A Numerical Analyst Looks at the "Cutoff Phenomenon" in Card Shuffling and Other Markov Chains

Gudbjorn F. Jonsson and Lloyd N. Trefethen

Abstract Diaconis and others have shown that certain Markov chains exhibit a "cutoff phenomenon" in which, after an initial period of seemingly little progress, convergence to the steady state occurs suddenly. Since Markov chains are just powers of matrices, how can such effects be explained in the language of applied linear algebra? We attempt to do this, focusing on two examples: random walk on a hypercube, which is essentially the same as the problem of Ehrenfest urns, and the celebrated case of riffle shuffling of a deck of cards. As is typical with transient phenomena in matrix processes, the reason for the cutoff is not readily apparent from an examination of eigenvalues or eigenvectors, but it is reflected strongly in pseudospectra—provided they are measured in the 1-norm, not the 2-norm. We illustrate and explain the cutoff phenomenon with MATLAB computations based in part on a new explicit formula for the entries of the $n \times n$ "riffle shuffle matrix," and note that while the normwise cutoff may occur at one point, such as $\frac{3}{2}\log_2 n$ for the riffle shuffle, weak convergence may occur at an equally precise earlier point such as $\log_2 n$.

1 Introduction

In the past fifteen years, Persi Diaconis and others have proved intriguing theorems showing that certain Markov chains exhibit a "cutoff phenomenon." One of these results, presented in a paper by Bayer and Diaconis in 1992, asserts that in a certain precise asymptotic sense as $n \to \infty$, it takes exactly $\frac{3}{2}\log_2 n$ riffle shuffles to randomize a deck of n cards [1]. The announcement of this result attracted a great deal of publicity, including a headline in the *New York Times* [17] as well as articles in *Newsweek, The Economist,* and *Seventeen.* For a survey of the cutoff phenomenon, see Diaconis's inaugural article [3] for the National Academy of Sciences.

For a number of years, the second author has been interested in phenomena of transient behavior in linear evolution processes—exponentials of matrices in the continuous case, powers of matrices in the discrete one. Asymptotically for large time, certain eigenvalues and eigenvectors dominate the behavior for these problems, but if the initial behavior is very different, that behavior typically has little to do with the eigenvalues and eigenvectors. Such effects arise most familiarly with matrices that are far from normal, that is, whose eigenvectors are far from orthogonal, and when eigenvalues fail to explain transient behavior, it has been observed that pseudospectra (defined in Section 5) often do better [21]. Applications in which linear transients are important include numerical stability and stiffness of discretizations of differential

equations [13], convergence of Krylov subspace matrix iterations [8], and hydrodynamic stability [23].

We set out to learn what connection there might be between these recent developments in Markov chains and in linear algebra. Can the cutoff phenomenon be explained in terms of norms of powers of matrices? Does it depend on non-normality? Is it reflected in lopsided pseudospectra? Can interesting examples be reduced to a small enough scale for experimentation in MATLAB? Do such experiments suggest new understanding?

As we shall explain, the answers to all these questions have proven to be yes—with one exception. The cutoff phenomenon does not depend on non-normality, and in fact, it can occur even when the matrix is normal. Normality is a notion relevant to behavior in the 2-norm, whereas for Markov chains, we are interested in the 1-norm.

2 Powers of the transition matrix P

To define a Markov chain, we begin with a finite state space, with N states. A *probability distribution* for this space is a row vector $u \in \mathbb{R}^N$ with the properties

$$u_j \geq 0 \quad (1 \leq j \leq N), \qquad \|u\|_1 = 1,$$

where u_j represents the probability that the Markov chain is in state j, and the 1-norm is defined by $\|u\|_1 = \sum_{j=1} |u_j|$. In linear algebra we are accustomed to the use of column vectors, but the use of row vectors is universal in the literature of Markov chains.

At each step of the chain, the probabilities evolve according to a fixed linear procedure. This corresponds to multiplication on the right (since u is row vector) by an $N \times N$ *transition matrix* P:

$$u^{(k+1)} = u^{(k)} P.$$

The entry p_{ij} is the probability that the chain will move to state j at the next step, if it is currently in state i, and thus we have $0 \leq p_{ij} \leq 1$ for any i and j. Repeated steps of the chain are governed by powers of P:

$$u^{(k)} = u^{(0)} P^k. \tag{2.1}$$

Since probability is conserved, each row sum of P must be equal to 1. In particular the maximum row sum of P is 1, a condition we shall write as

$$\|P\|_1 = 1.$$

We must be clear about what this formula means. According to the standard conventions in linear algebra, which are based on column vectors, the maximum row sum of a matrix P is $\|P\|_\infty$. However, in the present context it is appropriate to reverse the usual roles of $\|\cdot\|_1$ and $\|\cdot\|_\infty$ and define

$$\|P\|_1 \;=\; \max_{\|u\|_1 = 1} \|uP\|_1 \;=\; \max_i \sum_{j=1}^{N} |p_{ij}|. \tag{2.2}$$

In MATLAB, our 1-norm of a matrix P can accordingly be computed by the expressions `norm(P',1)` or `norm(P,inf)`. For the 2-norm, $\|P\|_2 = \|P^T\|_2$, so there is no need to modify the standard definition.

A *stationary probability distribution* is a row vector $\sigma \in \mathbb{R}^N$ that satisfies

$$\lim_{k\to\infty} uP^k = \sigma$$

for any nonnegative starting vector u with $\|u\|_1 = 1$. (We do not need to specify the norm defining this limit, as all norms on a finite-dimensional space are equivalent.) It is easily seen that σ is a stationary probability distribution if and only if the limit $P^\infty = \lim_{k\to\infty} P^k$ exists and is the $N \times N$ matrix whose rows are all equal to σ:

$$P^\infty = \begin{pmatrix} \text{—} & \sigma & \text{—} \\ \text{—} & \sigma & \text{—} \\ & \vdots & \\ \text{—} & \sigma & \text{—} \end{pmatrix} \tag{2.3}$$

This limiting behavior is guaranteed to occur under the conditions that P is *irreducible* and *aperiodic,* conditions that we assume from now on but shall not define. See any introduction to Markov chains, such as the appropriate chapters of [10] or [20]. For information about Markov chains more closely tied to the present discussion, see [2] and [18].

If a stationary probability distribution σ exists, then it satisfies

$$\sigma P = \sigma$$

and is thus a left eigenvector of P corresponding to the eigenvalue 1. Since the row sums of P are 1, a corresponding right eigenvector is $N^{-1}(1, 1, \ldots, 1)^T$. From the definition of a stationary probability distribution, it follows that σ must be the only properly normalized left eigenvector of P corresponding to the eigenvalue 1, and all other eigenvalues must be smaller in absolute value. In particular, if λ_2 denotes the second largest eigenvalue in absolute value (not necessarily unique), then $|\lambda_2| < 1$.

Here is an example with $N = 3$. Consider a random walk on the vertices of a triangle in which at each step, a particle moves with probability $\frac{1}{2}$ to each of the adjacent vertices. We have

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}, \qquad P^2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix},$$

indicating for example that for a particle at vertex 1, the probability of being again at vertex 1 after 2 steps is $\frac{1}{2}$. In the limit we get

$$P^\infty = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix},$$

with each row equal to the vector $\sigma = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, indicating that the probability is uniformly distributed on the vertices. Thus we see that for this matrix P, the principal left and right eigenvectors are identical, which is to be expected as the matrix is symmetric. For nonsymmetric P, σ need not be the uniform vector.

3 Powers of the decay matrix $A = P - P^\infty$

The powers P^k satisfy $\|P^k\|_1 = 1$ for all k, which does not tell us much. What are more interesting are the norms $\|P^k - P^\infty\|_1$. Let us define $A = P - P^\infty$. This matrix, which we shall call the *decay matrix,* represents the action of the Markov chain on the space spanned by the non-dominant eigenvectors. (Here and throughout, it is not necessary for our matrices to be diagonalizable, but our discussion assumes diagonalizability for simplicity.) By induction it is readily shown that

$$A^k = P^k - P^\infty \tag{3.1}$$

for each $k \geq 1$. Our interest is thus in the behavior of the norms $\|A^k\|_1$ as a function of k.

Curiously, this reduction to a matrix A whose powers converge to 0 is rarely mentioned in the literature of Markov chains, where explicit discussion of matrix norms is generally avoided. For a numerical analyst, however, this is certainly the natural language in which to frame a discussion of the cutoff phenomenon. Note that A is the rank-one modification of P obtained by subtracting off the matrix $P^\infty = e\sigma$, where $e = (1, 1, \ldots, 1)^T$. (The matrix P^∞ is the spectral projector associated with the eigenvalue 1.) In our applications below, since σ is known explicitly, this is how we compute A. Mathematically speaking, A might equivalently be constructed by computing an eigenvalue decomposition $P = VDV^{-1}$, letting $\tilde{D}$ be the rank one modification of D with the diagonal entry 1 replaced by zero, and setting $A = V\tilde{D}V^{-1}$. (If P is not diagonalizable, one could use a Jordan decomposition instead.)

For the example of random walk on a triangle, we find

$$A = \begin{pmatrix} -\frac{1}{3} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & -\frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & -\frac{1}{3} \end{pmatrix}, \qquad A^2 = \begin{pmatrix} \frac{1}{6} & -\frac{1}{12} & -\frac{1}{12} \\ -\frac{1}{12} & \frac{1}{6} & -\frac{1}{12} \\ -\frac{1}{12} & -\frac{1}{12} & \frac{1}{6} \end{pmatrix},$$

and in general, $A^k = (-\frac{1}{2})^{k-1}A$ for any $k \geq 1$, implying that A^k decreases to 0 at the rate exactly $|\lambda_2| = \frac{1}{2}$ at each step, independent of the choice of norm. Thus there is no cutoff phenomenon for random walk on a triangle, and the same is true for random walk on an n-gon [3].

The decay matrix A in this example is symmetric and hence *normal,* which means that it has a complete set of orthonormal eigenvectors. One might expect that this should be the reason why there are no transient effects in the convergence of A^k to 0. The kind of matrix behavior that might lead to such an expectation is illustrated in

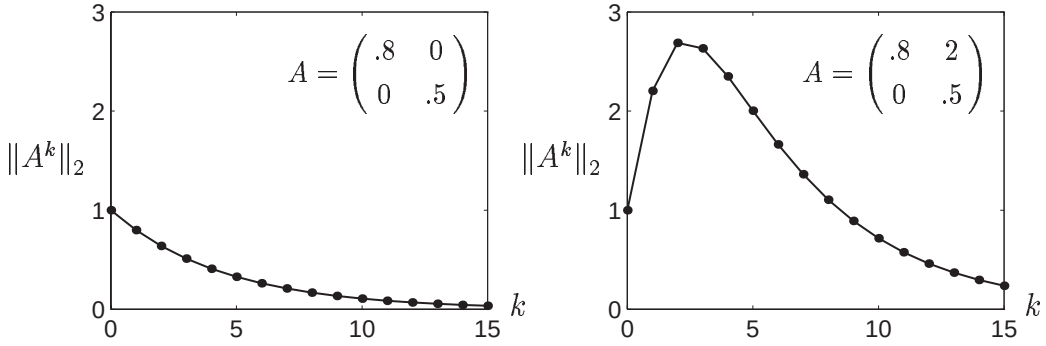

Figure 3.1: The matrix on the left is normal, while the matrix on the right is not. These curves of $\|A^k\|_2$ vs. k illustrate that in the 2-norm, transient effects in norms of matrix powers are associated with non-normality. These matrices A, however, do not correspond to Markov chains.

Figure 3.1, which shows the powers $\|A^k\|_2$ for a normal and a non-normal matrix with identical eigenvalues. It is the non-normal matrix that exhibits the transient effect; the normal matrix powers satisfy $\|A^k\|_2 = \|A\|_2^k$.

However, the correct explanation of the cutoff phenomenon lies elsewhere. Consider any reversible Markov chain whose stationary distribution σ is the uniform vector. (A Markov chain with transition matrix P and stationary probability distribution σ is *reversible* if DP is symmetric, where $D = \mathrm{diag}(\sigma)$.) Since the condition on σ implies that D is a multiple of the identity, P is symmetric, hence normal. Now, some examples of Markov chains that exhibit the cutoff phenomenon do not satisfy these conditions. The riffle shuffle Markov chain of [1], for example, is not reversible. Other Markov chains with cutoffs, however, do satisfy them, and the existence of such chains proves that non-normality is not essential for the existence of a cutoff. One example is shuffling a deck of n cards by random transpositions, which has a cutoff at $\frac{1}{2}n \log n$ [5]. Another is random walk on a hypercube, with a cutoff at $\frac{1}{4}n \log n$ [6, 4]. We now turn to this example.

4 Example 1: Random walk on a hypercube

The n-dimensional hypercube is suggested in Figure 4.1. A random walk on this graph can be defined as follows. At each step, a particle located at a particular vertex either moves to an adjacent vertex or remains where it is. Each event occurs with probability $1/(n+1)$, and the steps are independent.

The state space for this Markov chain is of dimension $N = 2^n$. The $N \times N$ transition matrix P is symmetric and sparse, with only $n+1$ nonzero entries in each row. To define the matrix fully, one would have to pick an ordering of the vertices, but we need not do this, as our observations all pertain to the 1-norm or the 2-norm, both of which are ordering-independent.

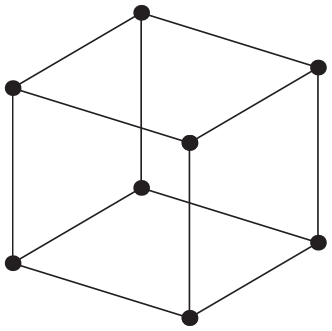

Figure 4.1: Schematic illustration of the hypercube of dimension n. At each step of the random walk, a particle at a vertex moves to one of the adjacent vertices or stays fixed, each with probability $1/(n+1)$.

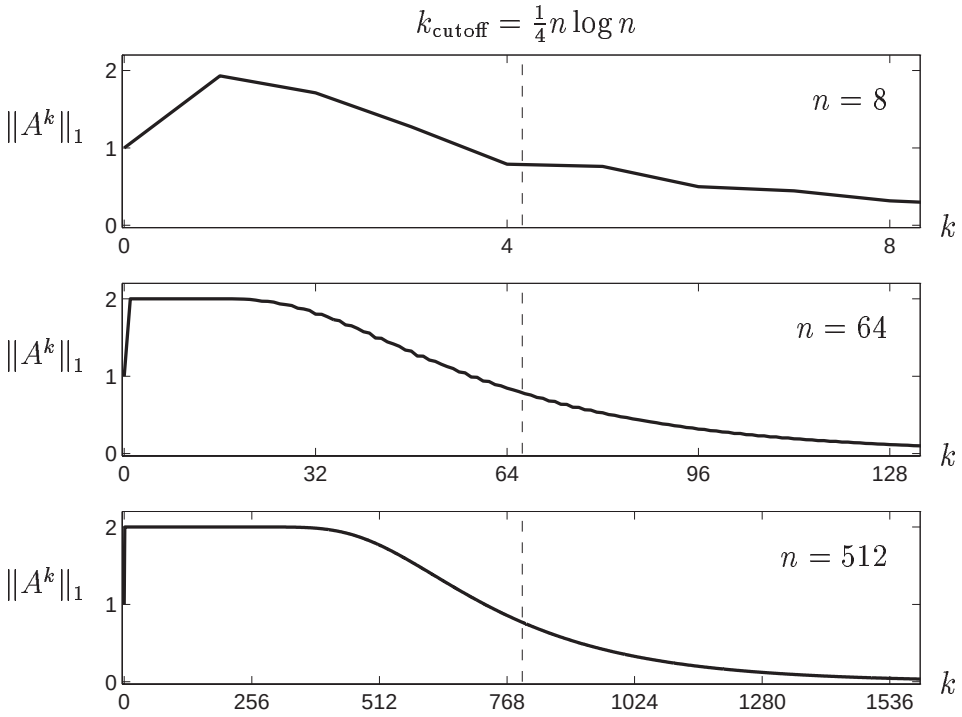

Figure 4.2: Computed illustration of the cutoff phenomenon for random walk on the n-cube. These results, including the chatter in the middle plot, are believed to be correct to plotting accuracy. As $n \to \infty$, the curve steepens to a step at $k_{\text{cutoff}} = \frac{1}{4} n \log n$. As explained in Section 6, exactly the same curves also describe the problem of Ehrenfest urns.

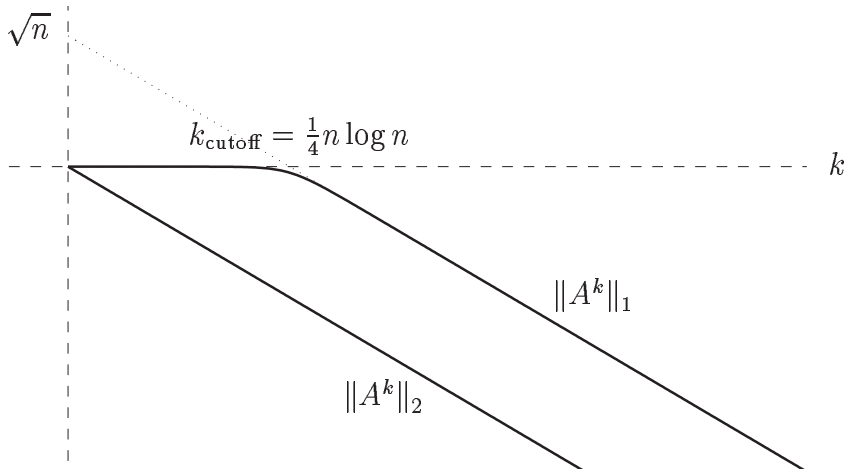

Figure 4.3: Schematic representation on a log scale of the cutoff phenomenon for random walk on a hypercube. Now in the 2-norm, there is no cutoff, and the same will be true of any symmetric Markov chain. For nonsymmetric chains, including those corresponding to the Ehrenfest urns and riffle shuffle problems, the 2-norm curve may be more complicated.

It is easy to imagine in a rough way what happens with this example. Suppose the particle begins at a particular vertex, corresponding to an initial probability distribution $u^{(0)}$ with the value 1 in one component and 0 elsewhere. At each step of the walk, the probability then diffuses around the hypercube. Obviously, n steps are needed before it is even possible to get to the diagonally opposite vertex, and it is plausible that significantly more steps than this will be needed before a uniform distribution of probability on the vertices is approached.

In fact, $\frac{1}{4} n \log n$ steps are needed, as has been proved by Diaconis, Graham, and Morrison [4]:

$$k_{\text{cutoff}} = \tfrac{1}{4} n \log n. \qquad \text{(hypercube/Ehrenfest)} \tag{4.1}$$

This theorem is illustrated in Figure 4.2, where $\|A^k\|_1$ is plotted as a function of k for $n = 8$, 64, and 512. The dashed line in each part of the figure represents $k = \frac{1}{4} n \log n$. For the larger two values of n, the curve shows a plateau of norms almost exactly equal to 2 for a long range of initial values of k. For example, $\|A^{15}\|_1 \approx 1.999976$ for $n = 64$. Then, around the dashed line, the values fall off exponentially to 0. Diaconis et al. proved that for $k = \alpha n \log n$ (more precisely, take k to be the nearest integer to $\alpha n \log n$), $2 - \|A^k\|_1$ decays to 0 as $n \to \infty$ for any fixed $\alpha < \frac{1}{4}$, and $\|A^k\|_1$ decays to 0 as $n \to \infty$ for any fixed $\alpha > \frac{1}{4}$. In other words, as n increases, the curves of Figure 4.2 steepen to a step function.

For this example A is normal, implying $\|A^k\|_2 = \|A\|_2^k$ for all $k \geq 0$. Thus the cutoff behavior that is so pronounced in the 1-norm must vanish completely in the 2-norm. Figure 4.3 illustrates this crucial norm dependence by plotting $\|A^k\|_1$ and $\|A^k\|_2$ schematically on a log scale. Eventually, both curves are straight, but for the

1-norm, there is a long flat section before the curve turns downward. Readers used to assuming that all norms are equivalent for practical purposes should bear in mind that the 1- and 2-norms of an $N \times N$ matrix may differ by as much as a factor of $\sqrt{N}$, and if $N = 2^{512}$ or 52!, this is a lot of room to play around!

It may be wondered, how were the data for Figure 4.2 computed, given that the matrices involved have dimensions as high as 2^{512}? The answer is that to generate that figure, as well as Figure 5.2 below, we reduced the $N \times N$ matrix problem to an equivalent problem of dimension $n+1$ by the method described in Sections 6 and 8. All computations were carried out in MATLAB®.

Even without any numerical computation, the eigenvalues and eigenvectors for random walk on the n-cube can be determined by the methods of Fourier analysis on groups [4]. The eigenvalues of P are the evenly spaced real numbers

$$1 - \frac{2j}{n+1} \qquad (0 \le j \le n), \tag{4.2}$$

with the jth eigenvalue having multiplicity $\binom{n}{j}$. The eigenvalues of A are the same, except with the eigenvalue 1 corresponding to $j = 0$ replaced by 0. Thus the asymptotic decay constant $|\lambda_2|$ in Figures 4.2 and 4.3 is $1 - 2/(n+1)$.

Our plots have shown that random walk on the hypercube exhibits a cutoff phenomenon, but they have not explained what is really going on. What causes the cutoff? Why does the plateau appear at the value $\|A^k\| \approx 2$? Both of these questions have simple answers, which we shall give in Section 6.

5 Eigenvalues and pseudospectra

In matrix problems where eigenvalues give an incomplete impression of behavior, it is often informative to examine the more general sets in the complex plane $\mathbf{C}$ known as *pseudospectra* [21, 22]. For each $\epsilon \ge 0$ and specified norm $\|\cdot\|$, the ϵ-pseudospectrum of A is the subset of $\mathbf{C}$ bounded by the ϵ^{-1} level curve of the norm of the resolvent:

$$\Lambda_\epsilon(A) \;=\; \{\, z \in \mathbf{C} :\; \|(zI - A)^{-1}\| \ge \epsilon^{-1} \,\}. \tag{5.1}$$

(We use the convention that $\|(zI-A)^{-1}\| = \infty$ if z is an eigenvalue of A.) Equivalently, $\Lambda_\epsilon(A)$ can be defined by eigenvalue perturbations:

$$\Lambda_\epsilon(A) \;=\; \{\, z \in \mathbf{C} :\; z \in \Lambda(A+E) \text{ for some } E \text{ with } \|E\| \le \epsilon \}, \tag{5.2}$$

where $\Lambda(A+E)$ denotes the spectrum of $A+E$. The equivalence of (5.1) and (5.2) implies that if the resolvent norm is large at a particular value $z \in \mathbf{C}$, then z is an eigenvalue of a slightly perturbed matrix—where the meanings of "large" and "slightly" depend on the choice of norm. This equivalence is proved for a general class of norms in [7].

Figure 5.1 shows 2-norm pseudospectra for random walk on the n-cube with $n = 39$. Since the matrix is normal, $\Lambda_\epsilon(A)$ is equal to the set of points at distance $\le \epsilon$ from

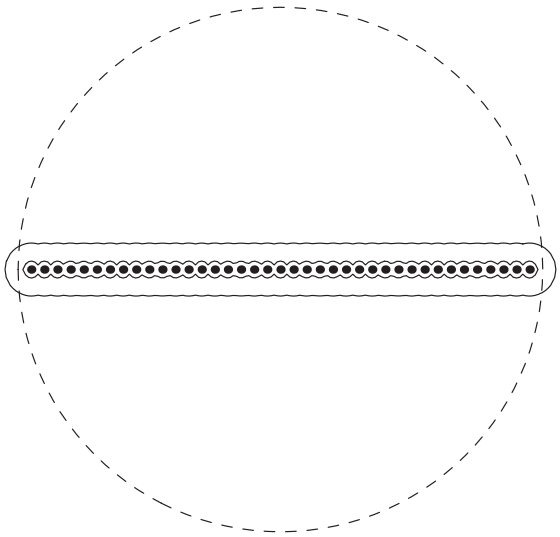

Figure 5.1: 2-norm pseudospectra $\Lambda_\epsilon(A)$ for random walk on the n-cube with $n = 39$ ($\epsilon = 10^{-1}, 10^{-1.5}$). The dashed curve marks the unit circle. The matrix A is normal, so the boundary of $\Lambda_\epsilon(A)$ is just the set of points at a distance ϵ from the spectrum; the 2-norm condition number of a matrix of eigenvectors is $\kappa_2(V) = \|V\|_2 \, \|V^{-1}\|_2 = 1$. For the Ehrenfest urns problem, described in Section 6, the picture would look different.

the spectrum of A. This trivial pseudospectral picture reflects the trivial behavior of $\|A^k\|_2$ as a function of k.

Figure 5.2, on the other hand, shows 1-norm pseudospectra for the same example. Now the picture is far more interesting. These pseudospectra bulge far away from the spectrum, an effect that grows more pronounced as n increases. Note that the curves displayed correspond to values of ϵ as low as 10^{-4}, i.e., resolvent norms as great as 10^4. Figures 5.1 and 5.2 imply, then, that although the eigenvalues of A are insensitive to matrix perturbations as measured in the 2-norm, they are highly sensitive to matrix perturbations in the 1-norm. In the next section we shall see that the eigenvectors associated with these sensitive eigenvalues have no natural connection to the "physics" of the transient behavior of this system.

One implication of large pseudospectra is that any matrix of eigenvectors of A—any matrix V in a diagonalization $A = V\Lambda V^{-1}$—must be ill-conditioned, with $\kappa(V) = \|V\| \, \|V^{-1}\| \gg 1$. This follows from the formula $(zI - A)^{-1} = V(zI - \Lambda)^{-1}V^{-1}$, which implies in any p-norm

$$\|(zI - A)^{-1}\| \;\leq\; \frac{\kappa(V)}{\mathrm{dist}(z, \Lambda(A))},$$

where $\mathrm{dist}(z, \Lambda(A))$ denotes the distance of z from the closest eigenvalue of A. In the 2-norm, any matrix of normalized eigenvectors of A has condition number 1; in fact, for an appropriate ordering of the vertices of the n-cube, V is the standard Hadamard matrix of dimension 2^n. In the 1-norm, however, condition numbers on the

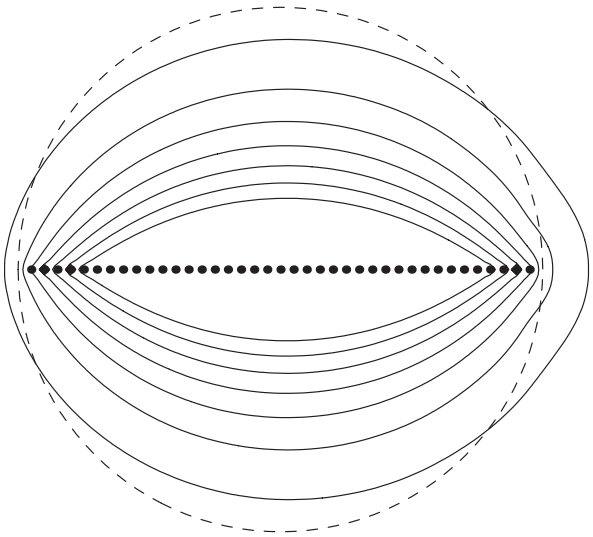

Figure 5.2: 1-norm pseudospectra for the same problem ($\epsilon = 10^{-1}, 10^{-1.5}, 10^{-2}, \ldots, 10^{-4}$). Now the pseudospectra lie far from the eigenvalues; the 1-norm condition number of at least one matrix of eigenvectors is $\kappa_1(V) = \|V\|_1 \, \|V^{-1}\|_1 \approx \times 10^6$, a figure that grows exponentially with n. This picture also applies without change to the Ehrenfest urns problem.

order of $2^{n/2}$ or 2^n are typical for this problem, depending how the eigenvectors are normalized. It follows that the expansion of an initial state in the basis of eigenvectors will typically involve coefficients many orders of magnitude larger than the signal itself, with the consequence that the early evolution will be determined more by patterns of cancellation among the coefficients than by decay of the individual eigenmodes.

By means of contour integrals and arguments related to the Kreiss matrix theorem [24], rigorous bounds can be developed that relate $\{\Lambda_\epsilon(A)\}$ and $\{\|A^k\|\}$, but we shall not pursue this matter here.

6 Example 1': Ehrenfest urns

The problem of random walk on a hypercube, with a state space of dimension 2^n, is essentially equivalent to the problem of Ehrenfest urns, whose state space is of dimension $n+1$. It is this reduction to a low-dimensional problem that makes possible some of the results of Diaconis, et al. [4] as well as our own computations for Figures 4.2 and 5.2.

Consider a Markov chain defined by the usual random walk on the hypercube, but instead of taking the state variable to be the vertex, take it to be just the *distance* from a distinguished vertex called vertex 0, as indicated in Figure 6.1. Since all the vertices lie at a distance from vertex 0 in the range from 0 to n, the new state space has dimension $n + 1$.

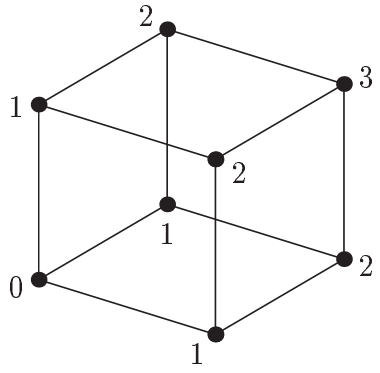

Figure 6.1: Compression of the n-cube, with state space of dimension 2^n, to a problem of dimension $n+1$. The numbers mark values of the new state variable, j, the distance along the graph from vertex 0.

The physics of the new chain is as follows. Consider a vertex of the n-cube at a distance j from vertex 0. This vertex has n neighbors, j of which are 1 step closer than it to 0 and $n-j$ of which are 1 step further away. Thus with probability $j/(n+1)$, it will move closer to 0 at the next step, with probability $(n-j)/(n+1)$, it will move further away, and with probability $1/(n+1)$, it will remain at the same distance. These formulas apply equally for all $\binom{n}{j}$ vertices at distance j from 0. It follows that the Markov process on the n-cube induces a Markov process on the state variable j, and if we order the states from 0 to n, the transition matrix looks like this:

$$P = \begin{pmatrix} \frac{1}{n+1} & \frac{n}{n+1} & & & & \\ \frac{1}{n+1} & \frac{1}{n+1} & \frac{n-1}{n+1} & & & \\ & \frac{2}{n+1} & \frac{1}{n+1} & \frac{n-2}{n+1} & & \\ & & \ddots & \ddots & \ddots & \\ & & & \frac{n-1}{n+1} & \frac{1}{n+1} & \frac{1}{n+1} \\ & & & & \frac{n}{n+1} & \frac{1}{n+1} \end{pmatrix}. \tag{6.1}$$

This is a tridiagonal, nonsymmetric matrix of dimension $n+1$. Its eigenvalues are the same as in (4.2), but now, each eigenvalue has multiplicity 1. The corresponding decay matrix A (not shown) is dense, with the same eigenvalues except with 1 replaced by 0. Most importantly, by combining all vertices with the same j into a single state, we have conserved probability, and the 1-norms $\|A^k\|_1$ and $\|(zI-A)^{-1}\|_1$ are the same for this matrix as for the hypercube (see Section 8.) The 2-norms are not conserved, but 2-norms are of little importance for Markov chains, and we shall not study them further.

Our new Markov chain can be interpreted as follows (Figure 6.2). Consider n indistinguishable balls, each located in one of two urns, Urn 0 or Urn 1. At each step

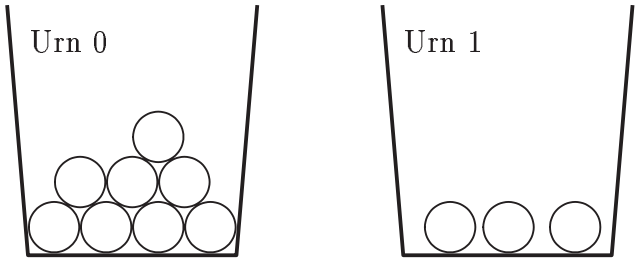

Figure 6.2: Ehrenfest urns containing n balls. At each step, a ball is selected at random and moved to the other urn. This corresponds to changing one coordinate of a particle at a vertex of the n-cube, thereby moving it to an adjacent vertex.

of a random process, a ball is selected at random and moved to the other urn (and with probability $1/(n+1)$, the ball is not moved). The state variable is j, the number of balls in Urn 1, taking values from 0 to n. How does the probability distribution on this state space evolve as more and more balls are moved? This is the problem of the *Ehrenfest urns,* which dates to 1907 [9]. (The physical motivation is very interesting, but we shall not discuss it [14].) Each ball represents one coordinate in the hypercube, and changing its urn corresponds to changing that coordinate from 0 to 1 or from 1 to 0.

We can now give an intuitive explanation of where the cutoff phenomenon comes from for the hypercube/Ehrenfest example. Consider Figure 6.3. This figure shows the evolution of the probability distribution as a function of step number k, assuming there are $n = 100$ balls and they begin in Urn 0 at step 0. At step 0, the probability distribution is a delta function of height 1 at position $j = 0$. At step 1, most of the probability, though not quite all, has moved to position $j = 1$. With further steps the distribution moves to the right and diffuses into approximately the shape of a Gaussian, and after 200 steps it is close to its final position centered at $j = 50$. The stationary distribution σ—the dominant left eigenvector of the matrix P—is known analytically [4]:

$$\sigma_j = 2^{-n}\binom{n}{j} \qquad (0 \le j \le n).$$

Here is the explanation of the cutoff. The crucial point is that the location of the initial distribution along the j axis is different from that of the stationary distribution. Evidently this particular Markov chain involves not just diffusion but also propagation of probability. Consequently, until the moving pulse gets near the middle of the interval, it has exponentially small overlap with its final position, and the 1-norm of the difference between the current and the asymptotic state is exponentially close to the sum of the 1-norms of those two states:

$$\|u^{(k)} - u^{(\infty)}\|_1 \;\approx\; \|u^{(k)}\|_1 + \|u^{(\infty)}\|_1 \;=\; 2. \tag{6.2}$$

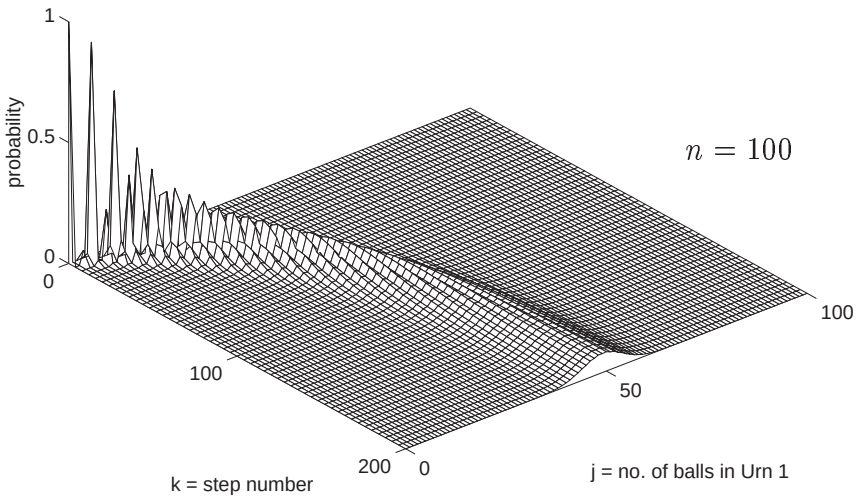

Figure 6.3: Evolution of the probability density function for the hypercube/Ehrenfest problem with $n = 100$. (The k axis here is undersampled for plotting reasons, so only every 4th step is visible.) The probability not only diffuses, but also propagates rightwards before eventually settling down in its stationary position in the middle of the interval. The cutoff phenomenon results from the fact that the early and final states have exponentially small overlap. For this value of n, $k_{\text{cutoff}} = \frac{1}{4} n \log n \approx 115$.

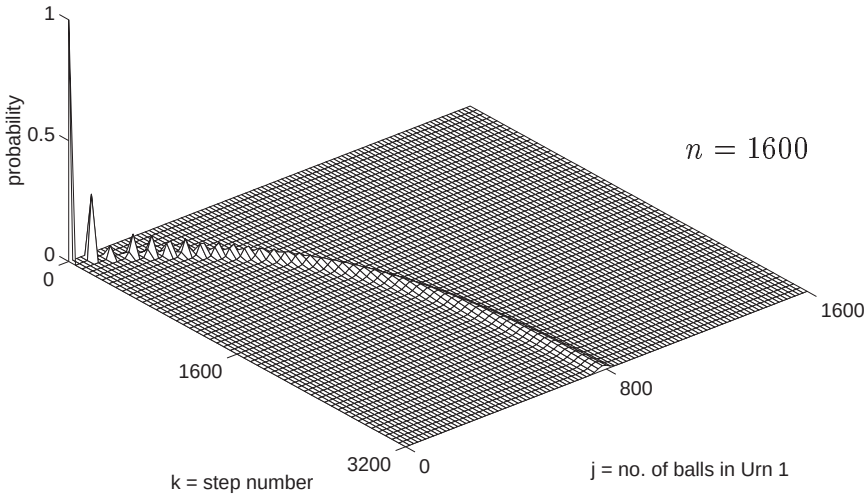

Figure 6.4: Same as Figure 6.3, but for $n = 1600$. The stationary distribution is 4 times broader, hence 4 times narrower relative to n, but the curve traced is the same. (The k and j axes are undersampled, so only an approximation to each 64th step is visible.) For this value of n, $k_{\text{cutoff}} = \frac{1}{4} n \log n \approx 2951$.

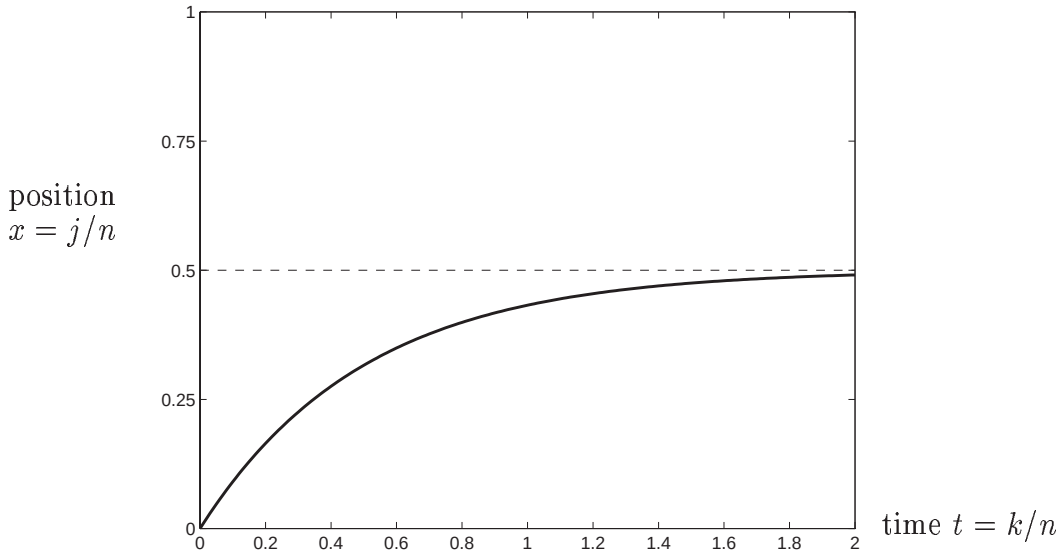

Figure 6.5: Position (6.3) of the center of the pulse for the hypercube/Ehrenfest problem in the limit $n \to \infty$. The cutoff point $k_{\text{cutoff}} = \frac{1}{4} n \log n$ lies off-scale to the right by a factor $O(\log n)$.

Figure 6.4 shows what happens when n is increased 16-fold and the j and k axes are rescaled by the same factor. The picture looks much the same, except that the wave pulse is narrower. Its standard deviation is $\sim \sqrt{n}/2$, and since the j axis is scaled by n, its relative width in the plot is $\sim 1/2\sqrt{n}$.

Figure 6.5 shows the trajectory that the waves in Figures 6.3 and 6.4 approximate, given by the formula

$$x = \tfrac{1}{2}(1 - e^{-2t}). \tag{6.3}$$

One way to derive this formula is to show by induction that the mean of the probability distribution at step k—that is, the expected number of balls in Urn 1—is exactly

$$\mu_k = \frac{n}{2}\left(1 - \left(\frac{n-1}{n+1}\right)^k\right), \tag{6.4}$$

or equivalently,

$$\frac{1}{2} - \frac{\mu_k}{n} = \frac{1}{2}\left(\frac{n-1}{n+1}\right)^k, \tag{6.5}$$

which is consistent with the result $|\lambda_2| = 1 - 2/(n+1)$ of (4.2). With $x = \mu_k/n$ and $t = k/n$, (6.4) reduces to (6.3) in the limit $n \to \infty$.

The careful reader may be puzzled at this point, as we once were. The arguments above, as well as Figures 6.3–6.5, suggest that the behavior of the hypercube/Ehrenfest problem scales in proportion to n. If there is a cutoff, it seems that it should occur at a step $k = O(n)$. In fact, however, the formula (4.1) is $k_{\text{cutoff}} = \frac{1}{4} n \log n$. Where does the factor $\log n$ come from?

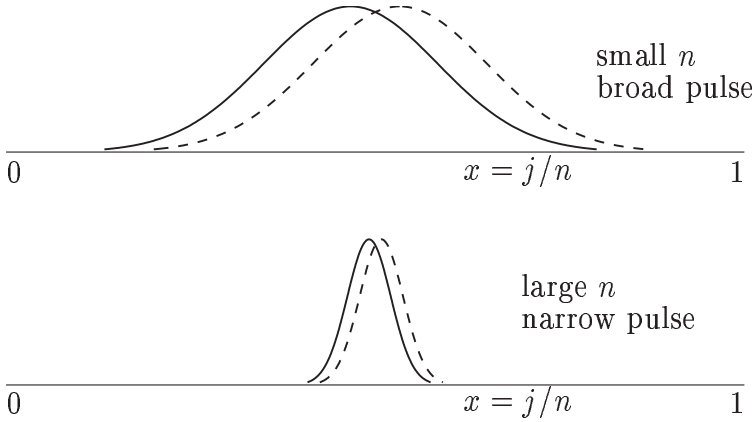

Figure 6.6: Since convergence in the hypercube/Ehrenfest problem is defined by the norm $\|u^{(k)} - u^{(\infty)}\|_1$, the "convergence criterion" depends on the width of the pulse, which is proportional to $n^{-1/2}$ in a relative sense as $n \to \infty$. It is this tightening convergence criterion that accounts for the $\log n$ factor in the formula $k_{\text{cutoff}} = \frac{1}{4} n \log n$.

The answer is suggested in Figure 6.6. As $n \to \infty$, the width of the stationary probability distribution grows narrower relative to n. Now as indicated in (6.2), the cutoff phenomenon depends on the norm of the difference $u^{(k)} - u^{(\infty)}$. For this norm to be small, the location of $u^{(k)}$ must be close to that of $u^{(\infty)}$ not in an absolute sense, but relative to their widths. In other words, *the convergence criterion that defines the cutoff grows stricter as* $n \to \infty$. The cutoff condition implied by (6.3) is

$$e^{-2t} = O(n^{-1/2}), \tag{6.6}$$

that is, $-2t = -\frac{1}{2} \log n$, hence $t \sim \frac{1}{4} \log n$, hence $k \sim \frac{1}{4} n \log n$. And so it is that we can derive the cutoff formula (4.1) solely from the knowledge that the probability wave is a pulse of width $O(n^{-1/2})$ at position (6.3).

Let us summarize where we stand. For the Ehrenfest urns problem, hence also for random walk on a hypercube, the large-time asymptotic behavior is essentially a process of one Gaussian sliding along the x axis until it coincides with another. The Gaussian is the dominant left eigenvector of P, and the error for large time, which is the second eigenvector of P or the dominant eigenvector of A, must look like the difference of two nearby Gaussians, as suggested in Figure 6.7, with the shape of an S. As more steps are taken, the two Gaussians align more closely and the amplitude of the S decays, but its width does not change.

This brings us to the question of the "physical significance" of eigenvectors. We have seen that the eigenvector whose decay governs the asymptotic behavior of the Ehrenfest problem is an S curve located at the center of the interval. For small time, the probability distribution has nothing whatever to do with this S curve. Nor does it have anything to do with the other eigenfunctions, which are higher-frequency

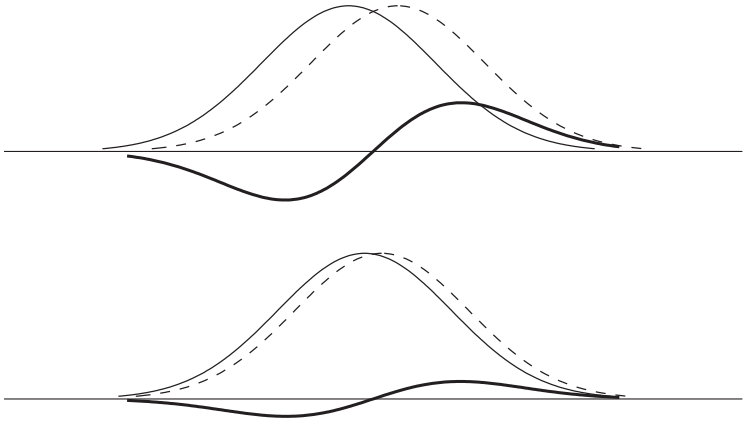

Figure 6.7: The difference of two nearby Gaussians is an S curve (heavy line), whose amplitude decays as the Gaussians slide on top of each other. This process describes the long-time convergence in the hypercube/Ehrenfest problem. The S curve is the dominant eigenvector of A, which determines the long-time shape of the distribution, but it has nothing to do with the behavior for short time.

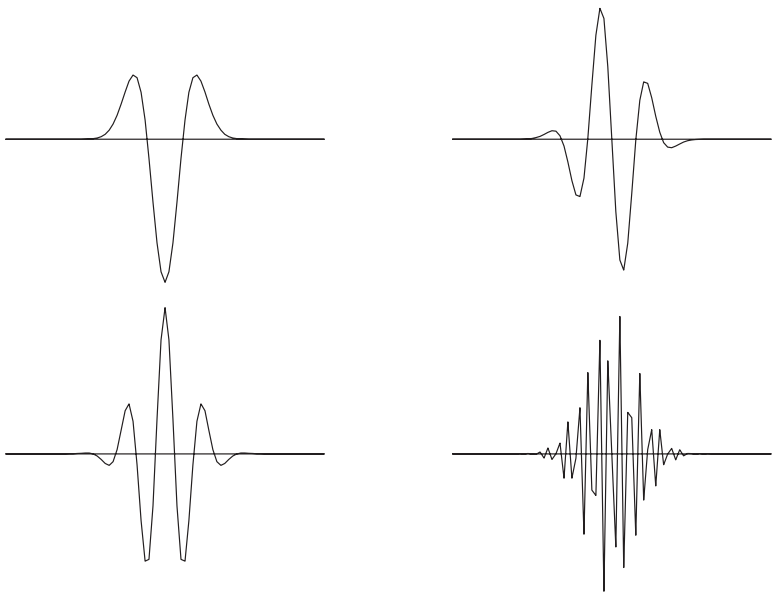

Figure 6.8: Some higher eigenfunctions for the hypercube/Ehrenfest problem with $n = 39$. Like the dominant eigenfunction of Figure 6.7, all are exponentially localized in the middle of the interval, and thus they have no natural connection to the evolution of initial transients.

oscillations (higher-order differences of Gaussians) that are also exponentially localized in the middle of the interval (Figure 6.8). In short, the system of eigenfunctions of this problem bears no natural relationship to the behavior of the system for short times. As we saw earlier with the investigation of condition numbers, the eigenvectors form an exponentially ill-conditioned basis in which to expand arbitrary states. These observations are the same as those that have been made in fluid mechanics [23] and in various other areas [21, 22].

In closing this section, let us return to the question of cutoff at $k = O(n)$ vs. $k = O(n \log n)$. Without a doubt, the $\log n$ factor is needed for convergence in the pointwise sense defined by $\|u^{(k)} - u^{(\infty)}\|_1$. However, *weak* convergence, as defined by the convergence to 0 of the discretizations of any integral

$$\int_0^1 [u^{(k)}(x) - u^{(\infty)}(x)]\,\varphi(x)\,dx$$

for smooth $\varphi(x)$, will require just $O(n)$ steps. (No particular constant in front of the n can be distinguished.) We might say that it takes $O(n)$ steps to satisfy a physicist, who is interested in integrated quantities along the lines of temperature or density, but $\sim \frac{1}{4} n \log n$ steps to defeat a gambler or a magician, who is prepared to beat the house by taking advantage of irregularities of probability on an arbitrarily fine scale.

7 Example 2: Riffle shuffle

Our treatment of riffle shuffling will be briefer. Delightfully, this seemingly complicated problem turns out to be closely analogous to the hypercube/Ehrenfest problem we have just discussed.

The famous Bayer–Diaconis paper treats a precisely defined shuffling model introduced by Gilbert and Shannon (1955) and Reeds (1981). Bayer and Diaconis describe the shuffle operation as follows [1]:

> A deck of n cards is cut into two portions according to a binomial distribution; thus the chance that k cards are cut off is $\binom{n}{k}/2^n$ for $0 \le k \le n$. The two packets are then riffled together in such a way that cards drop from the left or right heaps with probability proportional to the number of cards in each heap.

This shuffle model matches measurements of actual shuffling behavior reasonably well, and it is exceedingly elegant, more elegant than is perhaps apparent at first sight. We cannot detail its many properties, but urge the reader to look at [1].

This *riffle shuffle,* as we shall call it, defines a Markov chain on the state space of permutations of $\{1, \ldots, n\}$, of dimension $N = n!$. (For $n = 52$, $N \approx 8.1 \times 10^{67}$.) The associated matrices P and A are thus of dimensions $n! \times n!$ and nonsymmetric. It is for this problem that Bayer and Diaconis [1] proved the result (their Theorem 2) that there is a cutoff at

$$k_{\text{cutoff}} = \tfrac{3}{2} \log_2 n. \qquad \text{(riffle shuffle)} \tag{7.1}$$

It is remarkable that just as before, this Markov chain can be compressed to a problem of small dimension. Before, we introduced as a state variable the distance j from a distinguished vertex. Now, our new state variable will be r, *the number of rising sequences*. Once more we borrow the definition from Bayer and Diaconis [1]:

> A rising sequence is a maximal subset of an arrangement of cards, consisting of successive face values displayed in order. Rising sequences do not intersect, so each arrangement of a deck of cards is uniquely the union of its rising sequences. For example, the arrangement A,5,2,3,6,7,4 consists of the two rising sequences A,2,3,4 and 5,6,7, interleaved together.

Now suppose we start with a deck ordered from 1 to n. Bayer and Diaconis observe that the probability of being in any particular state after k shuffles depends only on r; this probability (their Theorem 1) is $\binom{2^k+n-r}{n}/2^{kn}$. What is more (their Corollary 2), the riffle shuffle induces a Markov chain on the space of numbers of rising sequences, that is, numbers r in the range $1 \le r \le n$. The dimension of the compressed Markov chain is thus $N = n$.

Bayer and Diaconis do not give a formula for the entries of the reduced $n \times n$ transition matrix P, though the entries have been determined in unpublished work by Gessel [1, 11]. We derived a formula ourselves and found

$$p_{ij} = 2^{-n}\binom{n+1}{2i-j}\frac{\alpha_j}{\alpha_i}, \tag{7.2}$$

where α_j is the number of permutations of $\{1,\dots,n\}$ that have j rising sequences. (Details can be obtained by contacting the first author.) These numbers α_j are known as *Eulerian numbers* and are given by the triangular recurrence

$$A_{1,1} = 1, \quad A_{1,k} = 0 \text{ for } k \ne 1, \quad A_{r,k} = kA_{r-1,k} + (r-k+1)A_{r-1,k-1} \tag{7.3}$$

with $\alpha_r = A_{n,r}$ [12, 16]. The stationary distribution for this Markov chain is

$$\sigma = (\alpha_1, \alpha_2, \dots, \alpha_n)/n!, \tag{7.4}$$

which gives us P^∞ by (2.3) and thence $A = P - P^\infty$. And, rather to our astonishment, we now find ourselves able to compute norms $\|A^k\|_1$ for the riffle shuffle problem with a 30-line MATLAB program, listed in the Appendix. We have checked that the results of this program (divided by 2, to match the "total variation norm" of the Markov chain literature) match all of the numbers reported in [1].

Figure 7.1 shows the results for card decks of size $n = 8$, 64, and 512. The dashed line represents $k_{\text{cutoff}} = \frac{3}{2}\log_2 n$. The cutoff phenomenon is pronounced, sharper than for the hypercube/Ehrenfest problem.

Figure 7.2 shows 1-norm pseudospectra of the decay matrix A with $n = 52$. The eigenvalues are known to be exactly 2^{-j} for $0 \le j \le n-1$ (with the eigenvalue 1 of P replaced by 0 for A), but the eigenvectors are now even more ill-conditioned, by a sizable margin, than for the hypercube/Ehrenfest problem.

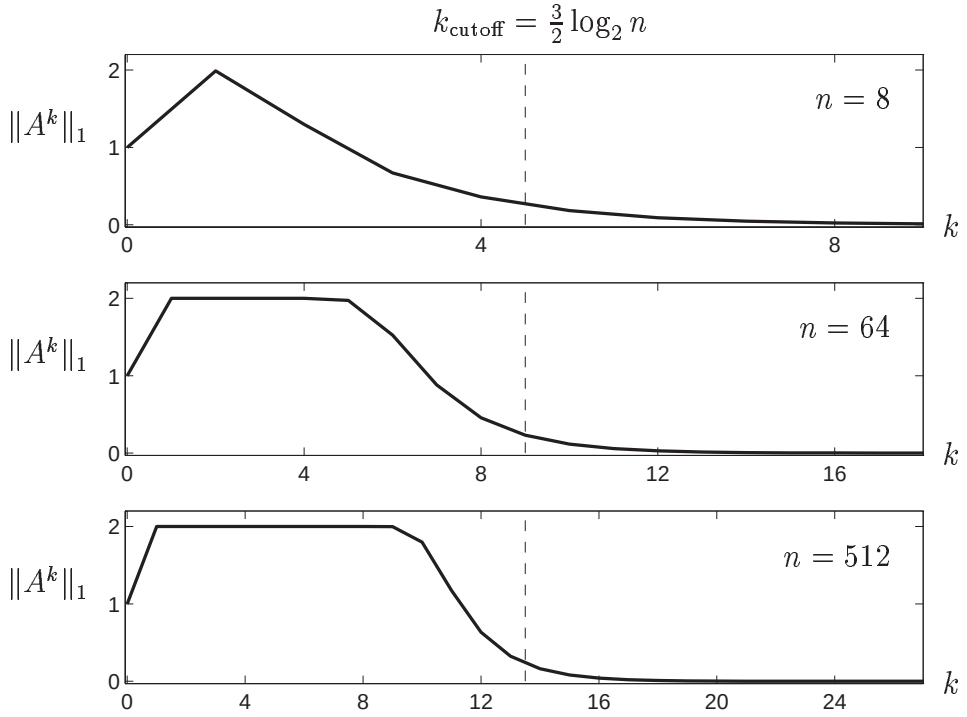

Figure 7.1: Like Figure 4.2, but for the riffle shuffle. As in Figure 4.2, these results are believed to be correct to plotting accuracy.

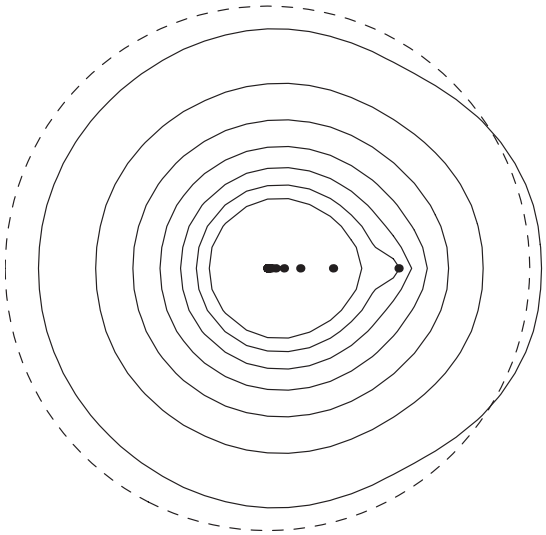

Figure 7.2: 1-norm pseudospectra as in Figure 5.2, but for the riffle shuffle with $n = 52$ ($\epsilon = 10^{-1}, 10^{-1.5}, 10^{-2}, \ldots, 10^{-4}$). The condition number of at least one matrix of eigenvectors is $\kappa_1(V) \approx \|V\|_1 \|V^{-1}\|_1 \approx 10^{60}$. The dashed curve again marks the unit circle.

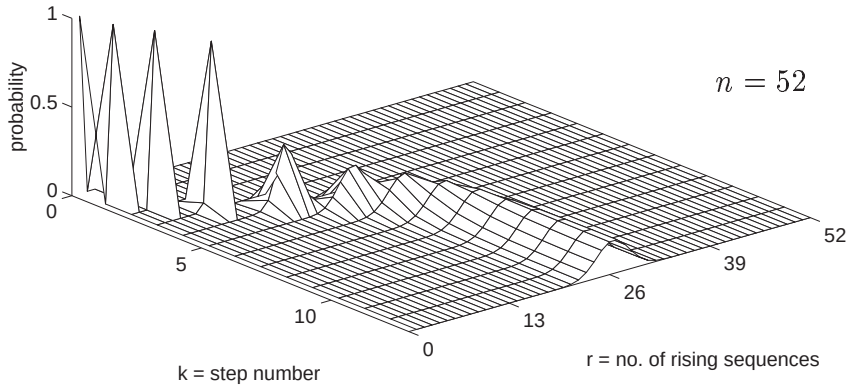

Figure 7.3: Like Figure 6.3, but for the riffle shuffle with $n = 52$. (There is no undersampling in this plot.) For this value of n, $k_{\mathrm{cutoff}} = \frac{3}{2}\log_2 n \approx 8.55$.

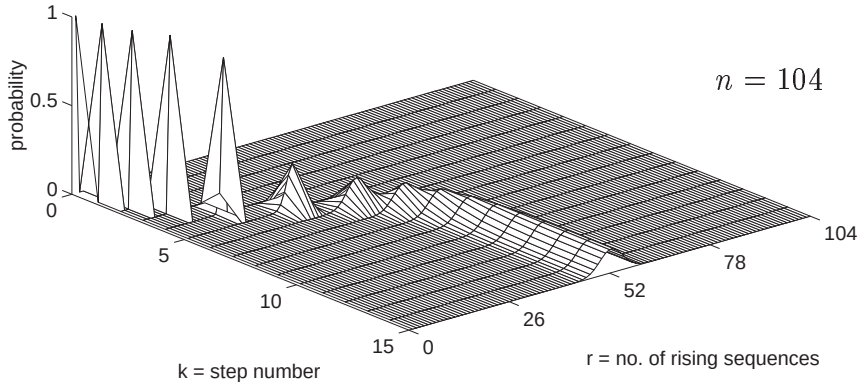

Figure 7.4: Same, except with $n = 104$ and $k_{\mathrm{cutoff}} = \frac{3}{2}\log_2 n \approx 10.1$.

Figures 7.3 and 7.4 are analogues of the earlier Figures 6.3 and 6.4. These figures reveal a closeness of the analogy between the hypercube/Ehrenfest and riffle shuffle problems that is striking indeed. The major difference is that whereas for the hypercube/Ehrenfest problem the wave propagates smoothly from one value of j to the next, for the riffle shuffle the value of r approximately doubles at each of the early steps. This is why there is no factor $O(n)$ in the formula for k_{cutoff}.

What happens to the trajectory as $n \to \infty$? What curve do Figures 7.3 and 7.4 approach?

Here, we get a surprise. The limit is not a smooth curve but a step, as shown in Figure 7.5. Moreover, the step occurs not at $\frac{3}{2}\log_2 n$ but at $\log_2 n$. Thus for the riffle shuffle, weak convergence occurs 33% sooner than norm convergence. It might be said that we need

- $\sim \log_2 n$ shuffles for physicists,
- $\sim \frac{3}{2}\log_2 n$ shuffles for gamblers and magicians.

(The physicists in question are presumed to be interested in integrated quantities with respect to the variable r, which may or may not cover many quantities of real interest.) We do not know if this distinction has been pointed out before.

Regrettably, the expression $\log_2 n$ is less dramatic than $\frac{3}{2}\log_2 n$, as it can be derived as a lower bound by very quick arguments, one of which (for a slightly different riffle shuffling model) is given in [15].

The formula $k_{\text{cutoff}} = \frac{3}{2}\log_2 n$ can be derived on the back of an envelope, by following the pattern of reasoning used for the hypercube/Ehrenfest problem. First, since r essentially doubles at each step initially until it becomes of size $O(n)$, $\sim \log_2 n$ steps are needed to achieve $r = O(n)$. From this point on, with $|\lambda_2| = \frac{1}{2}$ and the Gaussian pulse again of relative width $O(n^{-1/2})$, the convergence is governed by the analogue of (6.6),

$$2^{-k} = O(n^{-1/2}), \tag{7.5}$$

that is, $-k = -\frac{1}{2}\log_2 n$. Thus we need $k = \frac{1}{2}\log_2 n$ additional steps for norm convergence, bringing the total to $\frac{3}{2}\log_2 n$.

8 Compression of state spaces, and a multigrid analogy

In this section we give a few details of the compression process that makes our calculations feasible. For random walk on a hypercube, changing the state variable from the vertex to the distance of that vertex from vertex 0 reduces the dimensions of P and A from 2^n to $n+1$. For the riffle shuffle, changing the state variable from the permutation to the number of rising sequences in that permutation reduces the dimension from $n!$ to n. We now outline what these reductions look like from the point of view of linear algebra, and in particular, why they preserve $\|A^k\|_1$ and $\|(zI - A)^{-1}\|_1$.

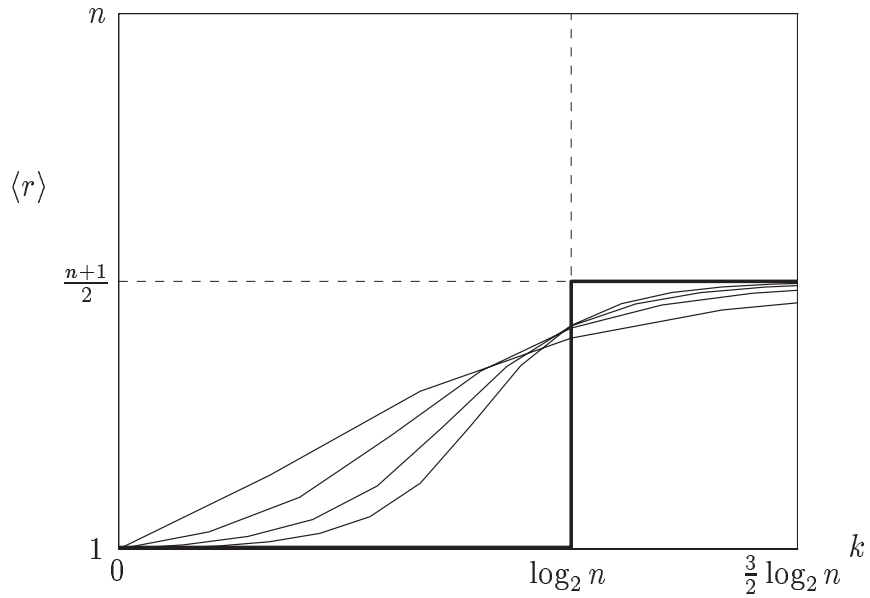

Figure 7.5: Analogue of Figure 6.5 for the riffle shuffle. The thinner curves show the expected number of rising sequences $\langle r \rangle$ as a function of step number for $n = 8, 32, 128, 512$. As $n \to \infty$, they converge to the step function marked as a thick curve. Thus for the riffle shuffle, weak convergence with respect to the variable r requires only $\sim \log_2 n$ steps, not $\sim \frac{3}{2} \log_2 n$.

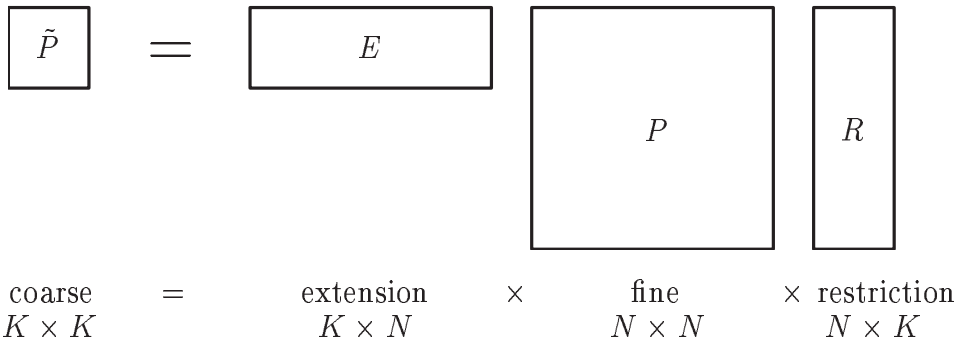

Figure 8.1: Sketch of the process of compression of certain Markov chains from dimension N to dimension K, which is analogous to the coarse-grid approximation used in multigrid iterations. Just as the transition matrices satisfy $\tilde{P} = EPR$, the decay matrices satisfy $\tilde{A} = EAR$ for the same E and R.

For this section only, we shall make use of special notation. We shall say that the original transition and decay matrices are P and A, of dimension N, and that the reduced matrices are $\tilde{P}$ and $\tilde{A}$, of dimension K ($= n+1$ for Ehrenfest urns, $= n$ for riffle shuffles).

Figure 8.1 summarizes the reduction from P to $\tilde{P}$. A high-level description is that E acts on the N rows of P or PR, combining them into K rows, while R acts on the N columns of P or EP, combining them into K columns. At the end we have a $K \times K$ matrix,

$$\tilde{P} = EPR. \tag{8.1}$$

To interpret this product as a composition of mappings, we must go from left to right, since Markov chain matrices act from the right on row vectors. Given a row vector in $\mathbb{R}^K$, extend it first to an N-vector by the "extension" matrix E. Then operate on that N-vector by the large matrix P. Finally, restrict the result back to $\mathbb{R}^K$ by the "restriction" matrix R. Readers familiar with multigrid iterations will recognize these as the same kind of operations used in that field to transfer functions between coarse and fine grids.

To see why the necessary 1-norms are preserved in this process, we must look at the structure of E, P, and R. For definiteness, we shall speak of the hypercube/Ehrenfest problem. The riffle shuffle problem is somewhat more complicated, but the arguments are still essentially the same.

Consider any row of P, corresponding to a particle beginning at one particular vertex of the n-cube. This is an N-vector representing probabilities that at the next step, the particle will be at each of the 2^n vertices. All of these numbers are 0, except for $n+1$ of them, which are equal to $1/(n+1)$.

The matrix R adds together all columns of P corresponding to vertices at the same distance j from vertex 0. Thus its effect on a particular row of P is to combine all the probabilities corresponding to vertices at a fixed distance j. In the case $n = 3$, for example, if the vertices are ordered by distance from vertex 0, we have

$$R = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & 1 & & \\ & 1 & & \\ & & 1 & \\ & & 1 & \\ & & 1 & \\ & & & 1 \end{pmatrix}.$$

For any n, R contains a single 1 in each row, implying $\|R\|_1 = 1$. If the columns are numbered from 0 to n, then column j contains $\binom{n}{j}$ entries equal to 1.

The matrix E averages together all rows of P corresponding to vertices at the same

distance j from vertex 0. For $n = 3$, it looks like this:

$$E \;=\; \begin{pmatrix} 1 & & & & & & & \\ & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & & & & \\ & & & & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \\ & & & & & & & 1 \end{pmatrix}.$$

If the rows are numbered from 0 to n, then row j contains $\binom{n}{j}$ nonzero entries, all equal to $1/\binom{n}{j}$, implying $\|E\|_1 = 1$. There is a single nonzero in each column.

The matrices E and R are in a sense dual, and indeed, we have $ER = I$ and $E = DR^T$ for an appropriate diagonal matrix D. A similar formula applies in multigrid iterations.

In the end, $\tilde{P}$ can be regarded as a transition matrix for the hypercube in which all vertices with fixed j have been identified both on input (by E) and on output (by R).

We now review some of the properties that follow from this construction of $\tilde{P}$, without giving details.

First, (8.1) generalizes to the formula

$$\tilde{P}^k \;=\; EP^kR \tag{8.2}$$

for any $k \geq 0$. The reason for this is that if a probability distribution on the hypercube depends only on the distance j from vertex 0, then that symmetry is preserved by repeated applications of P.

Second, $\tilde{A} = EAR$, and more generally, for any $k \geq 0$,

$$\tilde{A}^k \;=\; EA^kR. \tag{8.3}$$

This follows from (8.2) together with the definitions $A = P - P^\infty$, $\tilde{A} = \tilde{P} - \tilde{P}^\infty$.

Third, $\tilde{P}$ and P and likewise $\tilde{A}$ and A have the same eigenvalues, though with very different multiplicities.

Fourth, the compression process also applies to the resolvent matrix: for any z not equal to an eigenvalue,

$$(zI - \tilde{A})^{-1} \;=\; E(zI - A)^{-1}R. \tag{8.4}$$

For $|z| > 1$, this follows from the convergent Taylor series of the resolvent together with (8.3). The extension to general z can be justified by arguments of analytic continuation.

Finally, all of these reductions preserve 1-norms:

$$\|\tilde{P}^k\|_1 = \|P^k\|_1 = 1, \quad \|\tilde{A}^k\|_1 = \|A^k\|_1, \quad \|(zI - \tilde{A})^{-1}\|_1 = \|(zI - A)^{-1}\|_1. \tag{8.5}$$

For each of these equalities, the direction "$\leq$" is trivial since $\|E\|_1 = \|R\|_1 = 1$. The direction "$\geq$" follows from the sequence

$$\|\tilde{B}\|_1 \;\geq\; \|e_1\tilde{B}\|_1 \;=\; \|e_1B\|_1 \;=\; \|B\|_1, \tag{8.6}$$

where B is any of the matrices of interest above and e_1 is the N-vector of the form $(1, 0, 0, \dots)$. The inequality in (8.6) follows from (2.2). The first equality follows from the facts that the first row of E contains just a single number 1 and that the entries in any of the groups of entries in the first row of B added up by the action of a column of R are equal, so there is no cancellation in the addition. The second equality follows from the symmetry property that for any of the $N \times N$ matrices in question, all the rows have identical entries, though in permuted orders.

The identities (8.5) justify the claims of this paper that Figures 4.2, 5.2, 7.1 and 7.2 apply equally to the original or the reduced hypercube/Ehrenfest and riffle shuffle problems. We have also confirmed this numerically by various checks.

9 Summary

In conclusion, we wish to emphasize that we do not claim to have done anything that is mathematically new. The technical results we have presented are mainly due to Aldous, Bayer, Diaconis, Shahshahani, and others. On the other hand, the "hands on" style of this paper is entirely unlike what is usual in the Markov chain literature. If by presenting the cutoff phenomenon in this nonstandard way we have helped to increase the flow of ideas between probabilists and numerical analysts, we will be pleased.

We regret that we have not been able to state or prove theorems, as this would have made a paper with this broad scope impossibly long.

There is more to the study of transients in Markov chains than one finds in the literature related to the cutoff phenomenon. For an example of quite a different study in this general area, see [19].

We finish with an itemization of some of the points we have made.

1. The cutoff phenomenon is not usually expressed in the Markov chain literature in terms of norms of powers of the decay matrix A, but it can be readily reformulated in this way.

2. The dimension of A is often combinatorially large, but in some cases, at least, grouping of states may compress the problem to MATLAB size while preserving the crucial norms $\|A^k\|_1$ and $\|(zI - A)^{-1}\|_1$. The compression process is analogous to the coarse-grid approximation process that is the basis of multigrid iterations.

3. The cutoff phenomenon depends on the use of the 1-norm (for vectors, equal to half the probabilists' "total variation norm"). Non-normality is not always involved. On the other hand for non-normal examples, cutoffs can occur when the eigenvalues of the transition matrix are simple and well separated, so one cannot say that the cutoff phenomenon depends on multiple eigenvalues.

4. Cutoffs are associated with 1-norm pseudospectra that are far from the spectrum and with eigenvectors that have no natural relationship to the transient behavior.

5. The hypercube/Ehrenfest and riffle shuffle examples share an elementary ex-

planation: a probability wave must propagate from one place to another before convergence can occur. The asymptotic convergence process, where eigenvectors become relevant, involves alignment of Gaussians.

6. For the hypercube/Ehrenfest problem, weak convergence is achieved in $O(n)$ steps, and $\sim \frac{1}{4} n \log n$ steps are needed for norm convergence.

7. For the riffle shuffle example, weak convergence is achieved in $\sim \log_2 n$ steps, and $\sim \frac{3}{2} \log_2 n$ steps are needed for norm convergence.

10 Appendix: Matlab programs

```
  function [P,A] = ehr(n)

%   EHR   Ehrenfest urns matrices.
%
%         EHR(n) is the (n+1)x(n+1) transition matrix P corresponding
%         to n balls in two urns.
%
%         [P,A] = EHR(n) gives also the (n+1)x(n+1) decay matrix A = P-P^inf.

% Construct tridiagonal transition matrix:
  P = eye(n+1)/(n+1);
  for i = 1:n;
    P(i+1,i) = i/(n+1);
    P(i,i+1) = (n+1-i)/(n+1);
  end

% Compute stationary distribution and use it to construct P^inf:
  if nargout==2
    v = eye(1,n+1);
    vnew = eye(1,n+1);
    for j = 2:n+1
      vnew(1) = v(1);
      vnew(2:j) = v(2:j)+v(1:j-1);
      v = vnew/2;
    end
    Pinf = ones(n+1,1)*v;
    A = P - Pinf;
  end
```

```
  function [P,A] = riffle(n)

% RIFFLE  Riffle shuffle matrices.
%
%         RIFFLE(n) is the nxn transition matrix P corresponding
%         to the Gilbert-Shannon-Reeds riffle shuffle.
%
%         [P,A] = RIFFLE(n) gives also the nxn decay matrix A = P-P^inf.

% Compute logs of Eulerian numbers:
  a = zeros(1,n);
  anew = zeros(1,n);
  for j = 2:n
    anew(2:j-1) = log((2:j-1).*exp(a(2:j-1)-a(1:j-2))+(j-1:-1:2))+a(1:j-2);
    a = anew;
  end

% Compute logs of binomial coefficients:
  b = zeros(1,n+2);
  bnew = zeros(1,n+2);
  for j = 2:n+2
    bnew(2:j-1) = log(exp(b(2:j-1)-b(1:j-2))+1)+b(1:j-2);
    b = bnew;
  end

% Construct transition matrix P
  b = b-n*log(2);
  r = [b(1) -Inf*ones(1,n-1)];
  c = [b -Inf*ones(1,n-2)]';
  T = toeplitz(c,r);
  P = T(2:2:2*n,:);
  P = exp(P-a'*ones(1,n)+ones(n,1)*a);

% Compute stationary distribution and use it to construct P^inf:
  if nargout==2
    v = eye(1,n);
    vnew = eye(1,n);
    for j = 2:n
      vnew(1) = v(1);
      vnew(2:j) = (2:j).*v(2:j)+(j-1:-1:1).*v(1:j-1);
      v = vnew/j;
    end
    Pinf = ones(n,1)*v;
    A = P - Pinf;
  end
```

References

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

[2] P. Diaconis, *Group Representations in Probability and Statistics,* IMS, Hayward, CA, 1988.

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

[4] P. Diaconis, R. L. Graham, and J. A. Morrison, "Asymptotic analysis of a random walk on a hypercube with many dimensions," Random Struct. and Alg. **1**, 51–72 (1990).

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

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

[7] J. L. M. van Dorsselaer, J. F. B. M. Kraaijevanger, and M. N. Spijker, "Linear stability analysis in the numerical solution of initial value problems," Acta Numerica 1992, Cambridge U. Press, 1992, pp. 199–237.

[8] T. A. Driscoll, K.-C. Toh, and L. N. Trefethen, "From potential theory to matrix iterations in six steps," SIAM Review, to appear.

[9] P. Ehrenfest and T. Ehrenfest, "Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem," Phys. Zeit. **8**, 311–314 (1907).

[10] W. Feller, *An Introduction to Probability Theory and its Applications,* 3rd ed., Wiley, New York, 1968.

[11] I. Gessel, personal communication, August, 1997.

[12] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics,* Addison-Wesley, Reading, Massachusetts, 1989.

[13] D. J. Higham and L. N. Trefethen, "Stiffness of ODEs," BIT **33**, 285–303 (1993).

[14] M. Kac, "Random walk and the theory of Brownian motion," Amer. Math. Monthly **54**, 369–391 (1947).

[15] J. Keller, "How many shuffles to mix a deck?," SIAM Review **37**, 88–89 (1995).

[16] D. E. Knuth, *The Art of Computer Programming, v. 3: Sorting and Searching,* Addison-Wesley, Reading, Mass., 1973.

[17] G. Kolata, "In shuffling cards, 7 is winning number," New York Times, sec. C, p. 1, January 9, 1990.

[18] J. S. Rosenthal, "Convergence rates for Markov chains," SIAM Review **37**, 387–405 (1995).

[19] G. W. Stewart, "On Markov chains with sluggish transients," to appear in Communication in Statistics, Stochastic Models.

[20] H. M. Taylor and S. Karlin, *An Introduction to Stochastic Modeling,* Academic Press, 1984.

[21] L. N. Trefethen, "Pseudospectra of matrices." In: D. F. Griffiths and G. A. Watson, eds., Numerical Analysis 1991, Longman Scientific and Technical, Harlow, Essex, UK, 1992.

[22] L. N. Trefethen, "Pseudospectra of linear operators," SIAM Review, to appear, 1998.

[23] L. N. Trefethen, A. E. Trefethen, S. C. Reddy, and T. A. Driscoll, "Hydrodynamic stability without eigenvalues," Science **261**, 578–584 (1993).

[24] E. Wegert and L. N. Trefethen, "From the Buffon needle problem to the Kreiss matrix theorem," Amer. Math. Monthly **101**, 132–139 (1994).

Acknowledgement

This paper was prompted in part by an inspiring course taught at Cornell by Persi Diaconis in the fall of 1996. For help and advice along the way we are grateful to Diaconis and to Ira Gessel, Jing Huang, Ilse Ipsen, Yohan Kim, Gilbert Strang, and Divakar Viswanath. It was Kim who first had the idea of drawing plots like those of Figures 6.3, 6.4, 7.3, and 7.4. Our work was supported by NSF Grant DMS-9500975CS.

Gudbjorn F. Jonsson
Center for Applied Mathematics
Cornell University
Ithaca, New York 14853
USA
jonsson@cam.cornell.edu

Lloyd N. Trefethen
Oxford University Computing Laboratory
Wolfson Building, Parks Road
Oxford OX1 3QD
UK
nick.trefethen@comlab.ox.ac.uk