

## The mathematics of mixing things up

Persi Diaconis

Received: date / Accepted: date

**Abstract** How long should a Markov chain Monte Carlo algorithm be run? Using examples from statistical physics (Ehrenfest urn, Ising model, hard discs) as well as card shuffling, this tutorial paper gives an overview of a body of mathematical results that can give useful answers to practitioners (viz: seven shuffles suffice for practical purposes). It points to new techniques (path coupling, geometric inequalities, and Harris recurrence). The discovery of phase transitions in mixing times (the cutoff phenomenon) is emphasized.

**Keywords** Markov chains · rates of convergence · cutoff phenomenon

### 1 Introduction

Suppose someone has an algorithm for generating random “things.” How can we understand if or when it works? For example, I was recently contacted by the world’s largest manufacturer of card shuffling machines. They had a new gadget, an actual large black box. A deck of 52 cards was dropped in and came out the other end in one pass. they wanted to know if the deck was suitably random.

For small decks (e.g., four cards) many of us would say, “Just try it out a thousand times and see if all 24 arrangements come out about equally likely.” This isn’t feasible for 52 cards. Standard practice here is to think up some natural tests, for example:

- the position of the original top card;
- the separation of cards originally adjacent;
- the relative order of small groups after the shuffle.

---

Research supported in part by National Science Foundation grant DMS 0804324.

P. Diaconis  
Departments of Mathematics and Statistics  
Stanford University  
E-mail: diaconis@math.stanford.edu

The main point of this article is that there is a third way: you can try to prove a general convergence result which shows that for any test (or many tests) the answer will be satisfactory after an appropriate number of steps.

More carefully, let  $\mathcal{X}$  be a finite set of configurations. These might be arrangements of a deck of  $n$  cards or Ising configurations on a given graph. We are given a probability distribution on  $\mathcal{X}$ ,  $\pi(x)$ . In the card-shuffling case this is  $1/n!$ . In the Ising case  $\pi(x) = z^{-1} e^{-\beta H(x)}$  with  $H(x)$  the energy of an Ising configuration.

The algorithms discussed here are driven by a Markov chain. Thus there is a transition matrix  $K(x,y)$  giving the chance of moving from  $x$  to  $y$  in one step,

$$\sum_y K(x,y) = 1, \quad \text{for } K(x,y) \geq 0,$$

which has  $\pi(x)$  as a stationary distribution:

$$\sum_x \pi(x) K(x,y) = \pi(y) \quad (\text{often from } \pi(x) K(x,y) = \pi(y) K(y,x)).$$

Iterates of the algorithm correspond to powers of the matrix

$$K^2(x,y) = \sum_z K(x,z) K(z,y), \quad K^l(x,y) = \sum_z K(x,z) K^{l-1}(z,y).$$

Under mild conditions, for any starting state  $x$ , the algorithm converges:

$$K^l(x,y) \longrightarrow \pi(y) \quad \text{as } l \text{ tends to infinity.}$$

The question studied here is ‘‘How fast?’’ For this, a distance to stationarity is specified. The usual distance is *total variation*,

$$\|K_x^l - \pi\| = \frac{1}{2} \sum_y |K^l(x,y) - \pi(y)| = \max_{A \subseteq \mathcal{X}} |K^l(x,A) - \pi(A)|. \quad (1.1)$$

The first equality in (1.1) is the definition. The second equality is a (tiny) theorem; it helps interpret the distance. Think about the card-shuffling case. Then,  $\mathcal{X}$  is all  $n!$  permutations. Let  $A$  be a subset of permutations (e.g., all permutations where the ace of spades is in the top half of the deck). Look at the chance that, starting from  $x$ , after  $l$  shuffles the deck is in an arrangement in  $A$ . Compare this with the stationary probability of  $A$  (here  $|A|/n!$ ). Take the difference between these two numbers and then take the maximum of these differences over all  $A$ . If this is small, then for any test question  $A$ , the algorithm produces something close to stationary.

With these ingredients specified, there is a well-defined math problem: given  $\varepsilon > 0$ , how large must  $l$  be so that

$$\|K_x^l - \pi\| < \varepsilon?$$

Note that we have kept track of the starting state  $x$ ; in natural examples this can affect the answer and we would like to understand how.

*Example 1 (Ehrenfest's urn)* The classical Ehrenfest's urn is a toy model of diffusion, introduced to help understand some of the paradoxes of statistical mechanics. Consider  $n$  unlabeled balls and two urns. At the start, all the balls are in the left-hand urn. At each stage, a ball is chosen at random and moved to the opposite urn. A configuration may be coded as the number of balls in the left-hand urn; thus  $\mathcal{X} = \{0, 1, 2, \dots, n\}$ . The stationary distribution is binomial:  $\pi(j) = \binom{n}{j}/2^n$ . (After a long time, each ball is equally likely to be in either urn.) The transition mechanism described has a parity problem: after an even number of steps the state differs from its start by an even number. One simple fix: pick a ball at random or do nothing, each with probability  $1/(n+1)$ . Then

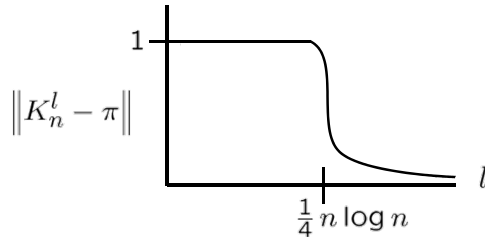
$$K(j, j-1) = \frac{j}{n+1}, \quad K(j, j) = \frac{1}{n+1}, \quad K(j, j+1) = \frac{n-j}{n+1}, \quad 0 \leq j \leq n. \quad (1.2)$$

The following theorem shows that, starting from  $n$  (or  $0$ ),  $\frac{1}{4}n \log n + cn$  steps are necessary and sufficient for convergence.

**Theorem 1 ([26])** *For the Ehrenfest Markov chain (1.2), if  $l = \frac{1}{4}n \log n + cn$ , for  $c > 0$ ,*

$$\|K_n^l - \pi\| \leq e^{-c}.$$

*There is a matching lower bound: if  $l = \frac{1}{4}n \log n + cn$ , for  $c < 0$ , the distance is exponentially close to 1.*



**Fig. 1** The sharp cutoff in convergence to stationarity at  $\frac{1}{4}n \log n$ , the cutoff phenomenon.

The theorem shows there is a sharp cutoff in convergence to stationarity at  $\frac{1}{4}n \log n$ . A graph appears as shown in Figure 1. Of course, the distance to stationarity decreases as  $l$  increases. However, the theorem shows it stays relatively flat, close to 1, before  $\frac{1}{4}n \log n$  and then abruptly cuts down, tending to zero exponentially fast. The precise shape of this cutoff curve is determined in [19] where it is shown to converge to one doubly exponentially fast in  $c < 0$ . This cutoff phenomenon [12] occurs in all of the examples of this section. The choice of distance doesn't affect this (though it can affect the place of the cutoff). Almost always, there is an extra  $\log n$  factor present (viz.  $n \log n$  versus  $n$ ). One of the problems of the emerging theory of Markov chain convergence is to explain and understand the existence, size, and shape of these

cutoffs. If the Ehrenfest process starts “balanced” with  $\lfloor n/2 \rfloor$  balls in the left-hand urn, it takes  $cn$  steps to converge and there is no cutoff.

I think of the cutoff phenomenon as a kind of phase transition; small changes in a parameter (here, the mixing time) make for large changes in a response variable (here, the distance to stationarity). Of course, the definition of “phase transition” (or even “phase”) is a contentious matter. For further discussion, see <http://www.aimath.org/pastworkshops/phasetransition.html>.

$l$	1	2	3	4	5	6	7	8	9
$\ K^l - \pi\ $	1.000	1.000	1.000	1.000	.924	.618	.32	.16	.08

**Table 1** Total variation to stationarity.

*Example 2 (Riffle shuffling)* A natural model for the usual method of riffle shuffling  $n$  cards was introduced by Gilbert, Shannon, and Reeds. Starting with the deck in order, cut off about half the deck. (The chance that the cut has  $c$  cards is  $\binom{n}{c}/2^n$ .) Then, the two packets are riffled together according to the following rule: if at some stage there are  $A$  cards in the left packet and  $B$  cards in the right packet, the chance of dropping the next card from the left packet is  $A/(A+B)$  (proportional to packet size). This is continued sequentially until all cards have been dropped. Experiments, reported in [14], show that this is a reasonable model for the way real people shuffle. It is also the maximum entropy method of shuffling; given the cut, all ways of interleaving are equally likely.

Here, the state space is  $\mathcal{X} = \{\text{all } n! \text{ permutations}\}$ . The stationary distribution is uniform ( $\pi(x) = 1/n!$ ). The chance of moving from arrangement  $x$  to  $y$  is  $K(x,y)$  as specified above; there is a formula for this but we do not need it. Joint work with David Bayer [7] gives a sharp analysis of riffle shuffling. When  $n = 52$ , the total variation to stationarity is given in Table 1. The distance stays close to its maximum, cutting down towards zero at about seven shuffles, after which it drops by a factor of two after each successive shuffle. The mathematics which follows shows this continues forever. The numbers in Table 1 are based on an exact computation, not asymptotics. For general-size decks, the following theorem shows that there is a sharp cutoff in convergence to stationarity at  $\frac{3}{2} \log_2 n$ .

**Theorem 2 ([7])** *For a deck of  $n$  cards, after  $l$  riffle shuffles, for any start  $x$ ,*

$$\left\| K_x^l - \pi \right\| = 1 - 2\Phi\left(\frac{-2^{-c}}{4\sqrt{3}}\right) + O\left(\frac{1}{\sqrt{n}}\right) \quad \text{for } l = \frac{3}{2} \log_2 n + cn$$

with  $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$ .

Here  $c$  can be any fixed real number so the theorem determines the shape of the cutoff. When  $n = 52$ ,  $\frac{3}{2} \log_2 n \doteq 8.5$  in agreement with the numbers above. However, the exact computations show that there is a reasonably sharp transition for  $n = 52$ .

There have been many variations of this analysis [13, 15] allowing different distances and methods of shuffling. For example, it is shown that cutting the deck doesn't

appreciably speed up mixing [31]. In the case of the casino card shuffling machine, a similar sharp mathematical model was possible [18] and the design was dropped by the company.

*Example 3 (Glauber dynamics for the Ising model)* Consider an  $n \times n$  periodic lattice and let  $\mathcal{X} = (\pm 1)^{n^2}$  be the configuration space. The stationary distribution is

$$\pi(\sigma) = z^{-1} e^{\beta \sum_{x,y} \sigma(x)\sigma(y) + h \sum_x \sigma(x)}.$$

The inverse critical temperature is  $\beta_c = \frac{1}{2} \log(1 + \sqrt{2})$ . Let  $K(x,y)$  be single-site Glauber dynamics with uniformly chosen random update site.

There is a large, healthy contingent on rates of convergence in the mathematical physics literature. Very roughly, for  $\beta \gg \beta_c$ , exponentially many steps are needed. For  $\beta < \beta_c$ , a polynomial number of steps suffice. A marvelous survey (and careful statements) is in [43]. The following sharp result was recently proved by Lubetzky and Sly [41].

**Theorem 3 ([41])** *For  $0 \leq \beta < \beta_c$  and any  $h$ , Glauber dynamics for the  $d = 2$  Ising model with periodic boundary conditions satisfies, for  $c > 0$ ,*

$$\left\| K_+^l - \pi \right\| \rightarrow 0 \quad \text{for } l = \frac{4n^2}{\lambda_\infty} (\log n + c \log \log n) \quad \text{as } n \rightarrow \infty.$$

Here  $\lambda_\infty$  is an explicit constant (the spectral gap for the dynamics on the infinite lattice). The result is sharp. For  $c < 0$ , the total variation distance tends to one.

This result is important as a first case of a sharp cutoff where there is no underlying group structure which allows an explicit diagonalization of the Markov chain. The arguments are somewhat general. They allow similar conclusions at high temperature for the Potts model, for higher dimensions, for anti-ferromagnetic models, and hard core gas models. The main requirement is a type of monotonicity. Of course, the exact form of the cutoff depends on the model. The proofs work for other local algorithms (e.g., Metropolis). They are accompanied by a host of related theorems:

- At  $\beta = \beta_c$ , Glauber dynamics mixes in polynomial time [42].
- For  $\beta > \beta_c$ , the mixing time is exponential ( $e^n$ ) with periodic boundary conditions. Surprisingly, it is conjectured to be polynomial with all positive boundary conditions (rigorously  $e^{(\log n)^2}$  [40]).

While we have emphasized the cutoff phenomena, this is not the most important point. For a practitioner, there is not a huge difference between a few hundred steps or a thousand steps. The main point is that there are new tools that give useful bounds for realistic problems and algorithms.

The next section gives a brief introduction to some of the methods used and pointers to useful, accessible literature so an outsider can begin to work on problems of interest.

## 2 Tricks and tools

Suppose you are a grown-up mathematician, physicist, or chemist and want to learn more about the subject of rates of convergence. There is a recent, best introductory book by Levin, Peres, and Wilmer [39]. This treats discrete spaces in a serious yet friendly way. There are also myriad surveys and overviews. Three good ones are Saloff-Coste [48] (aimed at the analytical side), Martinelli [43] (mathematical physics), and Montenegro and Tetali [45] (theoretical computer science). I also recommend [16]. All give rigorous results and contain extensive bibliographies. To give a flavor of what is available, this section sketches four sets of tools: spectral theory, functional inequalities, coupling, and Harris recurrence.

### 2.1 Spectral theory

Suppose that the Markov chain  $K(x, y)$  and stationary distribution  $\pi(x)$  satisfy the detailed balance condition  $\pi(x)K(x, y) = \pi(y)K(y, x)$  (probabilists call this reversibility). Let  $L^2(\mathcal{X})$  be the set of all functions  $f: \mathcal{X} \rightarrow \mathbb{R}$  (the real numbers). This is a Hilbert space with inner product  $\langle f, g \rangle = \sum_x f(x)g(x)\pi(x)$ . The transition matrix  $K$  operates on  $L^2(\mathcal{X})$  by  $Kf(x) = \sum_y K(x, y)f(y)$ . Detailed balance implies that  $K$  is self-adjoint  $\langle Kf, g \rangle = \langle f, Kg \rangle$ . Now, the spectral theorem implies that there is an orthonormal set of eigenfunctions  $\psi_x(y)$  with real eigenvalues  $\beta_x$ ; thus  $K\psi_x = \beta_x\psi_x$ . For a Markov chain  $-1 \leq \beta_x \leq 1$ . If for each  $x, y$ ,  $K^l(x, y) > 0$  for some  $l$  then there is a unique eigenvalue  $\beta_{x_0}$ , say with  $\beta_{x_0} = 1$  and  $\psi_{x_0}(y) \equiv 1$ . To bound rates of convergence, use the Cauchy–Schwarz inequality to show

$$\begin{aligned} \|K_x^l - \pi\| &= \frac{1}{4} \left( \sum_y |K^l(x, y) - \pi(y)| \right)^2 \\ &= \frac{1}{4} \left( \sum_y \frac{|K^l(x, y) - \pi(y)|}{\sqrt{\pi(y)}} \sqrt{\pi(y)} \right)^2 \\ &\leq \frac{1}{4} \sum_y \frac{(K^l(x, y) - \pi(y))^2}{\pi(y)} = \frac{1}{4} \sum_{z \neq x_0} \psi_z^2(x) \beta_z^{2l}. \end{aligned} \quad (2.1)$$

The last equality uses the Plancherel theorem.

Now knowledge of the eigenvalues and eigenvectors (and calculus estimates) can be used to bound the right-hand side.

*Example 4* For the Ehrenfest urn, Kac [34] determined the eigenvalues and eigenvectors as

$$\beta_j = \left( 1 - \frac{2j}{n+1} \right), \quad \psi_j(k) = \frac{1}{\binom{n}{j}^{1/2}} \sum_{m=0}^j (-1)^m \binom{k}{m} \binom{n-k}{j-m}$$

(Krawtchouck polynomials). To determine rates of convergence, one must choose  $l$  so that

$$\sum_{j=1}^n \left(1 - \frac{2j}{n+1}\right)^{2l} \psi_j^2(0) \leq \varepsilon.$$

We will not carry out this task here; see [14].

Of course, there is no reason to restrict to finite state spaces. However, even the simplest countable state spaces can lead to continuous spectra. For example, consider  $\mathcal{X} = \{0, 1, 2, \dots\}$ . Let  $\pi(j) = 1/2^{j+1}$ . A reversible Markov chain with  $\pi$  as stationary distribution is: from  $j \neq 0$ , move to  $j-1$   $2/3$  of the time and  $j+1$   $1/3$  of the time. From 0, stay or move to 1 with probability  $1/2$ . This chain has one eigenvalue equal to 1 (with eigenfunction identically equal to 1). All the rest of the spectrum is continuous. See Silver [49] for a detailed description and Anderson [3] for many further examples.

Very natural examples can have “interesting spectra.” As an example, consider the problem of placing  $n$  non-overlapping discs of radius  $\varepsilon$  randomly inside the unit square (in two dimensions). This hard disc model is a basic model of statistical physics. It was the original motivation for the Metropolis algorithm, Glauber dynamics, and molecular dynamics. See [16, 37, 38] for extensive motivation, algorithms, and discussion. Let  $\mathcal{X} \subseteq \mathbb{R}^{2n}$  be the configuration space for the centers of the discs. This is a complicated “cuspy” set described carefully in [22]. One algorithm to simulate from the uniform distribution is this: from a configuration  $x$ , fix a small parameter  $h$ . Pick a disc uniformly at random. Choose a point within  $h$  of this disc’s center uniformly at random. Try to move this center to the chosen point. If this results in an allowable configuration  $y$ , allow this move. If not, stay at  $x$ . This describes a transition measure  $K(x, dy)$  on the configuration space. This is symmetric and ergodic if  $n\varepsilon$  is small. We note that this chain is far from ergodic if only the natural condition  $n\varepsilon^2$  small is imposed; see [22]. The associated Markov chain has a uniform stationary distribution. The operator is self-adjoint and so the spectral theorem is in force. A detailed account of the spectrum appears in [21, 22, 23]. There is discrete, continuous, and embedded spectrum. Indeed, because of the “holding” in the Metropolis algorithm, the operator is a compact perturbation of a multiplier and continuous spectrum is forced. Fortunately, the spectrum near one is discrete. Using techniques of micro-local analysis, the spectrum (and eigenvectors!) are suitably approximable and a reasonably sharp analysis is possible in the limit as  $h \rightarrow 0$ . Here, the analysis is done for  $n$  and  $\varepsilon$  fixed (with  $\varepsilon$  smaller than  $1/n$ ). Determining the dependence when  $n$ ,  $\varepsilon$ , and  $h$  vary together is a natural (and open) problem. In later work [22], spectral techniques are combined with functional inequalities (Nash and Sobolev inequalities) resulting in more general results. These ideas are explained further in Section 2.2.

In summary, spectral analysis of Markov operators is sometimes possible. It has been most successful in problems where a group acts [14]. Recent work, importing tools from micro-local analysis, is promising but there is a long way to go to get to results useful to practitioners. The hard disc results involve “constants” which grow like  $(nd)^{nd}$  where  $d$  is the dimension of the discs (and  $n$  is the number of discs). Even when  $n = 100$  and  $d = 2$  this renders the results impractical.

## 2.2 Functional inequalities

The spectral analysis above uses all of the spectrum. It is natural to ask what can be learned from bounds on just the second eigenvalue  $\beta_1$  or the spectral gap  $1 - \beta_1$ . For definiteness suppose again that  $\mathcal{X}$  is finite and  $K(x, y), \pi(x)$  satisfy detailed balance. Then, using the inequality (2.1), bounding all the eigenvalues above by  $\beta_1$  and the easy fact  $\sum_x f_x^2(y) = 1/\pi(y)$  gives the bound

$$\left\| K_x^l - \pi \right\| \leq \frac{1}{\sqrt{\pi(x)}} \beta_1^l. \quad (2.2)$$

Choosing  $l$  to make the right side suitably small gives a bound on running time. Because of the presence of the factor  $1/\sqrt{\pi(x)}$ , this bound can be useful but “off.” For example, consider the Ehrenfest urn example. Then  $\beta_1 = (1 - \frac{2}{n+1})$ . With  $x = n$ ,  $\pi(x) = 1/2^n$  and  $l$  must be chosen large enough to make  $2^{n/2} (1 - \frac{2}{n+1})^l$  suitably small. This requires  $l$  of order  $n^2$  while the right answer is  $n \log n$ . Note that  $n^2$  isn't all that bad. It's much better than  $2^n$  or “no rate.” More generally, the answers that (2.2) gives are only off by  $\log |\mathcal{X}|$  and thus they can be useful if not perfect.

There is a host of techniques for bounding the spectral gap. These include Cheeger and Poincaré inequalities. Both of these use an energy form (analog of  $\int (\nabla f)^2$ ). For a discrete Markov chain, define the *Dirichlet form* for  $f \in L^2(x)$  as

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

The usual minimax characterization of eigenvalues shows that

$$1 - \beta_1 = \min \frac{\mathcal{E}(f, f)}{\text{var}(f)}$$

where

$$\text{var}(f) = \frac{1}{2} \sum (f(x) - f(y))^2 \pi(x) \pi(y) = \sum (f(x) - \bar{f})^2 \pi(x).$$

Here the minimum is over all non-constant  $f$  in  $L(\mathcal{X})$ . Thus if a constant  $A > 0$  can be found such that for all  $f$  the following Poincaré inequality holds:

$$\text{var} f \leq A \mathcal{E}(f, f). \quad (2.3)$$

Then

$$\beta_1 \leq 1 - \frac{1}{A}.$$

In [27], using an idea of Jerrum and Sinclair, a geometric approach to proving Poincaré inequalities is developed. For this, given  $\mathcal{X}$  and  $K(x, y)$ , associate a graph with vertex set  $\mathcal{X}$ , edge set  $\{(x, y) : K(x, y) > 0\}$ . For each  $x, y \in \mathcal{X}$ , choose a path  $\gamma_{xy}$  in this graph. Say the path is  $x_0 = x_1, x_2, \dots, x_l = y$  with  $K(x_i, x_{i+1}) > 0$ . Write  $|\gamma_{xy}| = l$

for the length of the path. The paths can be used to prove a Poincaré inequality as follows.

$$\begin{aligned} 2\text{var}(f) &= \sum_{x,y} (f(x) - f(y))^2 = \sum_{x,y} \left( \sum_{e \in \gamma_{xy}} (f(e^+) - f(e^-))^2 \pi(x)\pi(y) \right) \\ &\leq \sum_{x,y} |\gamma_{xy}| \sum_e (f(e^+) - f(e^-))^2 \pi(x)\pi(y) \\ &= \sum_e (f(e^+) - f(e^-))^2 \sum_{\gamma_{xy} \ni e} |\gamma_{xy}| \pi(x)\pi(y). \end{aligned}$$

Above,  $e = (e^+, e^-)$  is an edge in the graph. Let  $Q(e) = \pi(e^+)K(e^+, e^-)$ . Multiply the term  $(f(e^+) - f(e^-))^2$  above by  $Q(e)/Q(e)$  and bound the inner sum by

$$\max_e \frac{n}{Q(e)} \sum_{\gamma_{xy} \ni e} |\gamma_{xy}| \pi(x)\pi(y) = A. \quad (2.4)$$

This gives the following.

**Proposition 1 ([27])** *For a reversible Markov chain on a finite state space  $\mathcal{X}$ ,*

$$\beta_1 \leq 1 - 1/A \quad \text{with } A \text{ from (2.4).}$$

This gives meaning to the phrase “the geometry of Markov chains,” indeed, the quantity  $A$  can be usefully bounded in terms of things like the diameter of the graph, max degree, etc. The sum for  $A$  is over all paths containing a given edge  $e$ . This shows that  $A$  is a measure of bottlenecks: if it is possible to choose paths “disjointly” so that not too many use the same edge,  $A$  will be small, and  $\beta_1$  is far from 1. For many examples and much further development, see [27, 39]. Both of these sources also develop the geometric approach to Cheeger inequalities.

A host of extensions and refinements of these ideas leads to further extremely useful inequalities of Nash, Sobolev, and log-Sobolev type. These are harder to establish, but give better bounds on convergence. They are drawn from earlier developments in differential equations and all have useful extensions to continuous and infinite-dimensional settings.

In joint work with Saloff-Coste [24, 25], we worked through these topics in a discrete setting, including local path arguments as a way of establishing Nash inequalities. These papers or the more comprehensive [48] might be a useful place to start. Martinelli [43] develops Nash inequalities to study a variety of spin systems. In joint work with Lebeau and Michel [22], we study the hard disc problem using Nash and the closely related Sobolev inequalities. All of these papers contain extensive references.

The constants in Nash and Sobolev inequalities have bad dependence on dimension. A major advance comes in Gross’ log-Sobolev inequalities. These have the form

$$\mathcal{L}(f) \leq A\mathcal{E}(f|f) \quad \text{for all } f$$

where

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

Proving a log-Sobolev inequality is usually very hard work, but when it is done, the constant  $1/\sqrt{\pi(x)}$  in (2.2) can be replaced by  $\log(1/\pi(x))$ , a marked improvement that often results in the right answer. A splendid account of these developments appears in [4]. As a concrete example, the marvelous results of Lubetzky and Sly [41] described in Example 3 lean on log-Sobolev bounds.

### 2.3 Coupling

This is a probability technique which often gets near-optimal answers by pure thought. To get rates of convergence for a Markov chain  $K(x,y)$  starting at  $x$ , you consider a second Markov chain with the same transition matrix starting in the stationary distribution  $\pi$ . Each chain evolves according to  $K$  so the second chain is always in stationarity. Let  $T$  be the first time that the two chains are at the same state. Because the second chain is stationary and the first chain equals the second chain, at time  $T$  (and thus thereafter) the first chain is stationary. It seems like a magic trick but it works. Quantitative bounds are available via the coupling inequality:

$$\|K_x^l - \pi\| \leq P\{T > l\}. \quad (2.5)$$

Thus a bound on the coupling time gives a rate of convergence. Further, the maximal coupling theorem shows that a coupling exists so that there is equality in (2.5) for all  $l$ . Thus the method is sharp (at least in theory; couplings can be hard to find).

*Example 5* Here is a coupling for the Ehrenfest urn. First, it is useful to represent the process as simple random walk on the set of  $2^n$  binary vectors: the  $i$ th component is 1 or 0 as ball  $i$  is in the left or right urn. This treats the balls as labeled but the symmetry of the process shows that the labeling doesn't matter. Picking a ball at random and switching it to the opposite urn is the same as picking a coordinate at random and changing to its opposite mod 2. With this description, the process evolves as follows: with probability  $1/(n+1)$  do nothing. With probability  $n/(n+1)$  pick a coordinate at random and change to its opposite. Start the process off at the all-zero vector. Start a second process off at a uniformly chosen vector. Let the two processes run independently until they differ in an even number of places. Let  $T_1$  be the first time this happens. Thus  $T_1$  may equal 1 and  $T_1$  happens with high probability after  $cn$  steps for  $c$  large. After time  $T_1$ , the two processes evolve as follows: with probability  $1/(n+1)$  they stay fixed. If not, choose a coordinate at random. If the processes match in this coordinate, change both to the opposite (so the coordinate still matches). If the two processes do not match in the chosen coordinate, change the coordinate in the first process to its opposite, find the next non-matching coordinate (moving cyclically) and change this coordinate in the second process to its opposite. Note that this reduces the number of non-matching coordinates by two *and* from the perspective of each process, the rules of the original chain are followed. Let  $T$  be the first time, following

$T_1$ , that the two processes match in all coordinates. This  $T$  is a coupling time and coupling inequality (2.5) shows that

$$\|K_0^l - \pi\| \leq P\{T > l\}.$$

The classical coupon collectors problem now shows that for  $l = n \log n + cn$ ,  $P\{T > l\} \leq e^{-c}$ . This shows that  $n \log n$  steps suffice for mixing. It is slightly off from  $\frac{1}{4}n \log n$  provided by a complete spectral analysis, but oh, so much easier.

The best way to learn coupling is to see a collection of examples. Both [39] and [1] contain useful chapters on coupling. One very nice example of coupling in a statistical mechanics problem: Kannan, Mahoney, and Montenegro [35] give rates of convergence for an algorithm for the hard disc problem.

Three recent advances: first, path coupling [9] combines the geometry of paths as in Section 2.1 with coupling. The result is that instead of having to couple together two processes in completely different starting states, it is enough to couple two processes that differ by only a single step. This is often much simpler and quite complex; realistic Markov chains have been analyzed using path coupling. References [1, 39] contain details and examples.

A second important development is the coupling from the past algorithm of Propp and Wilson [47] and its variants by Fill [28]. This uses coupling as a practical algorithm to generate perfect samples that are exactly stationary. There is still some art in this but it has been successfully employed to generate exact samples from Ising and Potts models on  $2500 \times 2500$  grids and in dozens of other models (e.g., random tilings). Coupling from the past is easiest for monotone Markov chains where the state space has a partial order that is preserved under steps of the chain. Then, one need only wait until the smallest and largest states couple. In [47] it is shown that the coupling time is (at worst) logarithmically larger than the mixing time. David Wilson's web site contains a comprehensive list of references.

A third development is the skillful use of *non-Markovian* couplings to analyze Markov chains. These allow looking back further into the past in order to decide how the next step should proceed. A splendid example is Vigoda's analysis of the random coloring Markov chain (anti-ferromagnetic Potts model at zero temperature); see Frieze and Vigoda [30]. The idea has been used to give coupling proofs of the random transpositions Markov chain [8] and the Markov chain on the  $n$ -simplex which picks two coordinates at random and references them by two uniformly chosen coordinates with the same sum [50]. Coupling also appears as an ingredient of the next topic.

## 2.4 Harris recurrence

This is a technique for bounding rates of convergence for general state spaces. It does not require detailed balance. Perhaps the easiest approach is to start with a special case, the original method of proving convergence due to Doeblin. Let  $(S, \mathcal{S})$  be a general space with  $K(x, dy)$  a Markov operator with stationary distribution  $\pi(dx)$ .

Say that  $K, \pi$  satisfies a *Doebelin condition* if there is a constant  $c$ ,  $0 < c < 1$ , such that for all  $x$  and  $A$ ,

$$K(x, A) \geq c\pi(A). \quad (2.6)$$

If (2.6) holds, the chain may be represented as  $K(x, A) = c\pi(A) + (1 - c)R(x, A)$  with  $R(x, A) = (K(x, A) - c\pi(A))/(1 - c)$ . This last realization has a simple probabilistic interpretation: from  $x$ , flip a coin with probability of heads  $c$ . If the coin comes up heads, choose from  $\pi$ . If the coin comes up tails, choose from the Markov chain  $R(x, dy)$ . Clearly, the first time that the coin comes up heads, the chain is stationary. The coupling inequality gives

$$\|K_x^l - \pi\| \leq (1 - c)^l.$$

The Doebelin condition can be weakened to  $K^{l_0}(x, A) \geq c\pi(A)$  for some  $l_0$  yielding  $\|K_x^l - \pi\| \leq (1 - c)^{\lfloor l/l_0 \rfloor}$ . These bounds can be useful for Markov chains that take “big steps” such as the hit and run algorithm [2]. However, for local chains such as nearest neighbor walk on a graph,  $l_0$  can be large and the resulting bounds poor. Further, choosing a tradeoff between  $l_0$  and  $c$  requires understanding the rate of convergence. This is what we were trying to understand in the first place.

A refined version of the argument has been developed by Meyn, Tweedie, and Rosenthal. An accessible introduction is in [32]. The idea is to introduce a small set  $C$  and an auxiliary probability  $\sigma$  such that for some  $c, l_0$ ,

$$K^{l_0}(x, A) \geq c\sigma(A) \quad \text{for all } x \in C \text{ and all } A. \quad (2.7)$$

Now, whenever the chain is in  $C$ , it has a chance to couple using (2.7). One now has to bound the number of times that the chain hits  $C$  from its start at  $x$ . There are various ways of combining this information to get effective bounds; see [32]. At present writing, these Harris recurrence techniques are as close to useful, off-the-shelf tools as we have for routine problems. It is only fair to warn the reader that it is still hard work to get useful results in natural examples.

This was brought home to me recently while teaching a graduate course in Markov chains. One of the final projects that I assigned was a quantitative analysis of the following two-component Markov chain: let  $\Theta = [0, 1]$ ,  $\mathcal{X} = \{0, 1, 2, \dots, n\}$ . The chain proceeds from  $(j, \theta)$  as follows: choose  $j'$  from the binomial  $(n, \theta)$  distribution and then  $\theta'$  from a beta distribution with density

$$\frac{\Gamma(n+2)}{\Gamma(j)\Gamma(n-j)} \theta^j (1-\theta)^{n-j'}.$$

This chain has stationary distribution with density  $\binom{n}{j} \theta^j (1-\theta)^{n-j}$ . (Note: This is a probability density on  $\mathcal{X} \times \Theta$ ; sum in  $j$  to get 1 and integrate over  $\theta$  still giving 1.) The chain was introduced in the useful expository article [11]. Suppose  $n = 100$  and the chain starts at  $(0, \frac{1}{2})$ . How many steps are required to be within  $1/100$  of stationary? It is simple enough to simulate this chain, and empirically it seems to settle down after 100 to 200 steps. Three graduate students chose this project and tried to use Harris recurrence techniques to get honest bounds. The best they found is  $l$  of order  $10^{33}$ . More expert users fared no better [20]. Working with Khare and

Saloff-Coste [20], we were able to explicitly diagonalize this chain and prove that a few hundred steps suffice. This depended on a small miracle, and slight changes in the problem make the miracle go away.

## 2.5 Connections

It is natural to ask about connections between the various approaches to proof: if you know the spectrum (eigenvalues and eigenvectors) can you find a coupling [10, 44]? What is the connection between the small-sets Harris recurrence condition and Poincaré (or Nash, or log-Sobolev) inequalities [6], and what are the connections between the various functional inequalities [5]? These are active research problems in this area.

Perhaps the most important question has received less attention: how can the various techniques be combined to give useful rates of convergence in the kinds of Markov chain Monte Carlo problems in active use? An example of this are the use of Harris recurrence bounds to design perfect sampling algorithms [36].

This review doesn't cover all the bases. There are other techniques, for example, strong stationary times [17] or the curvature approach of Ollivier [46], which Joulin and Ollivier [33] use to get useful bounds on the variance. There are many problems other than rates of convergence: for example, hitting and cover times and most importantly, laws of large numbers and central limit theorems for functionals of interest. There are close relations between these last areas and rates of convergence [14, Sect. 4.6]. There are many more modern simulation techniques crying out for a careful analysis [29, 38].

I hope I have given a picture of a lively corner of research largely driven by applications in statistical mechanics but now applied in every area of scientific research and practice.

**Acknowledgements** This is a longer version of a talk given at the 12th Berkeley Statistical Mechanics Meeting, January 14–16, 2011. I thank a constructive, insightful referee for many useful comments which have been incorporated into this revision.

## References

1. Aldous, D., Fill, J.: Reversible Markov chains and random walks on graphs (2002). Monograph
2. Andersen, H.C., Diaconis, P.: Hit and run as a unifying device. *J. Soc. Fr. Stat. & Rev. Stat. Appl.* **148**(4), 5–28 (2007)
3. Anderson, W.J.: Continuous-time Markov chains. Springer Series in Statistics: Probability and its Applications. Springer-Verlag, New York (1991). An applications-oriented approach
4. Ané, C., Blachère, S., Chafaï, D., Fougères, P., Gentil, I., Malrieu, F., Roberto, C., Scheffer, G.: Sur les inégalités de Sobolev logarithmiques, *Panoramas et Synthèses [Panoramas and Syntheses]*, vol. 10. Société Mathématique de France, Paris (2000). With a preface by Dominique Bakry and Michel Ledoux

5. Bakry, D., Barthe, F., Cattiaux, P., Guillin, A.: A simple proof of the Poincaré inequality for a large class of probability measures including the log-concave case. *Electron. Commun. Probab.* **13**, 60–66 (2008)
6. Bakry, D., Cattiaux, P., Guillin, A.: Rate of convergence for ergodic continuous Markov processes: Lyapunov versus Poincaré. *J. Funct. Anal.* **254**(3), 727–759 (2008)
7. Bayer, D., Diaconis, P.: Trailing the dovetail shuffle to its lair. *Ann. Appl. Probab.* **2**(2), 294–313 (1992)
8. Bormashenko, O.: A coupling proof for random transpositions (2011). Preprint, Stanford University Department of Mathematics
9. Bubley, R., Dyer, M.: Path coupling: A technique for proving rapid mixing in Markov chains. In: *Proceedings. 38th Annual Symposium on Foundations of Computer Science (Cat. No.97CB36150)*, pp. 223–31. IEEE Comput. Soc., Miami Beach, FL (1997)
10. Burdzy, K., Kendall, W.S.: Efficient Markovian couplings: Examples and counterexamples. *Ann. Appl. Probab.* **10**(2), 362–409 (2000)
11. Casella, G., George, E.I.: Explaining the Gibbs sampler. *Amer. Statist.* **46**(3), 167–174 (1992)
12. Chen, G.Y., Saloff-Coste, L.: The cutoff phenomenon for ergodic Markov processes. *Electron. J. Probab.* **13**, no. 3, 26–78 (2008)
13. Conger, M.A., Howald, J.: A better way to deal the cards. *Amer. Math. Monthly* **117**(8), 686–700 (2010). DOI 10.4169/000298910X515758
14. Diaconis, P.: *Group representations in probability and statistics*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 11. Institute of Mathematical Statistics, Hayward, CA (1988)
15. Diaconis, P.: *Mathematical developments from the analysis of riffle shuffling*. In: *Groups, Combinatorics & Geometry (Durham, 2001)*, pp. 73–97. World Sci. Publ., River Edge, NJ (2003)
16. Diaconis, P.: The Markov chain Monte Carlo revolution. *Bull. Amer. Math. Soc. (N.S.)* **46**(2), 179–205 (2009). DOI 10.1090/S0273-0979-08-01238-X
17. Diaconis, P., Fill, J.A.: Strong stationary times via a new form of duality. *Ann. Probab.* **18**(4), 1483–1522 (1990)
18. Diaconis, P., Fulman, J., Holmes, S.: *Analysis of casino shelf shuffling machines* (2011). Preprint, Stanford University Department of Statistics
19. Diaconis, P., Graham, R.L., Morrison, J.A.: Asymptotic analysis of a random walk on a hypercube with many dimensions. *Random Structures Algorithms* **1**(1), 51–72 (1990)
20. Diaconis, P., Khare, K., Saloff-Coste, L.: Gibbs sampling, exponential families and orthogonal polynomials. *Statist. Sci.* **23**(2), 151–178 (2008). DOI 10.1214/07-STS252. With comments and a rejoinder by the authors
21. Diaconis, P., Lebeau, G.: Micro-local analysis for the Metropolis algorithm. *Math. Z.* **262**(2), 411–447 (2009). DOI 10.1007/s00209-008-0383-9
22. Diaconis, P., Lebeau, G., Michel, L.: *Geometric analysis for the Metropolis algorithm on Lipschitz domains* (2011). To appear, *Invent. Math.*
23. Diaconis, P., Neuberger, J.W.: Numerical results for the Metropolis algorithm. *Experiment. Math.* **13**(2), 207–213 (2004)

24. Diaconis, P., Saloff-Coste, L.: Logarithmic Sobolev inequalities for finite Markov chains. *Ann. Appl. Probab.* **6**(3), 695–750 (1996)
25. Diaconis, P., Saloff-Coste, L.: Nash inequalities for finite Markov chains. *J. Theoret. Probab.* **9**(2), 459–510 (1996)
26. Diaconis, P., Shahshahani, M.: Time to reach stationarity in the Bernoulli–Laplace diffusion model. *SIAM J. Math. Anal.* **18**(1), 208–218 (1987)
27. Diaconis, P., Stroock, D.: Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Probab.* **1**(1), 36–61 (1991)
28. Fill, J.A.: An interruptible algorithm for perfect sampling via Markov chains. *Ann. Appl. Probab.* **8**(1), 131–162 (1998). DOI 10.1214/aoap/1027961037
29. Frenkel, D., Smit, B.: Understanding Molecular Simulation: From Algorithms to Applications, *Computational Science Series*, vol. 1, second edn. Academic Press, San Diego (2002)
30. Frieze, A., Vigoda, E.: A survey on the use of Markov chains to randomly sample colourings. In: Combinatorics, complexity, and chance, *Oxford Lecture Ser. Math. Appl.*, vol. 34, pp. 53–71. Oxford Univ. Press, Oxford (2007). DOI 10.1093/acprof:oso/9780198571278.003.0004
31. Fulman, J.: Affine shuffles, shuffles with cuts, the Whitehouse module, and patience sorting. *J. Algebra* **231**(2), 614–639 (2000)
32. Jones, G.L., Hobert, J.P.: Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statist. Sci.* **16**(4), 312–334 (2001)
33. Joulin, A., Ollivier, Y.: Curvature, concentration and error estimates for Markov chain Monte Carlo. *Ann. Probab.* **38**(6), 2418–2442 (2010). DOI 10.1214/10-AOP541
34. Kac, M.: Random walk and the theory of Brownian motion. *Amer. Math. Monthly* **54**, 369–391 (1947)
35. Kannan, R., Mahoney, M.W., Montenegro, R.: Rapid mixing of several Markov chains for a hard-core model. In: Algorithms and Computation, *Lecture Notes in Comput. Sci.*, vol. 2906, pp. 663–675. Springer, Berlin (2003)
36. Kendall, W.S.: Geometric ergodicity and perfect simulation. *Electron. Comm. Probab.* **9**, 140–151 (electronic) (2004)
37. Krauth, W.: Statistical Mechanics. Oxford Master Series in Physics. Oxford University Press, Oxford (2006). Algorithms and computations, Oxford Master Series in Statistical Computational, and Theoretical Physics
38. Landau, D.P., Binder, K.: A Guide to Monte Carlo Simulations in Statistical Physics, 2 edn. Cambridge University Press, Cambridge (2005). DOI 10.1017/CBO9780511614460
39. Levin, D.A., Peres, Y., Wilmer, E.L.: Markov Chains and Mixing Times. American Mathematical Society, Providence, RI (2009). With a chapter by James G. Propp and David B. Wilson
40. Lubetzky, E., Martinelli, F., Sly, A., Lucio Toninelli, F.: Quasi-polynomial mixing of the 2D stochastic Ising model with “plus” boundary up to criticality. ArXiv e-prints (2010). URL <http://adsabs.harvard.edu/abs/2010arXiv1012.1271L>
41. Lubetzky, E., Sly, A.: Cutoff for the Ising model on the lattice. ArXiv e-prints (2009). URL <http://adsabs.harvard.edu/abs/2009arXiv0909.4320L>

- 
42. Lubetzky, E., Sly, A.: Critical Ising on the square lattice mixes in polynomial time. ArXiv e-prints (2010). URL <http://adsabs.harvard.edu/abs/2010arXiv1001.1613L>
  43. Martinelli, F.: Relaxation times of Markov chains in statistical mechanics and combinatorial structures. In: Probability on Discrete Structures, *Encyclopaedia Math. Sci.*, vol. 110, pp. 175–262. Springer, Berlin (2004)
  44. Matthews, P.: Strong stationary times and eigenvalues. *J. Appl. Probab.* **29**(1), 228–233 (1992)
  45. Montenegro, R., Tetali, P.: Mathematical aspects of mixing times in Markov chains. *Found. Trends Theor. Comput. Sci.* **1**(3), x+121 (2006)
  46. Ollivier, Y.: Ricci curvature of Markov chains on metric spaces. ArXiv Mathematics e-prints (2007)
  47. Propp, J.G., Wilson, D.B.: Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures Algorithms* **9**(1-2), 223–252 (1996). Proceedings of the Seventh International Conference on Random Structures and Algorithms (Atlanta, GA, 1995)
  48. Saloff-Coste, L.: Lectures on finite Markov chains. In: Lectures on Probability Theory and Statistics (Saint-Flour, 1996), *Lecture Notes in Math.*, vol. 1665, pp. 301–413. Springer, Berlin (1997)
  49. Silver, J.S.: Weighted Poincaré and exhaustive approximation techniques for scaled Metropolis-Hastings algorithms and spectral total variation convergence bounds in infinite commutable Markov chain theory. Ph.D. thesis, Harvard University Department of Mathematics (1996)
  50. Smith, A.: A Gibbs sampler on the  $n$ -simplex (2011). Preprint, Stanford University Department of Mathematics