Persi Diaconis
Department of Mathematics
Harvard University
Cambridge, Mass. 02138

Laurent Saloff-Coste
CNRS, Université Paul Sabatier
Lab. Statistique Probabilité
31062 Toulouse CEDEX, France

Abstract

The Metropolis algorithm is a widely used procedure for sampling from a specified distribution on a large finite set. We survey what is rigorously known about running times. This includes work from statistical physics, computer science, probability and statistics. Some new results are given as an illustration of the geometric theory of Markov chains.

1. Introduction.

Let $\mathfrak{X}$ be a finite set and $\pi(x) > 0$ a probability distribution on $\mathfrak{X}$. The Metropolis algorithm is a procedure for drawing samples from $\mathfrak{X}$. It was introduced by Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller [1953]. The algorithm requires the user to specify a connected, aperiodic Markov chain $K(x, y)$ on $\mathfrak{X}$. This chain need not be symmetric but if $K(x, y) > 0$, one needs $K(y, x) > 0$. The chain K is modified by auxiliary coin tossing to a new chain M with stationary distribution π. In words, if the chain is currently at x, one chooses y from $K(x, y)$. Let

$$A(x, y) = \frac{\pi(y)K(y, x)}{\pi(x)K(x, y)}. \tag{1}$$

If $A(x, y) \geq 1$, the chain moves to y. If $A(x, y) < 1$, flip a coin with probability of heads $A(x, y)$. If the coin comes up heads, the chain moves to y. If the coin comes up tails, the chain stays at x. Formally,

$$M(x, y) = \begin{cases} K(x, y) & \text{if } A(x, y) \geq 1 \text{ and } y \neq x \\ K(x, y)A(x, y) & \text{if } A(x, y) < 1 \\ K(x, y) + \sum_{z: A(x,z)<1} K(x, z)(1 - A(x, z)) & \text{if } y = x. \end{cases} \tag{2}$$

The following lemma says that the new chain has π as its stationary distribution:

LEMMA 1. *The chain $M(x, y)$ at (2) is an irreducible, aperiodic Markov chain on $\mathfrak{X}$ with*

$$\pi(x)M(x, y) = \pi(y)M(y, x) \qquad \text{for all } x, y. \tag{3}$$

In particular, for all x, y

$$\lim M^n(x, y) = \pi(y).$$

PROOF: Equation (3) is easily verified directly: if $A(x, y) > 1$,

$$\pi(x)M(x, y) = \pi(x)K(x, y) \quad \text{and} \quad A(y, x) < 1 \quad \text{so} \quad \pi(y)K(y, x)A(y, x) = \pi(x)M(x, y).$$

The same conclusion holds if $A(x, y) = 1$ and if $A(x, y) \leq 1$. The chain is clearly connected and is aperiodic by assumption. Now, the basic convergence theorem for Markov chains, see e.g., Karlin and Taylor [ ], implies the result. □

REMARK: In applications, $\mathfrak{X}$ is often a huge set and the stationary distribution π is given as $\pi(x) \propto e^{-H(x)}$ with $H(x)$ easy to calculate. The unspecified normalizing constant is usually impossible to compute. Note that this constant cancels out of the ratios $A(x,y)$ so that the chain $M(x,y)$ is easy to run.

The limit result (4) is unsatisfactory in applied work, one needs to know how large n should be to have $M^n(x,y)$ suitably close to $\pi(y)$. One standard quantification of "close to stationarity" is the total variation distance:

$$\|M_x^n - \pi\| = \max_{A \subset \mathfrak{X}} |M^n(x,A) - \pi(A)|, \quad \text{with} \quad \pi(A) = \sum_{y \in A} \pi(y).$$

If this distance is small, then the chance that the chain is in a set A is close to $\pi(A)$, uniformly. The techniques described below give fairly sharp bounds on convergence in terms of the size $|\mathfrak{X}|$, and the geometry of a graph with vertex set $\mathfrak{X}$ and an edge from x to y if $M(x,y) > 0$.

Section 2 describes a collection of examples where very sharp results are known. These include a chain on the symmetric group drawn from our joint work with Phil Hanlon, a variety of birth and death chains drawn from thesis work of Eric Belsley, and results for independence sampling base chains drawn from work by Jun Liu. The analysis requires a high degree of symmetry, but at least gives a collection of examples where the correct answer is known, so different bounds can be compared with the truth.

Section 3 describes work on statistical physics models widely used in image analysis. These include the Ising model and many variations. In low dimensions, away from "critical temperatures" and "phase transitions" the results show that order $n \log n$ steps are necessary and suffice where n is the number of lattices sites. In phase transition regions, the running time can be exponential in n. The main work here here is due to Martinelli, Schoneman, Stroock, Zegalinski, and their co-authors.

Section 4 gives an overview of the geometric theory. This consists of Poincaré, Cheeger, Nash, Sobolev, and Log Sobelev inequalities.

Section 5 describes work on sampling from log concave densities on convex sets. This work has been developed in computer science by Frieze, Kannen, Lovasz, Shimonovitz and their co-authors in connection with the celebrated problem of computing the volume of a convex set.

Section 6 describes some new work which allows sharp bounds for Metropolis chains on low-dimensional grids. This work is presented as an introduction to the geometric theory of Markov chains developed in [, , ,]. It gives matching upper and lower bounds (up to good constants) for problems like sampling from

$$\pi(i) \propto i(n-i) \quad \text{or} \quad \pi(i) \propto e^{-(i-\frac{n}{2})^2}$$

on $\{1, 2, \cdots n-1\}$, with the base chain being reflecting random walk.

The final section attempts to survey other literature, extensions to general state spaces, and some of the many improvements on the Metropolis algorithm which (currently) seem beyond rigorous analysis.

Real applications of the Metropolis algorithm are widespread. If the reader needs convincing, we recommend the three discussion papers in the journal of the Royal Statistical Society, Series B, **55** (1993), No. 3, these give many illustrations and pointers to the huge applied literature.

We have made no attempt to cover closely related work on annealing or the Gibbs sampler. We have attempted to give a complete picture of what is rigorously known about the Metropolis algorithm.

2. Examples. This section reports work on examples where symmetry allows careful analysis.

Example 1. A walk on the symmetric group. Let S_n denote the permutations of n items. In psychophysical experiments (rank these sounds for loudness), taste-testing, and preference studies, a variety of non uniform distributions on S_n are used. One family, the Mallows-model through the metric d, has form

$$\pi_\theta(\sigma) = \theta^{d(\sigma,\sigma_0)}/Z. \tag{2.1}$$

with $d(\cdot,\cdot)$ a metric on S_n, and σ_0 a centering permutation. We take $0 < \theta \leq 1$ and $Z(\theta, \sigma_0)$ a normalizing constant. Thus, if $\theta = 1$, π_θ is the uniform distribution. If $\theta < 1$, the distribution peaks at the permutation σ_0 and falls off exponentially as σ moves away from σ_0. A variety of metrics are in use. For example,

$$d(\sigma, \sigma_0) = \sum |\sigma(i) - \sigma_0(i)| \quad \text{(Spearman's footrule)} \tag{2.2a}$$

(2.2b) $d(\sigma, \sigma_0)$ = min. number of transpositions required Cayley distance to bring σ to σ_0.

Detailed discussion can be found in Diaconis [1988; Chapter 6], Critchlow [1985], or Fligner and Verducci [1993].

For n large (e.g., $n = 52$) the normalizing constant is impossible to calculate and samples from π_θ would routinely be drawn using the Metropolis algorithm from the base chain of random transpositions. Thus, if the chain is currently at σ, the chain proceeds by choosing i, j at random in $\{1, 2, \cdots, n\}$ and transposing, forming $\sigma' = (i, j)\sigma$. If $d(\sigma', \sigma_0) \leq d(\sigma, \sigma_0)$, the chain moves to σ'. If $d(\sigma', \sigma_0) > d(\sigma, \sigma_0)$ a coin is flipped with probability $e^{-d(\sigma', \sigma_0) - d(\sigma, \sigma_0)}$. If this comes out heads, the chain moves to σ'. Otherwise the chain stays at σ.

The running time of this chain for the Cayley distance was analyzed in [ ]. The following result shows that order $n \log n$ steps are necessary and suffice for convergence.

THEOREM 2.1. *For fixed $0 < \theta < 1$, let M^k be the k^{th} power of the Metropolis chain (2.1), starting at the identity, with the Cayley metric (2.2b). Suppose*

$$k = an \log n + cn, \quad \textit{with} \quad a = \frac{1}{2\theta} + \frac{1}{4\theta}\left(\frac{1}{\theta} - \theta\right), \quad \textit{and} \quad c > 0.$$

Then

$$\|M^k - \pi_\theta\| \leq f(\theta, c)$$

for $f(\theta, c)$ an explicit function, independent of n, with $f(\theta, c) \searrow 0$ as $c \nearrow \infty$.

Conversely, if $k = \frac{1}{2} n \log n - cn$,

$$\|M^k - \pi_\theta\| \nearrow 1 \quad \textit{as} \quad c \nearrow \infty.$$

REMARKS: 1. We conjecture that this chain has a sharp cutoff in its variation distance at $an \log n$. The result gives quite precise sense to "$n \log n$ steps are necessary and suffice".

2. The proof of Theorem (2.1) depends in crucial ways on the choice of d as Cayley's metric. It uses delicate estimates of all eigenvalues and eigenvectors, available through symmetric function theory.

3. We conjecture that order $n \log n$ steps are necessary and suffice for any reasonable metric (e.g., 2.1a). At present, the best that can be rigorously proved is that order $n!$ steps suffice and order $n \log n$ steps are necessary.

4. The paper with Hanlon gives several other special cases where such careful analysis can be carried out – Metropolis algorithms on the hypercube and families of matchings. In these cases, the Metropolis chains (as at 2.1) give a one-parameter family of deformations of transition matrix of the base chain having interesting special functions as eigen functions. Ross and Xu (1995) have made a fascinating connection between some of these twisted walks and convolution of hypergroups.

5. Belsley ( ) has carried out a delicate analysis of a related case: changing the base chain of random walk on a path to a geometric distribution. His results are described further in section 6.

Example 2. Independence base chains. Let π be a probability on the finite set $\mathfrak{X}$. Consider as the base chain repeated independent samples from a fixed probability $p(x)$ on $\mathfrak{X}$. Thus $K(x, y) = p(y)$ for all x. Jun Liu (1995) has explicitly diagonalized the Metropolis chain in this case. To describe his results, let $w(x) = \pi(x)/p(x)$. The chain can be written

$$M(x, y) = \begin{cases} p(y) \min\left\{1, \frac{w(y)}{w(x)}\right\} & \text{if} \quad y \neq x \\ p(x) + \sum_x p(z) \max\left\{0, 1 - \frac{w(z)}{w(y)}\right\} & \text{if} \quad y = x. \end{cases} \tag{2.3}$$

Such a chain arises naturally when comparing the widely used schemes of importance and rejection sampling with the Metropolis algorithm. In these schemes an independent sample is drawn from p. In importance sampling, averages of functions with respect to π are estimated by weighting the sample value x by $w(x)$. In

rejection sampling, sample values x are kept in the sample with probability $w(x)$ and thrown away otherwise. These are close cousins to the Metropolis algorithm.

To describe Liu's results, let the states be numbered (without loss) so $w(x_1) \geq w(x_2) \geq \cdots \geq w(x_{|\mathfrak{X}|})$. Write $w(i) = w(x_i), \pi(i) = \pi(x_i)$, etc. Let

$$S_\pi(k) = \pi(x_k) + \cdots + \pi(x_{|\mathfrak{X}|}), \quad S_p(k) = p(x_k) + \cdots + p(x_{|\mathfrak{X}|}).$$

THEOREM 2.2. *(Liu) The Metropolis chain (2.3) has eigenvalues* $1 = \beta_0 > \beta_1 \cdots > \beta_{|\mathfrak{X}|-1} > -1$ *with*

$$\beta_j = \sum_{i \geq j} \left(\frac{\pi(i)}{w(i) - w(j)}\right).$$

Further, the variation distance for the chain with starting distribution p *is bounded above by*

$$4\|M_x^k - \pi\|^2 \leq \sum_{j=1}^{|\mathfrak{X}|-1} \left\{ \frac{S_p(j)}{S_\pi(j)} - \frac{S_p(j+1)}{S_\pi(j+1)} + \frac{1}{w(j)} \right\} \beta_j^{2k}. \tag{2.4}$$

For the chain stated at x,

$$4\|M_x^k - \pi\|^2 \leq \pi(x)^{-1} \beta_1. \tag{2.5}$$

PROOF: With hindsight, it is quite straightforward to verify the result discovered by Liu: with states numbered as above, an eigenvector corresponding to eigenvalue β_k is

$$(0, \cdots, 0, \ S_\pi(k+1), -\pi(k), \cdots, -\pi(k))$$

where there are $(k-1)$ zero entries. For reversible Markov chains, the Cauchy-Schwartz inequality and the spectral theorem give

$$4\|M_x^k - \pi\|^2 \leq \|\frac{M_x^k}{\pi} - 1\|_2^2 = \sum_{j=1}^{|\mathfrak{X}|-1} \beta_j^{2k} f_j^2(x) \leq \frac{\beta_j^{2k}}{\pi(x)} \tag{2.6}$$

with f_j an orthonormal basis of right eigenfunctions for the matrix M. See e.g. [Lemma]. Normalizing the eigenfunctions and straightforward computation give (2.4) while (2.5) follows from the rightmost inequality of (2.6).

Here is a simple example for comparison with later examples: take
$\mathfrak{X} = \{0, 1, 2, \cdots n-1\}$, $\pi(j) = \theta^j / Z$, with $Z = \frac{1-\theta^n}{1-\theta}$ and $0 < \theta < 1$ fixed. Take the base chain uniform on $\mathfrak{X} : p(j) = 1/n$. Thus the states are naturally ordered and $w(j) = n\pi(j)$. From the theorem,

$$\beta_i = \sum_{j=1}^{n-1} \left(\frac{1}{n} - \frac{\pi(j)}{n\pi(j)}\right) = 1 - \frac{1}{n}\left(1 + \frac{(1-\theta^{n-2})}{1-\theta}\right)$$

the upper bound (2.5) gives

$$4\|M_{n-1}^k - \pi\|^2 \leq \theta^{-n}\left(1 - \frac{2}{n}\right)^{2k}.$$

This shows that k of order n^2 steps suffice to achieve stationarity. Use of all the eigenvalues, as at (2.4) shows that order n steps actually suffice for any starting state. It is clear that at least n steps are necessary: even if the chain starts at 0, the most likely state, it takes order n steps to have a good chance of moving once.

Liu uses the the results above to compare importance sampling, rejection sampling and the Metropolis algorithm for estimating expected values like $\Sigma h(x)\pi(x)$. Using the criterion of mean square error, he concludes, roughly speaking, that the Metropolis algorithm and rejection methods have essentially the same

efficiency, but importance sampling can show big gains. Of course, this application of the Metropolis algorithm is far from the original motivation: importance sampling assumes we can compute, or at least approximate, normalizing constants while the Metropolis algorithm can proceed without them.

3. Models from statistical physics.

Statistical physics has introduced a variety of models which are also used to analyze spatial data and model images in vision and image reconstruction. In this description, we restrict attention to binary spatial patterns in a portion of a lattice. For simplicity, we also restrict attention to the Ising model. The references cited apply to much more general situations.

Thus let Λ be a finite connected subset of the lattice $\mathbb{Z}^2$. Let $\mathfrak{X} = \{x : \Lambda \to \mathbb{Z}_2\}$. We think of $\mathbb{Z}_2 = \{\pm 1\}$ and $\mathfrak{X}$ as the set of two colorings of the sites in Λ. If $\{\pm\}$ is replaced by $\{0, 1\}$, we may think of an element of $\mathfrak{X}$ as a picture. Let s be a two-coloring of the boundary of Λ (points in $\mathbb{Z}^2 - \Lambda$ at distance 1 from points in Λ). This is a specified set of boundary conditions.

The Ising model is a probability distribution on $\mathfrak{X}$ specified by

$$\pi(x) \propto e^{\beta(\sum_{\langle i,j\rangle} x_i x_j + h \sum_i x_i)} \tag{3.1}$$

where the first sum is over neighboring pairs in $\mathbb{Z}^2$ with one or both of i, j in Λ and the second sum is over i in Λ. Here $\beta > 0$ is called inverse temperature and h, $-\infty < h < \infty$ is called the external field strength. With β, h, s fixed, (3.1) is a well specified probability measure on $\mathfrak{X}$. In applications, Λ is usually a square grid of size, e.g. 64×64 or 128×128 and it is clearly impossible to calculate the normalizing constant implicit in (3.1).

The Metropolis algorithm gives an easy way to generate from π; as base chain, let us take the following: pick i in Λ at random (uniformly) and change x_i to $-x_i$. This gives a connected chain on $\mathfrak{X}$. Call this random single site updating. This chain is periodic, but the Metropolis algorithm clearly has some holding probability so the chain $M^n(x, y)$ converges to $\pi(y)$.

There is a huge rigorous literature on properties of the stationary distribution π as a function of β, h, and s. Simon (1993) gives a careful extended discussion. We will not review this here but merely mention that there are regions of the β, h plane where the behavior of s matters (phase transitions occur) and regions where the behavior does not matter. Phase transitions occur for $h = 0$ and $\beta < \beta_c$ and not otherwise. As will be described, the Metropolis algorithm converges rapidly for (β, h) away from the critical values (order roughly $|\Lambda|^2 \log |\Lambda|$ steps suffice). It takes an exponential number of steps to converge for (β, h) critical values. The behavior of the constants involved as (β, h) approach the critical values is currently under active study. Schonmann (1993,1995) gives a review of this fascinating subject.

To state a precise result an annoying periodicity problem must be dealt with. Let

$$\widehat{M}(x, y) = \frac{1}{2}(I + M(x, y)) \tag{3.2}$$

be a modified Metropolis chain.

THEOREM 3.1. *(Martinelli-Olivieri-Schonmann (1993)) Let Λ be a square grid in $\mathbb{Z}^2$ with $|\Lambda| = n$. Then, for β, h not on the critical segment and any s, the Metropolis chain (3.2) for π defined at (3.1) based on random single site updating satisfies*

$$||M_x^k - \pi|| \leq Ae^{-Bk/(n^2 \log n)}$$

with A, B explicit functions of β, h which do ot depend on n or s.

REMARKS: 1. A very similar result was proved earlier by Stroock and Zegarlinski (1992). Their result holds for somewhat fewer values of β, h (e.g., $|h| \geq 4$ is required) but is stronger in holding uniformly for all Λ (not just square grids). They also give results which hold for larger dimensions while the techniques of Martinelli-Olivieri-Schonmann lean heavily on the assumption of $\mathbb{Z}^2$. A detailed comparison is in Frigessi, Martinelli-Stander (1993).

2. For β, h on the critical segment, things change drastically. Results of Martinelli (1993) and Thomas (1989) combine to show that the chain $\widetilde{M}$ takes order $e^{Bn^{1/2}}$ steps to converge. Again, B is a function of β, h and now s; indeed in the critical segment the stationary distribution π depends strongly on the boundary conditions which now do not wash away for large grids.

The proofs of the theorems above depend on detailed study of stationary distribution π and build on years of work by the statistical physics community. There is not much hope of carrying them over in any straightforward way to other high-dimensional uses of the Metropolis algorithm such as the permutation distributions of section 2. There is one very useful ingredient which is clearly broadly useful, the Log Sobolev inequality. The next section gives a brief description of this emerging technique.

4. Geometric techniques.

A hierarchy of technical tools have emerged for studying powers of Markov chains. At present, these go well beyond bounds on eigenvalues. The geometric tools are named after cousins from differential geometry and differential equations: inequalities of

Poincaré, Cheeger, Sobolev, Nash, Log Sobolev.

It is beyond the scope of this paper to give a thorough introduction to these; we give a brief outline and pointers to good expositions. Basic references are [, ,] with Sinclair [ ] a useful recent book.

For simplicity, we work in the context of reversible Markov chains although one of the exciting breakthroughs (see Fill [ ] and [,]) is that much can be pushed through in the non reversible case.

Let $\mathfrak{X}$ be a finite set, $K(x,y)$ an irreducible, aperiodic Markov matrix on $\mathfrak{X}$. Let $\pi(x)$ be the stationary distribution and suppose π, K is reversible (so $\pi(x)K(x,y) = \pi(y)K(y,x)$). Define an inner product on real functions from $\mathfrak{X}$ by $\langle f|g\rangle = \sum f(x)g(x)\pi(x)$. Then reversibility is equivalent to saying the operator K which takes f to $Kf(x) = \sum K(x,y)f(y)$ is self adjoint on ℓ^2 (so $\langle Kf|g\rangle = \langle f|Kg\rangle$). This implies that K has real eigenvalues

$$1 = \beta_0 > \beta_1 \cdots > \beta_{|\mathfrak{X}|-1} > -1$$

and an orthonormal basis of eigenvectors f_i (so $Kf_i = \beta_i f_i$).

One aim is to bound the total variation distance between $K^n(x,y)$ and $\pi(y)$. This is accomplished by using the Cauchy-Schwarz inequality to bound

$$4\|K_x^n - \pi\|_{TV}^2 \le \left\|\frac{K_x^n}{\pi} - 1\right\|_2^2 = \sum_{i=1}^{|\mathfrak{X}|-1} f_i^2(x)\beta_1^{2n} \le \frac{1}{\pi(x)}\beta_*^{2n} \tag{4.1}$$

with $\beta_* = \min(\beta_1, |\beta_{|\mathfrak{X}|-1}|)$.

This final bound is clearly proved by Jerrum and Sinclair [ ]. See also [, Sec. 6].

Thus one can get bounds on rates of convergence using eigenvalues. Next, one needs to get bounds on eigenvalues. This can be accomplished by using the minimax characterization. This involves the quadratic form

$$\mathcal{E}(f|f) = \langle (I-K)f|f\rangle = \frac{1}{2}\sum_{x,y}(f(x)-f(y))^2\pi(x)K(x,y).$$

Then

$$1 - \beta_1 = \min_f \frac{\mathcal{E}(f|f}{var(f)}, \quad var(f) = \sum (f(x)-\overline{f})^2\pi(x), \quad \overline{f} = \sum f(x)\pi(x). \tag{4.2}$$

Because of (4.2), bounds of $var(f)$ in terms of $\mathcal{E}(f|f)$

$$var(f) \le A\mathcal{E}(f|f)$$

or equivalently, bounds on the ℓ^2 norm on functions with $\overline{f} = 0$:

$$\|f\|_a^2 \le A\mathcal{E}(f|f)$$

give bounds $\beta_1 \leq 1 - \frac{1}{A}$. Such bounds are called *Poincaré inequalities.* An illustration of these techniques is given in Section 6A below.

In [ ] a simple technique for proving a Poincaré inequality is given using paths γ_{xy} from x to y in a graph with vertex set $\mathfrak{X}$ and an edge from z to w if $K(z,w) > 0$. These paths had been suggested earlier by work of Jerrum and Sinclair to bound eigenvalues using conductance (see our discussion of Cheeger's inequality below). It emerged that whenever paths were available, their direct use in Poincaré inequalities was preferable to their use via conductance. For example, Jerrum and Sinclair's pioneering work on approximation of the permanent used paths and conductance to give a bound for the second eigenvalue of the underlying chain

$$\beta_1 \leq 1 - \frac{c}{n^{12}}.$$

Using just their calculations, and replacing conductance by Poincaré, [ ] shows

$$\beta_1 \leq 1 - \frac{c}{n^5}.$$

Sinclair [ ] then went through several other arguments (problems of generating graphs with given degree, dimmer problems) and obtained substantial improvements in every case.

Cheeger's inequality bounds eigenvalues by considering

$$h = \min_{\pi(A) \leq \frac{1}{2}} \frac{\sum_{x \in A, y \in A^c} \pi(x) K(x,y)}{\pi(A)}. \tag{4.3}$$

Bounds on eigenvalues are obtained via

$$1 - 2h^2 \leq \beta_1 \leq 1 - \frac{h^2}{2}. \tag{4.4}$$

These ideas were introduced into combinatorial work by Alon and his coworkers for building expander graphs. There, the quantity h is of interest, one builds graphs where group theory can be used to bound β_1 and this gives bounds on h. The idea of getting bounds on β_1 by getting bounds on n directly is an important contribution of Jerrum and Sinclair.

An interesting class of problems where graphs can be embedded in Euclidean space and then tools from continuous geometry (Payne Weinberger inequalities) can be used to give direct bounds on h has been intensively studied in computer science by Dyer, Frieze, Kannan and Lovasz, Shimonovitz. This leads to remarkable bounds for problems like approximating the volume of convex sets. These seem unobtainable by other methods at present writing.

Roughly, these bounds proceed by taking a fine mesh (the underlying graph) in an ambient Euclidean space. Then, the eigenvalues of the graph Laplacian are shown to be close to the known eigenvalues of the combinatorial Laplacian. A superb survey was given by Kannan [ ]. A recent very interesting effort along these lines is given by Chung and Yau [ ].

Cheeger and Poincaré inequalities are fairly basic tools in modern geometry. More refined results are obtainable by using Nash, Sobolev, and Log Sobolev inequalities to which we now turn. Details for the following can be found in [,].

While Poincaré inequalities bound the ℓ^2 norm in terms of the quadratic form, Nash inequalities ask for more, a bound on a power of the ℓ^2 norm. In terms of the form, this appears as

$$\|f\|_2^{2+1/D} \leq B\{\mathcal{E}(f|f) + \frac{1}{N}\|f\|_2^2\} \tag{4.5}$$

where f is restricted to have ℓ^1 norm one. In (4.5), B, D and N are constants which enter into any conclusions.

In [ ] it is shown that (4.5) is *equivalent* to the conclusion that powers of the kernel M^n decay like C/n^D for $1 \leq n \leq N^2$. This gives crude bounds: "the N^{th} power is roughly flat" from which one can then use eigenvalue bounds. It is a little like finding the maximum of a function. One first uses a crude global tool to get into a neighborhood of the maximum and then uses something like Newton's method.

When applicable, Nash inequalities allow elimination of the $\pi(x)^{-1}$ term in the upper bound (4.1). Here is an example of what can be done. Let $\mathcal{C}$ be the lattice points inside a convex compact set in $\mathbb{R}^2$. Assume that two points in $\mathcal{C}$ can be connected by a lattice path within $\mathcal{C}$. A random walk proceeds by uniformly choosing one of the 4 possible neighbors of $x \in \mathcal{C}$. If the neighbor is inside $\mathcal{C}$, the walk moves to the chosen point. If the neighbor is outside $\mathcal{C}$, the walk stays at x. This gives a Markov matrix

$$K(x,y) = \begin{cases} 1/2d & \text{for } x \neq y \text{ neighboring points in } \mathcal{C} \\ g(x)/2d & \text{for } x = y \text{ in } \mathcal{C}. \end{cases} \tag{4.6}$$

where $g(x)$ is the number of neighbors of x that do not belong to $\mathcal{C}$. This is a connected, irreducible chain with the uniform distribution $\pi(x)$ as reversible measure. The following result is proved in [ ].

THEOREM 4.1. *For the chain (4.6), let $\gamma > 1$ be the euclidean diameter of $\mathcal{C}$. There are universal constants $a_1, a_2 > 0$ such that for any $x \in \mathcal{C}$,*

$$\|K_x^n - \pi\|_{TV} \leq a_1 e^{-a_2 c} \quad \text{for } n = c\gamma^2, \quad c > 0.$$

Further, there are universal constants $a_3, a_4 > 0$ such that

$$\|K_x^n - \pi\|_{TV} \geq a_3 e^{-a_4 c} \quad \text{for } n = c\gamma^2, \ c > 0$$

for some x in $\mathcal{C}$.

Roughly, Theorem 4.1 says order (diameter)^2 steps are necessary and suffice to approach stationarity.

In [ ] a very similar theorem is proved for the natural random walk on contingency tables, and natural walks on

$$\{x \in \mathbb{N}^\beta : Mx = y\}$$

with M an $a \times b$ totally unimodular matrix, y a given vector of integers. Further, preliminary calculations show that the same conclusions ((diameter)^2 steps are necessary and suffice) hold for the Metropolis algorithm with base chain nearest neighbor random walk on a low-dimensional grid if the stationary distribution is of polynomial type: e.g., proportional to $i(n-i)$ on $\{1, 2, \cdots, n-1\}$. Section 6B carries this out in some detail.

Techniques for proving such Nash inequalities that use paths in a local fashion (local Poincaré inequalities) are given in [ ]. These techniques seem broadly useful for problems in which the underlying graph has polynomial growth. One drawback; the constants involved grow exponentially in the dimension parameter D.

Sobolev inequalities are essentially equivalent to Nash inequalities. They ask for bounds of form

$$\|f\|_q^2 \leq C\{\mathcal{E}(f|f) + \frac{1}{T}\|f\|_2^2\}$$

where $q > 2$ and C, T are constants.

See [ ] for the equivalence of Sobolev and Nash inequalities. See Chung and Yau [ ] for a development of Sobolev inequalities on graphs.

Log Sobolev inequalities give a tool that is not plagued by the curse of dimension. Indeed, these inequalities were invented by analysts in trying to get results in infinite dimensions. A splendid introduction and survey to the continuous work is in Gross [ ]. The volume this is contained in has further useful articles. A careful account of Log Sobolev inequalities for finite problems is in [ ] from which the present account is drawn.

We say a chain K satisfies a *Log Sobolev inequality* if

$$\mathcal{L}(f) \leq c\mathcal{E}(f|f) \qquad \text{for some } c > 0 \quad \text{and all} \quad f$$

with

$$\mathcal{L}(f) = \sum_x f^2(x) \log\left(\frac{f^2(x)}{\|f\|_2^2}\right)\pi(x).$$

The best constant will be denoted c_* in the sequel.

If such an inequality is available, then

$$4\|K_x^n - \pi\|_{TV} \le (1 + 2e^2)^{1/2} e^{-c} \quad \text{for} \quad n \ge \frac{c_*}{4} \log\log \frac{1}{\pi(x)} + \frac{c}{1 - \beta_{|\mathcal{X}|-1}} + 1, \quad c > 0.$$

This inequality should be compared with (4.1) where the quantity $\log \frac{1}{\pi(x)}$ appears instead of $\log\log \frac{1}{\pi(x)}$. Above, $\beta_{|\mathcal{X}|-1}$ is needed to guard against periodicity problems. This is seldom a problem (see [, Sec.], for techniques for bounding $\beta_{|\mathcal{X}|-1}$).

Going from Poincaré/Cheeger inequalities to Nash/Sobolev inequalities necessitates more sophisticated use of available information: paths must be used locally and additional information such as polynomial growth of the underyling graph must be incorporated.

Good Log Sobolev inequalities are yet more difficult to prove. The situation is not all bad; Log Sobolev inequalities with poor constants can be extremely useful. Further, many mathematicians are working hard on these problems and there is much progress.

One of the key ideas is *hypercontractivity*. To explain, observe that applying a Markov kernel is a smoothing operation that flattens a point function into the stationary distribution after sufficiently many operations. One way to quantify this smoothing is to look at the norm of the kernel from one space of functions to another, e.g., for $1 \le p, q \le \infty$

$$\|K\|_{p \to q} = \min A \qquad \text{such that} \quad \|Kf\|_q \le A\|f\|_p.$$

For reversible chains, one can prove

THEOREM 4.2. *If* $\|K^n\|_{2 \to q} \le 1$ *for all* $n \ge 0$, $2 \le q < \infty$ *with* $e^{4\beta n} \ge q - 1$, *then*

$$\beta \mathcal{L}(f) \le \mathcal{E}(f|f).$$

There is also a fairly sharp converse. Thus smoothing bounds with interrelated n and q are equivalent to Log Sobolev inequalities.

In [ ] these ideas are used to give useful bounds on a variety of problems.

It is worth pointing out that it is extremely difficult to give the exact Log Sobolev constant. In fact, essentially the only non-trivial finite case where this value is known by direct argument is simple random walk on a two point space. See Gross [Ex. 2.6] for this not entirely trivial calculation. Getting the correct value for a path of length 3 is an open problem.

However, the Log Sobolev inequality for the direct product of two Markov chains follows easily from this inequality for the factors. This gives the Log Sobolev for the hypercube $\mathbb{Z}_2^d$.

All of the proofs for the Metropolis algorithm for Ising models cited in section 3 use Log Sobolev inequalities. In particular, Stroock and Zegalinski [ ] show that if a chain is built up step by step with an approximate product structure, e.g., mild dependence, then one can draw useful conclusions for the large chain.

5. Sampling from log concave densities and volume approximation.

Let K be a compact, convex set in Euclidean space $\mathbb{R}^d$. Let $f(x)$ be a probability density on K. Consider the problem of sampling from $f(x)$. This problem has been intensively studied in recent years in close connection with the problem of approximations to the volume of K. A comprehensive survey is given by Kannan [ ]. We here focus on the parts of the work having to do with the Metropolis algorithm.

A. Discrete algorithms. Frieze, Kannan, and Polson (1994) discretized the problem, dividing $\mathbb{R}^d$ into hypercubes of size δ, and running the Metropolis algorithm on a graph with vertices the centers of cubes intersecting K, with an edge between vertices if the cubes are adjacent. The weight at center x is the average of f over the cube containing x.

They assume available an approximation $\overline{f}(x)$ (defined only on the cube centers) which satisfies the following approximation and continuity requirements: for some $\alpha > 0$,

$$(1 + \alpha)^{-1} \overline{f}(x) \le \overline{f}(y) \le (1 + \alpha) \overline{f}(x) \quad \text{for adjacent points} \tag{5.1}$$

$$(1+\alpha)^{-1}\delta^d \overline{f}(x) \leq \int_{c(x)} f(z)dz \leq (1+\alpha)\delta^d \overline{f}(x) \tag{5.2}$$

$$(1+\alpha)\delta^{d-1}\overline{f}(x) \leq \int_{c(x)\cap c(y)} F(z)dz \leq (1+\alpha)\delta^{d-1}\overline{f}(x). \tag{5.3}$$

For $c(x), c(y)$ cubes having $c(x) \cap c(y)$ of dimension $d-1$.

With these assumptions, it is sufficient to analyze the Metropolis algorithm with weight $\overline{f}(x)$ at x.

We state here a special case of their result where $K = B(R)$, the Euclidean ball of radius R centered at 0, and where f satisfies the following assumption on its support: consider the half line $L_u = \{ru : ru\mathbb{R}^+\}$ with $u \in \mathbb{R}^d$. Let $h(r) = r^{n-1}f(ru)$ be defined on L_u. This is a log concave function of r if f is log concave. The following assumption says that the tails of f are at the boundary of $B(R)$.

$$\text{For all } R, \quad R \leq r \leq r', \quad h(0) \geq h(R) \geq h(r) \geq h(r'). \tag{5.4}$$

With these assumptions, the following result can be stated.

THEOREM 5.1. *(Frieze, Kannan, Polson) Let f be a log concave probability density which is positive on $\mathbb{R}^d$ and satisfies (5.1 - 5.4). Let $M(x,y)$ be the Metropolis algorithm on the centers of cubes of side δ which intersect the ball $B(R)$. Assume $\delta \leq R$. Then*

$$||M_x^k - \pi||_{TV} \leq \overline{f}(x)^{-\frac{1}{2}}(1-\lambda)^k$$

where

$$\lambda^{-1} = \max\{(1+\alpha)^3(K_0(K_1+1)+K_2)d'\delta, \frac{\gamma^2}{6\delta^2}\} \sim \frac{d}{4}(\frac{\gamma}{\delta})^2$$

for γ the Euclidean diameter of the set of cubes involved (the greatest distance between two such cubes),

$$K_0 = \frac{\gamma}{4\delta}(\gamma + 2\sqrt{d}\delta), K_1 = \frac{18\delta\sqrt{\lambda}}{\gamma}, K_2 = 9\sqrt{d}(\frac{\gamma}{\gamma} + \delta(\sqrt{d}+1)).$$

The final approximation holds as $\alpha \searrow 0, \delta\sqrt{d} \nearrow \gamma \to 0$.

REMARKS:

1. This result is remarkable even in fixed dimensions for a Gaussian density. Then it basically says that a natural algorithm converges exponentially fast, in a useful sense, that is, with good constants.
2. In high dimensions, observe that the constants do *not* get bad.
3. The above is a special case of the arguments. The restriction to balls or the restriction (5.4) are not required. The final result is more complicated to state.
4. In the end, the argument rests heavily on properties of convex sets in Euclidean space. It does not seem easy to adapt the tools involved to more general graphs. One interesting technical development which does seem broadly useful: a technique is introduced for dealing with a small "bad" set of the stated space where, e.g., $\pi(x)$ is very small. This should not affect things, since basically the chain does not visit small sets. However, the usual conductance approach involves an infimum over all sets. A different, useful approach for eliminating a small bad set appears in Lovász and Simonovitz [1993].

B. Continuous algorithms. Lovász and Simonovits have introduced a series of techniques for analyzing a Metropolis algorithm for sampling from a log concave density f on a compact convex set. A convenient recent reference is [ ]. Their work analyzes the following natural algorithm: suppose the chain is at x. Pick y from the uniform distribution on a ball of radius δ centered at x. If y is not in K, the walk stays at x. If y is in K, and $f(y)/f(x) \geq 1$, the walk moves to y. For $f(y)/f(x) < 1$, the usual Metropolis coin flip is executed. The chain moves to y or stays at x depending on the outcome.

This walk is analyzed without discretization. They extend the tools of conductance to general state spaces. This must prove useful. The heart of the argument is the same set of ideas about convex geometry in Euclidean spaces that are used by Frieze, Kannan and Polson. These have evolved from the original work of finding polynomial algorithms for volume computation due to Dyer, Frieze, and Kannan ( ).

One main focus of [ ] is getting good bounds on the complexity of volume computation (they get an order $n^7(\log n)^3$ algorithm). The Metropolis algorithm enters as a tool: for a convex body K, let $\varphi(x)$ be the smallest number t for which $x \in tK$. Set $f(x) = e^{-\varphi(x)}$. Then, $Vol(K) = \frac{1}{n!}\int_{\mathbb{R}^n} f(x)dx$. Further sampling from f gives an algorithm for approximating K.

Meyn and Tweedie [1994] have begun work on extending the tools of Harris recurrence to get useful quantitative results. They develop the theory for abstract spaces but do try a simple example of the Metropolis algorithm for sampling from the normal distribution on $\mathbb{R}$, the base chain being discrete time steps from a different normal.

6. Low-dimensional examples.

This section treats two classes of low-dimensional examples: probability distributions on a low-dimensional grid with nearest neighbor random walk "metropolized" to the given stationary distribution.

Recall that nearest neighbor walk in a grid of side length n takes order n^2 steps to reach stationarity in any fixed dimension. If the target distribution has an exponential (or faster) fall off from a central peak, our analysis shows that the Metropolis chain reaches stationarity in order n steps. This is the fastest possible: the chain has to travel order n steps to go between opposite corners of the grid. If the target distribution has polynomial fall off from a certain peak, the analysis shows that order n^2 steps are necessary and suffice to reach stationarity.

The analysis is described in some detail as an illustration of geometric methods described in section 4 above. In the exponential case, one novelty is the use of different weights in the Cauchy-Schwarz inequality. This suggestion of Alan Sokal is shown to give improved results. In the polynomial case, the Nash inequalities of [ ] are the driving tool.

A. Exponential fall off. To fix ideas, consider a one-dimensional grid: $\mathfrak{X} = \{0, 1, 2, \cdots, n-1\}$. Let the base chain be nearest neighbor random walk with holding $\frac{1}{2}$ at both ends. Represent the stationary distribution as

$$\pi(i) = z(a)a^{h(i)} \quad 0 < a < 1, \quad z = z(a) \quad \text{the normalizing constant.} \tag{6.1}$$

We assume

$$h(i+1) - h(i) \geq c \geq 1, \qquad 0 \leq i \leq n-2. \tag{6.2}$$

Thus $\pi(i)$ falls off exponentially from 0. Example are $h(i) = i^b$, for $b \geq 1$. Here, the Metropolis chain becomes

$$\begin{array}{lll} M(i,i-1) = \frac{1}{2} & M(i,i) = \frac{1}{2} - \frac{a^{h(i+1)-h(i)}}{2} & M(i,i+1) = \frac{a^{h(i+1)-h(i)}}{2}, \ 1 \leq i \leq n-2 \\ M(0,0) = 1 - \frac{a^{h(2)-h(0)}}{2} & M(0,1) = \frac{a^{h(2)-h(0)}}{2} & M(n-1,n-2) = M(n-1,n-1) = \frac{1}{2}. \end{array} \tag{6.3}$$

The main result is the following bound for the second eigenvalue of the chain:

PROPOSITION 6.1. *Assume (6.1-6.3). Then, the second eigenvalue of the chain satisfies*

$$\beta_1 \leq 1 - \frac{(1 - a^{c/2})^2}{\sqrt{2}}.$$

REMARKS: Thus the eigenvalue is bounded away from 1 uniformly in the size of the state space. This will be used to show that order n steps are necessary and suffice for total variation convergence.

PROOF: The argument uses the path techniques of [ ] in a novel way. We have

$$1 - \beta_1 = \min_f \frac{\mathcal{E}(f|f)}{var(f)}$$

with the min taken over non-constant f, $varf = \frac{1}{2}\sum_{x,y}(f(x) - f(y))^2\pi(x)\pi(y)$ and the Dirichlet form $\mathcal{E}(f|f) = \frac{1}{2}\sum_{x,y}(f(x) - f(y))Q(x,y)$ for $Q(x,y) = \pi(x)M(x,y)$. Choose paths: for $x < y$, $\gamma_{xy} = (x, x+1, x+2, \cdots, y)$. The same path is used backwards to connect y to x. Then,

$$2var(f) = \sum_{x,y} |f(x) - f(y)|^2\pi(x)\pi(y) = \sum_{x,y}\left(\sum_{e\in\gamma_{xy}} f(e^+) - f(e^-)\right)^2 \pi(x)\pi(y). \tag{6.4}$$

The inner sum will be bounded by the Cauchy-Schwarz inequality. Usually, this is done with weights taken as 1 which gives a factor of $|\gamma_{xy}|$. The novelty here is to use weights depending on the stationary distribution. For the edge e, the weights are chosen as $Q(e)^\theta$. Subsequent calculations show that any fixed θ in $(0, \frac{1}{2})$ will do, e.g., $\theta = \frac{1}{4}$. To bring this out, we keep θ as a parameter. Multiply and divide $f(e^+) - f(e^-)$ in (6.4) by $Q(e)^\theta$. Writing $|\gamma_{xy}|_\theta = \sum_{e\in\gamma_{xy}} Q(e)^{-2\theta}$, we have

$$\begin{aligned} 2var(f) &\le \sum_{x,y} |\gamma_{xy}|_\theta \sum_{e\ni\gamma_{xy}} Q(e)^{2\theta}(f(e^+) - f(e^-))^2\pi(x)\pi(y) \\ &= \sum_e (f(e^+) - f(e^-))^2 Q(e)Q(e)^{2\theta-1} \sum_{\gamma_{xy}\ni e} \pi(x)\pi(y)|\gamma_{x,y}|_\theta \\ &\le 2A\mathcal{E}(f|f) \end{aligned}$$

with

$$A = \max_e Q(e)^{2\theta-1} \sum_{\gamma_{xy}\ni e} \pi(x)\pi(y)|\gamma_{xy}|_\theta.$$

To bound A, observe first that $Q(i, i+1) = Q(i+1, i) = \frac{\pi(i+1)}{2}$. Next, the dominant term in $|\gamma_{xy}|_\theta$ is $Q(y-1, y)$. Pull this out and bound the ratio with the other terms using (6.2):

$$|\gamma_{x,y}|_\theta \le \frac{\pi(y)^{-2\theta}}{1 - a^{2c\theta}}.$$

Suppose that $e = (i, i+1)$. The quantity to be bounded is

$$(1 - a^{2c\theta})^{-1} Q(e)^{2\theta-1} \sum_{\substack{0\le j\le i \\ i+1\le k\le n}} \pi(j)\pi(k)^{1-2\theta}.$$

The sum in k is bounded above by $\dfrac{\pi(i+1)^{1-2\theta}}{1 - a^{c(1-2\theta)}}$. The sum in j is bounded above by 1. Combining bounds, we have

$$A \le 2^{2\theta-1}(1 - a^{c(1-2\theta)})(1 - a^{2c\theta})$$

choosing $\theta = \frac{1}{4}$ gives the bound announced. □

REMARKS: 1. In Proposition 6.1 the stationary distribution was chosen to have its maximum at 0. The same argument works if the maximum is taken on at any point in $\mathfrak{X}$. Thus $h(i)$ decreases up to x_0 and increases past x_0; the analog of (6.2) is assumed.

2. The easiest upper bound for total variation using the eigenvalue bound of Proposition 6.1 is as follows. First, bound the smallest eigenvalue $\beta_{n-1} \ge -1 + 2\min(M(i,i) \ge -1 + 2(\frac{1}{2} - \frac{a^c}{2}) = a^c$. Thus

$$\beta_* = \min(\beta_1, |\beta_{n-1}|) \le \min(1 - \frac{(1 - a^{c/2})^2}{\sqrt{2}}, a^c).$$

Now, the upper bound at (4.1) gives, for any $x \in \mathfrak{X}$

$$\|M_x^k - \pi\| \leq \pi(x)^{-1/2}\beta_*^k.$$

This is correct (up to constants) when $h(x) = x$; it says order n steps are necessary and suffice to reach stationarity for any starting position. If h grows faster, e.g., $h(x) = x^2$, the bound shows that for a walk starting at $x = 0$, order n steps suffice. For walks starting at $n-1$ the bound shows order $h(n-1)$ steps are sufficient. This is off. The following argument shows how to conclude that for total variation convergence order n steps suffice for any starting position provided h satisfies (6.2). The walk started at $n-1$. Essentially stays still or goes left. It is straightforward to show that the chance that the walk hits 0 in the first $3n$ steps is exponentially close to 1, with constants depending only on $(1-a)$. Once the walk hits zero, the argument above shows it is close to stationarity in at most order n further steps. This shows order n steps suffice for variation distance convergence. We do not know how many steps are required to make the ℓ^2 norm small but suspect it may be order $h(n-1)$.

3. The argument for Proposition 4.1 is fairly robust and will handle many variations. It does depend on the roughly unimodal nature of π. There are techniques in Deuschel and Maza (1994) and Ingrassia (1994) for bounding essentially arbitrary π. While these bounds are sharp, in the sense that there are examples where they cannot be improved in nice examples, such as those of Proposition 4.1, they can be very far off, suggesting that exponentially many steps are needed. Much remains to be done in giving useful tools for natural examples.

4. When $h(i) = i$, Belsley ( ) has given very sharp upper and lower bounds on the second eigenvalue. We are presently surprised that the general path techniques give results quite close to the truth for this case. Belsley works out much sharper asymptotics for variation convergence for this case. He shows that $2n + c(a)\sqrt{n}$ steps are necessary and suffice: if $c(a)$ is large and positive, the variation distance is close to zero. If $c(a)$ is large and negative, the variation distance is close to one.

5. The restriction $c \geq 1$ in (6.2) is made for simplicity. If $h(i+1) - h(i) \geq c$, then $\frac{h(i+1)}{c} - \frac{h(i)}{c} \geq 1$ and the chain with a replaced by a^c and h replaced by h/c satisfies the conditions. This leads to the bound

$$\beta_1 \leq 1 - \frac{(1 - a^{c^2/2})^2}{\sqrt{2}} \quad \text{for} \quad c > 0.$$

6. The argument goes through more or less as above for two-dimensional versions with $h(i,j)$ falling at least linearly from a single peak. Here one chooses paths which move from x to y, first making the first coordinates equal, then the second coordinates equal, and so on. We hope to carry out a detailed analysis of the multimodal case on grids in low dimension.

7. For the one-dimensional case, it is worth pointing out that Cheegers inequality can be used to give results similar to those in Propositon 6.1. See [ ], section 3 for details. For higher-dimensional grids, we find paths much easier to work with.

8. In light of the results for sampling from log concave distributions in the continuous case (section 3 above), it is natural to inquire how this type of condition works in Proposition 6.1. While natural examples ar easy to treat, the following shows that some care is needed. Consider the symmetric binomial distribution $\pi(i) = \binom{n-1}{i}/2^{n-1}$ on $\{0, 1, 2, \cdots, n-1\}$, with base chain reflecting random walk, the Metropolis chain is easily comparable to the classical Ehrenfest chain. The analysis shows the Metropolis chain has $\frac{c_1}{n} \leq 1 - \beta_1 \leq \frac{c_2}{n}$ for explicit constants c_1, c_2. The difference is this: the binomial falls off from its peak at $\frac{n}{2}$ exponentially, but at scale $\sqrt{n}$. It is (roughly) flat in a $\sqrt{n}$ neighborhood of $\frac{n}{2}$. The exponentials treated by Proposition 4.1 fall off exponentiation at scale 1.

B. Polynomial Fall Off. Consider $\mathfrak{X} = \{1, 2, \cdots, n\}$, with the base chain of nearest neighbor random walk with holding $\frac{1}{2}$ at both ends. Take the stationary distribution

$$\pi(i) = zi, \quad 1 \leq i \leq n, \quad z^{-1} = n(n+1)/2. \tag{6.5}$$

Thus $\pi(i)$ rises linearly from 1.

The Metropolis chain becomes

$$
\begin{aligned}
&M(i,i-1)=\tfrac{i-1}{2i} \qquad M(i,i)=\tfrac{1}{2}-\tfrac{i-1}{2i} \qquad M(i,i+1)=\tfrac{1}{2} \qquad 2\le i\le n-1\\
&M(1,1)=M(1,2)=\tfrac{1}{2} \qquad M(n,n-1)=\tfrac{n-1}{2n} \quad M(n,n)=1-\tfrac{n-1}{2n}.
\end{aligned} \tag{6.6}
$$

The following result shows that the walk (6.6) reaches stationarity in order n^2 steps. This is the same rate as the base chain.

PROPOSITION 6.2. *There are explicit positive constants A, B, C, D such that the Metropolis chain (6.6) satisfies*

$$Ae^{-Bk/n^2} \le \|M_x^k - \pi\| \le ce^{-Dk/n^2}$$

for all positive integer k, n.

PROOF: We apply the geometric tools of [ ]. Consider $\mathfrak{X}$ as a graph with an edge from i to $j+1$, $1 \le i \le n-1$. Write $|x-y|$ for the graph distance between x and y. Let $B(x,r) = \{y : |x-y| \le r\}$ and $V(x,r) = \sum_{y\in B(x,r)} \pi(y)$. The diameter of $\mathfrak{X}$ is $\gamma = n$.

A graph and stationary distribution have $(A;d)$ *moderate growth.* If $V(x,i) \ge \frac{1}{A}(\frac{r+1}{\gamma})^d$ for all $x \in \mathfrak{X}$, and $r = \{0,1,\cdots,\gamma\}$. An elementary verification shows that the Metropolis chain has (2,2) moderate growth.

For a real function f defined on $\mathfrak{X}$ and integer r, set

$$f_r(x) = \frac{1}{V(x,r)} \sum_{y\in B(x,r)} f(y)\pi(y).$$

We will verify below that the chain satisfies a local Poincaré inequality:

$$\|f - f_i\|_2^2 \le ar^2\mathcal{E}(f|f) \qquad \text{with} \quad a = 4. \tag{6.7}$$

Finally, the smallest eigenvalue satisfies $\beta_- \ge -1 + 2\min(M|_{i,i}) \ge -1 + \frac{1}{n}$.

For reversible chains satisfying moderate growth and local Poincaré inequalities, order (diameter)^2 steps are necessary and suffice for convergence. The following [Theorem 5.10] makes this precise.

THEOREM 6.3. *Let K, π be a reversible Markov chain on a finite set $\mathfrak{X}$. Assume that (K,π) has $(A;d)$ moderate growth and satisfies a local Poincaré inequality with constant a. Assume further that $\beta_- \ge -1 + \frac{1}{a\gamma^2}$. Then*

$$a_2 e^{-a_3 k/\gamma^2} \le \sup_x \|K_x^k - \pi\| \le a_1 e^{-k/a\gamma^2}$$

with $a_1 = (2e(1+d)A)^{1/2}(2+d)^{d/4}$ and a_2, a_3 explicit constants depending only on (A,d,a).

This result gives Proposition 4.2. Thus it only remains to verify (6.7). For this, using paths locally, one has (see [Lemma 5.2])

$$\|f - f_i\|_2^2 \le \eta(r)\mathcal{E}(f|f)$$

with $\eta(r) = \max_e \frac{2}{Q(e)} \sum_{\substack{\gamma_{xy}\ni e \\ |x-y|\le r}} |\gamma_{xy}| \frac{\pi(x)\pi(y)}{V(x,r)}$.

¿From (4.6) for the edge $e = (i, i+1)$, $Q(e) = \pi(i)M(i,i+1) = \frac{\pi(i)}{2}$. We must bound, for all i, r,

$$\frac{2}{Q(i,i+1)} \sum_{|k-j|\le r, j\le i} |k-j| \frac{\pi(j)\pi(k)}{V(j,r)}. \tag{6.8}$$

We may bound $|k-j|$. By r, consider two cases, $i < 2r, i \ge 2r$.

Case 1. $i < 2r$. Then, for $j \le i$,

$$V(j,r) \ge \frac{2}{n(n+1)} \sum_{e \le r} e = \frac{r(r+1)}{n(n+1)}.$$

Using this, the quantity in (4.8) is bounded above by

$$\frac{2}{ir} \sum_{\substack{j \le i \\ i+1 \le k \le i+r-1}} jk = \frac{i+1}{4r}(r^2 + 2ir) \le 4r^2.$$

Case 2. $i \ge 2r$. Then, for $j \le i$,

$$V(j,r) \ge V(i-r,r) \ge \frac{2}{n(n+1)} \sum_{e=i-2r}^{i-r} e \ge \frac{4ir + 5r^2}{n(n+1)}.$$

Using this we bound (4.8) by

$$\frac{2r}{4i(ir+r^2)} \sum_{\substack{i-r \le j \le i \\ i+1 \le h \le i+r-1}} = \frac{r}{8i(ir+r^2)}(2ir - r^2)(2ir + r^2) \le \frac{4i^2r^3}{8i(ir+r^2)} \le \frac{r^2}{2}.$$

□

REMARKS: 1. Very similar bounds can be obtained for stationary distributions of form $\pi(i) = zp(i)$, for positive polynomial p.

2. Preliminary computations indicate that similar bounds hold for higher-dimensional grids. Even for multimodal polynomials. As is evident from Proposition 6.2, the constants need to be kept track of carefully. Nevertheless, it appears that for multimodal densities with polynomial peaks and valleys on low-dimensional square grids order (diameter)^2 steps are necessary and sufficient to reach stationarity.

7. Final Remarks.

The Metropolis algorithm is one way of carrying out Monte Carlo Markov chain techniques for sampling from a given stationary distribution. Hastings [ ] gave the following extensions: they have the same form (1.1) as the Metropolis algorithm, but the extra coin flip $A(x,y)$ has form

$$A(x,y) = \frac{S(x,y)}{1 + T(x,y)}$$

where $T(x,y) = \pi(x)K(x,y)/\pi(y)K(y,x)$, and $S(x,y) \ge 0$ is symmetric, subject to the sole requirement that $0 \le A(x,y) \le 1$. Two special choices:

$$S(x,y) = \begin{cases} 1 + T(x,y) & \text{if } T(y,x) \ge 1 \\ 1 + T(y,x) & \text{if } T(y,x) \le 1 \end{cases}$$

gives the usual Metropolis algorithm. Choosing $S(x,y) = 1$ for all x, y gives a method called Barker dynamics.

It is natural to ask which of these procedures works best. Peskun [ ] gives an elegant extremal characterization of the Metropolis algorithm in this class of chains. For $f : \mathfrak{X} \to \mathbb{R}$ a function of interest, the limiting variance of the usual estimate of the mean value of f is

$$\sigma^2(f) = \lim_n nvar\{\frac{1}{n}\sum_{i=1}^{n} f(X_i)\}$$

where $X_1, X_2, \cdots$ is a realization of the chain.

Consider two chains P_1, P_2 with the stationary distribution π. Call P_1 better than P_2 if $\sigma^2(f, P_1) \leq \sigma^2(f, P_2)$ for all f. Peskun (1973) proves that the Metropolis algorithm is best in Hastings class of chains. His proof uses the following elegant theorem: Let P_1, P_2 be irreducible, reversible Markov chains with respect to π. If $P_1(x, y) \leq P_2(x, y)$ for all $x \neq y$ then P_1 is better than P_2. This is a careful way of saying that an algorithm that holds less gets random faster.

It is natural to compare the various dynamics in simple examples to see how their rates of convergence compare. As an example, in unpublished work, Jeff Silver has shown that any of Hastings variations can be analyzed from the base chain of simple random walk on an n-point path. The analysis of section 6A goes through without criminal difficulties to give the bound

$$\|P_x^k - \pi\| \leq$$

We thus see that the convergence is (roughly) as quick for any of these chains.

Heuristically, one wants to choose the base chain K so that its stationary distribution is close to π. It is natural to try to estimate π, and change the base chain as information about π comes in. Gilks, Best, and Tan is an early interesting effort in this direction. There is much to do here.

We have not attempted to survey other, closely related algorithms for sampling from π. To begin with, for low-dimensional examples such as those of section 6, there is a large body of competitive technology. See Devroy [ ] for a comprehensive survey. In high dimensions, Glauber dynamics (know as the Gibbs sampler) is a closely related method that is beginning to have some useful finite sample convergence result. See Rosenthat [ ] and the references cited there. There are *many* further ideas in the statistical physics literature. Goodman and Sokal (1989) develop multigrid Monte Carlo methods as well as giving a useful set of pointers to the physics literature. Browsking through recent years of the Journal of Statistical Physics will reveal hundreds of other methods and variations.

All of these are fair game for careful mathematical analysis.

Acknowledgements

We thank Jeffrey Silver for several technical contributions to the present paper.

Bibliography

[1] Belsley, E. (1993) Rates of convergence of Markov chains related to association schemes. Ph.D. dissertation, Dept. of Mathematics, Harvard University.

[2] Chung, F. and Yan, S.T. (1995) Eigenvalue inequalities for graphs and convex subgraphs. Technical Report, Dept. of Mathematics, Harvard University.

[3] Chung, F. and Yan, S.T. (1995) Eigenvalues of graphs and Sobolev inequalities. To appear, Combinatorics, Probability, and Computing.

[4] Critchlow, D. (1985) *Metric Methods for Analyzing Partially Ranked Data*, Springer Lecture Notes in Statistics. **B4** Springer-Verlag, Berlin.

[5] Deuschel, J.D. and Mazza, C. (1994) L^2 convergence of time nonhomogeneous Markov processes: I. Special estimates, *Ann. Appl. Prob.* **4**, 1012-1056.

[6] Diaconis, P. (1988) *Group representations in probability and statistics*, Institute of Mathematical Statistics, Hayward, CA.

[7] Diaconis, P. and Hanlon, P. (1992) Eigenanalysis for some examples of the Metropolis algorithm, *Contemp. Math.* **138**, 99-117.

[8] Diaconis, P. and Saloff-Coste, L. (1993) Comparison techniques for reversible Markov chains, *Ann. Appl. Prob.* **3**, 696-730.

[9] Diaconis, P. and Saloff-Coste, L. (1995) Nash inequalities for finite Markov chains. Technical Report, Dept. of Statistics, Stanford, CA.

[10] Diaconis, P. and Saloff-Coste, L. (1995) Logarithmic Sobolev inequalities and finite Markov chains. Preprint. Dept. of Mathematics, Harvard University.

[11] Diaconis, P. and Saloff-Coste, L. (1995) Random walk on contingency tables with fixed row and column sums. Technical Report. Dept. of Mathematics, Harvard University.
[12] Diaconis, P. and Stroock, D. (1991) Geometric bounds for eigenvalues for Markov chains, *Ann. Appl. Prob.* **1**, 36-61.
[13] Dyer, M., Frieze, A. and Kannan, R. (1991) A random polynomial time algorithm for approximating the volume of convex bodies, Jour. Assoc. Comp. Mach. **38**, 1-17.
[14] Fill, J. (1991) Eigenvalue bounds on convergence to stationarity for non-reversible Markov chains with an application to the exclusion process, *Inv. Appl. Prob.* **1**, 62-87.
[15] Fligner, M. and Verducci, J. (1993) Probability models and statistical analysis for ranking data, Springer Lecture Notes in Statistics **80**, Springer, New York.
[16] Frieze, A., Kannan, R. and Polson, N. (1994) Sampling from log concave distributions, *Ann. Appl. Prob.* **4**, 812-837.
[17] Frigessi, A., Hwang, C., Sheu, S. and Di Stefano, P. (1993) Convergence rates of the Gibb sample, the Metropolis algorithm and other single-site updating dynamics, *Jour. Roy. Stat. Soc.* **B 55**, 205-219.
[18] Frigessi, A., Martinelli, F. and Stander, J. (1993) Computational complexity of Markov chain Monte Carlo methods. Technical Report, IAC Rome.
[19] Gilks, W., Best, W., and Tan, K. (1992) Adapted rejection Metropolis sampling. Technical Report.
[20] Goodman, J. and Sokal, A. (1989) Multigrid Monte Carlo method. Conceptual foundations. *Phys. Rev.* **D 40**, 2035-2071.
[21] Gross, L. (1995) Logarithmic Sobolev inequalities and contractivity properties of semigroups. Springer Lecture Notes in Mathematics.
[22] Hastings, W. (1970) Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**, 97-109.
[23] Ingrassia, S. (1994) On the rate of convergence of the Metropolis algorithm and Gibbs sampler by geometric bounds, *Ann. App. Prob.* **4**, 347-389.
[24] Jerrum, M. and Sinclair, A. (1989) Approximating the permanent, *Siam J. Comput.* **18**, 19-78.
[25] Jerrum, M. and Sinclair, A. (1989) Approximate counting, uniform generation and rapidly mixing Markov chains, *Infor. and Comput.* **82**, 93-133.
[26] Jerrum, M. (1992) Large cliques elude the Metropolis process, *Random structures and algorithms* **3**, 347-359.
[27] Kannen, R. (1994) Monte Carlo Markov chains.
[28] Karlin, S. and Taylor, H. (1975) *A first course in stochastic processes*, 2nd ed. Academic Press, New York.
[29] Liu, J. (1995) A comparison of importance sampling, rejection methods, and the Metropolis algorithm. To appear, *Statistics and Computing.*
[30] Lovasz, L. and Simonovits, M. (1993) Random walks in a convex body and an improved volume algorithm, *Random Structures and Algorithms* **4**, 359-412.
[31] Martinelli, F., Olivieri, E. and Schonmann, R. (1994) For $2-D$ lattice spin systems, weak mixing implies strong mixing. *Comm. Math. Physics* **165**, 33-47.
[32] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953) Equations of state calculations by fast computing machines, *J. Chem. Phys.* **21**, 1087-1092.
[33] Meyn, S. and Tweedie, R. (1994) Computable bounds for geometric convergence rates of Markov chains, *Ann. Appl. Prob.* **4**, 981-1011.
[34] Peskun, P. (1973) Optimum Monte Carlo sampling using Markov chains, *Biometrika* **60**, 607-612.
[35] Rosenthal, J. (19)
[36] Ross, K. and Xu, D. (1994) Hypergroup deformations and Markov chains, *Jour. Theoret. Prob.* **7**, 813-830.
[37] Schonmann, R. (1994) Slow droplet-driven relaxation of stochastic Ising models in the vicinity of the phase coexistence region, *Commun. Math. Phys.* **161**, 1-49.
[38] Schonmann, R. (1995) Theorems and conjectures on the droplet-driven relaxation of stochastic Ising models. Technical Report, Dept. of Mathematics, UCLA.
[39] Simon, B. (1993) The statistical mechanics of lattice gases, I. Princeton Uniiversity Press, Princeton.

[40] Sinclair, A. (1992) Improved bounds for mixing rates of Markov chains and multicommodity flow, *Combinatorics, Prob. Comput.* **1**, 351-370.

[41] Sinclair, A. (1993) *Algorithms for random generation and counting: a Markov chain approach*, Birkhauser, Boston.

[42] Stroock, D. and Zegarlinski, B. (1992) The logarithmic Sobolev inequality for spin systems on a lattice, *Comm. Math. Physics* **149**, 175-194.

[43] Thomas, L. (1989) Bound on the mass gap for finite volume stochastic Ising models at low temperatures, *Comm. Math. Physics* **126**, 1-11.