SIAM REVIEW
Vol. 37, No. 3, pp. 387–405, September 1995

002

CONVERGENCE RATES FOR MARKOV CHAINS*

JEFFREY S. ROSENTHAL†

Abstract. This is an expository paper that presents various ideas related to nonasymptotic rates of convergence for Markov chains. Such rates are of great importance for stochastic algorithms that are widely used in statistics and in computer science. They also have applications to analysis of card shuffling and other areas.

In this paper, we attempt to describe various mathematical techniques that have been used to bound such rates of convergence. In particular, we describe eigenvalue analysis, random walks on groups, coupling, and minorization conditions. Connections are made to modern areas of research wherever possible. Elements of linear algebra, probability theory, group theory, and measure theory are used, but efforts are made to keep the presentation elementary and accessible.

Key words. Markov chain, eigenvalue, coupling, random walk on group

AMS subject classifications. 60J10, 60B15

1. Introduction and motivation.

Imagine 1000 lilypads arranged in a circle, numbered 0 through 999. Suppose a frog begins on lilypad number 0 and proceeds as follows. Each minute, she jumps either to the pad immediately to her right, or to the pad immediately to her left, or to the pad she's already on, each with probability 1/3. Thus, after one minute she is equally likely to be at pad 999, pad 0, or pad 1. After two minutes, she has probability 1/9 of being at pad 998 or pad 2, probability 2/9 of being at pad 999 or pad 1, and probability $3/9 = 1/3$ of being at pad 0.

It is intuitively clear that if we wait for a very large number of minutes, then our frog will have approximately equal probability of being at any of the 1000 pads. But how might we prove this assertion? More importantly, how long do we have to wait until this approximate equality of probabilities occurs? Is 1000 minutes enough? How about 10,000 minutes?

These questions are closely connected to an exciting area of modern mathematical research, the study of quantitative convergence rates for Markov chains. This research has many important applications. Perhaps the most important of these is to Markov chain Monte Carlo algorithms, where a Markov chain is defined in such a way that it will (hopefully) converge to a certain probability distribution of interest. Knowledge of the time required until satisfactory convergence takes place is crucial to the proper implementation of the algorithm. However, such knowledge is often very difficult to obtain in a rigorous manner.

Examples of such algorithms in applied settings include the Gibbs sampler in statistics (Gelfand and Smith (1990), Smith and Roberts (1993), Tierney (1994)), approximation algorithms in computer science (e.g., Jerrum and Sinclair (1989)), and various stochastic algorithms used in physics (see the review article Sokal (1989)). This is a very active applied area. However, much of the work suffers from lack of results about convergence rates of the algorithms being used.

Convergence rates for Markov chains also have applications to card shuffling, in which the arrangements of the cards have various probabilities of occurring at each step. Analysis of the underlying Markov chain (in this case a random walk on the symmetric group) gives information about how many of a given type of shuffle are required to make the distribution of the card arrangement be approximately uniform. The most famous result of this type is the result of Bayer and Diaconis (1992) that seven ordinary "riffle" shuffles are required to properly mix a deck of 52 cards. For additional background see Diaconis (1988).

*Received by the editors April 27, 1994; accepted for publication (in revised form) January 6, 1995.

†Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 1A1 (jeff@utstat.toronto.edu).

Finally, Markov chains (in particular random walks on groups) have been proposed as a method for generating random matrices to be used for encryption algorithms; see, for example, Sloane (1983).

Convergence rates of Markov chains are thus a very important topic in a variety of applied settings. It is the goal of this paper to present, in an accessible and straightforward way motivated by simple examples, some of the methods that have been used to obtain such rates.

It is our hope that this paper will stimulate further research. Indeed, there are many opportunities for further work in this area. One way to proceed is to examine some of the applied algorithms that use Markov chains (see, for example, Smith and Roberts (1993)), and attempt to apply the methods presented here (or develop new methods) to get bounds on their convergence rates. Rigorous bounds are usually not known for these applied algorithms, and any results of this kind are of great interest. The tighter and more general the bounds are, the more use they will be to applied researchers.

In this paper, we present some of the basic results about convergence rates for finite (and infinite) Markov chains. We attempt to make connections to modern research, but at the same time to keep the presentation elementary and accessible. After the preliminary material, we present the basic connection between Markov chains and eigenvalues (§4). We then explore the subject of random walks on groups (§5), for which tremendous progress has been made, which includes our frog's travels and also includes most models of card shuffling. Finally, we discuss coupling and minorization conditions (§6), which are robust techniques that have been used to study various stochastic algorithms in a variety of settings.

Along the way, we prove that we would have to wait over 120,000 minutes (over two months!) for our frog to have approximately equal probability of being at any of the 1000 pads.

None of the results presented here are new. Connections and references to the relevant literature are given where possible.

2. Basic definitions. Our frog process above is an example of a (discrete-time) *Markov chain*. In general, a Markov chain consists of a (measurable) state space $\mathcal{X}$, an initial distribution (i.e., probability measure) μ_0 on $\mathcal{X}$, and *transition probabilities* $P(x, dy)$ which give, for each point $x \in \mathcal{X}$, a distribution $P(x, \cdot)$ on $\mathcal{X}$ (which represents the probabilities of where the Markov chain will go one step after being at the point x). It is assumed that $f_A(x) = P(x, A)$ is a measurable function of $x \in \mathcal{X}$, for each fixed measurable subset $A \subseteq \mathcal{X}$. If $\mathcal{X}$ is a discrete space (e.g., a finite space), then the initial distribution can be specified by the vector of nonnegative real numbers $\mu_0(x)$ for $x \in \mathcal{X}$, where $\sum_x \mu_0(x) = 1$. Similarly, the transition probabilities can be specified by the matrix of nonnegative real numbers $P(x, y)$ for $x, y \in \mathcal{X}$, where $\sum_y P(x, y) = 1$ for each $x \in \mathcal{X}$.

In our frog example above, $\mathcal{X}$ consists of the integers $0, 1, 2, \ldots, 999$. Since the frog starts at the point 0 with probability 1, the initial distribution is specified by $\mu_0(0) = 1$, and $\mu_0(x) = 0$ for $x \neq 0$. Finally, the transition probabilities are specified by $P(x, y) = 1/3$ if $x = y$ or x and y are adjacent in the circle, and $P(x, y) = 0$ otherwise.

Given the initial distribution μ_0 and transition probabilities $P(x, dy)$, we can inductively define distributions μ_k on $\mathcal{X}$, representing the probabilities of where the Markov chain will be after k steps, by

$$\mu_k(A) = \int_{\mathcal{X}} P(x, A)\, \mu_{k-1}(dx), \qquad k = 1, 2, 3, \ldots.$$

On a discrete space, this can be written more directly as

$$\mu_k(y) = \sum_x P(x, y)\mu_{k-1}(x).$$

If we write μ_k as a row-vector, and P as a matrix with $[P]_{xy} = P(x, y)$, then this can be written even more directly as

$$\mu_k = \mu_{k-1}P = \cdots = \mu_0 P^k.$$

There is nothing mysterious about these formulae. They simply say that to be at the point y at time k, we must have been at *some* point x at time $k-1$ (with probability $\mu_{k-1}(x)$), and then jumped from x to y on the next step (with probability $P(x, y)$).

Thus, in our frog example, we would have that $\mu_2(998) = \mu_2(2) = 1/9$, $\mu_2(999) = \mu_2(1) = 2/9$, and $\mu_2(0) = 1/3$.

Once we have defined μ_k for all nonnegative integers k, we can ask about convergence properties. To be quantitative, we define the *total variation distance* between probability measures ν_1 and ν_2 by $\|\nu_1 - \nu_2\| := \sup_{A\subseteq\mathcal{X}} |\nu_1(A) - \nu_2(A)|$ (where the supremum is taken over measurable subsets A). For later reference, we mention two easily verified facts. First, if $\mathcal{X}$ is finite, then $\|\nu_1 - \nu_2\| = \frac{1}{2}\sum_x |\nu_1(x) - \nu_2(x)|$. Second, for any $\mathcal{X}$, we have

$$\|\nu_1 - \nu_2\| = \frac{1}{2} \sup_{\substack{f:\mathcal{X}\to\mathbf{C} \\ |f(x)|\le 1}} |E_{\nu_1}(f) - E_{\nu_2}(f)| = \sup_{\substack{f:\mathcal{X}\to\mathbf{R} \\ 0\le f(x)\le 1}} |E_{\nu_1}(f) - E_{\nu_2}(f)|,$$

where E stands for expected value. We may now state our fundamental questions.

A. Does there exist a probability distribution π on $\mathcal{X}$ such that $\|\mu_k - \pi\| \to 0$ as $k \to \infty$?

B. If so, then given $\epsilon > 0$, how large should k be to ensure that $\|\mu_k - \pi\| \le \epsilon$?

It is these questions which are the focus of this paper.

3. The simplest nontrivial example. To get a sense of what convergence properties a Markov chain can have, we consider what might be called the "simplest nontrivial example." We consider the state space $\mathcal{X} = \{0, 1\}$ consisting of just two points. Setting $p = P(0, 1)$, and $q = P(1, 0)$, we may write P in matrix form as

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

(We leave p and q as arbitrary numbers between 0 and 1.) We further suppose that the initial distribution is given by $\mu_0(0) = 1$, $\mu_0(1) = 0$, meaning that we start in state 0 with probability 1.

This example is simple enough that we can solve for μ_k explicitly. It is verified by induction (see Hoel, Port, and Stone (1972, §1.2)) that (assuming $p + q > 0$)

$$\mu_k(0) = \frac{q}{p+q} + \left(1 - \frac{q}{p+q}\right)(1-p-q)^k.$$

It immediately follows that

$$\mu_k(1) = 1 - \mu_k(0) = \frac{p}{p+q} - \left(1 - \frac{q}{p+q}\right)(1-p-q)^k.$$

(Naturally, if $p = q = 0$, then $\mu_k(0) = 1$ and $\mu_k(1) = 0$ for all k.)

We wish to make a number of observations about this example, since they will generalize considerably. First, note that assuming $|1 - p - q| < 1$, we will indeed have convergence. Setting $\pi(0) = \frac{q}{p+q}$ and $\pi(1) = \frac{p}{p+q}$, we have that

$$\|\mu_k - \pi\| = \left(1 - \frac{q}{p+q}\right)|1-p-q|^k,$$

which decreases exponentially quickly to 0, with rate governed by the quantity $1 - p - q$.

Second, note that this limiting distribution π is a *stationary distribution* in the sense that $\pi P = \pi$, and thus corresponds to a left eigenvector of the matrix P with eigenvalue 1. It is easily seen (by taking the limit $k \to \infty$ in the equation $\mu_k = \mu_{k-1} P$) that *any* limiting distribution π for any Markov chain must be stationary in this sense.

Third, note that the only time this convergence *fails* to take place is if $p = q = 0$ or $p = q = 1$. If $p = q = 0$ the Markov chain is *decomposable*, meaning that the state space $\mathcal{X}$ contains two nonempty disjoint subsets $\mathcal{X}_1$ and $\mathcal{X}_2$ which are closed, i.e., such that $P(x, \mathcal{X}_1) = 1$ for all $x \in \mathcal{X}_1$ and $P(y, \mathcal{X}_2) = 1$ for all $y \in \mathcal{X}_2$. If $p = q = 1$ the Markov chain is *periodic*, meaning that the state space $\mathcal{X}$ contains disjoint nonempty subsets $\mathcal{X}_1, \ldots, \mathcal{X}_d$ (with $d \geq 2$) such that for $x \in \mathcal{X}_j$, $P(x, \mathcal{X}_{j+1}) = 1$ (where if $j = d$, then $j + 1$ is taken to mean 1). The quantity d is the *period* of the Markov chain; in this example $d = 2$. However, if our Markov chain is *indecomposable* and *aperiodic*, then it converges exponentially quickly. We see in the next section that all finite Markov chains follow this rule.

Fourth, it is easily computed that the *eigenvalues* of the matrix P are 1 and $1 - p - q$. The eigenvalue 1, of course, corresponds to the eigenvector π. This computation suggests that the "nontrivial" eigenvalue $1 - p - q$ is intimately connected with convergence of the chain. We shall develop this connection in the next section.

Fifth, and perhaps even more intriguing, we compute that the quantity β defined by

$$\beta := \sum_y \min_x P(x, y)$$

satisfies $\beta = \min(p + q,\ 2 - p - q)$, so that $1 - \beta = |1 - p - q|$ is the absolute value of the nontrivial eigenvalue as above. This suggests that the convergence of the chain might be related to the quantity $1 - \beta$, with β defined as above; this relationship is explored in §6 via the method of "coupling."

Sixth, we compute that $\pi(0)P(0, 1) = \pi(1)P(1, 0)$. This implies that this chain is "reversible," meaning that when started from the stationary distribution π, then for any $x, y \in \mathcal{X}$, the amount of probability $\pi(x)P(x, y)$ that moves from x to y, is the same as the amount of probability $\pi(y)P(y, x)$ that moves from y to x. Among other things, this guarantees that its eigenvalues will all be real. However, not all Markov chains have this property. This issue is discussed briefly in §7.

Finally, we consider the even more specialized case in which $p = q$. This corresponds to a *random walk* on the *group* $\mathbf{Z}/(2)$ of integers modulo 2, because we step in the "same manner" no matter where on $\mathcal{X}$ we are. Here the "step distribution" is given by $Q(0) = 1 - p$, $Q(1) = p$ (corresponding to the group element that we will add (modulo 2), at each step, to our present position). We compute that $E_Q((-1)^x) = 1 - p - q$, the eigenvalue of the matrix P. (Here $(-1)^x$ equals 1 when $x = 0$, and equals -1 when $x = 1$.) This suggests that for random walks on groups, the eigenvalues can be computed simply by taking certain expected values with respect to the step distribution $Q(\cdot)$. This is discussed further in §5.

4. The eigenvalue connection. We let $\mathcal{X} = \{0, 1, \ldots, n - 1\}$ be a finite state space, μ_0 an initial distribution on $\mathcal{X}$, and P be a matrix of transition probabilities on $\mathcal{X}$. The fact that $\mu_k = \mu_0 P^k$ suggests that we need to control high powers of the transition matrix P. This in turn suggests that the eigenvalues of P will play an important role. We develop this idea here, drawing heavily on work of Diaconis and Shashahani (1981), Diaconis (1988), and Belsley (1993). For additional background, see Feller (1968, Chap. XVI) and Issacson and Madsen (1976).

In studying these eigenvalues, we make use of the fact that P has the same eigenvalues whether it operates on vectors from the right side or the left side. We begin with the following fact.

FACT 1. *Any stochastic matrix P has an eigenvalue equal to* 1.

Proof. Define the vector u by $u(x) = 1$ for all $x \in \mathcal{X}$, then it is easily verified that $Pu = u$. □

We now write the (generalized) eigenvalues of P (counted with algebraic multiplicity) as $\lambda_0, \lambda_1, \ldots, \lambda_{n-1}$. Without loss of generality we take $\lambda_0 = 1$. We further set $\lambda_* = \max_{1 \le j \le n-1} |\lambda_j|$, the largest absolute value of the nontrivial eigenvalues of P.

FACT 2. *We have $\lambda_* \le 1$. Furthermore, if $P(x, y) > 0$ for all $x, y \in \mathcal{X}$, then $\lambda_* < 1$.*

Proof. Suppose $Pv = \lambda v$. Choose an index x so that $|v(x)| \ge |v(y)|$ for all $y \in \mathcal{X}$. Then

$$|\lambda v(x)| = |(Pv)_x| = |\sum_y P(x, y)v(y)| \le \sum_y |v(y)|P(x, y) \le \sum_y |v(x)|P(x, y) = |v(x)|,$$

so that $|\lambda| \le 1$. Hence $\lambda_* \le 1$.

Now suppose $P(x, y) > 0$ for all x and y. It is then easily seen that the inequality above can only be equality if v is a constant vector, i.e., $v(0) = v(1) = \cdots = v(n-1)$. This shows that $\lambda_0 = 1$ is the only eigenvalue of absolute value 1 in this case. Hence if P is diagonalizable we are done.

If P is not diagonalizable, then we still need to prove that the eigenvalue $\lambda_0 = 1$ is not part of a larger Jordan block. If it were, then for some vector v we would have $Pv = v + u$, where $u = (1, 1, \ldots, 1)^t$. But then, choosing $x \in \mathcal{X}$ with $\Re e\, v(x) \ge \Re e\, v(y)$ for all $y \in \mathcal{X}$, we have that

$$1 + \Re e\, v(x) = \Re e\, (Pv)_x = \Re e \sum_y P(x, y)v(y) \le \Re e \sum_y P(x, y)v(x) = \Re e\, v(x),$$

a contradiction. □

The importance of eigenvalues for convergence properties comes from the following.

FACT 3. *Suppose P satisfies $\lambda_* < 1$. Then, there is a unique stationary distribution π on $\mathcal{X}$ and, given an initial distribution μ_0 and point $x \in \mathcal{X}$, there is a constant $C_x > 0$ such that*

$$|\mu_k(x) - \pi(x)| \le C_x k^{J-1} (\lambda_*)^{k-J+1},$$

where J is the size of the largest Jordan block of P. (It follows immediately that $\|\mu_k - \pi\| \le Ck^{J-1}(\lambda_*)^{k-J+1}$, where $C = \frac{1}{2}\sum C_x$.) In particular, if P is diagonalizable (so that $J = 1$), then

$$|\mu_k(x) - \pi(x)| \le \sum_{m=1}^{n-1} |a_m v_m(x)| |\lambda_m|^k \le \left(\sum_{m=1}^{n-1} |a_m v_m(x)| \right) (\lambda_*)^k,$$

where $v_0, \ldots, v_{n-1}$ are a basis of right eigenvectors corresponding to $\lambda_0, \ldots, \lambda_{n-1}$, respectively, and where a_m are the (unique) complex coefficients satisfying

$$\mu_0 = a_0 v_0 + a_1 v_1 + \cdots + a_{n-1} v_{n-1}.$$

If the eigenvectors v_j are orthonormal in $L^2(\pi)$, i.e., if $\sum_x v_i(x)\overline{v_j(x)}\pi(x) = \delta_{ij}$, then we get the further bound

$$\sum_x |\mu_k(x) - \pi(x)|^2 \pi(x) = \sum_{m=1}^{n-1} |a_m|^2 |\lambda_m|^{2k} \le \left(\sum_{m=1}^{n-1} |a_m|^2 \right) (\lambda_*)^k.$$

Proof. We begin by assuming that P is diagonalizable. Then, using that $\mu_k = \mu_0 P^k$, that $v_m P = \lambda_m v_m$, and that $\lambda_0 = 1$, we have that

$$\mu_k = a_0 v_0 + a_1 v_1 (\lambda_1)^k + \cdots + a_{n-1} v_{n-1} (\lambda_{n-1})^k.$$

Since $\lambda_* < 1$, we have $(\lambda_m)^k \to 0$ as $k \to \infty$ for $1 \le m \le n-1$, so that $\mu_k \to a_0 v_0$. It follows that $\pi = a_0 v_0$ must be a probability distribution. Hence, in particular, $a_0 = (\sum_y v_0(y))^{-1}$ so it does not depend on the choice of μ_0. Thus,

$$\mu_k(x) - \pi(x) = a_1 v_1(x)(\lambda_1)^k + \cdots + a_{n-1} v_{n-1}(x)(\lambda_{n-1})^k.$$

The stated bound on $|\mu_k(x) - \pi(x)|$ now follows from the triangle inequality. The expression for the $L^2(\pi)$ norm of $\mu_k - \pi$ follows immediately from orthonormality.

For nondiagonalizable P, we must allow some of the vectors v_m to be *generalized eigenvectors* in the sense that we may have $v_m P = \lambda_m v_m + \lambda_{m+1}$. The only difference from the previous argument is that now μ_k may contain some additional terms. If $v_j, v_{j+1}, \ldots, v_{j+\ell-1}$ form a Jordan block of size ℓ, corresponding to the value λ_m, then we may have to add to μ_k extra terms of the form $a_r v_s (\lambda_m)^{k_0}$, with $j \le r < s \le j + \ell - 1$ and $k_0 \ge k - \ell + 1$. Keeping track of these extra terms, and bounding their number by k^{J-1}, the stated conclusion follows. □

We illustrate these ideas with a concrete example.

Example. Consider the Markov chain on the state space $\mathcal{X} = \{1, 2, 3, 4\}$, with transition probabilities

$$P = \begin{pmatrix} 0.4 & 0.2 & 0.3 & 0.1 \\ 0.4 & 0.4 & 0.2 & 0 \\ 0.6 & 0.2 & 0.1 & 0.1 \\ 0.7 & 0.1 & 0 & 0.2 \end{pmatrix}.$$

Suppose the Markov chain starts in the state 1, so that $\mu_0 = (1, 0, 0, 0)$.

We compute numerically that the matrix P has eigenvalues $\lambda_0 = 1$, $\lambda_1 = 0.2618$, $\lambda_2 = 0.0382$, $\lambda_3 = -0.2$, with corresponding left eigenvectors

$$v_0 = (0.4671, 0.2394, 0.2089, 0.0846),$$
$$v_1 = (-0.4263, 0, 0.4263, 0),$$
$$v_2 = (-0.0369, 0.2301, -0.5656, 0.3724),$$
$$v_3 = (-0.2752, 0.4854, 0.0898, -0.3).$$

In terms of these eigenvectors, the initial state $\mu_0 = (1, 0, 0, 0)$ can be written as

$$\mu_0 = v_0 - 1.031\, v_1 - 0.4518\, v_2 - 0.2791\, v_3.$$

Now, we have taken v_0 to be a probability vector, so we immediately have $\pi(\cdot) = v_0(\cdot)$. Also, by the eigenvector properties, we have for example that

$$\begin{aligned} \mu_k(3) &= v_0(3) - 1.031(\lambda_1)^k v_1(3) - 0.4518(\lambda_2)^k v_2(3) - 0.2791(\lambda_3)^k v_3(3) \\ &= (0.2089) - 1.031(0.2618)^k(0.4263) \\ &\quad - 0.4518(0.0382)^k(-0.5656) - 0.2791(-0.2)^k(0.0898). \end{aligned}$$

Thus, noting that $|(1.031)(0.4263) + (0.4518)(0.5656) + (0.2791)(0.0898)| < 0.8$, and that $\lambda_* = 0.2618$, we have that

$$|\mu_k(3) - \pi(3)| < 0.8\,(0.2618)^k,$$

from which we can deduce values of k that make $\mu_k(3)$ arbitrarily close to $\pi(3)$. Other points in the state space (besides 3) are handled similarly.

Fact 3 gives a nice picture of a Markov chain converging geometrically quickly to a unique stationary distribution π. However, many Markov chains will not satisfy the condition that $P(x, y) > 0$ for all x and y. This raises the question of necessary and sufficient conditions to have $\lambda_* < 1$. The answer is as follows.

FACT 4. *A finite Markov chain satisfies* $\lambda_* < 1$ *if and only if it is both indecomposable and aperiodic.*

Proof. If the Markov chain is decomposable, with disjoint closed subspaces $\mathcal{X}_1$ and $\mathcal{X}_2$, define vectors u_1 and u_2 by $u_j(x) = 1$ if $x \in \mathcal{X}_j$, 0 otherwise. Then it is easily seen that $Pu_j = u_j$, for $j = 1, 2$, so that P has multiple eigenvalues 1, and $\lambda_* = 1$.

If the Markov chain is periodic, there are subspaces $\mathcal{X}_1, \ldots, \mathcal{X}_d$ with $P(x, \mathcal{X}_{j+1}) = 1$ for $x \in \mathcal{X}_j$, $1 \le j \le d-1$, and $P(x, \mathcal{X}_1) = 1$ for $x \in \mathcal{X}_d$. Define the vector v by $v(x) = e^{2\pi ij/d}$ for $x \in \mathcal{X}_j$. Then it is easily verified that $Pv = e^{2\pi i/d}v$. Thus, $e^{2\pi i/d}$ is an eigenvalue of P, so that again $\lambda_* = 1$.

For the converse, assume the Markov chain is indecomposable and aperiodic. Assume first that the Markov chain contains no transient states, i.e., there is positive probability of getting from any point x to any other point y (in some finite number of steps). We argue that some power of P has all its entries positive, so that the result follows from our Fact 4.

Fix $x \in \mathcal{X}$, and let $S_x = \{k \mid P^k(x, x) > 0\}$. Our assumptions imply that S_x is infinite and has greatest common divisor (g.c.d.) 1. The set S_x is also *additive*, in the sense that if $a, b \in S_x$ then $a + b \in S_x$. It is then a straightforward exercise to verify that there must be some $k_x > 0$ such that $k \in S_x$ for all $k \ge k_x$.

Find such k_x for each $x \in \mathcal{X}$, and set $k_0 = (\max_x k_x) + n$. We claim that $P^{k_0}(x, y) > 0$ for all $x, y \in \mathcal{X}$. Indeed, given x and y, by assumption there exists r_{xy} such that $P^{r_{xy}}(x, y) > 0$, and we may clearly take $r_{xy} \le n$. But then $P^{k_0}(x, y) \ge P^{k_0 - r_{xy}}(x, x)P^{r_{xy}}(x, y) > 0$, as desired.

It remains only to consider transient elements of the Markov chain. Suppose $x \in \mathcal{X}$ is transient. Then there exists $y \in \mathcal{X}$ and $r > 0$ such that $P^r(x, y) = \epsilon > 0$, but $P^m(y, x) = 0$ for all $m \ge 0$. Set $T = \{j \in \mathcal{X} \mid P^m(j, x) > 0 \text{ for some } m \ge 0\}$, so $y \notin T$. It is then easily computed that

$$\sum_{j \in T} |(vP^r)_j| \le \sum_{j \in T} |v(j)| - \epsilon |v(x)|.$$

It follows that if $vP = \lambda v$ with $|\lambda| = 1$, then we must have $v(x) = 0$, so that λ is an eigenvalue of the Markov chain restricted to $\mathcal{X} - \{x\}$. This reduces the problem to the previous case. □

We close by observing that this discussion has relied heavily on the fact that the state space $\mathcal{X}$ is *finite*. On infinite spaces, P is a linear operator but not a finite matrix, and the notion of eigenvalues must be replaced by the more general notion of *spectrum* of an operator. Conclusions about convergence rates are much more difficult in this case, but some progress has been made. See for example Belsley (1993) for countable state spaces, and Schervish and Carlin (1992) and Baxter and Rosenthal (1994) for general (uncountable) state spaces.

5. Random walks on groups. There is a particular class of Markov chains for which the eigenvalues and eigenvectors are often immediately available, namely, random walks on groups. Here $\mathcal{X}$ is a group (finite for most of the present discussion), and $Q(\cdot)$ is a probability distribution on $\mathcal{X}$ (to be referred to as the "step distribution"). The transition probabilities are then defined by $P(x, y) = Q(x^{-1}y)$; this has the interpretation that at each step we are multiplying our previous group element x on the right by a new group element, chosen according to the distribution $Q(\cdot)$; the probability that this brings us to y is the probability that we multiplied by the group element $x^{-1}y$.

Typically we take $\mu_0(id) = 1$. Then $\mu_1 = Q$, and $\mu_{k+1} = \mu_k * Q$, where $*$ stands for the *convolution* of measures.

These random walks on groups are much easier to analyze in terms of convergence to stationarity than are general Markov chains. The ideas presented here were pioneered by Diaconis and Shashahani (1981), and were greatly advanced by Diaconis (1988) and many others. This section draws heavily upon Chapter 3 of Diaconis (1988); in particular, many of our examples are taken from there. The interested reader is urged to consult this reference for a deeper treatment of this subject.

We begin with an elementary fact.

FACT 5. *Any random walk P on a finite group $\mathcal{X}$ satisfies $\pi P = \pi$, where π is defined by $\pi(x) = 1/n$ for all $x \in \mathcal{X}$ (and where $n = |\mathcal{X}|$). In words, the uniform distribution is stationary for any random walk on any finite group.*

Proof. We have

$$(\pi P)_y = \sum_x \pi(x)P(x, y) = (1/n)\sum_x Q(x^{-1}y) = (1/n)\sum_z Q(z) = 1/n = \pi(y),$$

as desired. □

We begin our investigation with the abelian case, in which there are very complete and satisfying results.

5.1. Finite abelian groups. All finite abelian (i.e., commutative) groups $\mathcal{X}$ are of the form

$$\mathcal{X} = \mathbf{Z}/(n_1) \times \mathbf{Z}/(n_2) \times \ldots \times \mathbf{Z}/(n_r),$$

a direct product of cyclic groups. That is, they consist of elements of the form $x = (x_1, \ldots, x_r)$, with the group operation being addition, done modulo n_j in coordinate j. A random walk on such $\mathcal{X}$ is defined in terms of a probability distribution $Q(\cdot)$ on $\mathcal{X}$. This induces transition probabilities defined by $P(x, y) = Q(y - x)$. (We write $y - x$ instead of $x^{-1}y$ here simply because we are writing the group operation using additive notation, as is standard for abelian groups.)

Example 1. Let $\mathcal{X} = \mathbf{Z}/(2)$, the two-element group, and set $Q(1) = p$, $Q(0) = 1 - p$. This corresponds exactly to our "simplest nontrivial example" with $q = p$.

Example 2. Frog's Walk. Let $\mathcal{X} = \mathbf{Z}/(n)$, the integers mod n, and set $Q(-1) = Q(0) = Q(1) = 1/3$. This corresponds to our frog's walk from the Introduction, in which there are n points arranged in a circle, and the frog either moves one step to the right, one step to the left, or stays where she is, each with probability $1/3$.

Example 3. Bit flipping. Let $\mathcal{X} = (\mathbf{Z}/(2))^d$, a product of d copies of the two-element group. Set $Q(0) = Q(e_1) = \cdots = Q(e_r) = 1/(d + 1)$, where e_r is the vector with a 1 in the rth spot and 0 elsewhere. This corresponds to a "bit-flipping" random walk on binary d-tuples, where at each stage we do nothing (with probability $1/(d + 1)$) or change one of the d coordinates (chosen uniformly) to its opposite value.

The usefulness of random walks on finite abelian groups comes from the fact that we can explicitly describe their eigenvalues and eigenvectors. To do this, we need to introduce *characters*. For $m = (m_1, \ldots, m_d) \in \mathcal{X}$, define

$$\chi_m(x) = e^{2\pi i[(m_1x_1/n_1)+\cdots+(m_dx_d/n_d)]}, \qquad x \in \mathcal{X}.$$

Thus, χ_m is a function from the state space $\mathcal{X}$ to the complex numbers. The following facts are easily verified.

1. $\chi_m(x + y) = \chi_m(x)\chi_m(y)$.
2. $\chi_m(0) = 1$. $|\chi_m(x)| = 1$. $\chi_m(-x) = \overline{\chi_m(x)}$.

3. $\langle \chi_m, \chi_j \rangle = \delta_{mj}$, where the inner product is defined by $\langle f, g \rangle = (1/n) \sum_x f(x)\overline{g(x)}$. In words, the characters are *orthonormal* in $L^2(\pi)$. In particular, they form a basis for all functions on $\mathcal{X}$.

4. $\sum_m \chi_m(x) = n\,\delta_{x0}$.

These properties imply the following key fact.

FACT 6. *For each $m \in \mathcal{X}$, we have*

$$\overline{\chi_m}\, P = \lambda_m\, \overline{\chi_m},$$

where

$$\lambda_m = E_Q(\chi_m).$$

In words, for each m, $\overline{\chi_m}$ is an eigenvector of P corresponding to the eigenvalue $E_Q(\chi_m)$.

Proof. We have

$$(\overline{\chi_m}\, P)_y = \sum_x \overline{\chi_m(x)}\, P(x, y) = \sum_x \chi_m(-x)\, Q(y - x) = \sum_z \chi_m(z - y)\, Q(z)$$
$$= \sum_z \chi_m(z)\chi_m(-y)\, Q(z) = E_Q(\chi_m)\, \overline{\chi_m(y)},$$

as desired. □

This fact immediately gives us all of the eigenvalues of the random walk, which is a significant achievement. (For example, in the simplest nontrivial example with $q = p$, it correctly predicts the eigenvalue $E_Q((-1)^x) = 1 - 2p$.) Combining this with Fact 3, and recalling that the characters are orthonormal in $L^2(\pi)$, we have the following fact.

FACT 7. *A random walk on a finite abelian group satisfies*

$$\|\mu_k - \pi\| \le \frac{1}{2}\sqrt{\sum_{m \neq 0} |\lambda_m|^{2k}} \le (\sqrt{n}/2)(\lambda_*)^k,$$

where $\lambda_m = E_Q(\chi_m)$.

Proof. We have from Fact 3 (since the χ_m are orthonormal) that

$$\sum_x |\mu_k(x) - \pi(x)|^2 \pi(x) = \sum_{m \neq 0} |a_m|^2\, |\lambda_m|^{2k},$$

where $\lambda_m = E_Q(\chi_m)$ as in Fact 6. Recalling that $\pi(x) = 1/n = a_m$, this reduces to

$$\sum_x |\mu_k(x) - \pi(x)|^2 = (1/n) \sum_{m \neq 0} |\lambda_m|^{2k}.$$

The result now follows from

$$4\,\|\mu_k - \pi\|^2 = \left(\sum_x |\mu_k(x) - \pi(x)|\right)^2 \le n \sum_x |\mu_k(x) - \pi(x)|^2,$$

by the Cauchy–Schwarz inequality. □

Let us now apply this bound to the second and third examples above. For the frog's walk, we have

$$\lambda_m = E_Q(\chi_m) = (1/3) + (2/3)\cos(2\pi m/n).$$

It follows that $\lambda_* = (1/3) + (2/3)\cos(2\pi/n)$. Using just λ_* in our bound above, we have (assuming $n \geq 3$, and using that $\cos(x) \leq 1 - x^2/4$ for $0 \leq x \leq \sqrt{6}$, and that $1 - x \leq e^{-x}$ for any x) that

$$\|\mu_k - \pi\| \leq (\sqrt{n}/2)(\lambda_*)^k \leq (\sqrt{n}/2)e^{-\frac{2\pi^2}{3n^2}k}.$$

This bound is small if k is large compared to $n^2 \log n$. We can actually get rid of the $\log n$ term by using the stronger bound with all the eigenvalues:

$$\begin{aligned}
\|\mu_k - \pi\|^2 &\leq \frac{1}{4}\sum_{m=1}^{n-1}(\lambda_m)^{2k} \\
&\leq \sum_{m=1}^{\lceil \frac{n-1}{4} \rceil} e^{-\frac{4\pi^2 m^2}{3n^2}k} \\
&\leq \sum_{m=1}^{\infty} e^{-\frac{4\pi^2 m}{3n^2}k} \\
&= \frac{e^{-\frac{4\pi^2}{3n^2}k}}{1 - e^{-\frac{4\pi^2}{3n^2}k}}.
\end{aligned}$$

This last expression is small if k is large compared to n^2.

One might wonder if the order n^2 can be reduced still further. In fact, it cannot. To see this, we produce a *lower bound* as follows. First note that

$$E_{\mu_k}(\chi_m) = \big(E_Q(\chi_m)\big)^k.$$

(This is similar to the fact that the characteristic function of a sum of independent random variables is the product of the individual characteristic functions.) This statement can easily be proved by induction. It can also be seen directly by noting that the quantity on the left is the eigenvalue of P^k corresponding to the eigenvector $\overline{\chi_m}$, and is thus the kth power of the corresponding eigenvalue for P.

It is further seen directly (or from the fact that χ_m is orthonormal to $\chi_0 \equiv 1$) that $E_\pi(\chi_m) = 0$ for $m \neq 0$. It now follows from our third equivalent definition of variation distance that

$$\|\mu_k - \pi\| \geq \frac{1}{2}|E_{\mu_k}(\chi_1)| = \frac{1}{2}|E_Q(\chi_1)|^k = \frac{1}{2}\left|\frac{1}{3} + \frac{2}{3}\cos\left(\frac{2\pi}{n}\right)\right|^k.$$

(We could have chosen any other character χ_m with $m \neq 0$ in place of χ_1.)

Taking $n = 1000$ and $k = 10{,}000$, this equals 0.438. Thus, our frog would have to take considerably more than 10,000 steps to have an approximately equal chance of being at any of her 1000 lily pads. To make this less than 0.1, we need $k = 122,302$.

More generally, for $n \geq 5$ this lower bound implies that

$$\|\mu_k - \pi\| \geq \frac{1}{2}\left(1 - \frac{1}{3}\left(\frac{2\pi}{n}\right)^2\right)^k \geq \frac{1}{2}\left(1 - \frac{k}{3}\left(\frac{2\pi}{n}\right)^2\right).$$

It is easily seen that this quantity will be far from 0 unless k is large compared to n^2. Thus, $O(n^2)$ iterations are, for large n, both necessary and sufficient to converge to uniformity for this process.

For the "bit-flipping" process, Example 3 above, we have $\chi_m(x) = (-1)^{m \cdot x}$, where $m \cdot x = m_1 x_1 + \cdots + m_d x_d$. It is easily computed that

$$\lambda_m = E_Q(\chi_m) = 1 - \frac{2N(m)}{d+1},$$

where $N(m)$ stands for the number of 1's in the binary d-tuple m. Hence, $\lambda_* = 1 - \frac{2}{d+1}$. Using this directly, and recalling that $n = |\mathcal{X}| = 2^d$, we have

$$\|\mu_k - \pi\| \leq 2^{d-1}\left(1 - \frac{2}{d+1}\right)^k \leq 2^d e^{-2k/(d+1)},$$

which is small provided k is large compared to d^2.

As in the previous example, we can do better by using all the eigenvalues. Indeed, there are $\binom{d}{j}$ choices for m, which have $N(m) = j$. Hence, we have (cf. Diaconis (1988, §3C)) that

$$\begin{aligned}
\|\mu_k - \pi\|^2 &\leq \frac{1}{4}\sum_{j=1}^{d}\binom{d}{j}\left|1 - \frac{2j}{d+1}\right|^{2k} \\
&\leq \frac{1}{2}\sum_{j=1}^{\lceil \frac{d+1}{2} \rceil}\binom{d}{j}\left(1 - \frac{2j}{d+1}\right)^{2k} \\
&\leq \frac{1}{2}\sum_{j=1}^{\infty}\frac{d^j}{j!}e^{-\frac{4j}{d+1}k} \\
&= \frac{1}{2}\left(e^{de^{-\frac{4k}{d+1}}} - 1\right).
\end{aligned}$$

This last expression is small if k is of the form $\frac{1}{4}d\log d + Cd$ with C large. This result is in fact the "correct" answer. Indeed, it can be shown (Diaconis (1988)) that to first order in d, *precisely* $\frac{1}{4}d\log d$ iterations are required to get close to uniform. Such a sharp result as this, giving the number of iterations *exactly* to first order in the size of the group, is the essence of the "cut-off phenomenon"; see Diaconis and Shashahani (1981), Aldous and Diaconis (1987), Diaconis (1988), and Rosenthal (1994c).

5.2. Finite nonabelian groups. For nonabelian groups, the situation is more complicated, but we can still make use of the "characters" of the group to find eigenvalues, at least under the additional assumption that our step distribution is "conjugate invariant."

Let $\mathcal{X}$ be a finite, nonabelian group (such as the symmetric group S_ℓ, which corresponds to shuffling a deck of cards). Such a group has associated with it *irreducible representations* $\rho_0, \rho_1, \ldots, \rho_r$, where $\rho_m : \mathcal{X} \to M_{d_m}(\mathbf{C})$ is a function taking the group $\mathcal{X}$ into the set of $d_m \times d_m$ complex matrices, which is multiplicative in the sense that $\rho_m(xy) = \rho_m(x)\rho_m(y)$ and that $\rho_m(id) = I_{d_m}$. (Here the multiplication on the left is in the group, while the multiplication on the right is matrix multiplication.)

It is known that these irreducible representations satisfy $\sum_m (d_m)^2 = |\mathcal{X}|$, i.e., that there are as many "representation entries" as there are elements of the group. Furthermore we may assume that $\rho_m(x^{-1}) = \rho_m(x)^*$, the conjugate transpose of $\rho_m(x)$. (In words, we may assume the matrices $\rho_m(x)$ are unitary.) It is then true that these "representation entries" are *orthogonal* under the appropriate inner product.

The connection with the abelian case comes as follows. The *characters* of the group are given by $\chi_m = \operatorname{tr} \rho_m$, the trace of the matrix. For *abelian* groups, we have $d_m = 1$ for all m,

so that the character and the representation are essentially the same; in that case, the current situation reduces to the previous one. In general, we have that $\sum_m d_m \chi_m(s) = n\delta_{s,id}$; again, if $d_m = 1$ for all m, this reduces to the previous case. Also, once again, the characters are orthonormal in $L^2(\pi)$.

In this generality, one cannot obtain simple formulas for the eigenvalues of the transition matrix P. Indeed, the matrix for P need not even be diagonalizable. However, let us assume that the step distribution $Q(\cdot)$ is *conjugate invariant*, in the sense that $Q(x^{-1}yx) = Q(y)$ for all $x, y \in \mathcal{X}$. That is easily seen to imply that $\rho_m(x^{-1})E_Q(\rho_m)\rho_m(x) = E_Q(\rho_m)$ for all m and for all $x \in \mathcal{X}$. In words, the matrix $E_Q(\rho_m)$ commutes with every matrix of the form $\rho_m(x)$, for $x \in \mathcal{X}$. A well-known result from group representation theory, Schur's Lemma, then implies that $E_Q(\chi_m)$ is a *scalar* matrix, i.e., a multiple of the identity. It follows by taking traces that

$$E_Q(\rho_m) = \left(E_Q(\chi_m)/d_m\right) I_{d_m},$$

where I_{d_m} is the $d_m \times d_m$ identity matrix.

Under this "conjugate invariant" assumption, we have the following fact.

FACT 8. *Let P correspond to a conjugate invariant random walk on a finite group $\mathcal{X}$ as above. For $0 \le m \le r$, and $1 \le i, j \le d_m$, we have*

$$\overline{\rho_{m(ij)}}\, P = (E_Q(\chi_m)/d_m)\, \overline{\rho_{m(ij)}}.$$

In words, the vector whose value at the point $x \in \mathcal{X}$ is the complex conjugate of the ij entry of the matrix $\rho_m(x)$, is an eigenvector for P, with eigenvalue $E_Q(\chi_m)/d_m$.

Proof. For $g \in \mathcal{X}$, we have

$$\begin{aligned}
\left(\overline{\rho_{m(ij)}}\, P\right)_g &= \sum_{x\in\mathcal{X}} \overline{\rho_{m(ij)}(x)}\, P(g,x) \\
&= \sum_{x\in\mathcal{X}} \overline{\rho_{m(ij)}(x)}\, Q(x^{-1}g) \\
&= \sum_{y\in\mathcal{X}} \overline{\rho_{m(ij)}(gy^{-1})}\, Q(y) \\
&= \sum_{y} Q(y) \sum_{z} \overline{\rho_{m(iz)}(g)\rho^*_{m(jz)}(y)} \\
&= \sum_{z} \overline{(\rho_m(g))_{iz}\, \left(E_Q(\rho_m)^*\right)_{jz}} \\
&= (E_Q(\chi_m)/d_m)\, \left(\overline{\rho_m(g)}\right)_{ij},
\end{aligned}$$

where we have used that $E_Q(\rho_m)$ is diagonal, with diagonal entries $E_Q(\chi_m)/d_m$. □

It follows immediately that the eigenvalues of P are precisely $E_Q(\chi_m)/d_m$, each repeated $(d_m)^2$ times. It also follows that the vector $\overline{\chi_m}$ is an eigenvector with this same eigenvalue, which directly generalizes the abelian case. Furthermore, as mentioned above, the characters χ_m are again orthonormal in $L^2(\pi)$. By exact analogy with our discussion there, we have the following fact.

FACT 9. *The variation distance to the uniform distribution π satisfies*

$$\|\mu_k - \pi\| \le \frac{1}{2}\sqrt{\sum_{m\neq 0}(d_m)^2|\lambda_m|^{2k}} \le (\sqrt{n}/2)(\lambda_*)^k,$$

with $n = |\mathcal{X}|$ and with $\lambda_m = E_Q(\chi_m)/d_m$.

Example. Random transpositions. Consider the symmetric group S_ℓ, with step distribution given by $Q(id) = 1/\ell$, $Q((ij)) = 2/\ell^2$ for all $i \neq j$. This corresponds to shuffling a deck of cards by choosing a random card uniformly with the left hand, choosing a random card uniformly with the right hand, and interchanging their positions in the deck (and doing nothing if we happened to pick the same card with both hands). Bounds on the distance to stationarity then correspond to bounds on how long the deck of cards must be shuffled until it is well mixed.

This was the example that motivated Diaconis and Shashahani (1981) to develop the modern, quantitative study of random walks on groups. To do a careful analysis of this model requires detailed knowledge of the representation theory of the symmetric group, which is rather involved. We note here simply that χ_1 for the symmetric group is the function that assigns to each group element, one less than the number of points in $\{1, 2, \ldots, \ell\}$ that it leaves fixed. Thus, $\chi_1(id) = \ell - 1$, and $\chi_1((ij)) = \ell - 3$. Also, $d_1 = \ell - 1$. Hence, the eigenvalue corresponding to χ_1 is given by

$$\lambda_1 = E_Q(\chi_1)/d_1 = \frac{(1/\ell)(\ell-1) + (1-(1/\ell))(\ell-3)}{\ell-1} = 1 - \frac{2}{\ell} \leq e^{-2/\ell}.$$

Now, it so happens (though we cannot prove it here) that for this random walk, $\lambda_* = \lambda_1$. Thus, using our bound developed above, we have that

$$\|\mu_k - \pi\| \leq \sqrt{\ell!/2}(\lambda_*)^k \leq e^{\ell \log \ell} e^{-2k/\ell},$$

which is small if k is large compared to $\ell^2 \log \ell$. Diaconis and Shashahani (1981) did a much more careful analysis of this process, using all the eigenvalues, and proved that to first order in ℓ, $\frac{1}{2}\ell \log \ell$ steps were necessary and sufficient, again proving a cut-off phenomenon.

A number of other random walks on finite groups have been considered and shown to exhibit a cut-off phenomenon, including random transvections (Hildebrand (1992)) and rank-one deformations (Belsley (1993)). Bayer and Diaconis (1992) analyzed ordinary "riffle" card shuffles on the symmetric group, and proved a cut-off phenomenon at $(3/2)\log_2 \ell$ iterations. In particular, for $\ell = 52$, they showed that about seven such shuffles were required to get close to stationarity. This shuffle is *not* conjugate invariant; thus, their methods were somewhat different from the above, and involved deriving exact expressions for μ_k for this random walk.

Finally, we mention that similar analyses to the above have been carried out for conjugate-invariant random walks on (infinite) compact Lie groups, such as those proposed for encryption algorithms by Sloane (1983). In Rosenthal (1994a), a process of "random rotations" on the orthogonal group $SO(n)$ was shown to converge to Haar measure with a cut-off at $\frac{1}{2}n \log n$. In Porod (1993), generalizations of a process of "random reflections" were shown to exhibit the cut-off phenomenon on all of the classical compact Lie groups (orthogonal, unitary, and symplectic). The basic method of proof in these examples is the same as for finite groups. However, here the number of eigenvalues is *infinite*, so there is the additional complication that bounds required are infinite sums.

6. Coupling and minorization conditions. Often, Markov chains of interest will not have the restrictive structure of a random walk on group. Thus, it is necessary to consider other approaches to bounding their convergence. In this section, we present an approach that does not use eigenvalues at all. Rather, it uses probabilistic ideas directly.

6.1. Coupling. The basic idea of coupling is the following. Suppose we have two *random variables* X and Y, defined jointly on some space $\mathcal{X}$. If we write $\mathcal{L}(X)$ and $\mathcal{L}(Y)$ for their respective probability distributions, then we can write

$$\begin{aligned}\|\mathcal{L}(X) - \mathcal{L}(Y)\| &= \sup_A |P(X \in A) - P(Y \in A)| \\ &= \sup_A |P(X \in A, X = Y) + P(X \in A, X \neq Y) \\ &\quad - P(Y \in A, Y = X) - P(Y \in A, Y \neq X)| \\ &= \sup_A |P(X \in A, X \neq Y) - P(Y \in A, Y \neq X)| \\ &\leq P(X \neq Y).\end{aligned}$$

Thus, *the variation distance between the laws of two random variables is bounded by the probability that they are unequal.*

We make use of this fact as follows. Given a Markov chain P on a space $\mathcal{X}$, with initial distribution μ_0, suppose we can find a new Markov chain (X_k, Y_k) on $\mathcal{X} \times \mathcal{X}$ with

(i) $X_0 \sim \mu_0$;
(ii) $Y_0 \sim \pi$;
(iii) $P(X_{k+1} \in A \mid X_k) = P(X_k, A)$;
(iv) $P(Y_{k+1} \in A \mid Y_k) = P(Y_k, A)$.
(v) There is a random time T such that $X_k = Y_k$ for all $k \geq T$.

In words, the chain X_k starts in the distribution μ_0 and proceeds according to the transitions $P(\cdot, \cdot)$. The chain Y_k starts in the distribution π and proceeds according to the same transitions $P(\cdot, \cdot)$. However, the *joint* law of (X_k, Y_k) is arbitrary, except that after some time T (called the *coupling time*), the two processes become equal.

The benefit of the above "coupling" is as follows. Since X_k is updated from $P(\cdot, \cdot)$, we have $\mathcal{L}(X_k) = \mu_k$. Also, since Y_k is also updated from $P(\cdot, \cdot)$, and since the distribution π is stationary, we have $\mathcal{L}(Y_k) = \pi$ for all k. It follows that

$$\|\mu_k - \pi\| = \|\mathcal{L}(X_k) - \mathcal{L}(Y_k)\| \leq P(X_k \neq Y_k) \leq P(T > k).$$

Thus, if we can find a coupling as above, we get an immediate bound on $\|\mu_k - \pi\|$ simply in terms of the tail probabilities of the coupling time T.

There is a huge literature on coupling, and it has a long history in Markov chain theory. See for example Aldous (1983), Lindvall (1992), and references therein. We shall concentrate here on a particularly simple and elegant use of coupling, related to minorization conditions.

6.2. Uniform minorization conditions. Suppose a Markov chain satisfies an inequality of the form

$$P^{k_0}(x, A) \geq \beta \zeta(A), \qquad x \in R, \quad A \subseteq \mathcal{X},$$

where k_0 is a positive integer, R is a subset of the state space $\mathcal{X}$, $\beta > 0$, and $\zeta(\cdot)$ is some probability distribution on $\mathcal{X}$.

Such an inequality is called a *minorization condition* for a Markov chain, and says that the transition probabilities from a set R all have common *overlap* of at least size β. Minorization conditions were developed by Athreya and Ney (1978), Nummelin (1984), and others. We shall see that they can help us define a coupling to get bounds on the chain's rate of convergence.

We consider here the *uniform* case in which $R = \mathcal{X}$, i.e., in which the minorization condition holds on the entire state space. (This is sometimes called the Doeblin condition.) We further assume for simplicity that $k_0 = 1$.

We shall now use this minorization condition to define a coupling. First define (X_k, Z_k) jointly as follows. Choose $X_0 \sim \mu_0$ and $Z_0 \sim \pi$ independently. Then, given X_k and Z_k, choose X_{k+1} and Z_{k+1} by flipping an independent coin that has probability β of coming up

heads, and then (a) if the coin is heads, choose a point $z \in \mathcal{X}$ distributed independently according to $\zeta(\cdot)$, and set $X_{k+1} = Z_{k+1} = z$. (b) If the coin is tails, then choose X_{k+1} and Z_{k+1} independently with

$$P(X_{k+1} \in A) = \frac{P(X_k, A) - \beta\zeta(A)}{1 - \beta};$$

$$P(Z_{k+1} \in A) = \frac{P(Z_k, A) - \beta\zeta(A)}{1 - \beta}.$$

(Note that the minorization condition ensures that these choices are in fact from probability distributions.)

These probabilities have been chosen precisely so that $P(X_{k+1} \in A \mid X_k) = P(X_k, A)$ (and similarly for Z_{k+1}). The point is, option (a) forces X_{k+1} to be equal to Z_{k+1}, and this chance of becoming equal is good for getting coupling bounds.

Let T be the first time the coin comes up heads. Then define Y_k by

$$Y_k = \begin{cases} Z_k, & k \leq T; \\ X_k, & k > T. \end{cases}$$

Thus, Y_k is essentially the same as Z_k, except that after the Markov chains become equal at time T, they will *remain* equal forever.

The combined chain (X_k, Y_k) is now a coupling with coupling time T. Also, since we had probability β of choosing option (a) each time, we see that $P(T > k) = (1 - \beta)^k$. Our above inequality immediately gives the following.

FACT 10. *Suppose a Markov chain satisfies $P(x, A) \geq \beta\zeta(A)$, for all $x \in \mathcal{X}$ and for all measurable subsets $A \subseteq \mathcal{X}$, for some probability distribution $\zeta(\cdot)$ on $\mathcal{X}$. Then given any initial distribution μ_0 and stationary distribution π, we have*

$$\|\mu_k - \pi\| \leq (1 - \beta)^k.$$

This fact goes back to Doob (1953) and has been used in Roberts and Polson (1994), Rosenthal (1993a), and many other places. It is quite powerful. For example, it immediately generalizes our earlier result that, on a finite state space, if all entries of the matrix P are positive then the chain converges geometrically quickly. In fact, now we require only that some column of P be all positive (and furthermore we immediately obtain a quantitative bound on convergence in that case).

It is easily seen that, given a Markov chain $P(x, \cdot)$, the largest value of β that we can use as above should be given by

$$\beta = \int_{\mathcal{X}} \inf_{x \in \mathcal{X}} P(x, dy),$$

which on a discrete space reduces to

$$\beta = \sum_{y \in \mathcal{X}} \min_{x \in \mathcal{X}} P(x, y).$$

In words, we may take β to be the sum of the minimum values of the entries in each column of P. (Note that $\beta = 1$ if and only if $P(x, \cdot)$ does not depend on x, in which case the

Markov chain converges *exactly* after a single step.) We can then immediately conclude that $\|\mu_k - \pi\| \leq (1 - \beta)^k$. Note that this was precisely our finding in the simplest nontrivial example of §3.

Example. Consider the Markov chain on $\mathcal{X} = \{1, 2, 3, 4, 5\}$ with transition matrix

$$P = \begin{pmatrix} 0.2 & 0.2 & 0.3 & 0.3 & 0 \\ 0.4 & 0 & 0.3 & 0.3 & 0 \\ 0.2 & 0.2 & 0.4 & 0.1 & 0.1 \\ 0.2 & 0.1 & 0.3 & 0.1 & 0.3 \\ 0.2 & 0 & 0.5 & 0.3 & 0 \end{pmatrix}.$$

We see by inspection that the column minimums are 0.2, 0, 0.3, 0.1, 0, respectively. Thus we may take $\beta = 0.2 + 0.3 + 0.1 = 0.6$, and immediately conclude that $\|\mu_k - \pi\| \leq (0.4)^k$. (Note that here $Q(1) = 1/3$, $Q(3) = 1/2$, and $Q(4) = 1/6$.)

Example. Let $\mathcal{X} = [0, 1]$ (the interval from 0 to 1), and set

$$P(x, dy) = \frac{1 + x + y}{\frac{3}{2} + x}\, dy.$$

We see by inspection that $P(x, dy) \geq \frac{2}{3}\, dy$ for all x and y, so that we may take $\beta = \frac{2}{3}$ to conclude that $\|\mu_k - \pi\| \leq (1/3)^k$. We can do even better by finding the best β as above:

$$\beta = \int_0^1 \left(\inf_{0 \leq x \leq 1} \frac{1 + x + y}{\frac{3}{2} + x} \right) dy = \int_0^{1/2} \frac{2}{3}(1 + y)dy + \int_{1/2}^1 \frac{2}{5}(2 + y)dy = \frac{29}{30}.$$

Hence, we actually have $\|\mu_k - \pi\| \leq (1/30)^k$. (Note that here $Q(\cdot)$ has density (with respect to dy) given by $\frac{30}{29}\frac{2}{3}(1 + y)$ for $0 \leq y \leq \frac{1}{2}$, and by $\frac{30}{29}\frac{2}{5}(2 + y)$ for $\frac{1}{2} < y \leq 1$.)

These two examples will probably convince the reader that the minorization method is sometimes very powerful. On the other hand, the best value of β above will often be 0; for example, this is certainly true for the frog's walk discussed in the introduction. One way to get around this difficulty is to replace P by P^{k_0} in the minorization condition, which requires replacing k by $[k/k_0]$ in the conclusion. In principle this approach should usually work well, but in practice it may be very difficult to compute or estimate quantities related to P^{k_0}. See Rosenthal (1993a) for one attempt in this direction.

Another method is to restrict the values of x in the minorization condition to being in some subset $R \subseteq \mathcal{X}$, as we now discuss.

6.3. Minorization conditions on subsets. Suppose that instead of the uniform minorization condition as above, we have a minorization condition which holds only on a subset $R \subseteq \mathcal{X}$. Then our bound above, which was based on coupling with probability β at each step, cannot be applied. Various other approaches have been used in this case. We very briefly outline them here.

If one allows the subset R to be arbitrarily large (in fact, to grow as a function of k), then it may be possible to bound the probability of escaping from R, and draw conclusions about $\|\mu_k - \pi\|$ in that way; see Rosenthal (1991).

In any case, each time our coupled process (X_k, Y_k) visits the subset $R \times R$, it has probability β of coupling. Using "drift conditions," it may be possible to bound the number of such returns to $R \times R$, and then use coupling as in the uniform case; see Rosenthal (1993b).

A related approach is presented in Meyn and Tweedie (1993), who use minorizations, drift conditions, splittings, and careful bounding to obtain bounds on $\|\mu_k - \pi\|$ directly, without introducing a second, coupled chain.

Instead of trying to bound $\|\mu_k - \pi\|$ directly, or use coupling, another approach is as follows. Consider a single Markov chain X_k, and each time it is in the subset R, with probability β update it according to $\zeta(\cdot)$. Call the times of such updatings *regeneration times*. Then, it is easily seen that the distribution of X_k depends only on the time since the last regeneration time. Thus, if these "times since the last regeneration" converge, then the original chain X_k must also converge. This is the essential idea behind convergence results in Athreya and Ney (1978), Nummelin (1984), Asmussen (1990), Mykland, Tierney, and Yu (1992), and elsewhere.

7. Other approaches. There are many other methods of bounding convergence rates of Markov chains, many of which have been applied to a number of examples of interest. We briefly mention some of these methods here.

For certain Markov chains including birth-death chains (i.e., Markov chains on the integers, which can move at most distance 1 on a given step), the eigenvalues and eigenvectors are related to the "orthogonal polynomials." Classically known results can be used to get good bounds on convergence rates. See Belsley (1993) and references therein.

Related to the coupling and minorization bounds presented herein is the method of strong stopping times (Aldous and Diaconis (1986), (1987)). Essentially, if the reference measure $\zeta(\cdot)$ in the minorization condition happens to be the stationary distribution $\pi(\cdot)$, then one can construct a random time τ such that the law of X_τ is precisely $\pi(\cdot)$, and such that X_τ is independent of τ. Such a time τ is a strong stopping time, and it is easily seen that $\|\mu_k - \pi\| \leq P(\tau > k)$. Another method of constructing strong stopping times is by constructing a dual Markov chain that keeps track of "how stationary" the Markov chain has become; see Diaconis and Fill (1990).

A different and very beautiful method of bounding convergence to certain specific distributions (e.g., normal, Poisson) is the method of Stein (1971) and Chen (1975). This involves characterizing the distribution of interest through some "identity" that it satisfies, and then seeing to what extent the distribution μ_k approximately satisfies that identity. In certain cases the technique has been simplified to the point where it is very usable. See Arratia, Goldstein, and Gordon (1989) and Barbour, Holst, and Janson (1992).

Finally, geometric arguments involving "paths" on graphs have recently been used to bound eigenvalues of Markov chains with great success in certain examples; see Jerrum and Sinclair (1989) and Diaconis and Stroock (1991). Geometric approaches have also been used to allow different Markov chains to be "compared" to each other, so that known information about one Markov chain can be used to obtain information about related chains; see Diaconis and Saloff-Coste (1993).

Some of these approaches use *reversibility* of a Markov chain, meaning that the identity $\pi(dx)P(x, dy) = \pi(dy)P(y, dx)$ holds for all $x, y \in \mathcal{X}$. This is equivalent to saying that, if the chain starts in the stationary distribution $\pi(\cdot)$, it has the same law whether time runs forwards or backwards. This immediately implies that P is a selfadjoint operator on $L^2(\pi)$ (and hence its eigenvalues are all real). Such structure is discussed and exploited in Diaconis and Stroock (1991), Keilson (1979), and elsewhere. In Fill (1991), it is shown how to make use of reversibility to obtain bounds on convergence, even if the original Markov chain P is nonreversible.

Most of the above work has been concerned primarily with convergence in total variation distance (or the related separation distance). There are, of course, many other notions of distance between probability measures that could be used, such as relative entropy. See Su (1994) for a start in this direction.

Naturally, these few words scarcely begin to cover the depth of work that has been applied to convergence questions. The reader is strongly encouraged to consult these and other references for further information.

Acknowledgments. I thank Eric Belsley for comments and corrections, and thank Persi Diaconis for introducing me to this subject and teaching me so much.

REFERENCES

D. ALDOUS (1983), *Random walk on finite groups and rapidly mixing Markov chains*, Séminaire de Probabilités XVII, Lecture Notes in Math., 986, Springer-Verlag, New York, pp. 243–297.

D. ALDOUS AND P. DIACONIS (1986), *Shuffling cards and stopping times*, Amer. Math. Monthly, 93, pp. 333–348.

——— (1987), *Strong stopping times and finite random walks*, Adv. Appl. Math., 8, pp. 69–97.

R. ARRATIA, L. GOLDSTEIN, AND L. GORDON (1989), *Two moments suffice for Poisson approximations: the Chen–Stein method*, Ann. Probab., 17, pp. 9–25.

S. ASMUSSEN (1987), *Applied Probability and Queues*, Wiley & Sons, New York.

K. B. ATHREYA AND P. NEY (1978), *A new approach to the limit theory of recurrent Markov chains*, Trans. Amer. Math. Soc., 245, pp. 493–501.

A. D. BARBOUR, L. HOLST, AND S. JANSON (1992), *Poisson Approximation*, Oxford University Press, Oxford.

J. R. BAXTER AND J. S. ROSENTHAL (1994), *Rates of convergence for everywhere-positive Markov chains*, Statist. Probab. Lett., 23, to appear.

D. BAYER AND P. DIACONIS (1992), *Trailing the dovetail shuffle to its lair*, Ann. Appl. Probab., 2, pp. 294–313.

E. D. BELSLEY (1993), *Rates of convergence of Markov chains related to association schemes*, Ph.D. thesis, Dept. of Mathematics, Harvard University, Cambridge, MA.

L. H. Y. CHEN (1975), *Poisson approximation for dependent trials*, Ann. Probab., 3, pp. 534–545.

P. DIACONIS (1988), *Group Representations in Probability and Statistics*, I.M.S. Lecture Series, 11, Hayward, CA.

——— (1991), *Finite Fourier methods: Access to tools*, Proc. Symposia Appl. Math., 44, pp. 171–194.

P. DIACONIS AND J. A. FILL (1990), *Strong stationary times via a new form of duality*, Ann. Probab., 18, pp. 1483–1522.

P. DIACONIS AND L. SALOFF-COSTE (1993), *Comparison theorems for reversible Markov chains*, Ann. Appl. Probab., 3, pp. 696–730.

P. DIACONIS AND M. SHASHAHANI (1981), *Generating a random permutation with random transpositions*, Z. Wahrsch. Ver. Geb., 57, pp. 159–179.

P. DIACONIS AND D. W. STROOCK (1991), *Geometric bounds for reversible Markov chains*, Ann. Appl. Probab., 1, pp. 36–61.

J. I. DOOB (1953), *Stochastic Processes*, Wiley, New York.

W. FELLER (1968), *An Introduction to Probability Theory and Its Applications*, Vol. I, third ed., Wiley & Sons, New York.

J. A. FILL (1991), *Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process*, Ann. Appl. Probab., 1, pp. 64–87.

A. E. GELFAND AND A. F. M. SMITH (1990), *Sampling based approaches to calculating marginal densities*, J. Amer. Statist. Assoc., 85, pp. 398–409.

M. HILDEBRAND (1992), *Generating random elements in $SL_n(F_q)$ by random transvections*, J. Algebraic Comb., 1, pp. 133–150.

P. G. HOEL, S. C. PORT, AND C. J. STONE (1972), *Introduction to Stochastic Processes*, Waveland Press, Prospect Heights, IL.

D. L. ISSACSON AND R. W. MADSEN (1976), *Markov Chains: Theory and Applications*, Wiley & Sons, New York.

M. JERRUM AND A. SINCLAIR (1989), *Approximating the permanent*, SIAM J. Comput. 18, pp. 1149–1178.

J. KEILSON (1979), *Markov Chain Models—Rarity and Exponentiality*, Springer-Verlag, New York.

T. LINDVALL (1992), *Lectures on the Coupling Method*, Wiley & Sons, New York.

S. P. MEYN AND R. L. TWEEDIE (1993), *Computable bounds for convergence rates of Markov chains*, Tech. Report, Dept. of Statistics, Colorado State University, Ft. Collins.

P. MYKLAND, L. TIERNEY, AND B. YU (1992), *Regeneration in Markov chain samplers*, Tech. Report 583, School of Statistics, University of Minnesota, Minneapolis.

E. NUMMELIN (1984), *General Irreducible Markov Chains and Nonnegative Operators*, Cambridge University Press, Cambridge.

U. POROD (1993), *The Cut-Off Phenomenon for Random Reflections*, Ph.D. thesis, Division of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD.

G. O. ROBERTS AND N. G. POLSON (1994), *On the geometric convergence of the Gibbs sampler*, J. Royal Statist. Soc. Ser. B, 56, pp. 377–384.

J. S. ROSENTHAL (1991), *Rates of convergence for Gibbs sampler for variance components models*, Tech. Report 9322, Dept. of Statistics, University of Toronto; Ann. Statist., to appear.

J. S. ROSENTHAL (1993a), *Rates of convergence for data augmentation on finite sample spaces*, Ann. Appl. Probab., 3, pp. 319–339.

——— (1993b), *Minorization conditions and convergence rates for Markov chain Monte Carlo*, Tech. Report 9321, Dept. of Statistics, University of Toronto, Ontario; J. Amer. Statist. Assoc., to appear.

——— (1994a), *Random rotations: Characters and random walks on* SO(N), Ann. Probab., 22, pp. 398–423.

——— (1994b), *Random walks on discrete and continuous circles*, J. Appl. Probab., 30, pp. 780–789.

——— (1994c), *On generalizing the cut-off phenomenon for random walks on groups*, Adv. Appl. Math., to appear.

M. J. SCHERVISH AND B. P. CARLIN (1992), *On the convergence of successive substitution sampling*, J. Comput. Graph. Statist., 1, pp. 111–127.

N. J. A. SLOANE (1983), *Encrypting by random rotations, in Cryptography, Lecture Notes in Computer Science*, T. Beth, ed., 149, pp. 71–128.

A. F. M. SMITH AND G. O. ROBERTS (1993), *Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion)*, J. Royal Statist. Soc. Ser. B, 55, pp. 3–24.

A. D. SOKAL (1989), *Monte Carlo methods in statistical mechanics: foundations and new algorithms*, Dept. of Physics, New York University, New York. Cours de Troisième Cycle de la Physique en Suisse Romande, Lausanne, Switzerland.

C. STEIN (1971), *A bound for the error in the normal approximation to the distribution of a sum of dependent random variables*, in Proc. Sixth Berkeley Symp. Math. Statist. Probab., 3, pp. 583–602.

F. SU (1994), *An upper bound for relative entropy of Markov chains*, manuscript.

L. TIERNEY (1994), *Markov chains for exploring posterior distributions*, Ann. Statist., to appear.