Bernoulli **19**(4), 2013, 1294–1305
DOI: 10.3150/12-BEJSP09

Some things we've learned (about Markov chain Monte Carlo)

PERSI DIACONIS

390 Serra Mall, Stanford, CA 94305-4065, USA. E-mail: diaconis@math.stanford.edu

This paper offers a personal review of some things we've learned about rates of convergence of Markov chains to their stationary distributions. The main topic is ways of speeding up diffusive behavior. It also points to open problems and how much more there is to do.

Keywords: Markov chains; nonreversible chains; rates of convergence

1. Introduction

Simulation, especially Markov chain Monte Carlo, is close to putting elementary probability (Feller Volume I-style) out of business. This was brought home to me recently in an applied study: Lauren Banklader, Marc Coram, and I were studying "smooshing cards," a widely used mixing scheme where a deck of cards is slid around on the table by two hands. How long should the sliding go on to adequately mix the cards? To gather data, we mixed 52 cards for a minute and recorded the resulting permutations 100 times. Why wouldn't these permutations be random? Our first thoughts suggested various tests: perhaps there would be too many cards that started adjacent that were still adjacent; perhaps the cards originally close to the top would stay close to the top; .... We listed about ten test statistics. To carry out tests requires knowing the null distributions. I could see how to derive approximations using combinatorial probability, for example, for a permutation π, consider $T(\pi) = \#\{i : |\pi_i - \pi_{i+1}| = 1\}$. This has an approximate Poisson(2) distribution with a reasonable error available using Stein's method [6,13]. For $T(\pi)$ the length of the longest increasing subsequence, some of the deepest advances in modern probability [5] allow approximation.

Marc and Lauren looked at me as though I was out of my mind: "But we can trivially find null distributions by simulations and know useful answers in an hour or two that are valid for $n = 52$." Sigh, of course they are right, so what's a poor probabilist to do?

One way I have found to go forward has been to study the algorithms used in simulation. This started with an applied problem: to investigate the optimal strategy in a card game, a programmer had generated millions of random permutations (of 52) using 60 random transpositions. I was sure this was too few (and the simulated results looked funny). This suggests the math question, "how many random transpositions are needed to mix n cards." With Mehrdad Shahshahani [29] we proved that $\frac{1}{2}n \log n + cn$ are necessary and suffice to get e^{-c} close to random. For $n = 52$, it takes 400–500. In retrospect, this is indeed using probability to investigate properties of an algorithm. I've never worried about finding worthwhile problems since then.

The literature on careful analysis of Markov chain mixing times is large. A splendid introduction [41], the comprehensive [1], and the useful articles by Laurent Saloff-Coste [48,49] give a

good picture. There are many other schools that study these problems. Statistical examples (and theorems) can be found in [40,47]; computer science examples are in [43]; statistical physics examples can be accessed via [42]. I have written a more comprehensive survey in [19].

The preceding amounts to hundreds of long technical papers. In this brief survey I attempt to abstract a bit and ask "What are some of the main messages?" I have tried to focus on applied probability and statistics problems. Topics covered are

- Diffusive mixing is slow: Section 2
- There are ways of speeding things up (deterministic doubling, nonreversible chains): Section 3
- Some speed-ups don't work (cutting the cards, systematic scans): Section 4.

Of course, the problems are not all solved and Section 5 gives a list of open questions I hope to see answered.

2. Diffusive mixing

Many Markov chains wander around, doing random walk on a graph. The simplest example is shown in Figure 1, a simple random walk on an n-point path.

Example 1. This chain has transition matrix $K(i, j) = 1/2,\ |i - j| = 1,\ K(1, 1) = K(n, n) = 1/2$. It has stationary distribution $\pi(i) \equiv 1/n$. Powers of the kernel are denoted K^l,

$$K^2(i, j) = \sum_k K(i, k)K(k, j), \qquad K^l(i, j) = \sum_k K(i, k)K^{l-1}(k, j).$$

It is not hard to show that there are universal, positive, explicit constants a, b, c such that for all i, n,

$$ae^{-bl/n^2} \le \left\| K_i^l - \pi \right\| \le ce^{-bl/n^2} \tag{1}$$

with $\|K_i^l - \pi\| = \frac{1}{2}\sum_j |K^l(i, j) - \pi(j)|$.

In situations like (1), we say order n^2 steps are necessary and sufficient for mixing. The n^2 mixing time is familiar from the central limit theorem which can indeed be harnessed to prove (1). The random walk wanders around taking order n^2 steps to go distance n. This is diffusive behavior.

The same kind of behavior occurs in higher dimensions. Fix a dimension d and consider the d-dimensional lattice $\mathbb{Z}^d$. Take a convex set $\mathcal{C}$ in $\mathbb{R}^d$ and look at $\mathcal{X}_\mathcal{C}$, the lattice points inside $\mathcal{C}$.

Figure 1. Simple random walk on an n point path with $1/2$ holding at both ends.

A random walk proceeds inside $\mathcal{X}_\mathcal{C}$ by picking a nearest neighbor uniformly at random (probability $1/2d$). If the new point is inside $\mathcal{X}_\mathcal{C}$ the walk moves there. If the new point is outside $\mathcal{X}_\mathcal{C}$ the walk stays. This includes a standard algorithm for generating a random contingency table with fixed row and column sums: from a starting table T, pick a pair of rows and a pair of columns. This delineates four entries. Try to change these by adding and subtracting 1 in pattern $\begin{smallmatrix} + & - \\ - & + \end{smallmatrix}$ or $\begin{smallmatrix} - & + \\ + & - \end{smallmatrix}$. This doesn't change the row or column sums. If it results in a table with nonnegative entries, make the change; otherwise stay at T. See [20,30] for more on tables.

Returning to the lattice points inside a general convex set, one expects a bound such as (1) with l/n^2 replaced by $l/(\text{diam})^2$ for diam the diameter of $\mathcal{C}$ (length of longest line inside $\mathcal{C}$). Theorems like this are proved in [23,28]. Note that the constants a, b, c depend on the dimension d. They can be as bad as d^d, so the results are not useful for high-dimensional problems. The techniques used are Nash and Sobolev inequalities. There are extensions of these called *log-Sobolev inequalities* [3,27] which give good results in high-dimensional problems. Unfortunately, it is hard to bound the log-Sobolev constant in natural problems.

It is natural to wonder about the choice of the total variation norm $\|K_i^l - \pi\|$ in (1). A variety of other norms are in active use:

- $\chi_i^2(l) = \sum \big(K^l(i,j) - \pi(j)\big)^2/\pi(j)$ $\quad l^2$-norm
- $\max_j 1 - \dfrac{K^l(i,j)}{\pi(j)}$ $\quad$ separation
- $\max_j \left|1 - \dfrac{K^l(i,j)}{\pi(j)}\right|$ $\quad l^\infty$-norm
- $\sum_j \pi(j) \log \dfrac{K^l(i,j)}{\pi(j)}$ $\quad$ Kullback–Liebler.

One of the things I feel I contributed is this: the choice of distance doesn't matter; just choose a convenient one and get on with it. Once you have figured out how to solve the problem with one distance, you usually have understood it well enough to solve it in others. There are inequalities that bound one distance in terms of others [36,48]. The standard choice, total variation, works well with coupling arguments. Indeed, the maximal coupling theorem says that there exist coupling times T so that

$$\left\|K_i^l - \pi\right\|_{\text{TV}} = P\{T > l\} \qquad \text{for all } l.$$

The l^2 distance works well with eigenvalues. Indeed, for reversible chains, on a state space of size n,

$$\chi_i^2(l) = \sum_{j=1}^{n} \lambda_j^{2l} \psi_j^2(i)$$

where λ_j, ψ_j are the eigenvalues and vectors. Furthermore, l^2 distances allow comparison while total variation doesn't; see [33,48]. Here "comparison" refers to a set of techniques where a sharp analysis of one chain can be effectively harnessed to give a useful analysis of a second chain of

interest. For example, on the symmetric group S_n, the random transpositions chain was given a sharp analysis using character theory to show that $\frac{1}{2}n\log n$ steps are necessary and sufficient for mixing. From this, the nonreversible chain "either switch the top two cards or cut the top card to the bottom" was shown to mix in $n^3\log n$ steps. Comparison uses l^2 tools of Dirichlet forms and eigenvalues.

In summary, diffusive behavior occurs for simple random walk Markov chains on low-dimensional spaces. It leads to unacceptably slow mixing. The next section suggests some fixes.

3. Methods of speeding things up

The main point made here is that it is often possible to get rid of diffusive behavior by inserting some simple deterministic steps in the walk. This is not a well developed area but the preliminary results are so striking that I hope this will change.

Example 2 (Uniform distribution on p points). Let p be a prime and C_p be the integers modulo p. Simple random walk goes from $j \in C_p$ to $j \pm 1$. It is convenient to change this to $j \to j, j+1, j-1$ with probability $1/3$. From the arguments in 1 this Markov chain has a uniform stationary distribution $\pi(j) = 1/p$ and from any starting state, order p^2 steps are necessary and sufficient to be close to random. There is diffusive behavior.

Consider the following variation: set $X_0 = 0$ and

$$X_{n+1} = 2X_n + \varepsilon_{n+1} \pmod{p}$$

with $\varepsilon_n = 0, +1, -1$ with probability $1/3$. This has the same amount of randomness but intersperses deterministic doubling. Let $K_n(j) = P\{X_n = j\}$. In [16] it is shown that the doubling gives a remarkable speed-up: order $\log p$ steps are necessary and sufficient for almost all p. One version of the result follows.

Theorem ([16]). *For any $\varepsilon > 0$, and almost all odd p, if $l > (C^* + \varepsilon)\log_2 p$ then $\|K_l - \pi\| < \varepsilon$ where $C^* = (1 - \log_2(\frac{5+\sqrt{17}}{9}))^{-1} = 1.01999186\ldots$.*

In a series of extensions, Martin Hildebrand [37,39] has shown this result is quite robust to variations: p need not be prime, the probability distribution of ε_i can be fairly general, the multiplier 2 can be replaced by a general a and even a_{n+1} chosen randomly (e.g., 2 or $1/2 \pmod{p}$ with probability $1/2$). The details vary and the arguments require new ideas.

Once one finds such a phenomenon, it is natural to study things more carefully. For example, is "almost all p" needed? In [16] it is shown that the answer is yes: there are infinitely many primes p such that $\log(p)\log\log(p)$ steps are necessary and sufficient. Hildebrand [38] shows that one cannot replace C^* by 1 in the theorem. In [21] similar walks are studied on other groups.

I have heard several stories about how adding a single extra move to a Markov chain speeded things up dramatically. This seems like an important area crying out for development. For example, in the "lattice points inside a convex set $\mathcal{X}_C$" of Section 1, is there an analog of deterministic

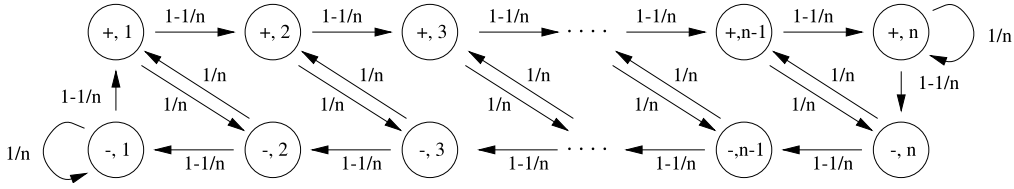

Figure 2. A discrete version of hybrid Monte Carlo.

doubling which speeds up the $(\text{diam})^2$ rate? The reflection walks of [10] for the original Metropolis problem of random placement of non-overlapping hard discs in a box is an important speed-up of local algorithms. Can it be abstracted?

Example 3 (Getting rid of reversibility). Consider again generating a random point in $\{1, 2, 3, \ldots, n\}$ by a local algorithm. In joint work with Holmes and Neale [22] the algorithm of Figure 2 was suggested. Along the top, bottom, and side edges of the graph, the walk moves in the direction shown with probability $1 - (1/n)$. On the diagonal edges the walk moves (in either direction) with probability $1/n$. The loops indicate holding with probability $1/n$. While this walk is definitely not reversible, it is doubly stochastic and so has a uniform stationary distribution. Intuitively, it moves many steps in one direction before switching directions (with probability $1/n$). In [22] it is shown that this walk takes just n steps to reach stationarity (and this is best possible for such a local algorithm). The analysis shows that this is a hidden version of the $X_{n+1} = a_{n+1}X_n + \varepsilon_n$ walk with $a_{n+1} = 1$ or -1 with probability $1 - (1/n)$ and $1/n$. The walk was developed as a toy version of the hybrid Monte Carlo algorithm of lattice field theory [31]. See [45] for its developments in statistics. This is a general and broadly useful class of algorithms that have resisted analysis. Someone should take up this challenge!

There has been some further development of the ideas in [22]. Chen, Lovász and Pak [14] abstracted the idea to a "lifting" of general Markov chains. They showed that the square-root speed-up (order n^2 to order n in the example) was best possible for their class of algorithms. Hildebrand [38] studied the lifted version of the Metropolis algorithm (based on nearest neighbor random walk on $\{1, 2, \ldots, n\}$) for a general stationary distribution. The algorithm of Figure 2 chooses to reverse with probability $1/n$. What about θ_n/n? Evidence in [22] suggests that $\theta_n = \sqrt{\log n}$ is better. Gade and Overton [35] set this up as an optimization problem, seeking to find the value of θ_n that maximizes the spectral gap. In a final important development, Neal [44] has shown that any reversible Markov chain can be speeded up, at least in terms of spectral gap, by a suitable nonreversible variant. See [24] for further developments, to spectral analysis for 2^d-order Markov chains.

In summary, the results of this section show that real speed-ups of standard algorithms are possible. These results should have practical consequences: even if it is hard to prove, it is usually easy to find a few "big moves" that preserve the stationary distribution. For a survey of approaches to designing algorithms that avoid diffusion, see [2].

4. Not all speed-ups work

One of the joys of proving things is that, sometimes, things that "everybody knows" aren't really true. This is illustrated with three examples: systematic vs. random scans, cutting the cards, and cooking potatoes.

Example 4 (Systematic vs. random scans). Consider applying the Gibbs sampler to a high-dimensional vector, for example, generating a replication of an Ising model on an $n \times n$ grid. The Gibbs sampler proceeds by updating one coordinate at a time. Is it better to be systematic, ordering the coordinates and visiting each in turn, or is choosing a random coordinate (i.i.d. uniform choices) better? "Everybody knows" that systematic scans are better. Yet, in the only cases where things can be proved, random scan and systematic scan have the same rates of convergence.

Two classes of examples have been studied. Diaconis and Ram [25] studied generation of a random permutation on n letters from the Mallow's model,

$$P_\theta(\sigma) = z^{-1}(\theta)\theta^{I(\sigma)}, \qquad 0 < \theta \le 1,$$

with $I(\sigma)$ the number of inversions. Here $\sigma = (\sigma(1), \sigma(2), \ldots, \sigma(n))$ is a permutation of n and $I(\sigma)$ is the number of $i < j$ with $\sigma(i) > \sigma(j)$. This is "Mallow's model through Kendal's tau." For $0 < \theta < 1$ fixed, it has $\sigma =$ identity most likely and falls away from this exponentially. The Metropolis algorithm forms a Markov chain, changing the current σ to $(i, i+1)\sigma$ if this decreases the number of inversions and by a coin flip with probability θ if $I((ij)\sigma) > I(\sigma)$; otherwise the chain stays at σ. Here, the systematic scan proposes $(1, 2)$, then $(2, 3), \ldots, (n-1, n), (n-2, n-1), \ldots, (1, 2)$, say. The random scan chooses t uniformly and independently each time. Benjamini, Berger, Hoffman and Mossel [8] show that order n^2 random scan steps suffice for random scan. Diaconis and Ram show that order n systematic scan steps suffice. Since each systematic scan costs $2n$ steps, the algorithms are comparable. A number of other scanning strategies and walks on different groups confirm the finding: being systematic doesn't help to change the order of magnitude of the number of steps needed for convergence. Two notable features: the analysis of [25] uses Fourier analysis on the Hecke algebra. The random scan analysis uses deep results from the exclusion process. Both are fairly difficult. See [11] for a different approach to proof.

A different set of examples is considered by Dyer–Goldberg–Jerrum [32]. They studied the standard algorithm for generating a random proper coloring of a graph with c colors (adjacent vertices must have different colors). The algorithm picks a vertex and replaces the color by a randomly chosen color. This step is accepted if the coloring is proper. How should vertices be chosen to get rapid mixing? Systematic scan periodically cycles through the vertices in a fixed order. Random scan chooses vertices uniformly. Intuitively, systematic seems better. However, their careful mathematical analysis shows the two approaches have the same convergence rates.

For Glauber dynamics, for Ising and Potts models on graphs, Yuval Peres (in personal communication) conjectures that random updates are never faster than systematic scan, and systematic scan can be faster than random updates by at most a factor of $\log n$ on an n-vertex graph. A speed-up of $\log n$ is attained at infinite temperature where systematic scan needs one round of n updates

and random scan needs $n \log n$ updates; see the opening example of [25]. Partial results in the monotone case are in [46], Thm. 3.1, 3.2, 3.3.

The results above are tentative because only a few classes of examples have been studied and the conclusion contradicts common wisdom. It suggests a research program; a survey of the literature on scanning strategies is in [25]. At least, someone should find one natural example where systematic scan dominates.

Example 5 ("Put your faith in Providence but always cut the cards?"). Does cutting the cards help mixing? I find it surprising that the answer is "Not really and it can even slow things down." To say things carefully, work on S_n the group of all $n!$ permutations. A probability on S_n is $Q(\sigma) \geq 0$, $\sum_\sigma Q(\sigma) = 1$. Repeated mixing is modeled by convolution,

$$Q^{*2}(\sigma) = \sum_\eta Q(\eta) Q\big(\sigma \eta^{-1}\big), \qquad Q^{*k}(\sigma) = \sum Q(\eta) Q^{*k-1}\big(\sigma \eta^{-1}\big).$$

The uniform distribution is $U(\sigma) = 1/n!$. A random cut C puts mass $1/n$ on each of the n-cycles $\begin{smallmatrix} 1 & 2 & \cdots & n \\ i & i+1 & \cdots & i-1 \end{smallmatrix}$, $1 \leq i \leq n$. It is easy to see, for any of the distances in Section 1, $d(C * Q, U) \leq d(Q, U)$. So, in this sense, cutting doesn't hurt (stay tuned!). But does it help? The answer depends on Q. For Q the usual Gilbert–Shannon–Reeds measure for riffle shuffling Q^{*k} is close to U for $k = \frac{3}{2} \log_2 n + c$ [7]. This is "about 7" when $n = 52$. For general n, Fulman [34] proves that applying C after Q^{*k} does not change the $\frac{3}{2} \log_2 n$ rates of convergence.

However, Diaconis and Shahshahani [17] construct a probability measure Q on S_n such that $Q * Q = U$ (but $Q \neq U$). For this Q, $(CQ) * (CQ) \neq U$. Thus, shuffling twice with this Q gives perfect mixing but interspersing random cuts fouls things up. Of course, this Q is not a naturally occurring mixing process. Still, it shows the need for proof.

An example where cutting helps (at least a bit) is in [18]. Here, Q is the random transpositions measure studied by [9,12,29]. In [29] it is shown that $\frac{1}{2} n \log n + cn$ steps are necessary and sufficient for randomness: if $c > 0$, $\|Q^{*k} - U\| \leq 2e^{-c}$; if $c < 0$, the distance is bounded away from 0 for all n. In [18], it is shown that the mixing time of $C * Q$ is $\frac{3}{8} n \log n + cn$. These are subtle differences. Hard work and good luck are required to get the lead term constants and cut-off accurately.

Example 6 (Cooking potatoes). When we stir food in a frying pan, e.g., sliced-up potatoes, some ill-defined ergodic theorem helps to explain why they get (roughly) evenly browned. One pale mathematical version of this problem considers n circular discs of potato arranged around the edge of a frying pan as shown in Figure 3. Imagine the discs have two sides, heads and tails. They start with all sides heads-up. At each step, a spatula of radius d potatoes is inserted at random and all potatoes over the spatula are turned over in place. For simplicity, assume that d and n are relatively prime. It is intuitively clear (and not hard to prove) that with repeated flips, the up/down pattern becomes random; all 2^n patterns are equally likely in the limit.

How long does it take to get close to random, and how does it depend on d? I am surprised that the answer doesn't depend on d; a tiny spatula of diameter 1 or a giant spatula of diameter $n/2$ all

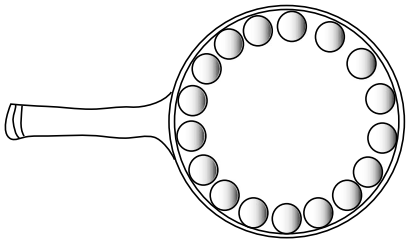

Figure 3. 16 circular discs inside a pan.

require $\frac{1}{4}n\log n + cn$ steps (necessary and sufficient) to mix. The result even holds for "combs," a spatula with teeth that turns over every other potato among d (or more general patterns).

To see why, regard the potatoes as a binary vector and write C_2^n for the state space. The spatula is a second binary vector, V. The probability measure Q adds a randomly chosen cyclic shift of V to the current state. Addition is coordinate-wise, mod 2. For $V = e_1 = (1, 0, \ldots, 0)$, this is just nearest neighbor random walk on the hypercube, also known as the Ehrenfest urn. The $\frac{1}{4}n\log n + cn$ answer is well known [29]. Consider general V. Let $V_1 = V, V_2, \ldots, V_n$ be the n-cyclic shifts of V. Relatively prime d and n ensures that $V_1, V_2, \ldots, V_n$ form a basis of the space of binary n-tuples. From linear algebra, there is an invertible matrix A ($n \times n$ mod 2 entries) taking V_i to e_i, $1 \le i \le n$. If $0 = X_0, X_1, X_2, \ldots$ is the Ehrenfest walk (spatula of size 1) and $0 = Y_0, Y_1, Y_2, \ldots$ is the walk based on V, then $P\{Y_k \in S\} = P\{X_k \in A^{-1}S\}$ for any set S. It follow that the total variation distance to uniformity is the same for the two processes. The same argument works for any basis $V_1, V_2, \ldots, V_n$ and any distance.

Suppose we allow a larger generating set $V_1, V_2, \ldots, V_N$ say with $N > n$. How should the $\{V_i\}_{i=1}^N$ be chosen to get rapid mixing? David Wilson [50] developed some elegant theory for this question.

Theorem (Wilson). *For all sufficiently large n and $N > n$, and $V_1, V_2, \ldots, V_N \in C_2^n$, the random walk based on repeatedly adding a uniformly chosen V_i satisfies*

1. *for any choice of $V_1, \ldots, V_N$, if $k < (1-\varepsilon)T(n,N)$ then $\|Q^{*k} - U\| > 1 - \varepsilon$;*
2. *for almost all choices of $V_1, \ldots, V_N$, if $k > (1+\varepsilon)T(N)$ then $\|Q^{*k} - U\| < \varepsilon$ provided the Markov chain is ergodic.*

Here $T(n,N) = \frac{N}{2}\frac{1}{1-H^{-1}(n/N)}$, $H(x) = x\log_2\frac{1}{x} + (1-x)\log_2\frac{1}{1-x}$, $0 \le x < 1$. Note that almost all choices in item 2 of the theorem will be ergodic when $N - n$ is sufficiently large. For example, when $N = 2n$, $T(n,N) \doteq 0.24853n$ steps are required. Further details are in [50].

5. Open questions

Question 1. In item 2 of Wilson's theorem (Example 6), the result holds for almost all choices $V_1, V_2, \ldots, V_N$. Can an explicit set be found, e.g., for $N = 2n$?

Question 2. The same set of problems can be considered for any group G. If a generating set S is chosen at random, what is the typical rate of convergence? This is the topic of random random walks. Hildebrand [37] gives a survey. Babai, Beals, and Seress [4] give the best bounds on the diameter of such random Cayley graphs. These may be turned into (perhaps crude) rates of convergence via bounds in [26]. I cannot resist adding mention of one of my old conjectures. For the alternating group A_n, it is known that a randomly chosen pair of elements generate A_n with probability approaching 1. I conjecture that the random walk based on any generating pair gets random in at most $n^3 \log n$ steps.

Question 3. Fix a generating set $S \subseteq G$. What element should be added to S to best speed up mixing? For example, suppose $G = S_n$ (for some odd n) and $S = \{(1,2), (1,2,3,\ldots,n)\}$, a transposition and an n-cycle. It is known that order $n^3 \log n$ steps are necessary and suffice for randomness [26,50]. Is there a choice of σ to be added that appreciably speeds this up? For S_n, it is conjectured that all such walks have a sharp cutoff [14].

Question 4. One may ask a similar question for random walk on any graph. To be specific, consider a connected d regular graph with n even. Thus, nearest neighbor random walk has a uniform stationary distribution. Add in $n/2$ edges forming a perfect matching. This gives a $(d+1)$ regular graph. What choice of edges give fastest mixing? If the original graph is an n-cycle and thus 2-regular, [15] shows that a random matching improves the diameter to $\log_2 n + o(1)$. She gives an explicit construction of a matching that has diameter $2\log_2 n + o(1)$. These diameter bounds translate into eigenvalue bounds and so bounds on rates of convergence using standard tools. However, something is lost in these translations and it would be worthwhile to know accurate rates of convergence to the uniform distribution.

An important variation: consider a reversible Markov chain $K(x,y)$ on a finite set $\mathcal{X}$ with stationary distribution $\pi(x)$. Suppose a weighted edge is to be added to the underlying graph and the resulting Markov chain is "Metropolized" so that it still has stationary distribution $\pi(x)$. What edges best improve mixing, or best improve the spectral gap? These questions are closely related to Section 3.

References

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

[2] Andersen, H.C. and Diaconis, P. (2007). Hit and run as a unifying device. *J. Soc. Fr. Stat. & Rev. Stat. Appl.* **148** 5–28. MR2502361

[3] Ané, C., Blachère, S., Chafaï, D., Fougères, P., Gentil, I., Malrieu, F., Roberto, C. and Scheffer, G. (2000). *Sur les Inégalités de Sobolev Logarithmiques*. *Panoramas et Synthèses* [*Panoramas and Syntheses*] **10**. Paris: Société Mathématique de France. MR1845806

[4] Babai, L., Beals, R. and Seress, Á. (2004). On the diameter of the symmetric group: Polynomial bounds. In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 1108–1112 (electronic). New York: ACM. MR2291003

[5] Baik, J., Deift, P. and Johansson, K. (1999). On the distribution of the length of the longest increasing subsequence of random permutations. *J. Amer. Math. Soc.* **12** 1119–1178. MR1682248

[6] Barbour, A.D., Holst, L. and Janson, S. (1992). *Poisson Approximation*. *Oxford Studies in Probability* **2**. New York: The Clarendon Press Oxford Univ. Press. MR1163825
[7] Bayer, D. and Diaconis, P. (1992). Trailing the dovetail shuffle to its lair. *Ann. Appl. Probab.* **2** 294–313. MR1161056
[8] Benjamini, I., Berger, N., Hoffman, C. and Mossel, E. (2005). Mixing times of the biased card shuffling and the asymmetric exclusion process. *Trans. Amer. Math. Soc.* **357** 3013–3029 (electronic). MR2135733
[9] Berestycki, N., Schramm, O. and Zeitouni, O. (2011). Mixing times for random k-cycles and coalescence-fragmentation chains. *Ann. Probab.* **39** 1815–1843. MR2884874
[10] Bernard, E.P. and Krauth, W. (2012). Event-driven Monte Carlo algorithm for general potentials. *Phys. Rev. E*. To appear.
[11] Bhakta, P., Miracle, S., Randall, D. and Pascoe Streib, A. (2012). Mixing times of self-organizing lists and biased permutations. Available at arXiv:1204.3239.
[12] Bormashenko, O. (2011). A coupling argument for the random transposition walk. Available at arXiv:1109.3915.
[13] Chatterjee, S., Diaconis, P. and Meckes, E. (2005). Exchangeable pairs and Poisson approximation. *Probab. Surv.* **2** 64–106. MR2121796
[14] Chen, F., Lovász, L. and Pak, I. (1999). Lifting Markov chains to speed up mixing. In *Annual ACM Symposium on Theory of Computing* (*Atlanta*, *GA*, 1999) 275–281 (electronic). New York: ACM. MR1798046
[15] Chung, F.R.K. (1989). Diameters and eigenvalues. *J. Amer. Math. Soc.* **2** 187–196. MR0965008
[16] Chung, F.R.K., Diaconis, P. and Graham, R.L. (1987). Random walks arising in random number generation. *Ann. Probab.* **15** 1148–1165. MR0893921
[17] Diaconis, P. (1988). Applications of noncommutative Fourier analysis to probability problems. In *École D'Été de Probabilités de Saint-Flour XV–XVII*, 1985–87. *Lecture Notes in Math.* **1362** 51–100. Berlin: Springer. MR0983372
[18] Diaconis, P. (1991). Finite Fourier methods: Access to tools. In *Probabilistic Combinatorics and Its Applications* (*San Francisco*, *CA*, 1991). *Proc. Sympos. Appl. Math.* **44** 171–194. Providence, RI: Amer. Math. Soc. MR1141927
[19] Diaconis, P. (2009). The Markov chain Monte Carlo revolution. *Bull. Amer. Math. Soc.* (*N.S.*) **46** 179–205. MR2476411
[20] Diaconis, P. and Gangolli, A. (1995). Rectangular arrays with fixed margins. In *Discrete Probability and Algorithms* (*Minneapolis*, *MN*, 1993). *IMA Vol. Math. Appl.* **72** 15–41. New York: Springer. MR1380519
[21] Diaconis, P. and Graham, R. (1992). An affine walk on the hypercube. *J. Comput. Appl. Math.* **41** 215–235. MR1181722
[22] Diaconis, P., Holmes, S. and Neal, R.M. (2000). Analysis of a nonreversible Markov chain sampler. *Ann. Appl. Probab.* **10** 726–752. MR1789978
[23] Diaconis, P., Lebeau, G. and Michel, L. (2011). Geometric analysis for the metropolis algorithm on Lipschitz domains. *Invent. Math.* **185** 239–281. MR2819161
[24] Diaconis, P. and Miclo, L. (2012). On the spectral analysis of second-order Markov chains. Unpublished manuscript.
[25] 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. MR1786485
[26] Diaconis, P. and Saloff-Coste, L. (1993). Comparison techniques for random walk on finite groups. *Ann. Probab.* **21** 2131–2156. MR1245303

[27] Diaconis, P. and Saloff-Coste, L. (1996). Logarithmic Sobolev inequalities for finite Markov chains. *Ann. Appl. Probab.* **6** 695–750. MR1410112
[28] Diaconis, P. and Saloff-Coste, L. (1996). Nash inequalities for finite Markov chains. *J. Theoret. Probab.* **9** 459–510. MR1385408
[29] Diaconis, P. and Shahshahani, M. (1981). Generating a random permutation with random transpositions. *Z. Wahrsch. Verw. Gebiete* **57** 159–179. MR0626813
[30] Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.* **26** 363–397. MR1608156
[31] Duane, S., Kennedy, A., Pendleton, B. and Roweth, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B* **195** 216–222.
[32] Dyer, M., Goldberg, L.A. and Jerrum, M. (2008). Dobrushin conditions and systematic scan. *Combin. Probab. Comput.* **17** 761–779. MR2463409
[33] Dyer, M., Goldberg, L.A., Jerrum, M. and Martin, R. (2006). Markov chain comparison. *Probab. Surv.* **3** 89–111. MR2216963
[34] Fulman, J. (2000). Affine shuffles, shuffles with cuts, the Whitehouse module, and patience sorting. *J. Algebra* **231** 614–639. MR1778162
[35] Gade, K.K. and Overton, M.L. (2007). Optimizing the asymptotic convergence rate of the Diaconis–Holmes–Neal sampler. *Adv. in Appl. Math.* **38** 382–403. MR2301703
[36] Gibbs, A. and Su, F. (2002). On choosing and bounding probability metrics. *Int. Statist. Rev.* **70** 419–435.
[37] Hildebrand, M. (2005). A survey of results on random random walks on finite groups. *Probab. Surv.* **2** 33–63. MR2121795
[38] Hildebrand, M. (2009). A lower bound for the Chung–Diaconis–Graham random process. *Proc. Amer. Math. Soc.* **137** 1479–1487. MR2465674
[39] Hildebrand, M. and McCollum, J. (2008). Generating random vectors in $(\mathbb{Z}/p\mathbb{Z})^d$ via an affine random process. *J. Theoret. Probab.* **21** 802–811. MR2443637
[40] Jones, G.L. and Hobert, J.P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statist. Sci.* **16** 312–334. MR1888447
[41] Levin, D.A., Peres, Y. and Wilmer, E.L. (2009). *Markov Chains and Mixing Times*. Providence, RI: Amer. Math. Soc. With a chapter by James G. Propp and David B. Wilson. MR2466937
[42] Martinelli, F. (2004). Relaxation times of Markov chains in statistical mechanics and combinatorial structures. In *Probability on Discrete Structures*. *Encyclopaedia Math. Sci.* **110** 175–262. Berlin: Springer. MR2023653
[43] Montenegro, R. and Tetali, P. (2006). Mathematical aspects of mixing times in Markov chains. *Found. Trends Theor. Comput. Sci.* **1** 237–354. MR2341319
[44] Neal, R.M. (2004). Improving asymptotic variance of MCMC estimators: Non-reversible chains are better. Technical report, Dept. of Statistics, University of Toronto. Available at http://www.cs.toronto.edu/~radford/ftp/asymvar.pdf.
[45] Neal, R.M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 113–162. Boca Raton, FL: CRC Press. MR2858447
[46] Peres, Y. and Winkler, P. (2011). Can extra updates delay mixing? Available at arXiv:1112.0603.
[47] Rosenthal, J.S. (2002). Quantitative convergence rates of Markov chains: A simple account. *Electron. Commun. Probab.* **7** 123–128 (electronic). MR1917546
[48] Saloff-Coste, L. (1997). Lectures on finite Markov chains. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1996). *Lecture Notes in Math.* **1665** 301–413. Berlin: Springer. MR1490046
[49] Saloff-Coste, L. (2004). Random walks on finite groups. In *Probability on Discrete Structures*. *Encyclopaedia Math. Sci.* **110** 263–346. Berlin: Springer. MR2023654

[50] Wilson, D.B. (1997). Random random walks on $\mathbf{Z}_2^d$. *Probab. Theory Related Fields* **108** 441–457. MR1465637