not require any explicit

[4] The reader should observe that this argument is completely general and follows only from properties 1 and 2 of the encoding function η_e. The only place where we have used anything specific to the hypercube is in plugging in the value of $P(z,z')$ in the final step.

[5] In view of this fact, we could have dispensed with property 2 in this simple example. However, when π is non-uniform, as is often the case in more realistic examples, property 2 becomes significant. Actually, it is usually necessary to work with a slightly weaker property: namely, that $\pi(x)\pi(y) \le \alpha\pi(z)\pi(\eta_e(x,y))$, where α is not too large. Exactly the same argument still holds, except that the factor α appears in the final bound on $\rho(f)$.

enumerative information (such as knowledge of the partition function Z). The power of this argument has been demonstrated in several quite complex examples (see the next subsection).

3.3. Applications of flows in statistical physics. Historically, the earliest non-trivial bounds on mixing times for Markov chains in statistical physics were proved using flows (usually together with the above injective mapping trick), and this remains the only known approach for most of these examples.[6] The first application, which motivated the development of flows, was to monomer-dimer systems [17,36] — see section 12.4 of [19] for an improved version. This analysis was later extended to Monte Carlo algorithms for dense dimer systems on arbitrary lattices [23] and for the ice model [31]. As in the case of coupling, the bounds on the mixing time are of the form $\mathrm{O}(n^k)$, where n is the volume and k a small constant (which one can presumably make smaller with more work).

In [18], Jerrum and Sinclair introduced a Markov chain for the ferromagnetic Ising model (with arbitrary interaction topology) based on the high-temperature expansion, and proved a similar $\mathrm{O}(n^k)$ bound on the mixing time, at *all* temperatures. This remains the only known Markov chain for the Ising model with a mixing time that is provably polynomial in the volume at all temperatures. It is somewhat frustrating that Markov chains such as that of Swendsen-Wang [37], which appear in practice to have excellent convergence behavior (at least for lattices), still elude analysis. (Note that the Swendsen-Wang Markov chain definitely does *not* converge fast on arbitrary graphs, as was proved recently by Gore and Jerrum [14].)

Another model in statistical physics for which a Monte Carlo algorithm has been successfully analyzed using flows is the self-avoiding walk model for linear polymers. The Markov chain here is due to Berretti and Sokal [4], and the analysis can be found in [33]. An earlier analysis using similar geometric techniques appeared in [24].

Flows have also been used in the analysis of Markov chains with computational applications outside statistical physics. Three of the most notable examples are approximating the permanent of a 0-1 matrix [17,36], counting the bases of a class of matroids [12] and finding a maximum matching by "simulated annealing" [17,34]. A comprehensive recent survey of all these applications (including those in statistical physics) can be found in [19].

Finally, we should mention a related geometric technique for analyzing the mixing time based on a quantity known as the *conductance*, or *Cheeger constant*. For a reversible Markov chain, the conductance is defined as

$$\Phi = \min_{\substack{S\subseteq\Omega \\ 0<\pi(S)\leq\frac{1}{2}}} \frac{c(S,\overline{S})}{\pi(S)},$$

[6] The technique is often referred to as the method of "canonical paths," since in most cases a single path γ_{xy} carries all the flow from x to y, as in the above hypercube example.

where $c(S,\overline{S})$ denotes the capacity of the cut separating the set S of states from its complement $\overline{S}=\Omega-S$, i.e., $c(S,\overline{S})=\sum_{x\in S,y\in\overline{S}}c(x,y)$. Note that Φ is an explicit measure of bottlenecks in the Markov chain. An analog of Theorem 3.1 (see, e.g., Theorem 2 of [35]) gives the same bound on $\tau_x(\epsilon)$ with $\rho(f)\ell(f)$ replaced by $2\Phi^{-2}$; thus the chain has a small mixing time if its conductance is not too small.

It should come as no surprise to the reader that flows and conductance are very closely related, via a form of max-flow min-cut theorem. Indeed, historically flows were developed as an indirect means of estimating the conductance for use in the above bound (though the direct flow-based bound of Theorem 3.1 has since proved to be sharper in most cases). For a detailed exposition of this connection, see [35].

Conductance has been used directly (without the aid of flows) to analyze the mixing time of several important Markov chains with an inherently "geometric" flavor. These include the work of Dyer, Frieze and Kannan [11], Lovász and Simonovits [26] and others on computing volumes (see [21] for a survey) and Karzanov and Khachian on counting linear extensions of a partial order [22].

REFERENCES

[1] D. Aldous, Random walks on finite groups and rapidly mixing Markov chains, *Séminaire de Probabilités XVII*, Springer Lecture Notes in Mathematics 986, 1981/82, pp. 243–297.

[2] N. Alon, Eigenvalues and expanders, *Combinatorica* **6** (1986), pp. 83–96.

[3] N. Alon and V.D. Milman, λ_1, isoperimetric inequalities for graphs and superconcentrators, *Journal of Combinatorial Theory Series B* **38** (1985), pp. 73–88.

[4] A. Berretti and A.D. Sokal, New Monte Carlo method for the self-avoiding walk, *Journal of Statistical Physics* **40** (1985), pp. 483–531.

[5] R. Bubley and M.E. Dyer, Graph orientations with no sink and an approximation algorithm for a hard case of #SAT, in *Proceedings of the 8th Annual ACM/SIAM Symposium on Discrete Algorithms*, 1997, pp. 248–257.

[6] R. Bubley and M.E. Dyer, Path coupling, Dobrushin uniqueness and approximate counting, School of Computer Studies Research Report Series 97.04, University of Leeds, 1997.

[7] R. Bubley, M.E. Dyer, and M.R. Jerrum, A new approach to polynomial-time generation of random points in convex bodies, *Random Structures and Algorithms*, 1997, to appear.

[8] P. Diaconis and L. Saloff-Coste, Comparison techniques for reversible Markov chains, *Annals of Applied Probability* **3** (1993), pp. 696–730.

[9] P. Diaconis and L. Saloff-Coste, What do we know about the Metropolis algorithm?, in *Proceedings of the 27th Annual ACM Symposium on Theory of Computing*, 1995, pp. 112–129.

[10] P. Diaconis and D. Stroock, Geometric bounds for eigenvalues of Markov chains, *Annals of Applied Probability* **1** (1991), pp. 36–61.

[11] M.E. Dyer, A. Frieze, and R. Kannan, A random polynomial time algorithm for approximating the volume of convex bodies, *Journal of the ACM* **38** (1991), pp. 1–17.

[12] T. Feder and M. Mihail, Balanced matroids, in *Proceedings of the 24th Annual ACM Symposium on Theory of Computing*, 1992, pp. 26–38.

[13] W. Feller, *An introduction to probability theory and its applications, Volume* I (3rd ed.), Wiley, New York, 1968.

[14] V. Gore and M.R. Jerrum, The Swendsen-Wang algorithm does not always mix rapidly, Preprint, University of Edinburgh, 1996.

[15] D. Griffeath, Coupling methods for Markov processes, in *Studies in Probability and Ergodic Theory*, G.-C. Rota ed., Academic Press, 1978, pp. 1–43.

[16] M.R. Jerrum, A very simple algorithm for estimating the number of k-colourings of a low-degree graph, *Random Structures and Algorithms* **7** (1995), pp. 157–165.

[17] M. R. Jerrum and A.J. Sinclair, Approximating the permanent, *SIAM Journal on Computing* **18** (1989), pp. 1149–1178.

[18] M. R. Jerrum and A.J. Sinclair, Polynomial-time approximation algorithms for the Ising model, *SIAM Journal on Computing* **22** (1993), pp. 1087–1116.

[19] M. R. Jerrum and A.J. Sinclair, The Markov chain Monte Carlo method: an approach to approximate counting and integration, in *Approximation algorithms for NP-hard problems*, D.S. Hochbaum ed., PWS Publishing, Boston, 1997, pp. 482–520.

[20] N. Kahale, A semidefinite bound for mixing rates of Markov chains, in *Proceedings of the 5th Integer Programming and Combinatorial Optimization Conference*, Springer Lecture Notes in Computer Science Vol. 1084, 1996, pp. 190–203.

[21] R. Kannan, Markov chains and polynomial time algorithms, in *Proceedings of the 35th Annual IEEE Symposium on Foundations of Computer Science*, 1994, pp. 656–671.

[22] A. Karzanov and L. Khachiyan, *On the conductance of order Markov chains*, Technical Report DCS 268, Rutgers University, June 1990.

[23] C. Kenyon, D. Randall, and A.J. Sinclair, Approximating the number of dimer coverings of a lattice, *Journal of Statistical Physics* **83** (1996), pp. 637–659.

[24] G.F. Lawler and A.D. Sokal, Bounds on the L^2 spectrum for Markov chains and Markov processes: a generalization of Cheeger's inequality, *Transactions of the American Mathematical Society* **309** (1988), pp. 557–580.

[25] T. Lindvall, *Lectures on the coupling method*, Wiley, New York, 1992.

[26] L. Lovász and M. Simonovits, Random walks in a convex body and an improved volume algorithm, *Random Structures and Algorithms* **4** (1993), pp. 359–412.

[27] M. Luby, D. Randall, and A.J. Sinclair, Markov chain algorithms for planar lattice structures, in *Proceedings of the 36th Annual IEEE Symposium on Foundations of Computer Science*, 1995, pp. 150–159.

[28] M. Luby and E. Vigoda, Approximately counting up to four, in *Proceedings of the 29th Annual ACM Symposium on Theory of Computing*, 1997, to appear.

[29] N. Madras and D. Randall, Factoring graphs to bound mixing rates, in *Proceedings of the 37th Annual IEEE Symposium on Foundations of Computer Science*, 1996, pp. 194–203.

[30] M. Mihail and C.H. Papadimitriou, On the random walk method for protocol testing, in *Proceedings of the 6th International Conference on Computer Aided Verification*, Springer Lecture Notes in Computer Science 818, 1994, pp. 132–141.

[31] M. Mihail and P. Winkler, On the number of Eulerian orientations of a graph, in *Proceedings of the 3rd Annual ACM-SIAM Symposium on Discrete Algorithms*, 1992, pp. 138–145.

[32] J.G. Propp and D.B. Wilson, Exact sampling with coupled Markov chains and applications to statistical mechanics, *Random Structures & Algorithms* **9** (1996), pp. 223–252.

[33] D. Randall and A.J. Sinclair, Testable algorithms for self-avoiding walks, in *Proceedings of the 5th Annual ACM-SIAM Symposium on Discrete Algorithms*, 1994, pp. 593–602.

[34] G.H. Sasaki and B. Hajek, The time complexity of maximum matching by simulated annealing, *Journal of the ACM* **35** (1988), pp. 387–403.

[35] A.J. Sinclair, Improved bounds for mixing rates of Markov chains and multicommodity flow, *Combinatorics, Probability and Computing* **1** (1992), pp. 351–370.

[36] A.J. Sinclair, *Randomised algorithms for counting and generating combinatorial structures*, Advances in Theoretical Computer Science, Birkhäuser, Boston, 1993.

[37] R.H. Swendsen and J-S. Wang, Nonuniversal critical dynamics in Monte Carlo simulations, *Physical Review Letters* **58** (1987), pp. 86–88.

[38] H. Thorisson, Coupling methods in probability theory, *Scandinavian Journal of Statistics* **22** (1995), pp. 159–182.