The Markov Chain Monte Carlo Revolution

Persi Diaconis*

Abstract

The use of simulation for high dimensional intractable computations has revolutionized applied mathematics. Designing, improving and understanding the new tools leads to (and leans on) fascinating mathematics, from representation theory through micro-local analysis.

1 Introduction

Many basic scientific problems are now routinely solved by simulation: a fancy "random walk" is performed on the system of interest. Averages computed from the walk give useful answers to formerly intractable problems. Here is an example drawn from course work of Stanford students Marc Coram and Phil Beineke.

Example 1 (Cryptography)**.** Stanford's Statistics Department has a drop-in consulting service. One day, a psychologist from the state prison system showed up with a collection of coded messages. Figure 1 shows part of a typical example.

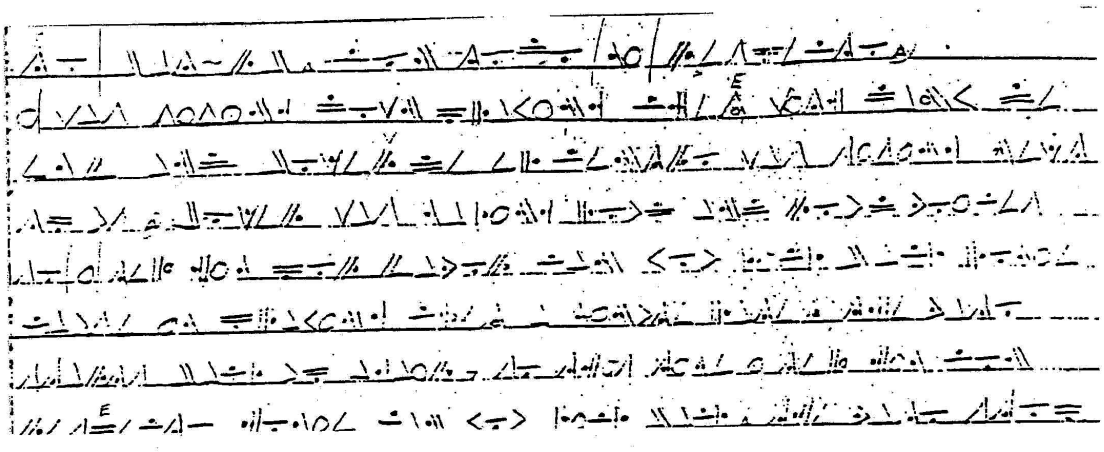


Figure 1

The problem was to decode these messages. Marc guessed that the code was a simple substitution cipher, each symbol standing for a letter, number, punctuation mark or space. Thus, there is an unknown function f

$$f : \{\text{code space}\} \longrightarrow \{\text{usual alphabet}\}.$$

One standard approach to decrypting is to use the statistics of written English to guess at probable choices for f, try these out, and see if the decrypted messages make sense.

To get the statistics, Marc downloaded a standard text (e.g., *War and Peace*) and recorded the first-order transitions: the proportion of consecutive text symbols from x to y. This gives a matrix $M(x, y)$ of transitions. One may then associate a plausibility to f via

$$\mathrm{Pl}(f) = \prod_i M\left(f(s_i), f(s_{i+1})\right)$$

* *Departments of Mathematics and Statistics, Stanford University*

where s_i runs over consecutive symbols in the coded message. Functions f which have high values of $\mathrm{Pl}(f)$ are good candidates for decryption. Maximizing f's were searched for by running the following Markov chain Monte Carlo algorithm:

- Start with a preliminary guess, say f.
- Compute $\mathrm{Pl}(f)$.
- Change to f_* by making a random transposition of the values f assigns to two symbols.
- Compute $\mathrm{Pl}(f_*)$; if this is larger than $\mathrm{Pl}(f)$, accept f_*.
- If not, flip a $\mathrm{Pl}(f_*)/\mathrm{Pl}(f)$ coin; if it comes up heads, accept f_*.
- If the coin toss comes up tails, stay at f.

The algorithm continues, trying to improve the current f by making random transpositions. The coin tosses allow it to go to less plausible f's, and keep it from getting stuck in local maxima.

Of course, the space of f's is huge (40! or so). Why should this Metropolis random walk succeed? To investigate this, Marc tried the algorithm out on a problem to which he knew the answer. Figure 2 shows a well-known section of Shakespeare's *Hamlet*.

```
ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS
NOBLER IN THE MIND TO SUFFER THE SLINGS AND ARROWS OF OUTRAGEOUS
FORTUNE OR TO TAKE ARMS AGAINST A SEA OF TROUBLES AND BY OPPOSING END
```

Figure 2

The text was scrambled at random and the Monte Carlo algorithm was run. Figure 3 shows sample output.

```
 100 ER ENOHDLAE OHDLO UOZEOUNORU O UOZEO HD OITO HEOQSET IUROFHE HENO ITORUZAEN
 200 ES ELOHRNDE OHRNO UOVEOULOSU O UOVEO HR OITO HEOQAET IUSOPHE HELO ITOSUVDEL
 300 ES ELOHANDE OHANO UOVEOULOSU O UOVEO HA OITO HEOQRET IUSOFHE HELO ITOSUVDEL
 400 ES ELOHINME OHINO UOVEOULOSU O UOVEO HI OATO HEOQRET AUSOWHE HELO ATOSUVMEL
 500 ES ELOHINME OHINO UODEOULOSU O UODEO HI OATO HEOQRET AUSOWHE HELO ATOSUDMEL
 600 ES ELOHINME OHINO UODEOULOSU O UODEO HI OATO HEOQRET AUSOWHE HELO ATOSUDMEL
 900 ES ELOHANME OHANO UODEOULOSU O UODEO HA OITO HEOQRET IUSOWHE HELO ITOSUDMEL
1000 IS ILOHANMI OHANO RODIORLOSR O RODIO HA OETO HIOQUIT ERSOWHI HILO ETOSRDMIL
1100 ISTILOHANMITOHANOT ODIO LOS TOT ODIOTHATOEROTHIOQUIRTE SOWHITHILOTEROS DMIL
1200 ISTILOHANMITOHANOT ODIO LOS TOT ODIOTHATOEROTHIOQUIRTE SOWHITHILOTEROS DMIL
1300 ISTILOHARMITOHAROT ODIO LOS TOT ODIOTHATOENOTHIOQUINTE SOWHITHILOTENOS DMIL
1400 ISTILOHAMRITOHAMOT OFIO LOS TOT OFIOTHATOENOTHIOQUINTE SOWHITHILOTENOS FRIL
1600 ESTEL HAMRET HAM TO CE OL SOT TO CE THAT IN THE QUENTIOS WHETHEL TIN SOCREL
1700 ESTEL HAMRET HAM TO BE OL SOT TO BE THAT IN THE QUENTIOS WHETHEL TIN SOBREL
1800 ESTER HAMLET HAM TO BE OR SOT TO BE THAT IN THE QUENTIOS WHETHER TIN SOBLER
1900 ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER
2000 ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER
```

Figure 3

After 100 steps, the message is a mess. After two thousand steps, the decrypted message makes sense. It stays essentially the same as further steps are tried. I find it remarkable that a few thousand steps of this simple optimization procedure work so well. Over the past few years, friends in math and computer science

courses have designed homework problems around this example [17]. Students are usually able to successfully decrypt messages from fairly short texts; in the prison example, about a page of code was available.

The algorithm was run on the prison text. A portion of the final result is shown in Figure 4. It gives a useful decoding that seemed to work on additional texts.

```
  to bat-rb. con todo mi respeto. i was sitting down playing chess with
danny de emf and boxer de el centro was sitting next to us. boxer was
making loud and loud voices so i tell him por favor can you kick back
homie cause im playing chess a minute later the vato starts back up again
so this time i tell him con respecto homie can you kick back.  the vato
stop for a minute and he starts up again so i tell him check this out shut
the f**k up cause im tired of your voice and if you got a problem with it
we can go to celda and handle it. i really felt disrespected thats why i
told him. anyways after i tell him that the next thing I know that vato
slashes me and leaves. dy the time i figure im hit i try to get away but
the c.o. is walking in my direction and he gets me right dy a celda. so i
go to the hole. when im in the hole my home boys hit doxer so now "b" is
also in the hole. while im in the hole im getting schoold wrong and
```

Figure 4

I like this example because a) it is real, b) there is no question the algorithm found the correct answer, and c) the procedure works despite the implausible underlying assumptions. In fact, the message is in a mix of English, Spanish and prison jargon. The plausibility measure is based on first-order transitions only. A preliminary attempt with single-letter frequencies failed. To be honest, several practical details have been omitted: we allowed an unspecified "?" symbol in the deviation (with transitions to and from "?" being initially uniform). The display in Figure 4 was 'cleaned up' by a bit of human tinkering. I must also add that the algorithm described has a perfectly natural derivation as Bayesian statistics. The decoding function f is a parameter in a model specifying the message as the output of a Markov chain with known transition matrix $M(x, y)$. With a uniform prior on f, the plausibility function is proportional to the posterior distribution. The algorithm is finding the mode of the posterior.

In the rest of this article, I explain Markov chains and the Metropolis algorithm more carefully in Section 2. A closely related Markov chain on permutations is analyzed in Section 3. The arguments use symmetric function theory, a bridge between combinatorics and representation theory.

A very different example — hard discs in a box — is introduced in Section 4. The tools needed for study are drawn from analysis, micro-local techniques (Section 5) along with functional inequalities (Nash and Sobolev inequalities).

Throughout, emphasis is on analysis of iterates of self-adjoint operators using the spectrum. There are many other techniques used in modern probability. A brief overview, together with pointers on how a beginner can learn more, is in Section 6.

2 A Brief Treatise on Markov Chains

2.1 A finite case

Let $\mathcal{X}$ be a finite set. A *Markov chain* is defined by a matrix $K(x, y)$ with $K(x, y) \geq 0$, $\sum_y K(x, y) = 1$ for each x. Thus each row is a probability measure so K can direct a kind of random walk: from x, choose y with probability $K(x, y)$; from y choose z with probability $K(y, z)$, and so on. We refer to the outcomes $X_0 = x, X_1 = y, X_2 = z, \ldots$ as a run of the chain starting at x. From the definitions $P(X_1 = y|X_0 = x) = K(x, y)$, $P(X_1 = y, X_2 = z|X_0 = x) = K(x, y)K(y, z)$. From this, $P(X_2 = z|X_0 = x) = \sum_y K(x, y)K(y, z)$, and so on. The nth power of the matrix has x, y entry $P(X_n = y|X_0 = x)$.

All of the Markov chains considered in this article have *stationary distributions* $\pi(x) > 0$, $\sum_x \pi(x) = 1$ with π satisfying

$$\sum_x \pi(x)K(x,y) = \pi(y). \tag{2.1}$$

Thus π is a left eigenvector of K with eigenvalue 1. The probabilistic interpretation of (2.1) is "pick x from π and take a step from $K(x,y)$; the chance of being at y is $\pi(y)$." Thus π is stationary for the evolution. The fundamental theorem of Markov chains (a simple corollary of the Peron–Frobenius theorem) says, under a simple connectedness condition, π is unique and high powers of K converge to the rank one matrix with all rows equal to π.

Theorem 1 (Fundamental Theorem of Markov Chains)**.** *Let $\mathcal{X}$ be a finite set and $K(x,y)$ a Markov chain indexed by $\mathcal{X}$. If there is n_0 so that $K^n(x,y) \geq 0$ for all $n > n_0$, then K has a unique stationary distribution π and, as $n \to \infty$,*

$$K^n(x,y) \to \pi(y) \quad \text{for each } x, y \in \mathcal{X}.$$

The probabilistic content of the theorem is that from any starting state x, the nth step of a run of the Markov chain has chance close to $\pi(y)$ of being at y if n is large. In computational settings, $|\mathcal{X}|$ is large, it is easy to move from x to y according to $K(x,y)$ and it is hard to sample from π directly.

Consider the cryptography example in the introduction. There, $\mathcal{X}$ is the set of all 1–1 functions f from code space to the usual alphabet $\{\mathrm{A}, \mathrm{B}, \ldots, \mathrm{Z}, 1, 2, \ldots, 9, 0, *, ., ?, \ldots\}$. Assume there are m distinct code symbols and n symbols in the alphabet space. The stationary distribution is

$$\pi(f) = z^{-1} \prod_i M\left(f(s_i), f(s_{i+1})\right) \tag{2.2}$$

where M is the (assumed given) first-order transition matrix of English and the product ranges over consecutive coded symbols in the fixed message. The normalizing constant z is defined by

$$z = \sum_f \prod_i \left(M\left(f(s_i), f(s_{i+1})\right)\right).$$

Note that z is unknowable practically.

The problem considered here is to sample f's repeatedly from $\pi(f)$. This seems daunting because of the huge size of $\mathcal{X}$ and the problem of unknown z. The Metropolis Markov chain $K(f, f_*)$ solves this problem.

2.2 Metropolis algorithm

Let $\mathcal{X}$ be a finite state space and $\pi(x)$ a probability on $\mathcal{X}$ (perhaps specified only up to an unknown normalizing constant). Let $J(x,y)$ be a Markov matrix on $\mathcal{X}$ with $J(x,y) > 0 \leftrightarrow J(y,x) > 0$. At the start, J is unrelated to π. The Metropolis algorithm changes J to a new Markov matrix $K(x,y)$ with stationary distribution π. It is given by a simple recipe:

$$K(x,y) = \begin{cases} J(x,y) & \text{if } x \neq y,\ A(x,y) \geq 1 \\ J(x,y)A(x,y) & \text{if } x \neq y,\ A(x,y) < 1 \\ J(x,y) + \displaystyle\sum_{z: A(x,z) < 1} J(x,z)(1 - A(x,z)) & \text{if } x = y \end{cases}. \tag{2.3}$$

In (2.3), the acceptance ratio is $A(x,y) = \pi(y)J(y,x)/\pi(x)J(x,y)$. The formula (2.3) has a simple interpretation: from x, choose y with probability $J(x,y)$; if $A(x,y) \geq 1$, move to y; if $A(x,y) < 1$, flip a coin with this success probability and move to y if success occurs; in other cases, stay at x. Note that the normalizing constant for π cancels out in all calculations. The new chain satisfies

$$\boxed{\pi(x)K(x,y) = \pi(y)K(y,x)}$$

and thus

$$\boxed{\sum_x \pi(x)K(x,y) = \sum_x \pi(y)K(y,x) = \pi(y)\sum_x K(y,x) = \pi(y),}$$

so that π is a left eigenvector with eigenvalue 1. If the chain (2.3) is connected, Theorem 1 is in force. After many steps of the chain, the chance of being at y is approximately $\pi(y)$, no matter what the starting state $\mathcal{X}$. Textbook treatments of the Metropolis algorithm are in [44] or [62]. A literature review can be found in [31].

In the cryptography example $\mathcal{X}$ is all 1–1 functions from symbol space (say of size m) to alphabet space (say of size $n \geq m$). Thus $|\mathcal{X}| = n(n-1)\dots(n-m+1)$. This is large if, e.g., $m = n = 50$. The stationary distribution is given in (2.2). The proposal chain $J(f, f^*)$ is specified by a random switch of two symbols,

$$J(f, f^*) = \begin{cases} \frac{1}{n(n-1)(m-n+2)(m-n+1)} & \text{if } f, f^* \text{ differ in at most two places} \\ 0 & \text{otherwise} \end{cases}.$$

Note that $J(f, f^*) = J(f^*, f)$ so $A(f, f^*) = \pi(f^*)/\pi(f)$.

2.3 Convergence

A basic problem of Markov chain theory concerns the rate of convergence in $K^n(x, y) \to \pi(y)$. How long must the chain be run to be suitably close to π? It is customary to measure distances between two probabilities by total variation distance

$$\|K_x^n - \pi\|_{\text{TV}} = \frac{1}{2} \sum_y |K^n(x, y) - \pi(y)| = \max_{A \subseteq \mathcal{X}} |K^n(x, A) - \pi(A)|.$$

This yields the math problem: Given K, π, x and $\epsilon > 0$, how large n so

$$\|K_x^n - \pi\|_{\text{TV}} < \epsilon?$$

Sadly, there are very few practical problems where this question can be answered. In particular, no useful answer in known for the cryptography problem. In Section 3, a surrogate problem is set up and solved. It suggests that when $n \doteq m$, order $n \log n$ steps suffice for mixing.

Suppose, as is the case for the examples in this paper, that the Markov chain is reversible: $\pi(x)K(x, y) = \pi(y)K(y, x)$. Let $L^2(\pi)$ be $\{g : \mathcal{X} \to \mathbb{R}\}$ with inner product $\langle g, h \rangle = \sum_x g(x)h(x)\pi(x)$. Then K operates on L^2 by $Kg(x) = \sum g(y)K(x, y)$. Reversibility implies $\langle Kg, h \rangle = \langle g, Kh \rangle$, so K is self-adjoint. Now, the spectral theorem says there is an orthonormal basis of eigenvectors ψ_i and eigenvalues β_i (so $K\psi_i = \beta_i\psi_i$) for $0 \leq i \leq |\mathcal{X}| - 1$ and $1 = \beta_0 \geq \beta_1 \geq \cdots \geq \beta_{|\mathcal{X}|-1} \geq -1$. By elementary manipulations

$$K(x, y) = \pi(y) \sum_{i=0}^{|\mathcal{X}|-1} \beta_i \psi_i(x)\psi_i(y), \quad K^n(x, y) = \pi(y) \sum_{i=0}^{|\mathcal{X}|-1} \beta_i^n \psi_i(x)\psi_i(y).$$

Using the Cauchy–Schwartz inequality, we have

$$4\|K_x^n - \pi\|_{\text{TV}}^2 \leq \sum_y \frac{(K^n(x, y) - \pi(y))^2}{\pi(y)} = \sum_{i=1}^{|\mathcal{X}|-1} \beta_i^{2n} \psi_i^2(x). \tag{2.4}$$

The bound (2.4) is the basic eigenvalue bound used to get rates of convergence for the examples presented here. To get sharp bounds on the right hand side requires good control of both eigenvalues and eigenvectors. For more details and many examples, see [79]. A detailed example on the permutation group is given in Section 3 below. Examples on countable and continuous spaces are given in Section 5.

2.4 General state spaces

Markov chains are used to do similar calculations on Euclidean and infinite-dimensional spaces. My favorite introduction to Markov chains is the book by Bremaud [10], but there are many sources; for finite state spaces see [83]. For a more general discussion, see [7] and the references in Section 6.1.

Briefly if $(\mathcal{X}, B)$ is a measurable space, a Markov kernel $K(x, dy)$ is a probability measure $K(x, \cdot)$ for each x. Iterates of the kernel are given by, e.g.,

$$K^2(x, A) = \int K(z, A)K(x, dz).$$

A *stationary distribution* is a probability $\pi(dx)$ satisfying

$$\pi(A) = \int K(x, A)\pi(dx)$$

under simple conditions $K^n(x, A) \to \pi(A)$ and exactly the same problems arise.

Reversible Markov chains yield bounded self-adjoint operators and spectral techniques can again be tried. Examples are in Section 4, Section 5, and Section 6.

3 From Cryptography to Symmetric Function Theory

This section answers the question, "What does a theorem in this subject look like?" It also illustrates how even seemingly simple problems can call on tools from disparate fields: here, symmetric function theory, a blend of combinatorics and representation theory. This section is drawn from joint work with Phil Hanlon [21].

3.1 The problem

Let $\mathcal{X} = S_n$, the symmetric group on n letters. Define a probability measure on S_n by

$$\pi(\sigma) = z^{-1}\theta^{d(\sigma,\sigma_0)} \quad \text{for } \sigma, \sigma_0 \in S_n,\ 0 < \theta \leq 1. \tag{3.1}$$

In (3.1), $d(\sigma, \sigma_0)$ is a metric on the symmetric group, here taken to be

$$d(\sigma, \sigma_0) = \text{ minimum number of transpositions required to bring } \sigma \text{ to } \sigma_0.$$

This is called Cayley's distance in [20] because a result of A. Cayley implies that $d(\sigma, \sigma_0) = n - c(\sigma^{-1}\sigma_0)$ with $c(\sigma)$ the number of cycles in σ. The metric is bi-invariant: $d(\sigma, \sigma_0) = d(\tau\sigma, \tau\sigma_0) = d(\sigma\tau, \sigma_0\tau)$. The normalizing constant z is known in this example: $z = \sum_\sigma \theta^{d(\sigma,\sigma_0)} = \prod_{i=1}^n (1 + \theta(i-1))$.

If $\theta = 1$, $\pi(\sigma)$ is the uniform distribution on S_n. For $\theta < 1$, $\pi(\sigma)$ is largest at σ_0 and falls off from its maximum as σ moves away from σ_0. It serves as a natural non-uniform distribution on S_n, peaked at a point. Further discussion of this construction (called Mallows model through Cayley's metric) with examples from psychology and computer science is in [18, 19, 28]. The problem studied here is

How can samples be drawn from σ?

One route is to use the Metropolis algorithm, based on random transpositions. Thus, from σ, choose a transposition (i, j) uniformly at random and consider $(i, j)\sigma = \sigma^*$. If $d(\sigma^*, \sigma_0) \leq d(\sigma, \sigma_0)$ the chain moves to σ^*. If $d(\sigma^*, \sigma_0) > d(\sigma, \sigma_0)$, flip a θ-coin. If this comes up heads, move to σ^*; else stay at σ. In symbols,

$$K(\sigma, \sigma^*) = \begin{cases} 1/\binom{n}{2} & \text{if } \sigma^* = (i,j)\sigma, \quad d(\sigma^*, \sigma_0) < d(\sigma, \sigma_0) \\ \theta/\binom{n}{2} & \text{if } \sigma^* = (i,j)\sigma, \quad d(\sigma^*, \sigma_0) > d(\sigma, \sigma_0) \\ c\left(1 - \theta/\binom{n}{2}\right) & \text{if } \sigma^* = \sigma, \text{ with } c = \#\left\{(i,j) : d((i,j)\sigma, \sigma_0) > d(\sigma, \sigma_0)\right\} \\ 0 & \text{otherwise.} \end{cases} \tag{3.2}$$

Observe that this Markov chain is "easy to run." The Metropolis construction guarantees that:

$$\pi(\sigma)K(\sigma, \sigma^*) = \pi(\sigma^*)K(\sigma^*, \sigma)$$

so that the chain has stationary distribution π. When $n = 3$ and $\sigma_0 = \text{id}$, the transition matrix is

$$
\begin{array}{c} \\ \text{id} \\ (12) \\ (13) \\ (23) \\ (123) \\ (132) \end{array}
\begin{array}{c}
\begin{array}{cccccc} \text{id} & (12) & (13) & (23) & (123) & (132) \end{array} \\
\begin{pmatrix}
1-\theta & \frac{\theta}{3} & \frac{\theta}{3} & \frac{\theta}{3} & 0 & 0 \\
\frac{1}{3} & \frac{2}{3}(1-\theta) & 0 & 0 & \frac{\theta}{3} & \frac{\theta}{3} \\
\frac{1}{3} & 0 & \frac{2}{3}(1-\theta) & 0 & \frac{\theta}{3} & \frac{\theta}{3} \\
\frac{1}{3} & 0 & 0 & \frac{2}{3}(1-\theta) & \frac{\theta}{3} & \frac{\theta}{3} \\
0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\
0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0
\end{pmatrix}
\end{array}
$$

The stationary distribution is the left eigenvector proportional to $(1, \theta, \theta, \theta, \theta^2, \theta^2)$.

This example bears a passing resemblance to the cryptography example: the set of $1-1$ functions of an m-set to an n-set is replaced by the symmetric group. Presumably, the stationary distribution in the cryptography example is peaked at a point (the best decoding) and the algorithms are essentially the same.

To analyze the chain (3.2) using spectral theory requires knowledge of the eigenvalues and vectors. By what still seems like a miracle, these are available in closed form. When $\theta = 1$, the chain (3.2) reduces to the 'transpose at random chain', perhaps the first Markov chain given a sharp analysis [32]. Here is a typical result.

Theorem 2. *For $0 < \theta \leq 1$, the Markov chain $K(\sigma, \sigma^*)$ in (3.2) has stationary distribution π from (3.1). Let $k = an \log n + cn$ with $a = 1/2\theta + 1/4\theta(1/\theta - \theta)$ and $c > 0$. Then, with $\sigma_0 = id$ and starting from the identity*

$$\|K^k - \pi\|_{TV} \leq f(\theta, c)$$

with $f(\theta, c) \to 0$ for $c \to 0$.

Remarks. The result shows that order $n \log n$ steps suffice to make the distance to stationarity small. The function $f(\theta, c)$ is explicit but a bit of a mess. There is a matching lower bound showing that order $n \log n$ steps are necessary as well. In the theorem, σ_0 was chosen as the identity and the chain starts at σ_0. If the chain starts far from the identity, for example at an n-cycle, it can be shown that order $n^2 \log n$ steps suffice. When, e.g., $n = 52$, $n \log n \doteq 200$ while $n^2 \log n \doteq 11,000$. These numbers give a useful feel for the running time.

3.2 Tools from symmetric function theory

The first step of analysis is to reduce the state space from the full symmetric group to the set of conjugacy classes. (Recall these are indexed by partitions of n.) The matrix $K(\sigma, \sigma^*)$ commutes with the action of the symmetric group by conjugation, so only transitions between conjugacy classes are needed. When $n = 3$ the transition matrix becomes

$$
\begin{array}{c} \\ 1^3 \\ 1,2 \\ 3 \end{array}
\begin{array}{c}
\begin{array}{ccc} 1^3 & 1,2 & 3 \end{array} \\
\begin{pmatrix}
1-\theta & \theta & 0 \\
\frac{1}{3} & \frac{2}{3}(1-\theta) & \frac{2}{3}\theta \\
0 & 1 & 0
\end{pmatrix}
\end{array}
$$

with stationary distribution proportional to $(1, 3\theta, 2\theta^2)$. Let

$$M(\mu, \lambda), \quad m(\lambda) \tag{3.3}$$

be the transition matrix and stationary distribution indexed by partitions λ, μ.

Theorem 3. *For $0 < \theta \leq 1$, the Markov chain (3.3) has an eigenvalue β_λ for each partition $(\lambda_1, \lambda_2, \ldots, \lambda_r)$ of n with*

$$\beta_\lambda = (1-\theta) + \frac{\theta n(\lambda^t) + n(\lambda)}{\binom{n}{2}}, \quad n(\lambda) = \sum_{i=1}^{n} (i-1)\lambda_i.$$

The corresponding right eigenfunction, normed to be orthonormal in $L^2(m)$, is

$$\frac{c_\lambda(\cdot)}{m(\lambda)\{j_\lambda \pi/\theta^n n!\}^{1/2}}. \tag{3.4}$$

In (3.4), c_λ are the change of basis coefficients in expressing the Jack symmetric functions in terms of the power-symmetric functions. The normalizing constant in (3.4) involves closed form, combinatorially-defined terms which will not be detailed further.

Here is an aside on the $c_\lambda(\cdot)$. Classical combinatorics involves things like partitions, permutations, graphs, balls in boxes and so on. A lot of this has been unified and extended in the subject of algebraic combinatorics. A central theme here is the ring $\Lambda_n(x_1 \dots x_k)$ of homogeneous symmetric polynomials of degree n. There are various bases for this space. For example, if $P_i(x_1 \dots x_k) = \sum x_j^i$ and $P_\lambda = P_{\lambda_1} P_{\lambda_2} \dots P_{\lambda_n}$, the P_λ form a basis as λ runs through the partitions of n (fundamental theorem of symmetric functions). Other well-known bases are the monomial and elementary symmetric functions. The stars of the show are the Schur functions (character of the general linear group). The change of basis matrices between these codes up a lot of classical combinatorics. A two-parameter family of bases, the Macdonald polynomials, is a central focus of modern combinatorics. Definitive, inspiring accounts of this are in Macdonald [65] and Stanley [82].

The Jack symmetric functions $J_\lambda(\boldsymbol{x}; \alpha)$ is one of the many bases. Here $\boldsymbol{x} = (x_1 \dots x_k)$ and α is a positive real parameter. When $\alpha = 1$, the Jacks become the Schur functions. When $\alpha = 2$, the Jacks become the zonal polynomials (spherical functions of GL_n/O_n). Before the work with Hanlon, no natural use for other values of α was known. Denote the change of basis coefficients from Jacks to power sums by

$$J_\lambda(\boldsymbol{x}; \alpha) = \sum_{\mu \vdash n} c(\lambda, \mu) P_\mu(x).$$

The $c(\lambda, \mu)$ are rational functions of α. For example, when $n = 3$,

$$\begin{aligned} J_{1^3} &= P_1^3 - 3P_{12} - 2P_3 \\ J_{12} &= P_1^3 + (\alpha - 1)P_{12} - \alpha P_3 \\ J_3 &= P_1^3 - 3\alpha P_{12} + 2\alpha^2 P_3 \end{aligned}$$

The algebraic combinatorics community had developed properties of Jack symmetric functions "because they were there." Using this knowledge allowed us to properly normalize the eigenfunctions and work with them to prove Theorems 1 and 2. Many more examples of this type of interplay are in [14]. A textbook account of our work is in [48].

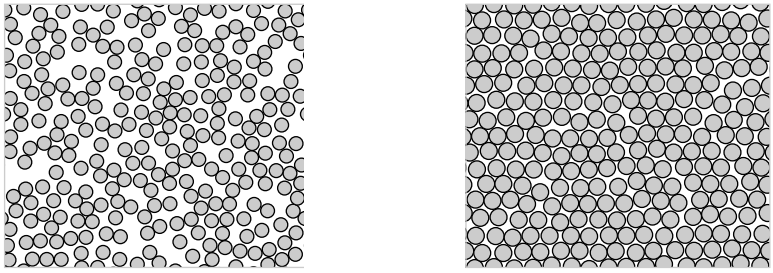
There is a fascinating research problem opened up by this analysis. When $\theta = 1$, the Jack symmetric functions are Schur functions and the change of basis coefficients are the characters of the symmetric group. The Markov chain becomes 'random transpositions'. This was analyzed in joint work with Shahshahani [32]. Adding in the deformation by the Metropolis algorithm deforms the eigenvalues and eigenvectors in a mathematically natural way. Is there a similar deformation that gets the coefficients of the Macdonald polynomials? This is ongoing joint work with Arun Ram. Changing the metric on S_n, using pairwise adjacent transpositions instead of all transpositions, gives a deformation to Hecke algebras. The Metropolis algorithm gives a probabilistic interpretation of the multiplication in these algebras. This again is joint work with Ram [28]. This affinity between the physically natural Metropolis algorithm and algebra is a mystery which cries out for explanation.

Turning back toward the cryptography example, how do things change if we go from the permutation group to the set of $1-1$ functions from an m-set to an n-set? When $\theta = 1$, this was worked out by Andrew Greenhalgh. The analysis involves the algebra of functions on S_n which are invariant under conjugation by the subgroup $S_m \times S_{n-m}$ and bi-invariant under the subgroup S_{n-m}. These "doubly invariant functions" form a commutative algebra discussed further in Ceccherini-Silberstein et al. [14], Sect. 9.8. Do things deform well when $\theta \neq 1$? It is natural to guess the answer is Yes.

It is important to end these fascinating success stories with the observation that any similarly useful analysis of the original cryptography example seems remote. Further, getting rates of convergence for the Metropolis algorithm for other metrics in (3.1) is a challenging open problem.

4 Hard Discs in a Box

Consider possible placements of n discs of radius ϵ in the unit square. The discs must be non-overlapping and completely contained in the unit square. Examples at low and high density (kindly provided by Werner Krauth from [57]) are shown in Figure 5.



$\eta = 0.48$ $\eta = 0.72$

(a) Low density **(b)** High density

Figure 5

In applications, n is fairly large (e.g., 100–10^6) and of course ϵ should be suitably small. The centers of the discs give a point in $\mathbb{R}^{2n}$. We know very, very little about the topology of the set $\mathcal{X}(n, \epsilon)$ of configurations: for fixed n, what are useful bounds on ϵ for the space to be connected? What are the Betti numbers? Of course, for ϵ "small" this set is connected but very little else is known. By its embedding in $\mathbb{R}^{2n}$, $\mathcal{X}(n, \epsilon)$ inherits a natural uniform distribution, Lebesgue measure restricted to $\mathcal{X}(n, \epsilon)$. The problem is to pick points in $\mathcal{X}(n, \epsilon)$ uniformly. If $X_1, X_2, \ldots, X_k$ are chosen from the uniform distribution and $f : \mathcal{X}(n, \epsilon) \to \mathbb{R}$ is a function, we may approximate

$$\int_{\mathcal{X}(n,\epsilon)} f(x)dx \quad \text{by} \quad \frac{1}{k}\sum_{i=1}^{k} f(X_i). \tag{4.1}$$

Motivation for this task and some functions f of interest will be given at the end of this section.

This hard discs problem is the original motivation for the Metropolis algorithm. Here is a version of the Metropolis algorithm for hard discs.

- Start with some $x \in \mathcal{X}(n, \epsilon)$.
- Pick a disc center at random (probability $1/n$).
- Pick a point in a disc of radius h, centered at the chosen disc center at random (from Lebesgue measure).
- Try to move the chosen disc center to the chosen point; if the resulting configuration is in $\mathcal{X}(n, \epsilon)$, accept the move; else, stay at x.

The algorithm continues, randomly moving coordinates. If $X_1, X_2, \ldots, X_k$ denotes the successive configurations, theory shows that X_k becomes uniformly distributed provided ϵ, k are small. For large k, the X_i can be used as in (4.1).

Motivation

The original motivation for this problem comes from the study of phase transition in statistical mechanics. For many substances (e.g., water), experiments produce phase diagrams such as that shown in Figure 6. Every aspect of such phase diagrams is intensely studied. The general picture, a finite length liquid–vapor phase transition line ending in a critical point, a triple point where all three forms co-exist and a solid–liquid phase line seemingly extending to infinity, seems universal. The physicist G. Uhlenbeck [87, p. 11] writes "Note that since these are *general* phenomena, they must have *general* explanation; the precise details of the molecular structure and of the intermolecular forces should not matter." In discussing the solid–liquid

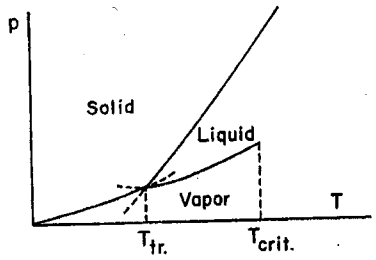


Figure 6

transition, Uhlenbeck (p. 18) notes that the solid–liquid transition seemingly occurs at any temperature provided the pressure is high enough. He suggests that at high pressure, the attractive intermolecular force does not play a role "...and that it is the sharp repulsive forces that are responsible for the solid–fluid transition. It is this train of thought which explains the great interest of the so-called *Kirkwood transition*. In 1941, Kirkwood posed the problem of whether a gas of hard spheres would show a phase transition...."

From then to now, chemists and physicists have studied this problem using a variety of tools. Current findings indicate a phase transition when the density of discs is large (about .71, still well below the close packing density). Below this transition density, the discs "look random"; above this density the discs look close to a lattice packing. These notions are quantified by a variety of functions f. For example

$$f(x) = \left| \frac{1}{N} \sum_{j=1}^{N} \frac{1}{N_j} \sum_{k} e^{6i\theta_{jk}} \right|$$

where the sum is over the N particles encoded by $x \in \mathbb{R}^{2N}$, the sum in k is over the N_j neighbors of the jth particle and θ_{jk} is the angle between the particles j and k in an arbitrary but fixed reference frame. If the configuration x has a local hexatic structure, this sum should be small. Typical values of f are studied by simulation. Different functions are used to study long-range order.

The above rough description may be supplemented by the useful survey of [64]. A host of simulation methods are detailed in [2]. An up-to-date tutorial on hard discs appears in [57, Chap. 2].

For purposes of this paper, the main points are i) the hard disc model is a basic object of study and ii) many key findings have been based on variants of the Metropolis algorithm. In the next section, we flush out the Metropolis algorithm to more standard mathematics.

5 Some Mathematics

Here is a generalization of the hard discs Metropolis algorithm. Let $\Omega \subseteq \mathbb{R}^d$ be a bounded connected open set. Let $\bar{p}(x) > 0, z = \int_\Omega \bar{p}(x)dx < \infty, p(x) = z^{-1}\bar{p}(x)$ specify a probability density on Ω. If required, extend p to have value 0 outside the closure of Ω. Many sampling problems can be stated thus:

$$\text{Given } \bar{p}, \text{ choose points in } \Omega \text{ from } p.$$

Note that the normalizing constant z may not be given and is usually impossible to usefully approximate. As an example, consider placing fifty hard discs in the unit square when $\epsilon = 1/100$. The set of allowable configurations is a complex, cuspy set. While $\bar{p} \equiv 1$ on Ω, it would not be practical to compute z. Here is one version of the Metropolis algorithm which samples from p. From $x \in \Omega$, fix a small, positive h.

- Choose $y \in B_x(h)$, from normalized Lebesgue measure on this ball.
- If $p(y) \geq p(x)$, move to y.
- If $p(y) < p(x)$, move to y with probability $p(y)/p(x)$.
- Else stay at x.

Note that this algorithm does not require knowing z. The transition from x to y yields a transition kernel

$$\begin{aligned} P(x,dy) =& m(x)\delta_x + \frac{h^{-d}}{\mathrm{Vol}(B_1)}\delta_{B_1}\left(\frac{x-y}{h}\right)\min\left(\frac{p(x)}{p(y)},1\right)dy \\ \text{with} \qquad & m(x) = 1 - \int_{\mathbb{R}^d}\frac{h^{-d}}{\mathrm{Vol}(B_1)}\delta_{B_1}\left(\frac{x-y}{h}\right)\min\left(\frac{p(x)}{p(y)},1\right)dy. \end{aligned} \tag{5.1}$$

This kernel operates on $L^2(p)$ via

$$P\cdot f(x) = \int_{\mathbb{R}^d} f(y)P(x,dy).$$

It is easy to see that $P(x,dy)$ is a bounded self-adjoint operator on $L^2(p)$. The associated Markov chain may be described 'in English' by

- Start at $X_0 = x \in \Omega$.
- Pick X_1 from $P(x,dy)$.
- Pick X_2 from $P(X_1,dy)$.
- And so on ....

Thus $P\{X_2\in A\} = P_x^2(A) = \int_{\mathbb{R}^d} P(z,A)P(x,dz)$, $P\{X_k\in A\} = P_x^k(A) = \int_{\mathbb{R}^d} P(z,A)P^{k-1}(x,dz)$. Under our assumptions (Ω connected, h small), for all $x\in\Omega$ and $A\subseteq\Omega$, the algorithm works:

$$P_x^k(A) \xrightarrow[\infty]{k} \int_A p(y)dy.$$

It is natural to ask how fast this convergence takes place: How many steps should the algorithm be run to do its job? In joint work with Gilles Lebeau and Laurent Michel, we prove the following.

Theorem 4. *Let Ω be a connected Lipshitz domain in $\mathbb{R}^d$. For p measurable (with $0<m\le p(x)\le M<\infty$ on Ω) and h fixed and small, the Metropolis algorithm* (5.1) *satisfies*

$$\left|P_x^k(A) - \int_A p(y)dy\right| \le c_1 e^{-c_2 k h^2} \quad \textit{uniformly in } x\in\Omega,\ A\subseteq\Omega. \tag{5.2}$$

In (5.2), $c_1,\ c_2$ are positive constants that depend on $\bar{p}$ and Ω but not on x, k or h. The result is sharp in the sense that there is a matching lower bound. Good estimates of c_2 are available (see the following section).

Note that the Metropolis algorithm (5.1) is based on steps in the full-dimensional ball $B_\epsilon(x)$ while the Metropolis algorithm for discs in Section 2 is based on just changing two coordinates at a time. With extra effort, a result like (5.2) can be shown for the hard disc problem as well. Details are in [25]. As a caveat, note that we do not have good control on c_1 in terms of the dimension d or smoothness of Ω. The results are explicit but certainly not sharp.

The Metropolis algorithm of this section is on a Euclidean space with basic steps driven by a 'ball walk.' None of this is particularly important. The underlying state space can be quite general, from finite (all $1-1$ functions from one finite set to another as in our cryptography example) to infinite-dimensional (Markov chains on spaces of measures). The proposal distribution needn't be symmetric. All of the introductory books on simulation discussed in Section 6 develop the Metropolis algorithm. In [8] it is shown to be the L^1 projection of the proposal distribution to the p self-adjoint kernels on Ω. A survey of rates of convergence results on finite state spaces with extensive references to the work of computer science theorists on "approximate counting" and mathematical physicists on "Ising models" is in [31]. Finally, there are *many* other classes of algorithms and proof techniques in active development. This is brought out in Section 6 below.

5.1 Ideas and tools

To analyze rates of convergence it is natural to try spectral theory, especially if the operators are self-adjoint. This sometimes works. It is sometimes necessary to supplement with tools such as comparison and extension theory, Weyl-type bounds on eigenvalues, bounds on eigenvectors and Nash–Sobolev inequalities. These are basic tools of modern analysis. Their use in a concrete problem may help some readers come into contact with this part of the mathematical world.

Spectral bounds for Markov chains

Let $\mathcal{X}$ be a set, $\mu(dx)$ a reference measure and $m(x)$ a probability density with respect to μ (so $m(x) \geq 0, \int m(x)\mu(dx) = 1$). Let $P(x, dy)$ be a Markov kernel on $\mathcal{X}$. This means that for each x, $P(x, \cdot)$ is a probability measure on $\mathcal{X}$. This P may be used to "run a Markov chain" $X_0, X_1, X_2, \ldots$, with starting state $X_0 = x$ say, by choosing X_1 from $P(x, \cdot)$ and then X_2 from $P(X_1, \cdot)$, and so on. The pair (m, P) is called *reversible* (physicists say "satisfies detailed balance") if P operating on $L^2(m)$ by $Pf(x) = \int f(y)P(x, dy)$ is self-adjoint: $\langle Pf, g\rangle = \langle f, Pg\rangle$. Often, $P(x, dy) = p(x, y)\mu(dy)$ has a kernel and reversibility becomes $m(x)p(x, y) = m(y)p(y, x)$ for all x, y. This says the chain run forward is the same as the chain run backward, in analogy with the time reversibility of the laws of mechanics. Here P operates on all of $L^2(m)$ so we are dealing with bounded self-adjoint operators.

Suppose for a moment that P has a square integrable kernel $p(x, y)$, so $Pf(x) = \int_{\mathcal{X}} p(x, y)f(y)\mu(dy)$. Then P is compact and there are eigenvectors f_i and eigenvalues β_i so

$$Pf_i = \beta_i f_i$$

under a mild connectedness condition $f_0 \equiv 1, \beta_0 = 1$ and $1 = \beta_0 > \beta_1 \geq \beta_2 \geq \cdots > -1$. Then

$$p(x, y) = m(x) \sum_{i=0}^{\infty} \beta_i f_i(x) f_i(y)$$

and the iterated kernel satisfies

$$p^n(x, y) = m(y) \sum_{i=0}^{\infty} \beta_i^n f_i(x) f_i(y).$$

If $f_i(x), f_i(y)$ are bounded (or at least controllable), since $f_0 \equiv 1$,

$$p^n(x, y) \to m(y) \quad \text{as } n \to \infty.$$

This is the spectral approach to convergence. Note that to turn this into a quantitative bound (From starting state x, how large must n be to have $\|P_x^n - m\| < \epsilon$?), the β_i and f_i must be well understood.

The Metropolis algorithm on the permutation group discussed in Section 3 gives an example on finite spaces. Here is an example with an infinite state space drawn from my work with Khare and Saloff-Coste [23] where this program can be usefully pushed through. Note that this example does not arise from the Metropolis construction. It arises from a second basic construction, Glauber dynamics.

Example 2 (Birth and Immigration)**.** The state space $\mathcal{X} = \{0, 1, 2, \ldots\}$. Let $\mu(dx)$ be counting measure and

$$m(x) = \frac{1}{2^{x+1}}, \qquad p(x, y) = \left(\frac{1}{3}\right)^{x+y+1} \binom{x+y}{x} \Big/ \left(\frac{1}{2}\right)^{x+1} \tag{5.3}$$

This Markov chain is used to model population dynamics with immigration. If the population size at generation n is denoted X then, given $X_n = x$,

$$X_{n+1} = \sum_{i=1}^{x} N_{i,n} + M_{n+1}$$

where $N_{i,n}$, the number of offspring of the ith member of the population at time n, are assumed to be independent and identically distributed with

$$p(N_{i,n} = j) = \frac{2}{3}\left(\frac{1}{3}\right)^j, \qquad 0 \leq j < \infty.$$

Here M_{n+1} is migration, assumed to have the same distribution as $N_{i,n}$. Note that $m(x)p(x,y) = m(y)p(y,x)$ so reversibility is in force.

In (5.3) the eigenvalues are shown to be $\beta_j = 1/2^j,\ 0 \le j < \infty$. The eigenfunctions are the orthogonal polynomials for the measure $1/2^{j+1}$. These are Meixner polynomials $M_j(x) = {}_2F_1\left(\binom{-j-x}{1}| -1\right)$. Now, the spectral representation gives the following.

Proposition 1. *For any starting state x for all $n \ge 0$*

$$\chi_x^2(n) = \sum_{y=0}^{\infty} \frac{(p^n(x,y) - m(y))^2}{m(y)} = \sum_{i=1}^{\infty} \beta_i^{2n} M_i^2(x) \frac{1}{2^i}.$$

Next, there is an analysis problem: Given the starting population x, how large should n be so that this chi-square distance to m is small? For this simple case, the details are easy enough to present in public.

Proposition 2. *With notation as in Proposition 1,*

$$\begin{aligned} &\chi_x^2(n) \le 2^{-2c} \qquad \text{for } n = \log_2(1+x) + c, \quad c > 0 \\ &\chi_x^2(n) \ge 2^{2c} \qquad \text{for } n\ = \log_2(x-1) - c, \quad c > 0. \end{aligned}$$

Proof. Meixner polynomials satisfy for all j and $x > 0$

$$|M_j(x)| = \left| \sum_{i=0}^{j \wedge x} (-1)^i \binom{j}{i} x(x-1)\dots(x-i+1) \right| \le \sum_{i=0}^{j} \binom{j}{i} x^i = (1+x)^j.$$

Thus, for $n \ge \log_2(1+x') + c$,

$$\begin{aligned} \chi_x^2(n) &= \sum_{j=1}^{\infty} M_j^2(x) 2^{-j(2n+1)} \le \sum_{j=1}^{\infty} (1+x)^{2j} 2^{-j(2n+1)} \\ &\le \frac{(1+x)^2 2^{-(2n+1)}}{1 - (1+x)^2 2^{-(2n+1)}} \le \frac{2^{-2c-1}}{1 - 2^{-2c-1}} \le 2^{-2c}. \end{aligned}$$

The lower bound follows from using only the lead term. Namely

$$\chi_x^2(n) \ge (1-x)^2 2^{-2n} \ge 2^{2c} \qquad \text{for } n = \log_2(x-1) - c. \qquad \square$$

The results show that convergence is rapid: order $\log_2(x)$ steps are necessary and sufficient for convergence to stationarity.

We were suprised and delighted to see classical orthogonal polynomials appearing in a natural probability problem. The account [23] develops this and gives dozens of other natural Markov chains explicitly diagonalized by orthogonal polynomials.

Alas, one is not always so lucky. The Metropolis chain of (5.1) has the form $Pf(x) = m(x)f(x) + \int h(x,y)f(y)dy$. The multiplier $m(x)$ leads to continuous spectrum. One of our discoveries [24, 25, 59] is that for many chains, this can be side-stepped and the basic outline above can be pushed through to give sharp useful bounds.

5.2 Some theorems

Return to the Metropolis algorithm of Theorem 4. We are able to prove the following.

Theorem 5. *For a bounded Lipshitz domain in $\mathbb{R}^d$, let $p(x)$ satisfy $0 < m \le p(x) \le M < \infty$ for all $x \in \Omega$. Let P_h be defined by (5.1). There are $h_0 > 0,\ \delta_0 \in (0,1)$ and $c_i > 0$ so that*

- *Spec* $P_h \subseteq [-1+\delta_0, 1]$ *for all* $h \le h_0$.
- 1 *is a simple eigenvalue of* P_h.
- *Spec* $P_h \cap [1-\delta_0, 1]$ *is discrete.*
- *The number of eigenvalues of* P_h *in* $[1-h^2\lambda, 1]$, $0 \le \lambda \le \delta_0 h^{-2}$ *(with multiplicity), is bounded above by* $c_1(1+\lambda)^{d/2}$.
- *The spectral gap* $G(h)$ *satisfies* $c_2 h^2 \le G(h) \le c_3 h^2$.
- *For all* $n \ge 1$ *and any* $x \in \Omega$, $\|P^n_{x,h} - p\|_{TV} \le c_4 e^{-nG(h)}$.

More precise evaluation of the gap is available if the boundary of the domain is 'quasi-regular'. Then consider the operator

$$Lf(x) = \frac{-1}{2(d+1)}\left(\triangle f + \frac{\nabla p}{p}\cdot\nabla f\right)$$

with domain $L = \{f \in H^2(p) : |\partial_n f|_{\partial\Omega} = 0\}$. This L has compact resovant with eigenvalues $0 = \nu_0 < \nu_1 < \nu_2 < \dots$.

Theorem 6. *If* $\partial\Omega$ *is quasi-regular and the density* $p(x)$ *is bounded and continuous on* $\bar{\Omega}$, *then*

$$\lim_{h\to 0} h^{-2} G(h) = \nu_1.$$

Reducing to the Neuman problem for L sometimes allows accurate evaluation of the gap [24, 59].

We are able to show that for the hard disc problem of Section 2, a suitable power of the operator of (5.3) satisfies the conditions of Theorems 5 and 6. The associated Ω for hard discs is a complex cuspy set and the extension of standard theory to Lipshitz domains is truly forced.

Again, several caveats are in order. The Theorems *are* satisfactory for a small number of discs but for questions of physical relevance (the dense case), our results have very little content. At present, we do not have good control over the dependence of the constants on the various Lipshitz constants or dimensions. Previous efforts to quantify such things [30] lead to results like $c \doteq (d/4)^{d/4}$. With 100 discs, $d = 200$ and the practical relevance of the results may be questioned. Further, the restriction to densities bounded below is a limitation. Of course, we hope to deal with such issues in future work.

A second caveat: the Metropolis algorithm is not cutting-edge simulation technology. There are block analysis techniques and ways of making non-local moves of several particles [37, 51] which seem useful and call out for analysis.

Finally, spectral techniques are only one of many routes to analysis. Marvelous theorems can be proved by coupling, and Harris recurrence techniques which combine Lyapounov functions and coupling are often useful. Coupling arguments for hard discs are in [26] and [54].

There is also a widely studied discrete version of the problem. There, n particles are placed on the vertices of a connected graph. At each step, a particle is chosen at random and a neighboring site is chosen at random. If the neighboring site is empty, the chosen particle moves there; otherwise the particle stays where it was. This is called "Kawasaki dynamics for the simple exclusion process". This process, with many variations, has a large literature usefully surveyed in [60]. Concrete rates of convergence can be found in [29], [40], [67], [89], ... It is only fair to warn the reader that the similar problem where particles move with a drift on a lattice subject to exclusion (asymmetric exclusion process) has an even larger literature and has evolved into quite a separate subject.

5.3 One idea

One contribution of the analysis which should be broadly useful is an approach to avoiding the continuous spectrum. A wide variety of techniques for bounding eigenvalues and decay of powers for stochastic (e.g., positive) kernels have been developed by the probability community over the past 25 years. These include inequalities of Poincaré, Nash, Sobolev and log-Sobolev type. A useful reference for this material is [79]. The new idea is to apply these techniques to pieces of the operators (which need not be stochastic). The discovery is that this can be pushed through.

In more detail, consider the kernel P_h of (5.1) operating on $L^2(p)$. Write

$$P_h = \Pi + P_h^1 + P_h^2 + P_h^3$$

with Π the orthogonal projection onto the constants

$$P_h^1(x,y) = \sum_{\beta_j \text{ close to } 1} \beta_j(h) f_{j,h}(x) f_{j,h}(y), \qquad P_h^2(x,y) = \sum_{\frac{1}{10} < \frac{1-\beta_j}{h^2} < \frac{9}{10}} \beta_{j,h} f_{j,h}(x) f_{j,h}(y),$$

and

$$P_h^3 = P_h - \Pi - P_h^1 - P_h^2.$$

The pieces are orthogonal, so powers of P_h equal the sum of powers of the pieces. Now P_h^1 and P_h^2 can be analyzed as above. Of course, bounds on eigenvalues and eigenfunctions are still needed. The continuous spectrum is hidden in P_h^3 and one easily gets the crude bounds $\|P_h^3\|_{L^\infty \to L^\infty} \le ch^{-3d/2}$, $\|P_h^3\|_{L^2 \to L^2} \le (1-\delta_0)$. These can be iterated to give $\|(P_h^3)^n\|_{L^\infty \to L^\infty} \le ce^{-\mu n}$ for a universal $\mu > 0$ and all $n > 1/h$. Thus P_h^3 is negligable.

The work is fairly technical but the big picture is fairly stable. It holds for natural walks on compact Riemannian manifolds [59] and in the detailed analysis of the one-dimensional hard disc problem [24, 27].

The main purpose of this section is to show how careful analysis of an applied algorithm can lead to interesting mathematics. In the next section, several further applications of Markov chain Monte Carlo are sketched. None of these have been analyzed.

6 Going Further, Looking Back; Contacts With Math, Contacts Outside Math

This section covers four topics. How someone outside probability can learn more about the present subject; a literature review on rates of convergence; a list of examples showing how a wide spectrum of mathematical tools have been used in analyzing Markov chains; and pointers to applications in various scientific applications.

6.1 Going further

Suppose you are a 'grown up mathematician' who wants to learn some probability. The problem is, probability has its own language and images. It's a little like learning quantum mechanics — the mathematical tools are not a problem but the basic examples and images are foreign. There are two steps. The first is elementary probability — the language of random variables, expectation, independence, conditional probability and the basic examples (binomial, Poisson, geometric, normal, gamma, beta) with their distribution theory. The second is mathematical probability — σ-algebras, laws of large numbers, central limit theory, martingales and brownian motion. Not to mention Markov chains.

The best procedure is to first sit in on an elementary probability course and then sit in on a first-year graduate course. There are hundreds of books at all levels. Two good elementary books are [39] and [78]. This last is a very readable classic (don't miss Chapter 3!). I use Billingsley's book [9] to teach graduate probability.

To learn about Monte Carlo, the classic book [44] is *short* and contains most of the important ideas. The useful books [15] or [62] brings this up to date. Two very good accounts of applied probability which develop Markov chain theory along present lines are [7] and [10]. The advanced theory of Markov chains is well covered by [3] (analytic theory), [38] (semi group theory) and [42] (Dirichlet forms). Two very useful survey articles on rigorous rates of convergence are [67] and [79]. The on-line treatise [1] has a wealth of information about reversible Markov chains. All of the cited works contain pointers to a huge literature.

6.2 Looking back

In this article, I have focused on using spectral theory to give rates of convergence for Markov chains. There are several other tools and schools. Two important ones are *coupling* and *Harris recurrence*. Coupling

is a 'pure probability' approach in which two copies of a Markov chain are considered. One begins in stationarity, the second at a fixed starting point. Each chain evolves *marginally* according to the given transition operator. However, the chains are also set up to move towards each other. When they hit a common point, they "couple" and then move on together. The chain started in stationarity is stationary at every step, in particular at the coupling time T. Thus, at time T, the chain starting from a fixed point is stationary. This approach transforms the task of bounding rates of convergence to bounding the coupling time T. This can sometimes be done by quite elementary arguments. Coupling is such a powerful and original tool that it must have applications far from its origins. A recent example is Robert Neel's proof [71] of Liouville theorems for minimal surfaces.

Book-length treatments of coupling are [61] and [86]. The very useful path coupling variant in [11] and [35] is developed into a marvelous theory of Ricci curvature for Markov chains by [73]. The connections between coupling and eigenvalues is discussed in [12]. The coupling from the past algorithm of Propp–Wilson [77] has made a real impact on simulation. It sometimes allows exact samples to be drawn from intractable probability distributions. It works for the Ising model. I clearly remember my first look at David Wilson's sample of a $2,000 \times 2,000$ Ising model at the critical temperature. I felt like someone seeing Mars through a telescope for the first time.

Harris recurrence is a sophisticated variant of coupling which has a well-developed 'user interface.' This avoids the need for clever, ad hoc constructions. The two chains can be exactly coupled for general state spaces when they hit a 'small set.' They can be driven to the small set by a Lyapounov function. A splendid introduction to Harris recurrence is in [53]. A book-length development is [68]. The topic is often developed under the name of "geometric ergodicity." This refers to bounds of the form $\|K_x^l - \pi\|_{\mathrm{TV}} \leq A(x)\gamma^l$ for $A(x) > 0$ and $0 < \gamma < 1$. Observe that usually, proofs of geometric ergodicity give no hold on $A(x)$ or on γ. In this form, the results are practically useless, saying little more than 'the chain converges for large l.' Bounds with explicit $A(x), \gamma$ are called "honest" in the literature [53]. The work of Jim Hobert and his collaborators is particularly rich in useful bounds for real examples. For further discussion and references, see [23] and [50].

In the presence of geometric ergodicity, a wealth of useful auxiliary results become available. These include central limit theorems and large deviations bounds for averages $1/N \sum f(X_i)$ [56]. the variance of such averages can be usefully estimated [46]. One can even hope to do 'perfect sampling' from the exact stationary distribution [55]. There has been a spirited effort to understand what the set-up required for Harris recurrence says about the spectrum [5, 6]. (Note that coupling and Harris recurrence do not depend on reversibility.)

6.3 Contact with math

The quest for sharp analysis of Markov chains has led to the use and development of tools from various areas of mathematics. Here is a personal catalog.

Group representations

Natural mixing schemes can sometimes be represented as random walk on groups or homogeneous spaces. Then, representation theory allows a useful Fourier analysis. If the walks are invariant under conjugation, only the characters are needed. If the walks are bi-invariant under a subgroup giving a Gelfand pair, the spherical functions are needed. A book-length discussion of this approach is in [14]. Sometimes, the probability theory calls for new group theory. An example is the random walk on the group of upper-triangular matrices with elements in a finite field: starting at the identity, pick a row at random and add it to the row above. The character theory of this group is wild. Carlos-Andre has created a cruder "super-character" theory which is sufficiently rich to handle random walk problems. The detailed use of this required new formula [4] and leads to an extension of the theory to algebra groups in joint work with Isaacs and Theme [22, 34]. This has blossomed into thesis projects [45, 84, 85]. This thread is a nice example of the way that applications and theory interact.

Algebraic geometry

The creation of Markov chains to efficiently perform a sampling task can lead to interesting mathematics. As an example, consider the emerging field of "algebraic statistics." I was faced with the problem of generating (uniformly) random arrays with given row and column sums. These arrays (called contingency tables in statistics) have non-negative integer entries. For two-dimensional arrays, a classical Markov chain Monte Carlo algorithm proceeds as follows. Pick a pair of rows and a pair of columns at random; this delineates four entries. Change these entries by adding and subtracting in one of the following patterns:

$$\begin{pmatrix} + & - \\ - & + \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} - & + \\ + & - \end{pmatrix}.$$

This doesn't change the row/column sums. If the resulting array still has non-negative entries, the chain moves there. Otherwise, the chain stays at the original array.

I needed to extend this to higher-dimensional arrays and similar problems on the permutation group and other structures where linear statistics are to be preserved. The problem is, the analog of the $\left(\begin{smallmatrix} + & - \\ - & + \end{smallmatrix}\right)$ moves that statisticians have thought of do not connect the space. Bernd Sturmfels recognized the original $\left(\begin{smallmatrix} + & - \\ - & + \end{smallmatrix}\right)$ moves as generators of a determinental ideal and suggested coding the problem up as finding generators of a toric ideal. All of the problems fit into this scheme and the emerging fields of computational algebra and Gröbner bases allow practical solutions. The story is too long to tell here in much detail. The original ideas are worked out in [33]. There have been many extensions, bolstered by more than a dozen Ph.D. theses. A flavor of this activity and references can be gathered from [49]. The suite of computational resources in the computer package `Latte` also contains extensive documentation. The subject of algebraic statistics has expanded in many directions. See [74] for its projection into biology and [76] for its projection into design of experiments. As usual, the applications call for a sharpening of algebraic geometric tools and raise totally new problems.

For completeness I must mention that despite much effort, the running time analysis of the original Markov chain on contingency tables has not been settled. There are many partial results suggesting that $(\text{diam})^2$ steps are necessary and sufficient, where diameter refers to the graph with an edge between two arrays if they differ by a $\left(\begin{smallmatrix} + & - \\ - & + \end{smallmatrix}\right)$ move. There are also other ways of sampling that show great promise [16]. Carrying either the analysis or the alternative procedures to the other problems in [33] is a healthy research area.

PDE

The analysis of Markov chains has a very close connection with the study of long time behavior of the solutions of differential equations. In the Markov chain context we are given a kernel $K(x, dy)$ with reversible stationary measure $\pi(dx)$ on a space $\mathcal{X}$. Then K operates as a self-adjoint contraction on $L^2(\pi)$ via $Kf(x) = \int f(y)K(x, dy)$. The associated quadratic form $\mathcal{E}(f|g) = \langle (I - K)f, g\rangle$ is called the Dirichlet form in probability. A Poincaré inequality for K has the form

$$\|f\|_2^2 \le A\mathcal{E}(f|f) \quad \text{for all } f \in L^2(\pi) \text{ with } \int f d\pi = 0.$$

Using the minimax characterization, a Poincaré inequality shows that there is no spectrum for the operator in $[1 - 1/A, 1)$ (Markov operators always have 1 as an eigenvalue). There is a parallel 'parity form' which allows bounding negative spectrum. If the spectrum is supported on $[-1 + 1/A, 1 - 1/A]$ and the Markov chain is started at a distribution σ with L^2 density g, then

$$\|K_\sigma^l - \pi\|_{\text{TV}}^2 \le \|g - 1\|_2^2 \left(1 - \frac{1}{A}\right)^{2l}.$$

This is a useful, explicit bound but it is often "off", giving the wrong rate of convergence by factors of n or more in problems on the symmetric group S_n. A host of more complex techniques can give better results. For example, K satisfies a Nash inequality if for all suitable f,

$$\|f\|_2^{2+1/D} \le A\left\{\mathcal{E}(f|f) + \frac{1}{N}\|f\|_2^2\right\}\|f\|_1^{1/D},$$

and a log-Sobolev inequality if

$$\mathcal{L}(f) \leq A\mathcal{E}(f|f), \qquad \mathcal{L}(f) = \int f^2(x) \log \frac{f(x)^2}{\|f\|_2^2} \pi(dx).$$

Here A, N and D are constants which enter into any conclusions. These inequalities are harder to establish and have stronger conequences. Related inequalities of Cheeger and Sobolev are also in widespread use. For surveys of this technology, see [69] or [79]. The point here is that most of these techniques were developed to study PDE. Their adaptation to the analysis of Markov chains requires some new ideas. This interplay between the proof techniques of PDE and Markov chains has evolved into the small but healthy field of functional inequalities [5, 6] which contributes to both subjects.

Modern PDE is an enormous subject with many more tools and ideas. Some of these, for example, the calculus of pseudo-differential operators and micro-local techniques, are just starting to make in-roads into Markov chain convergence [24, 25, 59].

A major omission in the discussion above are the contributions of the theoretical computer science community. In addition to a set of problems discussed in the final part of this section, a host of broadly useful technical developments have emerged. One way of saying things is this: How does one establish any of the inequalities above (from Poincaré through log-Sobolev) in an explicit problem? Mark Jerrum and Alistair Sinclair introduced the use of paths to prove Cheeger inequalities (called conductance in computer science). Dyer, Frieze, Lovász, Kannan and many students and coauthors have developed and applied these ideas to a host of problems, most notably the problems of approximating the permanent of a matrix and approximating the volume of a convex set. Alas this work suffers from the "polynomial time bug." The developers are often satisfied with results showing that n^{17} steps suffice (after all, it's a polynomial). This leads to theory of little use in practical problems. I believe that the ideas can be pushed to give useful results but at the present writing much remains to be done. A good survey of this set of ideas is in [69].

6.4 Contacts outside math

To someone working in my part of the world, asking about applications of Markov chain Monte Carlo is a little like asking about applications of the quadratic formula. The results are really used in every aspect of scientific inquiry. The following indications are wildly incomplete. I believe you can take any area of science, from hard to social, and find a burgeoning MCMC literature specifically tailored to that area. I note that *essentially none* of these applications is accompanied by any kind of practically useful running time analysis. Thus the following is really a list of open research problems.

Chemistry and physics

From the original application to hard discs through lattice gauge theory [66], MCMC calculations are a mainstay of chemistry and physics. I will content myself by mentioning four very readable books, particularly good at describing the applications to an outsider; I have found them useful ways to learn the science. For physics, [57] and [72]. For chemistry, [41] and [58]. A good feeling for the ubiquity of MCMC can be gleaned from the following quote from the introductory text of the chemist Ben Widom [88, p. 101]:

> "Now, a generation later, the situation has been wholly transformed, and we are able to calculate the properties of ordinary liquids with nearly as much assurance as we do those of dilute gases and harmonic solids .... What is new is our ability to realize van der Waal's vision through the intervention of high speeed digital computing."

Biology

One way to access applications of MCMC in various areas of biology is to look at the work of the statistical leaders of groups driving this research: Jun Liu (Harvard), Michael Newton (Wisconsin), Mike West (Duke) and Wing Wong (Stanford). The home pages of each of these authors contain dozens of papers, essentially all driven by MCMC. Many of these contain innovative, new algorithms (waiting to be studied). In addition, I mention the on-line resources "Mr. Bayes" and "Bugs." These give hundreds of tailored programs for MCMC biological applications.

Statistics

Statisticians work with scientists, engineers and businesses in a huge swathe of applications. Perhaps $10 - -15\%$ of this is driven by MCMC. An overview of applications may be found in the books [43] or [62]. For the very active area of particle filters and their many engineering applications (tracking, filtering), see [36]. For political science-flavored applications, see [43]. Of course, statisticians have also contributed to the design and analysis of these algorithms. An important and readable source is [13].

Group theory

This is a much smaller application. It seems surprising, because group theory (the mathematics of symmetry) seems so far from probability. However, computational group theory, as coded up in the on-line libraries Gap and Magma, makes heavy use of randomized algorithms to do basic tasks such as deciding whether a group (usually given as the group generated by a few permutations or a few large matrices) is all of $S_n(\mathrm{GL}_n)$. Is it solvable, can we find its lower central series, normal closure, Sylow(p) subgroups, etc.? Splendid accounts of this subject are in [47] or [80]. Bounding the running time of widely-used algorithms such as the meat axe or the product replacement algorithm [75] are important open problems on the unlikely interface of group theory and probability.

Computer science (theory)

The analysis of algorithms and complexity theory is an important part of computer science. One central theme is the polynomial/non-polynomial dichotomy. A large class of problems such as computing the permanent of a matrix or the volume of a convex polyhedron have been proved to be # p-complete. Theorists (Broder, Jerrum, Vazarani) have shown that while it may take exponential time to get an exact answer to these problems, one can find provably accurate approximations in a polynomial number of operations (in the size of the input) *provided* one can find a rapidly mixing Markov chain to generate problem instances at random. The above rough description is made precise in the readable book [81]. This program calls for methods of bounding the running time of Markov chains. Many clever analyses have been carried out in tough problems without helpful symmetries. It would take us too far afield to develop this further. Three of my favorite papers (which will lead the reader into the heart of this rich literature) are the [63] analysis of the hit-and-run algorithm, the [52] analysis of the problem of approximating permanents and the [70] analysis of knapsack problems. All of these contain deep, original mathematical ideas which seem broadly useful. As a caveat, recent results of Widgerson suggest a dichotomy: either randomness can be eliminated from these algorithms or $P = NP$. Since nobody believes $P = NP$, there is true excitement in the air.

To close this section, I reiterate that almost none of these applications comes with a useful running time estimate (and almost never with careful error estimates). Also, for computer science, the applications are to computer science *theory*. Here the challenge is to see if practically-useful algorithms can be made from the elegant mathematics.

Acknowledgments

This paper leans on 30 years of joint work with students and coauthors. It was presented at the 25th anniversary of the amazing, wonderful MSRI. I particularly thank Marc Coram, Susan Holmes, Kshitij Khare, Werner Krauth, Gilles Lebeau, Laurent Michel, John Neuberger, Charles Radin and Laurent Saloff-Coste for their help with this paper. A helpful referee and the patience of editor Susan Friedlander are gratefully acknowledged.

References

[1] Aldous, D. and Fill, J. (2002). Reversible Markov chains and random walks on graphs. Monograph.

[2] Allen, M. P. and Tildesely, D. J. (1987). *Computer simulation of liquids.* Oxford University Press, Oxford.

[3] Anderson, W. J. (1991). *Continuous-time Markov chains*. Springer Series in Statistics: Probability and its Applications. Springer-Verlag, New York. An applications-oriented approach.

[4] Arias-Castro, E., Diaconis, P., and Stanley, R. (2004). A super-class walk on upper-triangular matrices. *J. Algebra*, 278(2):739–765.

[5] Bakry, D., Cattiaux, P., and Guillin, A. (2008). Rate of convergence for ergodic continuous Markov processes: Lyapunov versus Poincaré. *J. Funct. Anal.*, 254(3):727–759.

[6] Barthe, F., Bakry, P., Cattiaux, P., and Guillin, A. (2008). Poincaré inequalities for logconcave probability measures: a Lyapounov function approach. *Elec. Comm. Probab.*, to appear.

[7] Bhattacharya, R. N. and Waymire, E. C. (1990). *Stochastic processes with applications*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.

[8] Billera, L. J. and Diaconis, P. (2001). A geometric interpretation of the Metropolis-Hastings algorithm. *Statist. Sci.*, 16(4):335–339.

[9] Billingsley, P. (1995). *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, third edition. A Wiley-Interscience Publication.

[10] Brémaud, P. (1999). *Markov chains*, volume 31 of *Texts in Applied Mathematics*. Springer-Verlag, New York. Gibbs fields, Monte Carlo simulation, and queues.

[11] Bubley, B. and Dyer, M. (1997). Path coupling: a technique for proving rapid mixing in Markov chains. *FOCS*, pages 223–231.

[12] Burdzy, K. and Kendall, W. S. (2000). Efficient Markovian couplings: examples and counterexamples. *Ann. Appl. Probab.*, 10(2):362–409.

[13] Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in hidden Markov models*. Springer Series in Statistics. Springer, New York. With Randal Douc's contributions to Chapter 9 and Christian P. Robert's to Chapters 6, 7 and 13, With Chapter 14 by Gersende Fort, Philippe Soulier and Moulines, and Chapter 15 by Stéphane Boucheron and Elisabeth Gassiat.

[14] Ceccherini-Silberstein, T., Scarabotti, F., and Tolli, F. (2008). *Harmonic analysis on finite groups*, volume 108 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge. Representation theory, Gelfand pairs and Markov chains.

[15] Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2000). *Monte Carlo methods in Bayesian computation*. Springer Series in Statistics. Springer-Verlag, New York.

[16] Chen, Y., Diaconis, P., Holmes, S. P., and Liu, J. S. (2005). Sequential Monte Carlo methods for statistical analysis of tables. *J. Amer. Statist. Assoc.*, 100(469):109–120.

[17] Conner, S. (2003). Simulation and solving substitution codes. Master's thesis, Department of Statistics, University of Warwick.

[18] Critchlow, D. E. (1985). *Metric methods for analyzing partially ranked data*, volume 34 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin.

[19] Diaconis, P. (1988). *Group representations in probability and statistics*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 11. Institute of Mathematical Statistics, Hayward, CA.

[20] Diaconis, P. and Graham, R. L. (1977). Spearman's footrule as a measure of disarray. *J. Roy. Statist. Soc. Ser. B*, 39(2):262–268.

[21] Diaconis, P. and Hanlon, P. (1992). Eigen-analysis for some examples of the Metropolis algorithm. In *Hypergeometric functions on domains of positivity, Jack polynomials, and applications (Tampa, FL, 1991)*, volume 138 of *Contemp. Math.*, pages 99–117. Amer. Math. Soc., Providence, RI.

[22] Diaconis, P. and Isaacs, I. M. (2008). Supercharacters and superclasses for algebra groups. *Trans. Amer. Math. Soc.*, 360(5):2359–2392.

[23] Diaconis, P., Khare, K., and Saloff-Coste, L. (2008a). Gibbs sampling, exponential families and orthogonal polynomials, with discussion. *Statist. Sci.*, to appear.

[24] Diaconis, P. and Lebeau, G. (2008). Micro-local analysis for the Metropolis algorithm. *Math. Z.*, to appear.

[25] Diaconis, P., Lebeau, G., and Michel, L. (2008b). Geometric analysis for the Metropolis algorithm on Lipshitz domains. Technical report, Department of Statistics, Stanford University, preprint.

[26] Diaconis, P. and Limic, V. (2008). Spectral gap of the hard-core model on the unit interval. Technical report, Department of Statistics, Stanford University, preprint.

[27] Diaconis, P. and Neuberger, J. W. (2004). Numerical results for the Metropolis algorithm. *Experiment. Math.*, 13(2):207–213.

[28] Diaconis, P. and Ram, A. (2000). Analysis of systematic scan Metropolis algorithms using Iwahori-Hecke algebra techniques. *Michigan Math. J.*, 48:157–190. Dedicated to William Fulton on the occasion of his 60th birthday.

[29] Diaconis, P. and Saloff-Coste, L. (1993). Comparison theorems for reversible Markov chains. *Ann. Appl. Probab.*, 3(3):696–730.

[30] Diaconis, P. and Saloff-Coste, L. (1996). Nash inequalities for finite Markov chains. *J. Theoret. Probab.*, 9(2):459–510.

[31] Diaconis, P. and Saloff-Coste, L. (1998). What do we know about the Metropolis algorithm? *J. Comput. System Sci.*, 57(1):20–36. 27th Annual ACM Symposium on the Theory of Computing (STOC'95) (Las Vegas, NV).

[32] Diaconis, P. and Shahshahani, M. (1981). Generating a random permutation with random transpositions. *Z. Wahrsch. Verw. Gebiete*, 57(2):159–179.

[33] Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.*, 26(1):363–397.

[34] Diaconis, P. and Thiem, N. (2008). Supercharacter formulas for pattern groups. *Trans. Amer. Math. Soc.*, to appear.

[35] Dobrushin, R. L. (1970). Prescribing a system of random variables by conditional distributions. *Theor. Probab. Appl.– Engl. Tr.*, 15:453–486.

[36] Doucet, A., de Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo in Practice*. Springer-Verlag, New York.

[37] Dress, C. and Krauth, W. (1995). Cluster algorithm for hard spheres and related systems. *J. Phys. A*, 28(23):L597–L601.

[38] Ethier, S. N. and Kurtz, T. G. (1986). *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York. Characterization and convergence.

[39] Feller, W. (1968). *An introduction to probability theory and its applications. Vol. I*. Third edition. John Wiley & Sons Inc., New York.

[40] Fill, J. A. (1991). Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process. *Ann. Appl. Probab.*, 1(1):62–87.

[41] Frenkel, D. and Smit, B. (2002). *Understanding molecular simulation: From algorithms to applications, 2nd edition*. Computational Science Series, Vol 1. Academic Press, San Diego.

[42] Fukushima, M., Ōshima, Y., and Takeda, M. (1994). *Dirichlet forms and symmetric Markov processes*, volume 19 of *de Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin.

[43] Gill, J. (2007). *Bayesian methods: a social and behavioral sciences approach, 2nd ed.* Statistics in the Social and Behavioral Sciences. Chapman & Hall/CRC. Second edition.

[44] Hammersley, J. M. and Handscomb, D. C. (1965). *Monte Carlo methods*. Methuen & Co. Ltd., London.

[45] Hendrickson, A. O. F. (2008). *Supercharacter theories of finite simple groups*. PhD thesis, University of Wisconsin.

[46] Hobert, J. P., Jones, G. L., Presnell, B., and Rosenthal, J. S. (2002). On the applicability of regenerative simulation in Markov chain Monte Carlo. *Biometrika*, 89(4):731–743.

[47] Holt, D. F., Eick, B., and O'Brien, E. A. (2005). *Handbook of computational group theory*. Discrete Mathematics and its Applications (Boca Raton). Chapman & Hall/CRC, Boca Raton, FL.

[48] Hora, A. and Obata, N. (2007). *Quantum probability and spectral analysis of graphs*. Theoretical and Mathematical Physics. Springer, Berlin. With a foreword by Luigi Accardi.

[49] Hosten, S. and Meek, C. (2006). Preface. *J. Symb. Comput.*, 41(2):123–124.

[50] Jarner, S. F. and Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stochastic Process. Appl.*, 85(2):341–361.

[51] Jaster, A. (2004). The hexatic phase of the two-dimensional hard disks system. *Phys. Lett. A*, 330(cond-mat/0305239):120–125.

[52] Jerrum, M., Sinclair, A., and Vigoda, E. (2004). A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *J. ACM*, 51(4):671–697 (electronic).

[53] Jones, G. L. and Hobert, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statist. Sci.*, 16(4):312–334.

[54] Kannan, R., Mahoney, M. W., and Montenegro, R. (2003). Rapid mixing of several Markov chains for a hard-core model. In *Algorithms and computation*, volume 2906 of *Lecture Notes in Comput. Sci.*, pages 663–675. Springer, Berlin.

[55] Kendall, W. S. (2004). Geometric ergodicity and perfect simulation. *Electron. Comm. Probab.*, 9:140–151 (electronic).

[56] Kontoyiannis, I. and Meyn, S. P. (2003). Spectral theory and limit theorems for geometrically ergodic Markov processes. *Ann. Appl. Probab.*, 13(1):304–362.

[57] Krauth, W. (2006). *Statistical mechanics*. Oxford Master Series in Physics. Oxford University Press, Oxford. Algorithms and computations, Oxford Master Series in Statistical Computational, and Theoretical Physics.

[58] Landau, D. P. and Binder, K. (2005). *A Guide to Monte Carlo Simulations in Statistical Physics*. Cambridge University Press, Cambridge.

[59] Lebeau, G. and Michel, L. (2008). Semiclassical analysis of a random walk on a manifold. *Ann. Probab.*, to appear(arXiv:0802.0644).

[60] Liggett, T. M. (1985). *Interacting particle systems*, volume 276 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, New York.

[61] Lindvall, T. (2002). *Lectures on the coupling method*. Dover Publications Inc., Mineola, NY. Corrected reprint of the 1992 original.

[62] Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer-Verlag, New York.

[63] Lovász, L. and Vempala, S. (2006). Hit-and-run from a corner. *SIAM J. Comput.*, 35(4):985–1005 (electronic).

[64] Löwen, H. (2000). Fun with hard spheres. In *Statistical physics and spatial statistics (Wuppertal, 1999)*, volume 554 of *Lecture Notes in Phys.*, pages 295–331. Springer, Berlin.

[65] Macdonald, I. G. (1995). *Symmetric functions and Hall polynomials*. Oxford Mathematical Monographs. The Clarendon Press Oxford University Press, New York, second edition. With contributions by A. Zelevinsky, Oxford Science Publications.

[66] Mackenzie, P. (2005). The fundamental constants of nature from lattice gauge theory simulations. *J. Phys. Conf. Ser.*, 16(doi:10.1088/1742-6596/16/1/018):140–149.

[67] Martinelli, F. (2004). Relaxation times of Markov chains in statistical mechanics and combinatorial structures. In *Probability on discrete structures*, volume 110 of *Encyclopaedia Math. Sci.*, pages 175–262. Springer, Berlin.

[68] Meyn, S. P. and Tweedie, R. L. (1993). *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag London Ltd., London.

[69] Montenegro, R. and Tetali, P. (2006). Mathematical aspects of mixing times in Markov chains. *Found. Trends Theor. Comput. Sci.*, 1(3):x+121.

[70] Morris, B. and Sinclair, A. (2004). Random walks on truncated cubes and sampling $0-1$ knapsack solutions. *SIAM J. Comput.*, 34(1):195–226 (electronic).

[71] Neel, R. W. (2008). A martingale approach to minimal surfaces. *J. Funct. Anal.*, (doi:10.1016/j.jfa.2008.06.033). arXiv:0805.0556v2 [math.DG] (in press).

[72] Newman, M. E. J. and Barkema, G. T. (1999). *Monte Carlo methods in statistical physics*. The Clarendon Press Oxford University Press, New York.

[73] Ollivier, Y. (2008). Ricci curvature of Markov chains on metric spaces. Preprint, submitted, 2008.

[74] Pachter, L. and Sturmfels, B., editors (2005). *Algebraic statistics for computational biology*. Cambridge University Press, New York.

[75] Pak, I. (2001). What do we know about the product replacement algorithm? In *Groups and computation, III (Columbus, OH, 1999)*, volume 8 of *Ohio State Univ. Math. Res. Inst. Publ.*, pages 301–347. de Gruyter, Berlin.

[76] Pistone, G., Riccomagno, E., and Wynn, H. P. (2001). *Algebraic statistics*, volume 89 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL. Computational commutative algebra in statistics.

[77] Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. In *Proceedings of the Seventh International Conference on Random Structures and Algorithms (Atlanta, GA, 1995)*, volume 9, pages 223–252.

[78] Ross, S. M. (2002). *A First Course in Probability, 7th Edition*. Cambridge University Press, Cambridge.

[79] Saloff-Coste, L. (1997). Lectures on finite Markov chains. In *Lectures on probability theory and statistics (Saint-Flour, 1996)*, volume 1665 of *Lecture Notes in Math.*, pages 301–413. Springer, Berlin.

[80] Seress, Á. (2003). *Permutation group algorithms*, volume 152 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge.

[81] Sinclair, A. (1993). *Algorithms for random generation and counting.* Progress in Theoretical Computer Science. Birkhäuser Boston Inc., Boston, MA. A Markov chain approach.

[82] Stanley, R. P. (1999). *Enumerative combinatorics. Vol. 2*, volume 62 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge. With a foreword by Gian-Carlo Rota and appendix 1 by Sergey Fomin.

[83] Taylor, H. M. and Karlin, S. (1984). *An introduction to stochastic modeling.* Academic Press Inc., Orlando, FL.

[84] Thiem, N. and Marberg, E. (2008). Superinduction for pattern groups. Technical report, Department of Mathematics, University of Colorado, Boulder.

[85] Thiem, N. and Venkateswaran, V. (2008). Restricting supercharacters of the finite group of unipotent uppertriangular matrices. Technical report, Department of Mathematics, University of Colorado, Boulder.

[86] Thorisson, H. (2000). *Coupling, stationarity, and regeneration.* Probability and its Applications (New York). Springer-Verlag, New York.

[87] Uhlenbeck, G. E. (1968). An outline of statistical mechanics. In Cohen, E. G. D., editor, *Fundamental Problems in Statistical Mechanics*, volume 2, pages 1–19. North-Holland Publishing Co., Amsterdam.

[88] Widom, B. (2002). *Statistical Mechanics: A Concise Introduction for Chemists.* Cambridge University Press, Cambridge.

[89] Yau, H.-T. (1997). Logarithmic Sobolev inequality for generalized simple exclusion processes. *Probab. Theory Related Fields*, 109(4):507–538.