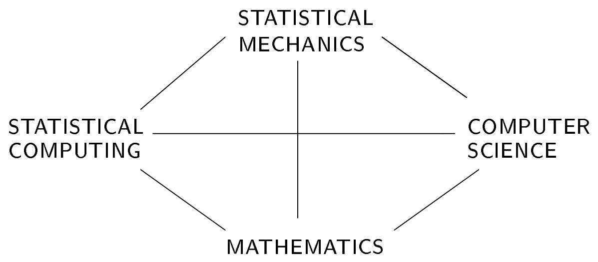
From Shuffling Cards to Walking Around the Building: An Introduction to Modern Markov Chain Theory

Persi Diaconis

Abstract. This paper surveys recent progress in the classical subject of Markov chains. Sharp rates of convergence are available for many chains. Examples include shuffling cards, a variety of simulation procedures used in physics and statistical work, and random walk on the chambers of a building. The techniques used are a combination of tools from geometry, PDE, group theory and probability.

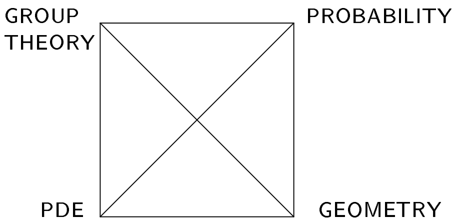
0 Introduction

The classical subject of Markov chains has seen spectacular progress in the past ten years. Progress is seen through theoretical advances and practical applications. These may be roughly depicted as the interactions between:



Briefly, Markov chain Monte Carlo is a mainstay of the computational side of statistical mechanics. There, one wants to draw samples from probability measures on high dimensional state spaces (e.g., the Ising model). One practical way to proceed is to run a fancy kind of random walk (the Metropolis algorithm or Glauber dynamics) [50] which reaches equilibrium at the desired measure. Similar procedures have created a revolution in the computational side of statistics (the Gibbs sampler)[22], [40]. In theoretical computer science, a slew of intractable problems (#-p–complete) problems like computing the permanent of a matrix or the volume of a convex polyhedron) have provably accurate approximations in polynomial time because simple Markov chains can be constructed and proved to converge rapidly [58]. All of this rests on new mathematical developments.

The new mathematics uses ideas from diverse areas.



Probabilistically, new ideas like coupling [48] and stopping time techniques [3], [19] give 'pure thought' solutions to previously intractable problems. Techniques from PDE and spectral geometry allow bounds on the eigenvalues of the basic operators in terms of the geometry of the underlying chain (bottleneck measures, discrete curvatures, and volume growth). Comparison techniques allow study of a chain of interst by comparison with a neat chain which can be analyzed through group representations. The various areas interact so there are probabilistic proofs of results in classical geometry and group theory and vice versa.

The present paper offers a thread through this maze by following the development of a single example: mixing n cards by repeatedly removing the top card and inserting it at random.

The example is studied in Section 1 which introduces basic notation, shows what a theorem in the subject looks like, and proves that $n \log n$ shuffles suffice to mix up n cards. Thus, when $n = 52$, about 200 shuffles are necessary and suffice. The argument introduces coupling arguments and shows that the underlying non self adjoint operators are explicitly diagonalizable. Section 2 offers a variety of extensions where a similar analysis obtains. These include the usual method of shuffling cards. Section 3 extends things to random walk on the chambers of a hyperplane arrangement and then to walks on the chambers of a building. These examples show an intimate connection between probability, algebra, and geometry.

Section 4 gives pointers to many topics not covered, a brief example of the geometric theory of Markov chains (again applied to shuffling cards), some open problems, and a beginner's guide to the literature.

1 Some Markov chains on permutations.

1.1 The Tsetlin library.

Picture a pile of file-folders which are used from time to time. The i^{th} folder is used with weight w_i with $w_i > 0$, $w_1 + \cdots + w_n = 1$. It is natural to want frequently used folders near the top. A scheme which achieves this, even if the w_i are unknown, is simply to replace the most recently used folder on top. To put this into a mathematical framework, label the folders $1, 2, \cdots, n$ and let an arrangement of these labels be denoted by a permutation σ say with $\sigma(i)$ the label at position i. Moving folder i to the top changes σ by a cycle $\big(1, 2 \cdots \sigma^{-1}(i)\big)$. The chance of moving from σ to ζ in one step is

$$K(\sigma, \zeta) = \begin{cases} w_i & \text{if } \zeta = \big(1, 2 \cdots \sigma^{-1}(i)\big)\ \sigma \\ 0 & \text{otherwise.} \end{cases} \tag{1.1}$$

It helps some of us to think of $K(\sigma, \zeta)$ as the (σ, ζ) entry of an $n!$ by $n!$ matrix. Then, making repeated moves is represented by matrix multiplication. Thus, the chance of going from σ to ζ in two steps is

$$K^2(\sigma, \zeta) = \sum_\eta K(\sigma, \eta)\ K(\eta, \zeta)$$

After all, to get from σ to ζ, one must go to some possible η and then from η to ζ. Similarly, $K^l(\sigma, \zeta)$ is defined.

A matrix of form (1.1) with $K(\sigma, \zeta) \geq 0$, $\sum_\zeta K(\sigma, \zeta) = 1$ is called a stochastic matrix and the process of successive arrangements is called a Markov chain. The Peron-Frobenious theorem implies that under mild regularity conditions (connectedness and aperiodicity, satisfied in all examples here), such a Markov chain has a unique stationary distribution $\pi(\sigma) > 0$, $\sum_{\sigma \varepsilon S_n} \pi(\sigma) = 1$. This is characterized as the unique left eigenvector of K with eigenvalue 1 (so $\sum_{\sigma \varepsilon S_n} \pi(\sigma) K(\sigma, \zeta) = \pi(\zeta)$). It is also characterized by the limiting result as l tends to infinity

$$\lim K^l(\sigma, \zeta) = \pi(\zeta) \text{ for all } \sigma \tag{1.2}$$

Algebraically, this says that if the matrix K is raised to a high power, all the rows are approximately equal to π. Probabilistically, this says that for any starting state σ, after many steps, the chance that the chain is in state ζ is approximately equal to $\pi(\zeta)$, no matter what the starting state is.

For the Tsetlin Library (1.1), the stationary distribution π is easy to describe. One description is "sample from the weights $\{w_i\}$ without replacement." That is, form a random permutation σ by choosing $\sigma(1) = j$ with probability w_j. This first choice being made, delete weight $w_{\sigma(1)}$, renormalize the remaining weights to sum to one, and sample from these to determine $\sigma(2)$. Continuing in this way gives σ. Formally:

$$\pi(\sigma) = \frac{w_{\sigma_1}}{1 - w_{\sigma_1}} \frac{w_{\sigma_2}}{1 - w_{\sigma_1} - w_{\sigma_2}} \cdots \frac{w_{\sigma_{n-1}}}{1 - w_{\sigma_1} - \cdots - w_{\sigma_{n-2}}}. \tag{1.3}$$

This natural probability measure arises in dozens of applied contexts from psychophysical experiments (as the Luce model) to oil and gas exploration [17].

The standard way of quantifying the rate of convergence of K^l to π is to use the total variation distance; let $K^l_\sigma(A) = \sum_{\zeta \varepsilon A} K^l(\sigma, \zeta)$,

$$||K^l_\sigma - \pi|| = \max_{A \in S_n} |K^l_\sigma(A) - \pi(A)| = \frac{1}{2} \sum_{\zeta \varepsilon S_n} |K^l(\sigma, \zeta) - \pi(\zeta)|$$

These equalities are easily proved.

As an example of the kind of theorem that emerges, we show

THEOREM 1.1. *For the Tsetlin library chain (1.1)*

$$||K^l_\sigma - \pi|| \leq \sum_{i=1}^{n} (1 - w_i)^l \tag{1.4}$$

Remark 1. Consider the simple case where all $w_i = \frac{1}{n}$. This gives a simple shuffling scheme: Cards are repeatedly removed at random and placed on top. This is the inverse of top to random described before (the rates are the same). The bound on the right side of (1.4) becomes $n(1 - \frac{1}{n})^l$. Using $1 - x \leq e^{-x}$, we see that when $l = n(\log n + c)$ with $c > 0$, $||K^l_\sigma - \pi|| \leq e^{-c}$. It is not hard to see this is sharp: If $l = n(\log n - c)$ with $c > 0$ the distance to stationarity is essentially at its maximum value of 1. A graph of the distance to stationarity versus l appears in Figure 1. The limiting shape of this graph is derived in [19]. This shows an example of the cutoff phenomenon [17]. While the distance $||K^l - \pi||$ is monotone decreasing in l, the transition from one to zero happens in a short interval centered at $n \log n$. In [17] similar cutoffs are proved for many other choices of weights, e.g., $w_i = \frac{c}{(i+1)^s}$.

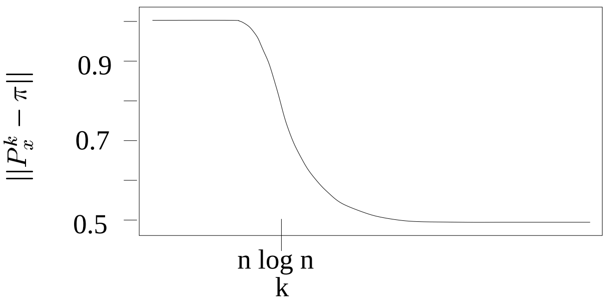


Figure 1: Distance to stationarity for top to random shuffle

Proof of Theorem 1.1 The proof uses a coupling argument. Picture two decks of cards. The first starts in order $1, 2, \cdots, n$. The second starts in random order drawn from the stationary distribution (1.2). At each time $t = 1, 2, 3 \cdots$ choose a label i with probability w_i and move card i to the top of both decks. Note that this forces the two decks to be in the same order in the top positions. Once cards labeled i are in the same position in the two decks, they stay that way under further moves. It follows that the first time T that all indices have been chosen at least once, the two decks are in the same order. The second deck began in stationarity, and repeated moves preserve stationarity. Thus at time T the first deck is distributed in the stationary distribution.

The time T is called a coupling time [48]. It is easy to prove the formal bound

$$||K_\sigma^l - \pi|| \le P\{T > l\}.$$

A further simple argument shows $P\{T > l\} \le \sum(1-w_i)^l$. $\square$

Remark 2. There is a large literature on the move to front scheme as a method of dynamic storage allocation. See [36].

Remark 3. The Markov chain (1.1) is not self-adjoint. Nonetheless, Phaterfod [52] shows that the matrix K has real eigenvalues $\beta_s = \sum_{i \varepsilon s} w_i$ with multiplicity the number of permutations with fixed-point set S. Here S runs over subsets of $[n] = \{1, 2, \cdots, n\}$. It is curious that we have no analytic tools to use these eigenvalues for deriving bounds such as (1.3).

2 MORE VIGOROUS SHUFFLES.

The Tsetlin library scheme can be varied by choosing a subset $S \subset [N]$ with weight w_s and moving the folders with labels in S to the top, keeping them in the same relative order.

THEOREM 2.1. *Suppose the weights w_s separate in the sense that for every i and j, $w_s > 0$ for some s with $i \in S, j \notin S$ or $i \notin S, j \in S$. Then the subset to top chain has a unique stationary distribution π and*

$$||K_\sigma^l - \pi|| \le \sum_{\substack{i \in s, j \notin s \\ j \in s, i \notin s}} (1-w_s)^l$$

Proof. Theorem 2.1 is proved by the following coupling: Let T be the first time every pair of labels, i, j have been separated at least once. This is a coupling time and theorem 1.2 follows. $\square$

Remark 4. Again, all the eigenvalues are real, known, and useless [9].

There is a special case of Theorem (2.1) that is of general interest. Suppose that all the weights $\{w_s\}_s$ are equal to $\frac{1}{2^n}$. The shuffling scheme amounts to choosing a random subset and moving these labels to the top. The inverse process is the Gilbert-Shannon-Reeds (G-S-R) distribution for riffle shuffling ordinary playing cards. Here, one cuts off the top j cards with probability $\binom{n}{j}/2^n$. The top and

l	1	2	3	4	5	6	7	8	9	10
$\|\|K^l - \pi\|\|$	1.000	1.000	1.000	1.000	.924	.614	.334	.167	.085	.043

Table 1: Total variation distance after riffling 52 cards l times

bottom halves are riffled together according to the following scheme: At a given time, if one half has A cards, the second half has B cards, drop the next card from the first half with probability $\frac{A}{A+B}$. It is not hard to see these are inverse descriptions. The G-S-R distribution is quite a realistic description of the way real people shuffle real cards. Of course, in this case $\pi(\sigma) = 1/n!$ is the uniform distribution.

The chance of separating i and j in one shuffle is evidently $1/2$. Thus the bound of theorem (1.2) is

$$||K_\sigma^l - \pi|| \leq \binom{n}{2}(\frac{1}{2})^l$$

The right side of this bound is small when l is larger than $2\log_2 n$.

In joint work with David Bayer [8], more accurate estimates of the distance are derived. We prove the sharp result that $\frac{3}{2}\log_2 n$ shuffles are necessary and suffice:

Theorem 2.2. *For the G-S-R model of riffle shuffles, let* $l = \dfrac{3}{2}\log_2 n + c$. *Then*

$$||K_\sigma^l - \pi|| = 1 - 2\Phi\left(\frac{-2^{-c}}{4\sqrt{3}}\right) + 0\left(\frac{1}{\sqrt{n}}\right).$$

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

When $n = 52$ we derive the following exact result shown in Table 1 above.
Theory shows that the total variation distance continues to decrease by a power of 2 for larger l. Evidently, there is a sharp threshold centered at about seven shuffles. Theorem 2.2 says that for large n, a graph of total variation versus l looks like Figure 1 with a cutoff at $3/2\log_2 n$.
Table 1 is derived from a simple closed form expression: The chance that the deck is in position σ after l shuffles equals $\binom{2^l+n-d}{n}/2^{nl}$ with d the number of descents in σ^{-1}. This close connection between descents and shuffling lends to new formulae in combinatorics—enumeration of permutations by descents and cycle structure [40], [23]. It is also closely connected to Hodge type decompositions for Hochschild homology [42]. This rich circle of interconnections feeds back into probability: While it takes $\frac{3}{2}\log_2 n$ shuffles to make all aspects of a permutation match the uniform distribution, features depending on long cycles are essentially random after one shuffle.

As a final generalization, consider shuffling driven by a block ordered partition $[B_1, B_2, \cdots, B_k]$. To shuffle, remove cards with labels in B_1, and place them on top, keeping them in their same relative order. These are followed by cards with labels in B_2, and so on. Choosing weights for each block ordered partition leads to a shuffling scheme that includes the Tsetlin library (weight w_i on $[i, [n]\backslash i]$) and the G-S-R model (weight $1/2^n$ on $[s, [n]\backslash s]$). These general shuffling schemes were suggested by Bidigare, Hanlon and Rockmore [9]. They permit an essentially complete analysis [12], [13]. As will be seen next, these shuffles too are a very special case of random walk on a hyperplane arrangement.

3 Random walks on the chambers of a hyperplane arrangement.

We work in $\mathbb{R}^d$. Let $A = \{H_1, \cdots, H_k\}$ be a finite collection of affine hyperplanes. These divide space into chambers C and faces F. For example, Fig. 2 shows three hyperplanes in $\mathbb{R}^2$.

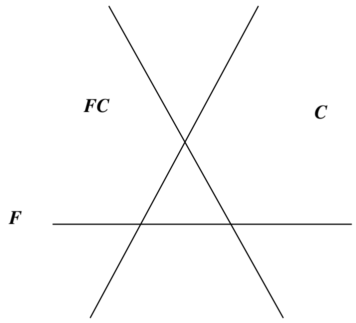


Figure 2: 3 lines in the plane

There are seven chambers, nine half-line faces (one labeled F), and three point faces. There is a natural action of faces on chambers denoted $F * C$: This is the unique chamber adjacent to F and closest to C (in the sense of crossing the fewest number of hyperplanes). For example, in Figure 2 the product of the chamber C, with the face F is the chamber $F * C$. This has distance two from C, while the other chamber adjacent to F is at distance three.

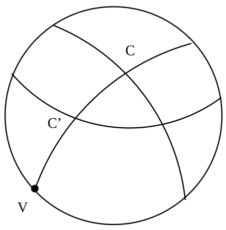
Bidigare, Hanlon, and Rockmore (B-H-R) [9] suggested choosing weights $\{w_F\}$ and defining a random walk defined on C by repeatedly multiplying by faces drawn from these weights. They found the eigenvalues of these chains were positive sums of the weights. Brown and Diaconis [13] showed the chains are diagonalizable, determined the stationary distribution, and gave reasonably sharp coupling bounds for convergence to stationarity.

The B-H-R results extend the shuffling results of Section 2 above: In $\mathbb{R}^n$, the *braid arrangement* has hyperplanes $\{H_{ij}\}_{i<j}$ with $H_{ij} = \{(x_1 \cdots x_n) : x_i = x_j\}$. The chambers of the braid arrangement are naturally labeled by the $n!$ permutations (the relative order of the coordinates inside the chamber). The faces of the braid arrangements are determined by various equalities among coordinates. They are easily seen to be labeled by block ordered partitions discussed in Section 2.

Moreover, the action of faces on chambers is just the shuffling scheme described in Section 2.

There are many hyperplane arrangements where the chambers can be labeled with natural combinatorial objects such as trees or tilings of various regions [13] [59]. The face walks give natural Markov chains on these spaces which permit a complete analysis. There is a useful description of the stationary distribution, the operators are diagonalizable with positive eigenvalues which are partial sums of the weights. Finally, there are good rates of convergence using a coupling argument.

For this expository account, I content myself with a single geometric example drawn from joint work with Louis Billera and Ken Brown [11]. Consider n planes (through zero) in $\mathbb{R}^3$. These are most easily pictured by their intersection with the unit sphere. For example, Figure 2 shows the northern hemisphere cut into chambers or regions by 4 planes—one being the plane of the picture.

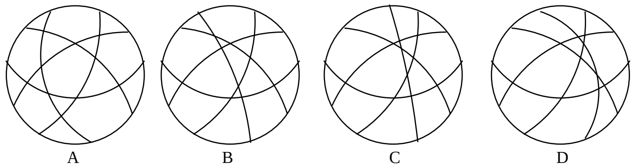


The projection of C on v.

Figure 3: 4 planes in $\mathbb{R}^3$.

The chambers are the open regions shown together with a matching set "under" the sphere. There are 14 chambers altogether ($\binom{n}{2} + 2$ for n planes in general position). Consider the random walk on chambers generated by picking a random vertex of the arrangement uniformly. The walk moves from its current chamber to the chamber adjacent to the chosen vertex. It is intuitively clear that the chance of winding up in a given region c depends on the number $i(c)$ of sides of the region, regions with large values of $i(c)$ being more likely. In [11] we showed that $\pi(c) = (i(c)-2)/2(f_0-2)$ with f_0 the total number of vertices in the arrangement. Thus in Fig. 2 $f_0 = 14$ and the 8 triangular regions have $\pi = 1/14$ while the 6 quadrilaterals have $\pi = 1/7$. We have no intuitive explanation for this; we just observed it was true in small cases and proved it beginning with a rather indirect description of the stationary distribution given by sampling from the vertices *without* replacement. We find the result surprising; for example, there are four essentially different configurations of six planes in $\mathbb{R}^3$. These are shown in Table 2 together with their vital statistics. In all cases, the stationary distribution for an i-gon is proportional to $i-2$. Note that some configurations don't have any i-gons. The eigenvalues and coupling for this example show that the walk reaches stationarity after two steps!

These examples show that hyperplane walks have some remarkable properties. They do not yet explain what makes things tick. The next two sections give



Six great circles in general position.

i	A	B	C	D
3	20	14	12	12
4	0	12	16	18
5	12	6	4	0
6	0	0	0	2

Figure 4: Table 2: Six great circles in general position with number of i-gons

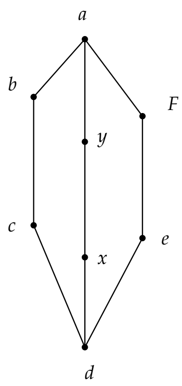
random walks on buildings where things start to break and random walks on semi-groups, the current ultimate generalization.

4 Random walks on the chambers of a building.

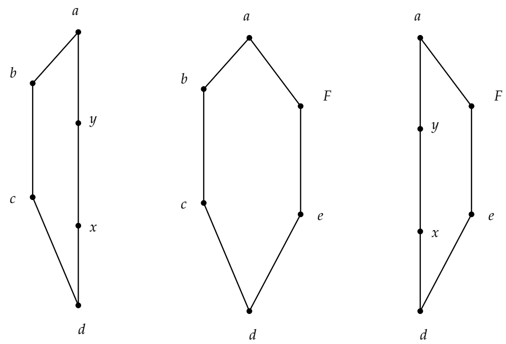
There is a natural extension of the walks of a hyperplane arrangement generated by a reflexion group such as the braid arrangement. This gives random walks on the chambers of a building; we will work with finite objects (spherical buildings). This section reports work of Ken Brown.

A building is a simplicial complex given with a set of subcomplexes called apartments. These apartments must be (isomorphic to) the chambers of a euclidean hyperplane arrangement generated by a finite reflexion group. The top dimensional cells of the complex are called chambers. As an example, the following complex is a building.

Figure 5: An A_2 building



The vertices (one-cells) are numbered a, b, c, d, e, f, x, y. The two-cells are the edges shown; these are the chambers. There are three apartments



Each of these may be identified with the braid arrangement in $\mathbb{R}^3$.

There is a natural action of a face of a building on the chambers. One of the building axioms says that any two faces are in an apartment. Thus any face and chamber are in an apartment and it makes sense to multiply them using the procedure described in Section 3. In the A_2 building pictured in Fig. 5, consider the chamber $\{a, b\}$ and the vertex d; $d * \{a, b\} = \{d, c\}$ because $\{c, d\}$ is the closest chamber to $\{a, b\}$ adjacent to d (distance 2). Any finite tree is a building, and the product of an edge with a vertex may be similarly defined.

A Markov chain on the chambers of a building may be defined by choosing an arbitrary system of weights on the lower dimensional simplicies. This generalizes the shuffling scheme of Section B but does not include general hyperplane arrangements.

For the A_2 building pictured in Figure 5, the walk may be pictured as a service discipline where a single server occupies an edge. Customers arrive at vertices with given propensities and the server slides over to the edge closest to the next customer. One may ask how much time the server spends on a given edge in the long run.

This class of examples introduces some new behavior: It is no longer true that the eigenvalues are positive or even sums of the weights. As an example, consider Figure 5 with equal weights on b, f, x. The eigenvalues are real and the matrix is diagonalizable, but the eigenvalues are algebraic numbers which are no longer linear in the weights. It is an open problem to find examples of random walks on buildings where the eigenvalues are complex. Despite all this, the following example shows that these walks have some elegant special cases where everything works out neatly:

Consider a vector space V which is n-dimensional over a finite field $\mathbb{F}_q$ with $q = p^a$ elements for some prime p. A flag is a maximal increasing sequence of subspaces. Thus it consists of a line in a plane in a three-space and so on up to an $n-1$ space. We will describe a simple random walk on the space of flags which is a direct analog of the random to top chains in Section 1 above. The walk is driven by a system of weights for each line $l : \{w_l\} l \in \mathbb{P}_{n-1}$. Here $w_l \geq 0, \sum\limits_l w_l = 1$. The walk proceeds as follows: Suppose it is currently at the flag $v_1 \subset v_2 \cdots \subset v_{n-1}$. Choose a line with probability w_l. Modify the flag to begin with l:

$$l \subset l + v_1 \subset l + v_2 \cdots \subset l + v_n$$

If $l \notin v_{i-1}$ but $l \in v_i$, the chain of subspaces repeats since $l + v_{i-1} = l + v_i$. Strike out this repetition to get a new maximal flag. This defines a random walk in maximal flags which "moves a random vector to the front." As $q \to 1$; a subspace becomes a subset and a flag becomes a permutation; the walk becomes move to front.

Brown [12] gives an elegant analysis of these chains which perfectly parallels the analysis of Section 3.

Theorem 4.1. *[Brown] For the random line to front with weights $\{w_l\}$, there is an eigenvalue for each subspace x (including $\phi, \mathbb{P}^n$)*

$$\lambda_x = \sum_{l \varepsilon x} w_l .$$

This has multiplicity $m_j(q) = \sum q^{MAJ(\pi)} = [j]! \sum_{k=0}^{j} \frac{(-1)^k q^{\binom{k}{2}}}{[k]!}$, *with* $j =$ *codim*(x), *and the first sum over derangements π in S_j. If w_j is uniform*

$$\|K_x^m - \pi\| \leq \frac{(q^n - 1)(q^{n-1} - 1)}{(q^2 + 1)(q - 1)} \left(\frac{q^{n-2} - 1}{q^n - 1} \right)^m .$$

Remark 5. The last bound shows that $m = n - 1$ steps suffice to achieve randomness when n is large and q is fixed. This is clearly the minimum by dimension arguments so the bound is sharp in this case.

4.1 What is the ultimate generalization?

The results in Sections 1–4 have a marked similarity; it is natural to try to derive a common generalization. In all cases one is "multiplying something" by an associative product (the one case where things went wrong for the A_2 building of Fig. 5, it turns out the product isn't associative). This suggests random walk on a semigroup as a possible general setting. Let $\mathcal{X}$ be a semigroup and w_x a probability on $\mathcal{X}$. Let ρ be an ideal in $\mathcal{X}$ (so $xc \in \rho$ for all $x \in \mathcal{X}, c \varepsilon \rho$). Then generate a random walk by repeatedly choosing elements from $\{w_x\}$ and multiplying. While there is some general theory for these random walks [53] [20], they are too general to hope that results such as real eigenvalues go through. Indeed, any Markov chain on a set S can be represented as a random walk on the set of all functions from S to S.

Ken Brown [12] has shown that results of Sections 1–4 above and many others are captured by semigroups which have all elements idempotent ($x^2 = x$) and further satisfy the cancellation property $xyx = xy$ for all x, y. These are called "left regular bands" in the semigroup business. Brown's proof introduces the semigroup algebra and studies its representations. The irreducible representations of a left regular band turn out to be one-dimensional, and this leads to a complete description of the eigenvalues and multiplicities. The coupling bound had been carried out earlier. [13]

Are these semigroup walks the ultimate generalization? It seems too early to tell. Further, the tools available for hyperplane, building, and semigroup walks are still not sharp enough to prove the cutoff phenomenon as in Theorems 1.1 and 2.2. These seem to need the more refined setting of the descent algebra. There is much yet to understand, but the above developments give a flavor of some exciting mathematics in progress.

5 Topics not covered

The results in Sections 1–5 show the developments of a single theme. There are many other themes that have led to exciting developments. This brief section gives pointers to the literature. It may be supplemented by browsing through the preprint service for Monte Carlo Markov chains: http://www.stats.bris.ac.uk/˜maspb/mcmc. Throughout, $\mathcal{X}$ is a finite set, $K(x, y)$ is a stochastic matrix, and π is the stationary distribution.

5.1 Coupling

Coupling techniques are available in some generality. In *principle* there is a maximal couping which is sharp in the sense the $\|K^l - \pi\| = P\{T > l\}$ for $l = 1, 2, 3, \cdots$. These are usually impossible to find. At present, finding useful couplings is an art. Lindvall [48] is a book-length introduction to coupling. Examples can be found in Aldous [1] and in [16]. Recently, couplings have been used to solve extremely tough problems. Finally, the coupling from the past method of Propp-Wilson [60] has been used to allow exact generation for several distributions of interest. There is a useful bound on the spectral gap given a coupling bound [43], [2]. All of this said, it is still quite difficult to generate useful bounds for many chains of interest using coupling. This is why the geometric theory of Markov chains has been actively developed.

5.2 The geometric theory of Markov chains

Suppose that the underlying chain is reversible: $\pi(x)K(x, y) = \pi(y)K(y, x)$ for all x, y. Form a graph with vertex set $\mathcal{X}$ and an undirected edge from x to y if $K(x, y) > 0$. The geometric theory relates geometric properties of this graph such as diameter, volume growth and various measures of bottlenecks (curvature) to the convergence rates of the chain. This borrows tools from spectral geometry and PDE such as the following inequalities

Poincaré, Cheeger, Nash, Sobolev, Log Sobolev.

Useful introductions to these ideas with many examples may be found in [5], [56], [58]. Along with others, I have written about these things in [35], [27], [29], [30]. Expositions in graph theoretic language appear in [10][15]. The computer science community has also written about such geometric bounds with [49], [46] being surveys with extensive references.

The theory has been adapted to nonreversible chains [38][29]. Here is a simple example which illustrates the geometric tools. Consider $\mathcal{X} = S_n$ all permutations. Let n be odd. Let $S = \{(1,2), (1,2,\cdots,n)\}$: A transposition and an n- cycle. Define $K(\sigma,\eta) = \begin{cases} \frac{1}{2} & \text{if } \eta\sigma^{-1} \in S \\ 0 \end{cases}$. This is the Markov chain described informally as "either transpose the top two cards or move the top card to the bottom." This simple chain should be easy to analyze , but no coupling bound is known. The results show that there are universal constants A_1, A, B such that

$$Ae^{-Bl/n^3 \log n} \leq ||K^l - \pi|| \leq Ae^{-Bl/n^3 \log n}$$

Roughly, this says order $n^3 \log n$ shuffles suffice. When $n = 52$, $n^3 \log n$ is more than half a million. Thus, this shuffling scheme is much slower than "top to random." Here is a brief outline of the argument:

Let $L^2 = \{f : \mathcal{X} \to \mathbb{R}$ with (f, g) $= \sum_x f(x)g(x)\pi(x)\}$. For this example $\pi(x) = \frac{1}{n!}$. Let K operate linearly on L^2 by $Kf(x) = \sum_y K(x,y)f(y)$. If K, π were reversible, K would be self-adjoint. In the present example, K is *not* self-adjoint. We first symmetrize K, forming $\tilde{K} = KK^*$. This is a self-adjoint operator with a simple description: Set $T = \{(1,2)(1\cdots n)^{-1}, (1\cdots n)(1,2)\}$. Then $\tilde{K}(\sigma,\sigma) = \frac{1}{2}, \tilde{K}(\sigma,\zeta) = \frac{1}{4}$ if $\zeta\sigma^{-1}\varepsilon T, \hat{K}(\sigma,\zeta) = 0$ otherwise.

The eigenvalues $\tilde{\beta}_i$ of $\tilde{K}$ can be characterized through the quadratic form $\tilde{E}(f,g) =< (I - \tilde{K})f, g >$. As shown in [29], Section 2, convergence rates for the original chain K can be expressed in terms of $\tilde{\beta}$.

$$||K^l - \pi||^2 \leq \frac{1}{4} \sum_{\beta=1}^{n!-1} \tilde{\beta}^{2l}$$

Here $\tilde{\beta}_0 = 1$ does not appear in the sum.

Finally, one can get good bounds on the eigenvalues $\tilde{\beta}_i$ by comparison with a third chain: random transpositions. $\tilde{\tilde{K}}(\sigma,\zeta) = \dfrac{1}{n}$ if $\sigma = \zeta, 2/n^2$ if $\sigma\zeta^{-1}$ is a transposition and zero otherwise.

For this third chain, a formula for the eigenvalues and their multiplicities is available using character theory [33]. To compare $\tilde{K}$ and $\tilde{\tilde{K}}$ one shows $\tilde{\tilde{E}} \leq An^2\tilde{E}$ for universal A. This in turn is accomplished by writing $(1,2)(1,\cdots,n)^{-1}$ and $(1\cdots n)(1,2)$ in terms of transpositions. Many examples of this sort appear in [25]. Details for the present example can be found in [29], Section 2.

The argument sketched above shows the interactions between probability, geometry, group theory, and PDE. More sophisticated examples appear in [37], [47], [55].

5.3 General state spaces

I have principally been involved with bounding Markov chains on finite state spaces. There has also been much work on general state spaces. At present writing very little of this is quantitative and what is available is often too crude to be useful to practitioners. Meyn and Tweedie [51] is a book-length development of the asymptotic theory, and Rosenthal [54] is a recent example of the quantitative theory with references to related work.

5.4 Some open problems

The present article does not do justice to perhaps the most exciting development; the infinite variety of tricks and techniques that practitioners develop to give believable answers in practical problems. Even the most basic techniques in widespread use—the Metropolis algorithm [32] and the Gibbs sampler are beyond current theoretical understanding. There has been spectacular progress in special cases such as the Ising model (work of Stroock-Zegarlinski, Martinelli, Schoneman, and others). However, the following kind of problem is completely open: On the symmetric group consider $\pi(\sigma) = Z(\theta)\theta^{d(\sigma,\sigma_0)}$. Here $0 < \theta \leq 1$, d is a metric on permutations such as $\sum |\sigma_0(i) - \sigma(i)|, \sigma_0$ is a fixed, known permutation, and Z is a normalization factor. The problem is to generate from π. One simple method: Use the Metropolis algorithm with base chain random transpositions. Analysis of the time to stationarity is beyond theory at present writing. It seems natural to conjecture that order $n \log n$ steps are necessary and suffice to reach stationarity. See [21] for such a result for a special choice of metric.

In a similar vein; trying to make mathematical sense out of any widely used Monte Carlo Markov chain procedure from umbrella sampling to hybrid Monte Carlo offers very challenging mathematics problems.

5.5 More open problems.

Section 5.2 showed how to bound the rate of convergence of a non-reversible chain in terms of the eigenvalues of its multiplicative reversibalization. These in turn were bounded by comparison with a random transpositions chain. Comparison only works for reversible chains. The problem is, find a way of using the explicit eigenvalues of the chains in Sections 2-4 above. Here are three explicit questions. *First*, is there any way of using the eigenvalues to derive explicit bounds on total variation. There are useful bounds for reversible chains [35]. *Second*, can one relate the eigenvalues of a non-reversible chain K to the eigenvalues of its multiplicative reversibilitization? For example, for random to top, the reversibilization becomes random to random. For riffle shuffles, the reversibilization becomes 'remove a random subset and shuffle it back at random.' This is a natural model of traffic where two lanes merge into one and then split into two. *Third*, in the hyperplane setting of Section 3 a local walk can be defined on the chambers. Choose a weight w_i for each hyperplane. From a region C choose one of its bounding hyperplanes with probability proportional to its weight and reflect to the adjacent chamber. This gives a reversible Markov chain with stationary distribution proportional to the sum of the weights of hyperplanes bounding a chamber. Such walks are used to generate random tilings and elsewhere. Is there any way to use the known eigenvalues of the chamber walks of Section 3 to analyze the local walks? These questions go through as stated for buildings.

REFERENCES

[1] Aldous D. (1983) *Random walks on finite groups and rapidly mixing Markov chains.* In Séminaire de Probabilités, XVII, LNM 986, Springer, Berlin.

[2] Aldous D. (1982) *Some inequalities for reversible Markov chains.* Jour. Lond. Math Soc. 25, 564-576.

[3] Aldous D. and Diaconis P. (1987) *Strong uniform times and finite random walks.* Adv. Appl. Math. 8, 69-97.

[4] Aldous D. and Diaconis P. (1986) *Shuffling cards and stopping times.* American Math Monthly, 93, 333-348.

[5] Aldous D. and Fill J. (1996) Preliminary version of a book on finite Markov chains available via homepage http://www.stat.berkeley.edu/users/aldous

[6] Alon N. and Milman V. (1985) λ_1*, isoperimetric inequalities for graphs and superconcentrators.* J. Comb. Th. B, 38, 78-88.

[7] Bakry D. (1994) *L'hypercontractivité et son utilisation en théorie des semigroups.* In Ecole d'été de Saint Flour 1992, Springer LNM 1581.

[8] Bayer D. and Diaconis P. (1992) *Trailing the dovetail shuffle to its lair.* Ann. Appl. Prob. 2, 294-313.

[9] Bidigare, P., Hanlon, P. and Rockmore, D. (1998) *A combinatorial description of the spectrum for the Tsetlin library.* Duke Math Journal. To appear.

[10] Biggs, N. (1993) *Algebraic graph theory*, 2nd ed. Cambridge, Cambridge University Press.

[11] Billera, L., Brown, K., Diaconis, P. (1998) *Geometry and probability in three dimensions.* American Mathematical Monthly. To appear.

[12] Brown, K. (1998) *Random walk on semi-groups.* Unpublished manuscript.

[13] Brown, K. and Diaconis, P. (1998) *Random walks and hyperplane arrangements.* Annals Probab. To appear.

[14] Cheeger J. (1970) *A lower bound for the smallest eigenvalue of the Laplacian.* Problems in Analysis, Synposium in Honor of S. Bochner. Princeton University Press. 195-199.

[15] Chung, F. (1997) *Spectral graph theory.* American Math. Soc. Providence.

[16] Diaconis P. (1986) *Group representations in probability and statistics.* IMS, Hayward.

[17] Diaconis P. (1996) *The cutoff phenomenon in finite Markov chains.* Proc. Natl. Acad. Sci. USA, 93, 1659-1664.

[18] Diaconis P. and Fill J. (1990) *Strong stationary times via a new form of duality.* Ann. Prob. 18, 1483-1522.

[19] Diaconis, P., Fill, J. and Pitman, J. (1992) *Analysis of top to random shuffles.* Combinatorics, Probability and Computing, 1, 135-155.

[20] Diaconis, P. and Freedman, D. (1998) *Iterated random functions.* SIAM Journal. To appear.

[21] Diaconis, P. and Hanlon, P. (1992) *Eigen analysis for some examples of the Metropolis algorithm.* Contemporary Math. 138, 989-117.

[22] Diaconis P. and Holmes S. (1995) *Three Examples of Monte-Carlo Markov Chains: at the Interface between Statistical Computing, Computer Science and Statistical Mechanics.* In Discrete Probability and Algorithms, (Aldous et al, ed.) 43-56. The IMA volumes in Mathematics and its Applications, Vol. 72, Springer-Verlag.

[23] Diaconis P., McGrath M, and Pitman J. (1993) *Riffle shuffles, cycles, and descents* Combinatorica 15, 1, 11-29.

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

[25] Diaconis P. and Saloff-Coste L. (1993) *Comparison techniques for random walk on finite groups.* Ann. Prob. 21, 2131-2156.

[26] Diaconis P. and Saloff-Coste L. (1994) *Moderate growth and random walk on finite groups.* G.A.F.A., 4, 1-36.

[27] Diaconis P. and Saloff-Coste L. (1995) *An application of Harnack inequalities to random walk on nilpotent quotients.*, J. Fourier Anal. Appl., Kahane special issue, 187-207.

[28] Diaconis P. and Saloff-Coste L. (1995) *Random walks on finite groups: a survey of analytical techniques.* In Probability on groups and related structures XI, H. Heyer (ed), World Scientific.

[29] Diaconis P. and Saloff-Coste L. (1996) *Nash inequalities for finite Markov chains.*, J. Th. Prob. 9, 459-510.

[30] Diaconis P. and Saloff-Coste L. (1996) *Logarithmic Sobolev inequalities and finite Markov chains.* Ann. Appl. Prob. 6, 695-750.

[31] Diaconis P. and Saloff-Coste L. (1995) *Walks on generating sets of Abelian groups.* Prob. Th. Rel. Fields. 105, 393-421.

[32] Diaconis P. and Saloff-Coste L. (1995) *What do we know about the Metropolis algorithm.* Jour. Comp. Sys. Sci. To appear.

[33] Diaconis D. and Shahshahani M. (1981) *Generating a random permutation with random transpositions.* Z. Wahrsch. Verw. Geb., 57, 159-179.

[34] Diaconis P. and Shahshahani M. (1987) *Time to reach stationarity in the Bernoulli–Laplace diffusion model.* SIAM Jour. Math. Anal., 18, 208-218.

[35] Diaconis P. and Stroock D. (1991) *Geometric bounds for eigenvalues for Markov chains.* Ann. Appl. Prob. 1, 36-61.

[36] Dobrow, R. and Fill, J. (1995) *The move to the front rule for self-organizing lists. With Markov dependent records.* In Discrete Probability and Algorithms, Springer, New York, pp. 57-80.

[37] Dyer M. and Frieze A. (1991) *Computing the volume of convex bodies: a case where randomness provably helps.* Probabilistic Combinatorics and its applications, Proceedings of the AMS Symposia in Applied Mathematics 44, 123-170.

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

[39] Frieze A., Kannan R. and Polson N. (1994) *Sampling from log concave distributions.* Ann. Appl. Prob. 4, 812-837.

[40] Gessel, I. and Reutenruer, C. (1993) *Counting permutations with given cycle structure and descent set.* Journal Combin. Th. A. 64, 184-215.

[41] Gross L. (1993) *Logarithmic Sobolev inequalities and contractivity properties of semigroups.* In Lecture Notes in Math. 1563. Springer.

[42] Hanlon, P. (1990). *The action of S_n on the components of the Hodge decompositions of Hochschild homology.* Mich. J. Math, 37, 105-124.

[43] Holley, R. (1985) *Rapid convergence to equilibrium in one-dimensional stochastic Ising models.* Ann. Prob. 13

[44] Jerrum M. and Sinclair A. (1993)*Polynomial time approximation algorithms for the Ising model,* SIAM Journal of Computing, 22, 1087-1116.

[45] Jerrum M. and Sinclair A. (1997) *The Markov chain Monte Carlo method: an approach to approximate counting and integration.* In Approximation algorithms for NP-hard problems, D.S. Hochbaum (Ed.), PWS Publishing, Boston.

[46] Kannan R. (1994) *Markov chains and polynomial time algorithms.* Proceedings of the 35th IEEE Symposium on Foundations of Computer Science, Computer Society Press, 656-671.

[47] Lawler G. and Sokal A. (1988) *Bounds on the L^2 spectrum for Markov chains and Markov processes: a generalization of Cheeger inequality.* Trans. AMS, 309, 557-580.

[48] Lindvall, T. (1992) *Lectures on the coupling method.* Wiley, N.Y.

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

[50] Metropolis N., Rosenbluth A., Rosenbluth M., Teller A. and Teller E. (1953) *Equations of state calculations by fast computing machines.* J. Chem. Phys. 21, 1087-1092.

[51] Meyn, S. and Tweedie, R. (1993) *Markov chains and stochastic stability.* Springer, New York.

[52] Phatarfod, R. (1991) *On the matrix occurring in a linear search problem.* J. Appl. Probab. 28, 336-346.

[53] Rosenblatt, M. (1986) *Random walk on semigroups* Contemporary Math, 50, 115-120.

[54] Rosenthal, J. (1995) *Minorization conditions and convergence rates for Markov chain Monte Carlo.* Jour. American Statis. Assoc., 90, 558-566.

[55] Saloff-Coste L. (1996) *Simple examples of the use Nash inequalities for finite Markov chains.* In Semstat III, Current Trends in Stochastic Geometry and its Applications. W.S. Kendall, O.E. Barndorff-Nielsen and MC. van Lieshout, Eds. Chapman & Hall.

[56] Saloff-Coste L. (1997) *Lectures on finite Markov chains.* In Springer Lecture Notes in Mathematics 1665, 301-408.

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

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

[59] Stanley, R. (1996) *Hyperplane arrangements, interval orders and trees.* Proc. Nat. Acad. Sci., USA, 93.

[60] Wilson, D. and Propp, J. (1996) *How to get an exact sample from a generic Markov chain and sample a random spanning tree from a directed graph.* Proc. Seventh Annual ACM-SIAM Symposium on Discrete Algorithms. ACM, NY, 448-457.

Persi Diaconis
Departments of Mathematics
and Statistics
Stanford University
Stanford, California 94305
USA