Chapter 3. Random Walks on Groups

A. Examples

A fair number of real world problems lead to random walks on groups. This section contains examples. It is followed by more explicit mathematical formulations and computations.

1. Random walk on the circle and random number generation

Think of Z_p (the integers mod p) as p points wrapped around a discrete circle. The simplest random walk is a particle that moves left or right, each with probability $\frac{1}{2}$. We can ask: how many steps does it take the particle to reach a given site? How many steps does it take the particle to hit every site? After how many steps is the distribution of the particle close to random? In Section C, we show that the answer to all of these questions is about p^2.

A class of related problems arises in computer generation of pseudo random numbers based on the recurrence $X_{k+1} = aX_k + b(\text{mod } p)$ where p is a fixed number (often 2^{32} or the prime $2^{31} - 1$) and a and b are chosen so that the sequence $X_0 = 0$, X_1, $X_2, \ldots,$ has properties resembling a random sequence. An extensive discussion of these matters is in Knuth (1981).

Of course, the sequence X_k is deterministic and exhibits many regular aspects. To increase randomness several different generators may be combined or "shuffled." One way of shuffling is based on the recurrence $X_{k+1} = a_k X_k + b_k$ (mod p) where (a_k, b_k) might be the output of another generator or might be the result of a "true random" source as produced by electrical or radioactive noise. We will study how a small amount of randomness for a and b spreads out to randomness for the sequence X_k.

If $a_k \equiv 1$ and b_k takes values ± 1 with probability $\frac{1}{2}$, we have a simple random walk. If $a_k \not\equiv 1$ is fixed but nonrandom, the resulting process can be analyzed by using Fourier analysis on Z_p. In Section C we show that if $a_k \equiv 2$, then about $\log p \log\log p$ steps are enough to force the distribution of X_k to be close to uniform (with b_k taking values 0, ± 1 uniformly). This is a great deal faster than the p^2 steps required when $a_k \equiv 1$. If $a_k \equiv 3$, then $\log p$ steps are enough.

What if a_k is random? Then it is natural to study the problem as a random walk on A_p - the affine group mod p. This is the set of pairs (a, b) with $a, b \in Z_p$, $a \neq 0$, $gcd(a, p) = 1$. Multiplication is defined by

$$(a, b)(c, d) = (ac,\ ad + b).$$

Some results are in Example 4 of Section C, but many simple variants are unsolved.

A different group arises when considering the second order recurrence $X_{k+1} = a_k X_k + b_k X_{k-1} \pmod p$ with a and b random. It is natural to define $Y_k = \binom{X_k}{X_{k-1}}$, then

$$Y_{k+1} = \begin{pmatrix} a_k & b_k \\ 1 & 0 \end{pmatrix} Y_k = \left[\Pi \begin{pmatrix} a_i & b_i \\ 1 & 0 \end{pmatrix}\right] Y_0, \text{ with say } Y_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

This leads to considering a product of random matrices, and so to a random walk on $GL_2(Z_p)$. See Diaconis and Shahshahani (1986a) for some results.

2. Card Shuffling

How many times must a deck of cards be shuffled until it is close to random? Historically, this was a fairly early application of probability. Markov treated it as one of his basic examples of a Markov chain (for years, the only other example he had was the vowel/consonant patterns in Eugene Onegin). Poincare devoted an appendix of his 1912 book on probability to the problem, developing methods similar to those in Section C. The books by Doob (1935) and Feller (1968) each discuss the problem and treat it by Markov chain techniques.

All of these authors show that any reasonable method of shuffling will *eventually* result in a random deck. The methods developed here allow explicit rates that depend on the deck size. As will be explained, these are much more accurate than the rates obtained by using bounds derived from the second largest eigenvalue of the associated transition matrix.

Some examples of specific shuffles that will be treated below:

a) *Random transpositions.* Imagine n cards in a row on a table. The cards start in order, card 1 at the left, card 2 next to it,. . ., and card n at the right of the row. Pairs of cards are randomly transposed as follows: the left had touches a random card, and the right hand touches a random card (so left = right with probability $\frac{1}{n}$). The two cards touched are interchanged. A mathematical model for this process is the following probability distribution on the symmetric group:

$$\begin{aligned} T(\text{id}) &= \frac{1}{n} \\ T(\tau) &= \frac{2}{n^2} \text{ for } \tau \text{ any transposition} \\ T(\pi) &= 0 \quad \text{otherwise.} \end{aligned}$$

Repeatedly transposing cards is equivalent to repeatedly convolving T with itself. It will be shown that the deck is well mixed after $\frac{1}{2}n \log n + cn$ shuffles.

Some variants will also be discussed: repeatedly transposing a random card with the top card (la Librairie de la Marguerite), or repeatedly interchanging a card with one of its neighbors.

b) *Borel's shuffle.* In a book on the mathematics of Bridge, Borel and Cheron (1955) discuss the mathematics of shuffling cards at length. They suggest several open problems; including the following shuffle: The top card of a deck is removed and inserted at a random position, then the bottom card is removed and inserted at a random position. This is repeated k times. We will analyze such procedures in Chapter 4, showing that $k = n \log n + cn$ "moves" are enough. The same techniques give similar rates for the shuffle that repeatedly puts a random card on top, or the shuffle that repeatedly removes a card at random and replaces it at random.

c) *Riffle shuffles.* This is the usual way that card players shuffle cards, cutting off about half the pack and riffling the two packets together. In Chapter 4 we will analyze a model for such shuffles due to Gilbert, Shannon, and Reeds. We will also analyze records of real riffle shuffles. The analysis suggests that 7 shuffles are required for 52 cards.

d) *Overhand shuffles.* This is another popular way of shuffling cards. The following mathematical model seems reasonable: the deck starts face down in the hand. Imagine random zeros and ones between every pair of cards with a zero under the bottom card of the deck. Lift off all the cards up to the first zero and place them on the table. Lift off all the cards up to the second zero and place this packet on top of the first removed packet. Continue until no cards remain. This is a single shuffle. It is to be repeated k times. Robin Pemantle (1988) has shown that about 2500 shuffles are required for 52 cards.

3. Random walk on the d-cube Z_2^d

Regard Z_2^d as the vertices of a cube in d dimensions. The usual random walk starts at a point and moves to one of the d neighbors with probability $\frac{1}{d}$. This is repeated k times. This is a nice problem on its own. It has a surprising connection with a classical problem in statistical mechanics: in the Ehrenfest's urn model, d balls are distributed in two urns. A ball is chosen at random and moved to the other urn. This is repeated k times and the problem is to describe the limiting distribution of the process. For a fascinating description of the classical approach see M. Kac (1947). Kac derives the eigenvalues and eigenvectors of the associated transition matrix by a tour de force. The following approach due to Siegert (1949) suggests much further research:

Let the state of the system be described by a binary vector of length d, with a 1 in the ith place denoting that ball i is in the right hand urn. The transition mechanism translates precisely to a random walk on the d cube! Indeed, the state changes by picking a coordinate and changing to its opposite mod 2. This changes the problem into analyzing the behavior of a random walk on an Abelian group. As we will see, this is straightforward; Fourier analysis gives all the eigenvalues and eigenvectors of the associated Markov chain.

Originally the state of the system in the Ehrenfest's urn was the number of balls in the right hand urn. The problem was "lifted" to a random walk on a group. That is, there was a group G (here Z_2^d) and a probability P on G (here move to the nearest neighbor) and a function $L: G \rightarrow$ state space (here the number of ones) such that the image law under L of the random walk was the given Markov chain. There has been some study of the problem of when the image of a Markov chain is Markov. Heller (1965) contains much of interest here. Mark Kac was fascinated with this approach and asked: When can a Markov chain be lifted to a random walk on a group? Diaconis and Shahshahani (1987b) give results for "Gelfand Pairs." The following exercise comes out of discussions with

Mehrdad Shahshahani.

EXERCISE 1. Let P be a probability on the symmetric group S_n. Think of the random walk generated by P as the result of repeatedly mixing a deck of n cards. For a permutation π, let $L(\pi) = \pi(1)$. The values of L are the result of following only the position of card 1. Show that the random walk induces a Markov chain for L. Show that this chain has a doubly stochastic transition matrix. Conversely, show that for any doubly stochastic matrix, there is a probability P on S_n which yields the given matrix for L.

Remark. It would be of real interest to get analogs of this result more generally. For example: find conditions on a Markov chain to lift to a random walk on an Abelian group. Find conditions on a Markov chain to lift to a random walk with a probability P that is constant on conjugacy classes. When can a Markov chain on the ordinary sphere be lifted to a random walk on the orthogonal group O_3?

Returning to the cube, David Aldous (1983b) has applied results from random walk on the d cube to solve problems in the theory of algorithms. Eric Lander (1986) gives a very clear class of problems in DNA gene mapping which really involves this process. Diaconis and Smith (1986) develop much of the fluctuation theory of coin-tossing for the cube. There is a lot going on, even in this simple example.

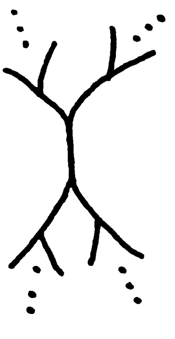
4. INFINITE GROUPS

For the most part, these notes treat problems involving finite groups. However, the techniques and questions are of interest in solving applied problems involving groups like the orthogonal group and p-adic matrix groups. Here is a brief description.

1. *The "Grand Tour" and a walk on O_n.* Statisticians often inspect high-dimensional data by looking at low-dimensional projections. To give a specific example, let $x_1, \ldots, x_{500} \in \mathbb{R}^{20}$ represent data on the Fortune 500 companies. Here x_1, the data for company 1, might have coordinates x_{11} = total value, x_{12} = number of women employed, etc. For $\gamma \in \mathbb{R}^{20}$, the projection in direction γ would be a plot (say a histogram) of the 500 numbers $\gamma \cdot x_1, \ldots, \gamma \cdot x_{500}$. Similarly the data would be projected onto various two-dimensional spaces and viewed as a scatterplot. Such inspection is often done interactively at a computer's display screen, and various algorithms exist for changing the projection every few seconds so that a scientist interested in the data can hunt for structured views.

Such algorithms are discussed by D. Asimov (1983). In one such, the direction γ changes by a small, random rotation. Thus, one of a finite collection Γ_i of 20×20 orthogonal matrices would be chosen at random, and the old view is rotated by Γ_i. This leads to obvious questions such as, how long do we have to wait until the views we have seen come within a prescribed distance (say 5 degrees) of any other view. A good deal of progress on this problem has been made by Peter Matthews in his Stanford Ph.D. thesis. Matthews (1988a) uses Fourier analysis on the orthogonal group and diffusion approximations to get useful numerical and theoretical results.

2. *Salmon fishing and* $GL_2(Q_2)$. Consider a binary tree



Let us describe a homogeneous random walk on such a tree. A particle starts at a fixed vertex. An integer distance is picked from a fixed measure, and the particle moves to one of the finite sets of vertices at this distance at random. The particle continues to move in this way. Questions involving recurrence (does the particle ever get back to where it started?) and the distribution of the distance of the walk from its starting position were raised by population geneticists studying life along a river system.

Sawyer (1978) gives background and much elegant analysis. It turns out that the tree is a coset space (homogeneous space) of the 2×2 matrices with entries in the 2-adic rationals, with respect to the subgroup of matrices with 2-adic integer entries. Number theorists have worked out enough of the representation theory to allow a dedicated probabilist to get elegant formulas and approximations.

3. *Other groups.* There is of course vast literature on random walks on $\mathbb{R}^n$. This is summarized in Feller (1971) or in Spitzer (1964). Much of this material has been generalized to non-commutative groups. Heyer (1977) contains a thorough survey. Recently there has been a really successful attack on random walk problems on Lie groups. The work of Furstenberg and Guivarclh is beautifully summarized in Bougerol-Lacroix (1985).

B. The basic setup

We now formally define what we mean by "close to random" and introduce an inequality that allows a good bound on the distance to uniformity in terms of Fourier transforms. Let G be a finite group. Let P and Q be probability distributions on G. Define the *variation distance* between P and Q as

$$||P - Q|| = \max_{A \subset G} |P(A) - Q(A)|.$$

Because we will use it heavily, we pause to discuss some basic properties.

Exercise 2. Prove that

$$||P - Q|| = \frac{1}{2} \sum_s |P(s) - Q(s)| = \frac{1}{2} \max_{||f|| \leq 1} |P(f) - Q(f)|, \tag{1}$$

where, in the last expression, f is a function from G to $\mathbb{R}$ with $|f(s)| \leq 1$, and $P(f) = \Sigma_s P(s)f(s)$ is the expected value of f under P. Also, prove the validity of the following interpretation of variation distance suggested by Paul Switzer: Given a single observation, coming from P or Q with probability $\frac{1}{2}$, you are to guess which. Your chance of being correct is $\frac{1}{2} + \frac{1}{2}||P - Q||$.

EXERCISE 3. Show that if U is uniform, and $h: G \to G$ is $1 - 1$, then

$$||P - U|| = ||Ph^{-1} - U|| \text{ where } Ph^{-1}(A) = P(h^{-1}(A)).$$

EXERCISE 4. Let $G = S_n$. Part (a): let P be defined by "card 1 is on top, all the rest are random." Thus, $P(\pi) = 0$ if $\pi(1) \neq 1$ and $P(\pi) = 1/(n-1)!$ otherwise. What is $||P-U||$? Part (b): suppose P is defined by "card 1 is randomly distributed in a fixed set A of positions, all the other cards are random?" What is $||P - U||$?

Further properties of the variation distance are given in the following remarks and in lemma 4 of Chapter 3, lemma 4 of Chapter 4 and lemma 5 of Chapter 4.

Remark 1. The variation distance can be defined for any measurable group. It makes the measures on G into a Banach space. For G compact, the measures are the dual of the bounded continuous functions and $||\ ||$ is the dual norm. For continuous groups, the variation distance is often not suitable, since the distance between a discrete and continuous probability is 1. In this case, one picks a natural metric on G, and uses this to metrize the weak-star topology. Of course, for finite groups, all topologies are equivalent and the variation distance is chosen because of the natural interpretation given by (1): if two probabilities are close in variation distance, they give practically the same answer for any question.

Remark 2. Consider a random walk on S_n. In the language of shuffling cards, it might be thought that the following notion would be a more suitable definition of when cards are close to uniformly well shuffled: suppose the cards are turned face up one at a time and we try to guess at the value of each card before it is shown. For the uniform distribution, as in Diaconis and Graham (1977), we expect to get $H_n = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}$ right on average. If the deck is not well mixed, the increase in the number of cards we can guess correctly seems like a useful measure of how far we are from uniform. Formally, one may define a guessing strategy for each possible history. Its value on a given permutation π defines a function $f(\pi)$ and (1) shows that, on average, $|P(f) - H_n| < n||P - U||$ no matter what guessing strategy is used. This may serve as a guide for how small a distance $||P - U||$ to aim for.

Remark 3. The variation distance is closely related to a variety of other metrics. For example, two other widely used measures of distance between probabilities are

$$d_H(P,Q) = \sum_s (P(s)^{\frac{1}{2}} - Q(s)^{\frac{1}{2}})^2 - \text{ Hellinger distance}$$

$$I(P,Q) = \sum_s P(s) \log[P(s)/Q(s)] - \text{ Kullback-Leibler separation.}$$

These satisfy

$$\frac{d_H}{2} \leq || \quad || \leq \sqrt{d_H(1 - d_H/4)} \leq \sqrt{d_H}$$

$$\sqrt{2}|| \quad || \leq \sqrt{I}.$$

It follows that when d_H or I are small, the variation distance is small. The converse can be shown to hold under regularity conditions.

Metrics on probabilities are discussed by Dudley (1968), Zolatorev (1983) and Diaconis and Zabell (1982). Rachev (1986) is a recent survey.

THE BASIC PROBLEM.

We can now ask a sharp mathematical question: Let P be a probability on G. Given $\varepsilon > 0$, how large should k be so that $||P^{*k} - U|| < \varepsilon$?

It is not hard to show that P^{*k} tends to uniform if P is not concentrated on a subgroup or a coset of a subgroup. Here is a version of the theorem due to Koss (1959):

Theorem 1. *Let G be a compact group. Let P be a probability on G such that for some n_0 and $c, 0 < c < 1$, for all $n > n_0$,*

$$(*) \qquad P^{*n}(A) > cU(A) \text{ for all open sets } A.$$

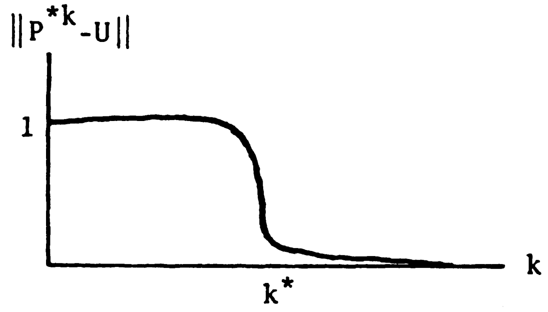
Then, for all k,

$$||P^{*k} - U|| \leq (1 - c)^{[k/n_0]}.$$

Remarks. Condition * rules out periodicity. The conclusion shows that eventually the variation distance tends to zero exponentially fast. The result seems quantitative, but it's hard to use it to get bounds in concrete problems: as an example, consider simple random walk on Z_n. How large must k be to have the distance to uniform less than $\frac{1}{10}$? To answer, we must determine a suitable n_0 and c. This seems difficult. A short proof of the theorem is given here in Chapter 4.

There is a huge literature relating to this theorem. Heyer (1977) contains an extensive survey. Bhattacharya (1972) and Major and Shlossman (1979) contain quantitative versions which are more directly useable. Csiszar (1962) gives a proof which indicates "why it is true": briefly, convolving increases entropy and the maximum entropy distribution is the uniform. Bondesson (1983) discusses repeated convolutions of different measures.

Remark. The following "cut off" phenomena occurs in most cases where the computations can be done: the variation distance, as a function of k, is essentially 1 for a time and then rapidly becomes tiny and tends to zero exponentially fast past the cut off. Thus a graph might appear



We will determine these cut off points k^* in most of the examples discussed in Theorem 1. In such a case we will say that k^* steps suffice.

One purpose of this chapter is to discuss several ways of approximating the variation distance that give sharp non-asymptotic results. The basic tool used in the analytical approach of this section is the following inequality first used in Diaconis and Shahshahani (1981):

LEMMA 1. *(Upper bound lemma). Let Q be a probability on the finite group G. Then*

$$||Q-U||^2 \le \frac{1}{4}\Sigma^* d_\rho \; Tr(\hat{Q}(\rho)\hat{Q}(\rho)^*)$$

where the sum is over all non-trivial irreducible representations.

Proof. From (1),

$$\begin{aligned}4||Q-U||^2 &= \{\Sigma_s|Q(s)-U(s)|\}^2 \le |G|\Sigma|Q(s)-U(s)|^2\\ &= \Sigma^* d_\rho \; \mathrm{Tr}(\hat{Q}(\rho)\hat{Q}(\rho)^*).\end{aligned}$$

The inequality is Cauchy-Schwarz. The final equality is the Plancherel Theorem, and $\hat{Q}(\rho)=1$ for ρ trivial, $\hat{U}(\rho)=0$ for ρ non-trivial. □

Remark 1. The Cauchy-Schwarz inequality is not as crude as it may first appear. It is applied when Q is close to uniform, so $|Q(s)-U(s)|$ is roughly constant. In the examples of Section II below, and in all other "real" examples, the lemma gives the right answer in the sense that the upper bound matches a lower bound to one or two terms. The following exercise gives a lower bound of similar form. For some groups it shows the rate is off by at worst $\log|G|$. Exercise 14 gives a natural example, and Exercise 6 a contrived example, where this occurs.

EXERCISE 5. With the notation of the upper bound lemma, show that

$$||Q-U|| \ge \frac{1}{2|G|}\Sigma^* d_\rho \; \mathrm{Tr}(\hat{Q}(\rho)\hat{Q}(\rho)^*).$$

Also show

$$||Q-U||^2 \ge \frac{1}{4|G|}\Sigma^* d_\rho \; \mathrm{Tr}(\hat{Q}(\rho)\hat{Q}(\rho)^*).$$

EXERCISE 6. Let G be a finite group. Define a probability P on G by

$$P(\mathrm{id}) = 1 - \frac{\varepsilon}{2},\ P(s) = \frac{\varepsilon}{2(|G|-1)} \text{ for } s \neq \mathrm{id},\ 0 \le \varepsilon \le 2.$$

Show that

$$P^{*k}(\mathrm{id}) = \frac{1}{|G|} + \frac{|G|-1}{|G|}(1 - \frac{\varepsilon}{2}\frac{|G|}{|G|-1})^k$$
$$P^{*k}(s) = \frac{1}{|G|} - \frac{1}{|G|}(1 - \frac{\varepsilon}{2}\frac{|G|}{|G|-1})^k.$$

Using this, show that $||P^{*k} - U|| = \frac{|G|-1}{|G|}\ |1 - \frac{\varepsilon}{2}\frac{|G|}{|G|-1}|^k$. Show that

$$\Sigma^* d_\rho \operatorname{Tr}(\hat{P}(\rho)^k \hat{P}) = (|G|-1)(1 - \frac{\varepsilon}{2}\frac{|G|}{|G|-1})^{2k}.$$

Remark 2. Lower bounds can be found by choosing a set $A \subset G$ and using $|Q(A) - U(A)| \le ||Q - U||$. Often A can be chosen so that it is possible to calculate, or approximate, both $Q(A)$ and $U(A)$, *and* show that the distance between them is large. Several examples are given in the next section.

Remark 3. Total variation is used almost exclusively for the next two chapters. It is natural to inquire about the utility of the mathematically tractable L^2 norm

$$||P - U||_2^2 = \Sigma(P(s) - \frac{1}{|G|})^2.$$

This has a fatal flaw: Suppose $|G|$ is even, and consider P uniformly distributed over half the points and zero on the others. $||P - U||_2 = \frac{1}{\sqrt{|G|}}$ is close to zero for $|G|$ large. Thus the interpretability of the L^2 norm depends on the size of the group. This makes it difficult to compare rates as the size of the group increases.

The norm $|G|^{\frac{1}{2}}||P - U||_2$ corrects for this. It seems somewhat artificial, and in light of the upper bound lemma and exercise 5, it is essentially the same as the variation distance.

C. SOME EXPLICIT COMPUTATIONS

Example 1. Simple random walk on the circle. Consider Z_p, the additive group of integers mod p. Define $P(1) = P(-1) = \frac{1}{2}$, $P(j) = 0$ otherwise. The following result shows that somewhat more than p^2 steps are required to get close to uniform.

Theorem 2. *For $n \ge p^2$, with p odd and greater than 7,*

$$||P^{*n} - U|| \le e^{-\alpha n/p^2} \text{ with } \alpha = \pi^2/2.$$

Conversely, for $p \geq 7$ and any n

$$||P^{*n} - U|| \geq \frac{1}{2}e^{-\alpha n/p^2 - \beta n/p^4} \alpha = \pi^2/2, \ \beta = \pi^4/11.$$

Proof. The Fourier transform of P is

$$\hat{P}(j) = \frac{1}{2}\left(e^{\frac{2\pi i j}{p}} + e^{\frac{-2\pi i j}{p}}\right) = \cos(2\pi j/p).$$

The upper bound lemma yields

$$||P^{*n} - U||^2 \leq \frac{1}{4}\sum_{j=1}^{p-1} \cos(2\pi j/p)^{2n} = \frac{1}{2}\sum_{j=1}^{(p-1)/2} \cos(\pi j/p)^{2n}.$$

To bound this sum, use properties of cosine. One neat way to proceed was suggested by Hansmartin Zeuner: use the fact

$$\cos x \leq e^{-x^2/2} \text{ for } x \in [0, \pi/2].$$

This follows from $h(x) = \log(e^{x^2/2}\cos x), h'(x) = x - \tan x \leq 0$; so $h(x) \leq h(0) = 0$, for $x \in [0, \pi/2]$.

This gives

$$\begin{aligned} ||P^{*n} - U||^2 &\leq \frac{1}{2}\sum_{j=1}^{(p-2)/2} e^{-\pi^2 j^2 n/p^2} \leq \frac{1}{2}e^{-\pi^2 n/p^2}\sum_{j=1}^{\infty} e^{-\pi^2 (j^2-1)n/p^2} \\ &\leq \frac{1}{2}e^{-\pi^2 n/p^2}\sum_{j=0}^{\infty} e^{-3\pi^2 jn/p^2} \\ &= \frac{1}{2}\frac{e^{-\pi^2 n/p^2}}{1 - e^{-3\pi^2 n/p^2}}. \end{aligned}$$

This works for any n and any odd p. If $n \geq p^2$, $[2(1 - e^{-3\pi^2})]^{-1} < 1$ and thus we have proved more than we claimed for an upper bound.

Observe that the sum in the upper bound is dominated by the term with $k = \frac{p-1}{2}$. This suggests using the function $\cos(2\pi kj/p)$ alone to give a function bounded by 1 which has expected value zero under the uniform distribution. Using the symmetry of P,

$$\Sigma P^{*n}(j) \cos(2\pi jk/p) = \widehat{P^{*n}}(k) = \cos(2\pi k/p)^n = (-1)^n \cos(\pi/p)^n.$$

Now, (1) in section B above yields

$$2||P^{*n} - U|| \geq |\cos(\pi/p)^n|.$$

If $x \leq \frac{1}{2}$, $\cos x \geq e^{-x^2/2-x^4/11}$ say. This yields the lower bound with no conditions on n, for $p \geq 7$. □

Remark 1. The same techniques work to give the same rate of convergence (modulo constants) for other simple measures P such as $P(0) = P(1) = P(-1) = \frac{1}{3}$ or P uniform on $|j| \leq a$. Use of primitive roots and the Chinese remainder theorem gives rates for the multiplicative problem $X_n = a_n X_{n-1}$ (mod p) where $X_0 = 1$, and a_i are i.i.d. random variables taking values mod p. For example, suppose p is a prime and a is a primitive root mod p. Then the multiplicative random walk taking values $a, 1$ or a^{-1} (mod p), each with probability 1/3, takes $c(p)p^2$ steps to become random on the non-zero residues (mod p).

Remark 2. If n and p tend to infinity so $n/p^2 \to c$, the sum in the upper bound lemma approaches a theta function, so

$$||P^{*n} - U||^2 \leq \frac{1}{2}\sum_{j=1}^{\infty} e^{-\pi^2 j^2 c} + o(1).$$

Spitzer (1964), pg. 244) gives a similar result. Diaconis and Graham (1985b) show a similar theta function is asymptotically equal to the variation distance.

Remark 3. There are two other approaches to finding a lower bound in Theorem 1. Both result in a set having the wrong probability if not enough steps are taken.

Approach 1. For any set A, $||P^{*n}-U|| \geq |P^{*n}(A)-U(A)|$. Take $A = \{j: |j| \leq p/4\}$. Use the inversion theorem directly to calculate (and then approximate) $P^{*n}(A)$.

Approach 2. Consider a random walk on the integers Z taking steps ±1 with probability $\frac{1}{2}$. Let S_n be the partial sum. The process considered in Theorem 1 is S_n (mod p). Using the central limit theorem, if n is small compared to p^2, S_n has only a small chance to be outside $\{j: |j| \leq p/4\}$. This can be made rigorous using the Berry-Esseen theorem.

Exercise 7. Write out an honest proof, with explicit constants, for one of the two approaches suggested above. Show $||P^{*n} - U|| \to 1$ if $n = c(p)p^2, c(p) \to 0$.

Remark 4. The random walk based on $P(j) = P(-j) = \frac{1}{2}$ where $(j,p) = 1$ converges at the same rate as when $j = 1$ because of the invariance of variation distance under $1-1$ transformations (Exercise 3 above). Andrew Greenhalgh has shown that it is definitely possible to put $2k+1$ points down carefully, so that the random walk based on $P(j_1) = \ldots = P(j_{2k+1}) = 1/(2k+1)$ converges much faster ($c(p)p^{1/k}$ steps needed) than the random walk based on $P(j) = 1/(2k+1)$ for $|j| \leq k$.

It would be of interest to know the rate of convergence for "most" choices of k points.

The following exercises give other results connected to random walk on Z_p.

EXERCISE 8. Consider the random walk on Z_p generated by $P(1) = P(-1) = \frac{1}{2}$. It is intuitively clear (and not hard to prove) that the walk visits every point. There must be a point which is the last point visited (the last virginal point). Prove that this last point is *uniform* among the $n - 1$ non-starting points.

I do not know how to generalize this elegant result. Avrim Blum and Ernesto Ramos produced computation-free proofs of this result. Both showed that it fails for simple random walk on the cube Z_2^3.

EXERCISE 9. (Fan Chung). Prove that the convolution of symmetric unimodal distributions on Z_n is again symmetric unimodal.

EXERCISE 10. Let n be odd. Consider the random walk on Z_n generated by $P(1) = P(-1) = \frac{1}{2}$. Prove that after an even number of steps, the walk is most likely to be at zero. More generally, show that the walk is monotone in the sense that $P^{*2n}(j) \geq P^{*2n}(j + 2i)$ where $0 \leq j \leq j + 2i \leq n/2$.

This exercise originated in a statistical problem posed by Tom Ferguson. A natural way to test if an X taking values in Z_p is drawn from a uniform distribution is to look in a neighborhood of a prespecified point and reject uniformity if the point falls in that neighborhood. Consider the alternative $H_1 \colon P = P^{*n}$ for simple random walk starting at the prespecified point. The exercise, combined with the Neyman-Pearson lemma implies classical optimality properties for this test.

Ron Graham and I showed that the same type of result holds for nearest neighbor walk on Z_2^n, but fails for nearest neighbor walks on a discrete torus like Z_{13}^3. Monotonicity also fails for the walk originally suggested by Ferguson, namely random transpositions in the symmetric group (see Section D of this chapter) with neighborhoods given by Cayley's distance — the minimum number of transpositions required to bring one permutation into another (see Chapter 6-B).

Example 2. Nearest neighbor walk on Z_2^d. Define $P(0) = P(0 \ldots 01) = P(0 \ldots 10) = \ldots = P(10 \ldots 0) = \frac{1}{d+1}$. The random walk generated by P corresponds to staying where you are, or moving to one of the d nearest neighbors, each with chance $\frac{1}{(d+1)}$. The following result is presented as a clear example of a useful lower bound technique.

Theorem 3. *For P as defined above, let $k = \frac{1}{4}(d + 1)[\log d + c]$,*

$$(1) \qquad \|P^{*k} - U\|^2 \leq \frac{1}{2}(e^{e^{-c}} - 1).$$

As $d \to \infty$, for any $\varepsilon > 0$ there is a $C < 0$ such that $c < C$ and k as above imply

$$(2) \qquad \|P^{*k} - U\| \geq 1 - \varepsilon.$$

Proof. There is a 1-dimensional representation associated to each $x \in Z_2^d$; $\hat{P}(x) = \sum_y (-1)^{x \cdot y} P(y) = 1 - \frac{2\omega(x)}{d+1}$ where $\omega(x)$ is the number of ones (or weight) of x.

The upper bound lemma gives

$$||P^{*k} - U||^2 \le \frac{1}{4}\sum_{x\neq 0}(\hat{P}(x))^{2k} = \frac{1}{4}\sum_{j=1}^{d}\binom{d}{j}(1-\frac{2j}{d+1})^{2k}$$
$$\le \frac{2}{4}\sum_{j=1}^{d/2}\frac{d^j}{j!}e^{-j\log_{d-jc}} = \frac{1}{2}\sum_{j=1}^{d/2}\frac{e^{-jc}}{j!} \le \frac{1}{2}(e^{e^{-c}}-1).$$

For the lower bound observe that the dominating terms in the upper bound come from x of weight 1. Define a random variable $Z\colon Z_2^d \to \mathbb{R}$ by $Z(x) = \sum_{i=1}^{d}(-1)^{x_i} = d - 2\omega(x)$. Under the uniform distribution, x_i are independent fair coin tosses so $E_U Z = 0$, $\mathrm{Var}_U(Z) = d$, and Z has an approximate normal distribution. Under the distribution P^{*k}, arguing as in Example 1 shows

$$E_k(Z) = d(1-\frac{2}{d+1})^k,\quad E_k(Z^2) = d + d(d-1)(1-\frac{4}{d+1})^k.$$
$$\mathrm{Var}_k(Z) = d + d(d-1)(1-\frac{4}{d+1})^k - d^2(1-\frac{2}{d+1})^{2k}.$$

With $k = \frac{1}{4}((d+1)\log d + cd + c)$, as $d \to \infty$,

$$E_k(Z) = \sqrt{d}e^{-c/2}(1 + 0(\frac{\log d}{d})),$$
$$d(d-1)(1-\frac{4}{d+1})^k = (d-1)e^{-c}(1+0(\frac{\log d}{d}))$$
$$d^2(1-\frac{2}{d+1})^{2k} = de^{-c}(1+0(\frac{\log d}{d})).$$

So $\mathrm{Var}_k(Z) = d + 0(e^{-c}\log d)$ uniformly for $c = o(\log d)$. Note that asymptotically $\mathrm{Var}_k(Z) \sim d$, independent of c of order $0(\log d)$. This is crucial to what follows.

For the lower bound, take $A = \{x\colon |Z(x)| \le \beta\sqrt{d}\}$. Then we have

$$||P^{*k} - U|| \ge |P^{*k}(A) - U(A)|.$$

From (3) and Chebychev,

$$U(A) \ge 1 - \frac{1}{\beta^2}$$

$$P^{*k}(A) \le P^{*k}\{|Z - E_k(Z)| \ge E_k(Z) - \beta\sqrt{d}\} \le \frac{\mathrm{Var}_k(Z)}{(E_k(Z)-\beta\sqrt{d})^2}$$
$$\sim \frac{1}{(e^{-c/2}-\beta)^2} \text{ as } d \to \infty.$$

Choosing β large, and c suitably large (and negative) results in $||P^{*k} - U|| \to 1$. □

Remark 1. In this example, the set A has a natural interpretation as the set of all binary vectors with weight close to $\frac{d}{2}$. Since the random walk starts at 0, if it doesn't run long enough, it will tend to be too close to zero.

Remark 2. It is somewhat surprising that $\frac{1}{4}d$ log d steps are enough: It takes $\frac{1}{2}d$ log d steps to have the right chance of reaching the opposite corner $(11\ldots1)$.

Example 3. Simple random walk with randomness multiplier. Let p be an odd number. Define a process on Z_p by $X_0 = 0$, $X_n = 2X_{n-1} + \varepsilon_n (\text{mod } p)$ where ε_i are independent and identically distributed taking values 0, ± 1 with probability $\frac{1}{3}$. Let P_n be the probability distribution of X_n. In joint work with Fan Chung and R. L. Graham it was shown that $n = c$ log p loglog p steps are enough to get close to uniform. Note that X_n is based on the same amount of random input as simple random walk discussed in Example 1. The deterministic doubling serves as a randomness multiplier speeding convergence from p^2 to log p loglog p steps.

Theorem 4. *For P_n defined above, if*

$$n \geq \log_2 p[\frac{\log\log_2 p}{\log 9} + s],$$

then

$$||P_n - U||^2 \leq \frac{1}{2}(e^{9^{-s}} - 1).$$

Proof. Since $X_0 = 0$, $X_1 = \varepsilon_1$, $X_2 = 2\varepsilon_1 + \varepsilon_2, \ldots, X_n = 2^{n-1}\varepsilon_1 + \ldots + \varepsilon_n (\text{mod } p)$. This reduces the problem to a computation involving independent random variables. The Fourier transform of P_n at frequency j is

$$\prod_{a=0}^{n-1} (\frac{1}{3} + \frac{2}{3} \cos \frac{2\pi 2^a j}{p}).$$

Since

$$(\frac{1}{3} + \frac{2}{3} \cos(2\pi x))^2 \leq h(x) \overset{\mathrm{d}}{=} \begin{cases} \frac{1}{9} & \text{if } x \in [\frac{1}{4}, \frac{3}{4}) \\ 1 & \text{otherwise.} \end{cases}$$

It will be enough to bound

$$\prod_{a=0}^{n-1} h(\{\frac{2^a j}{p}\}),$$

where $\{\cdot\}$ denotes fractional part. Observe that if the (terminating) binary expansion of x is $x = \cdot\alpha_1\alpha_2\alpha_3\ldots$, then

$$h(x) = \begin{cases} \frac{1}{9} & \text{if } \alpha_1 \neq \alpha_2 \\ 1 & \text{if } \alpha_1 = \alpha_2. \end{cases}$$

Let $A(x,n)$ denote the number of alternations in the first n binary digits of x: $A(x,n) = |\{1 \le i < n: \alpha_i \ne \alpha_{i+1}\}|$. Successively multiplying by powers of 2 just shifts the bits to the left. It follows that

$$\prod_{a=0}^{n-1} h(\{\frac{2^a j}{p}\}) \le 9^{-A(j/p,n)}.$$

Suppose first that p is of the special form $p = 2^t - 1$. The fractions j/p become

$$\begin{aligned} 1/p &= \overbrace{00\ldots01}^{t}\ \overbrace{00\ldots01}^{t}\ \ldots \\ 2/p &= 00\ldots10\ \ 00\ldots10\ \ldots \\ 3/p &= 00\ldots11\ \ 00\ldots11\ \ldots \\ p-1/p &= 11\ldots10\ \ 11\ldots10\ \ldots \end{aligned}$$

If $n = rt$, the number of alternations in the first n places of row j/p is no smaller than r times the number of alternations in the first t places of j/p. It follows that

$$\begin{aligned} \sum_{j=1}^{p-1}\prod_{a=0}^{n-1} h\{\frac{2^a j}{p}\} &\le \sum_{j=1}^{p-1} 9^{-rA(j/p,t)} \\ &\le 2\sum_{k=1}^{t}\binom{t}{k}9^{-kr} = 2[(1+9^{-r})^t - 1] \qquad (1) \\ &\le 2[e^{t9^{-r}} - 1]. \end{aligned}$$

The second inequality in (1) used the easily proved bound $|j: A(\frac{j}{p}, t) = k| \le 2\binom{t}{k}$. Now, if $n = rt$ with $r = \frac{\log t}{\log 9} + s$, the upper bound lemma gives

$$||P_n - U||^2 \le \frac{1}{2}[e^{9^{-s}} - 1]$$

as claimed.

For general odd p, define t by $w^{t-1} < p < 2^t$. For r as chosen above, partition the initial $n = rt$ digits in the binary expansion of j/p into r blocks of length t: $B(i,j)1 \le i \le r$:

$$j/p = \overbrace{\alpha_1\ldots\alpha_t}^{B(1,j)}\ \overbrace{\alpha_{t+1}\ldots\alpha_{2t}}^{B(2,j)}\ldots\ \overbrace{\alpha_{(r-1)t+1}\ldots\alpha_{rt}}^{B(r,j)}.$$

Thus,

$$\sum_{j=1}^{p-1}\prod_{a=0}^{n-1} h(\{\frac{2^a j}{p}\}) \le \sum_{j=1}^{p-1} 9^{-A(B(1,j))-\ldots-A(B(r,j))}. \qquad (2)$$

By the choice of t, all the blocks $B(1,j)$, $1 \leq j \leq p-1$ in the first column are distinct and have at least one alternation. Furthermore, since $(2,p)=1$, the set of blocks in the ith column does not depend on i. This information will be used together with the following interchange lemma: If $0<\alpha<1$, and $a \leq a'$, $b \leq b'$, then

$$\alpha^{a+b'} + \alpha^{a'+b} \leq \alpha^{a+b} + \alpha^{a'+b'}.$$

To prove this, simply expand $(\alpha^a - \alpha^{a'})(\alpha^b - \alpha^{b'}) \geq 0$. The lemma implies that collecting together terms with the same blocks in the exponent only increases the upper bound. Thus, the right side of (2) is no bigger than

$$\sum_{j=1}^{p-1} 9^{-r\ A(j/2^t-1,t)},$$

the sum that appears in equation (1) above! Using the bound there completes the proof. □

Remark 1. A more careful version of the argument implies that for any odd p, the cutoff is at $c^*\log_2 p\ \mathrm{loglog}_2 p$ with $c^* = 1/\log_2 \pi_1$ where

$$\pi_1 = \prod_{a=1}^{\infty} (\frac{1}{3} + \frac{2}{3}\cos(2\pi/2^a))^2.$$

Chung, Diaconis and Graham (1987) show that for p of form 2^t-1, $c^*t\log t$ steps are required. The lower bound technique again uses the "slow" terms in the upper bound lemma to define a random variable $Z(j) = \Sigma_{|k|=1} e^{2\pi ijk/p}$ where the sum is over k's with a single 1 in their binary expression. Under the uniform distribution Z has mean 0 and "variance" $(= E(Z\overline{Z})) = t$. Under P_n, Z has mean approximately $t\pi_1^{\frac{1}{2}}$ and variance of order $\sqrt{t}$.

Chung, Diaconis and Graham also prove that for most odd p, 1.02 $\log_2 p$ steps are enough. A curious feature of the proof is that we do not know single explicit sequence of p's such that $2\log p$ steps suffice to make the variation distance smaller than $\frac{1}{10}$.

Remark 2. There is a natural generalization of this problem which may lead to further interesting results. Let G be an Abelian group. Let $A\colon G \to G$ be an automorphism (so A is 1-1, onto and $A(s+t) = A(s) + A(t)$). Consider the process

$$X_n = A(X_{n-1}) + \epsilon_n$$

where $X_0 = \mathrm{id}$ and ϵ_i are iid. This can be represented as a convolution of independent random variables

$$X_n = A^{n-1}(\epsilon_1) + A^{n-2}(\epsilon_2) + \ldots + \epsilon_n.$$

If $A^k = \mathrm{id}$, these can be further grouped in blocks of k (when k divides n) to give a sum of iid variables. Then, methods similar to those used in the present example may provide rates.

It is not necessary to use an automorphism; $f(s) = A(s) + t$, with $t \varepsilon G$ fixed and A an automorphism works in a similar way. It is not necessary that G be Abelian. If the Law of ϵ_i is constant on conjugacy classes so is the law of $A(\epsilon_i)$ *and* the random variables commute in distribution (see exercise 2.7).

One natural example to try has $G = \mathbb{Z}_n^2$, A *a* 2×2 matrix, and ϵ_i the nearest neighbor random variable taking values $(00), (01), (0-1), (10), (-10)$ each with probability $\frac{1}{5}$.

Remark 3. Fourier techniques can be used to bound other distances. This remark gives a result for the maximum over all "intervals" of Z_p. The next remark discusses arbitrary compact groups. The techniques are close to work of Joop Kemperman (1975).

Let P and Q be probabilities on Z_p. Define $D(P,Q) = \sup_J P(J) - Q(J)$ where the sup is over all "intervals" in Z_p.

LEMMA. $D(P,Q) \leq \frac{1}{\sqrt{2}} \sum_{j=1}^{p-1} |\hat{P}(j) - \hat{Q}(j)|/j^*$ *where* $j^* =$ *min*$(j, p-j)$.

Proof. For J an interval on the circle, $|P(J) - Q(J)| = |P(J^c) - Q(J^c)|$, where of course J^c is an interval too. It follows that only intervals not containing zero need be considered. Let $[\ell_1, \ell_2]$ be such an interval, with $\ell_1 < \ell_2$ (clockwise). Then

$$P([\ell_1, \ell_2]) - Q([\ell_1, \ell_2]) = P([0, \ell_2]) - P([0, \ell_1)) - Q([0, \ell_2]) + Q([0, \ell_1)).$$

Now

$$P([0,\ell]) = \sum_{a=0}^{\ell} P(a) = \frac{1}{p} \sum_{a=0}^{\ell} \sum_{j=0}^{p-1} \hat{P}(j) e^{-\frac{2\pi i j a}{p}}$$

$$= \frac{1}{p} \sum_{j=0}^{p-1} \hat{P}(j)(1 - e^{-2\pi i(\ell+1)j/p})/(1 - e^{-2\pi i j/p}).$$

This implies that $P - Q$ equals

$$\frac{1}{p} \sum_{j=1}^{p-1} [\hat{P}(j) - \hat{Q}(j)][e^{-2\pi i \ell_1 j/p} - e^{-2\pi i(\ell_2+1)j/p}]/(1 - e^{-2\pi i j/p}).$$

Thus $D(P,Q)$ is bounded above by

$$\frac{\sqrt{2}}{p} \sum_{j=1}^{p-1} |\hat{P}(j) - \hat{Q}(j)| / \sqrt{1 - \cos(2\pi j/p)}.$$

Now for $0 \leq x \leq \frac{\pi}{2}$, $1 - \cos x \geq \frac{x^2}{3}$, so for $1 \leq j \leq p/4$, $\frac{\sqrt{2}}{p}(1 - \cos(2\pi j/p))^{-\frac{1}{2}} \leq \sqrt{6}/2\pi j \leq \frac{1}{j\sqrt{2}}$. For $\frac{\pi}{2} \leq x \leq \pi, 1 - \cos x \geq 1$, so for $p/4 \leq j \leq p/2, \frac{1}{p}(1 - \cos(1\pi j/p))^{-\frac{1}{2}} \leq \frac{1}{p} \leq \frac{1}{2j}$. For the rest, $\cos(2\pi j/p) = \cos(2\pi(p-j)/p)$. □

EXERCISE 11. Using this lemma, with P_n as defined in Example 3, show there are constants a and b such that for every odd p,

$$D(P_n, U) \leq ae^{-bn/\log p}.$$

Remark 4. There must be similar bounds for any compact group. To see the use of such results let T be the unit circle: $T = \{z \in C: |z| = 1\}$. Fix an irrational $\alpha \in T$ and consider simple random walk with step size α, thus $X_0 = 0$, and $X_n = X_{n-1} \pm \alpha$. Since X_n is concentrated on a discrete set, the variation distance to uniform is always 1. Nonetheless, the distribution of X_n converges to the uniform distribution in the weak star topology. To discuss rates, a metric must be chosen. A convenient one is

$$D(P, Q) = \sup_I |P(I) - Q(I)|$$

for I ranging over intervals of T. This metrizes weak star convergence.

Kemperman (1975) proves two useful inequalities that give bounds on D involving the Fourier transform for P a probability on T, and $m \epsilon Z$,

$$\hat{P}(m) = \int_0^1 e^{2\pi imx} P(dx).$$

(1) $D(P, U) = \sup_I P(I) - U(I)| \leq \{12 \sum_{m=1}^{\infty} |\hat{P}(m)|^2/(2\pi m)^2\}^{\frac{1}{3}}$.

(2) $D(P, U) \leq \frac{2}{\pi} \sum_{m=1}^{\infty} |\hat{P}(m)|/m$.

Niederreiter and Philipp (1973) discuss multivariate versions.

EXERCISE 12. Consider simple random walk on the unit circle, as in remark 3 above, with α a quadratic irrational. Use bounds (1) and (2) above to estimate rates of convergence. A direct combinatorial argument can be used to show that $D(P^{*n}, U) \leq c(\log n)/\sqrt{n}$.

It seems quite possible to carry over bounds like (2) in Remark 4 to any compact group G. Choose a bi-invariant metric $d(x, y)$ on G and consider $D(P, Q) = \sup|P(I) - Q(I)|$ where I ranges over all translates of balls centered at the identity. Then $D(P, Q)$ can be bounded as in remark 2; Lubotzky, Phillips, and Sarnak (1986) give results for the sphere. Their paper makes fascinating use of deep number theory which must be useful for other problems. Chapter 6 below discusses bi-invariant metrics.

Example 4. Random walks on the affine group A_p. (An elaborate exercise). Let p be a prime. Random numbers are often generated by the recursive scheme $X_n = aX_{n-1} + b(\text{mod } p)$. This sequence of exercises allows estimates of the rate of convergence when a and b are random. The transformation $x \rightarrow ax + b$ with a non-zero (mod p) will be written $T_{ab}(x)$. The set of such transformations form a finite group A_p. We write (a,b) for the typical group element. The product is $(a,b)(c,d) = (ac, ad + b)$, the identity is $(1,0)$ and $(a,b)^{-1} = (a^{-1}, -ba^{-1})$. This group has $p(p-1)$ elements. The subgroups $\{(1,b)\} \cong Z_p$ and $\{(a,0)\} \cong Z_p^*$ are useful.

(1) Identify the p distinct conjugacy classes. Explain why measures constant on conjugacy classes are not very interesting.

(2) From part (1) there are p distinct irreducible representations; $p-1$ of these are the 1-dimensional representations given by choosing a character χ_i of Z_p^* and defining $\rho_i(a,b) = \chi_i(a)$. Show that these are distinct irreducible representations. Show that there is one other irreducible representation ρ of degree $p-1$. Use Serre's exercise 2-6 to construct this representation by considering the action of A_p on Z_p. By explicitly choosing a basis, show

$$\chi_\rho(1,0) = p-1,$$
$$\chi_\rho(1,b) = -1, \quad b \neq 0$$
$$\chi_\rho(a,b) = 0, \quad a \neq 1.$$

(3) Let ρ^+ and ρ^* be the restriction of ρ in Part 2 to Z_p and Z_p^* respectively. Using the character of ρ and the inner product machinery, show that ρ^* is the regular representation of Z_p^* and ρ^+ contains each non-trivial irreducible representation of Z_p once.

(4) Let P^+ be a probability on Z_p and P^* a probability on Z_p^*. Let χ_i^+ and χ_i^* be characters of Z_p and Z_p^*. Let $P(a,b) = P^*(a) \cdot P^+(b)$. Show

(a) $\hat{P}(\rho) = \hat{P}^+(\rho^+) \cdot \hat{P}^*(\rho^*)$.

(b) The eigenvalues of the matrix $\hat{P}^*(\rho^*)$ are the p-1 numbers $\hat{P}^*(\chi_i^*)$; the eigenvalues of $\hat{P}^+(\rho^+)$ are the p-1 numbers $\hat{P}^+(\chi_i^+)$, χ_i^+ non-trivial.

(5) Let p be an odd prime such that 2 is a primitive root mod p. Consider the random walk on A_p which starts at 0 and is based on P^*, P^+, with $P^*(1) = P^*(2) = P^*((p+1)/2) = \frac{1}{3}$ and $P^+(0) = P^+(1) = P^+(-1) = \frac{1}{3}$. Show that $k = c(p)p^2$ log steps are enough to get arbitrarily close to random if $c(p) \to \infty$ as p does. Use this to argue that the random point $T_{X_n}(0)$ is close to uniformly distributed on Z_p after this many steps.

One way through the computations uses the following fact. Let $\tau(A)$ be the spectral radius (= max|eigenvalue|) of the matrix A. If A and B are diagonalizable matrices then $\tau(AB) \leq \tau(A)\tau(B)$.

(6) Show by considering the first coordinate of (a,b) that $k = cp^2$ steps are not enough if c is fixed.

Remark. The argument sketched above gives $c(p)p^2 \log p$. I presume that $c(p)p^2$ is the correct answer. Actually, numerical computation strongly suggests that the random walk $X_n = aX_{n-1} + b$, where (a,b) has the distribution described in part 5, becomes uniform in order $(\log p)^A$ steps for $A = 1$ or 2.

When p is composite there are more conjugacy classes. It is an interesting exercise to determine these. I have not succeeded in finding a natural "small" measure constant on conjugacy classes which permits analysis.

D. Random transpositions: an introduction to the representation theory of the symmetric group.

As described in Section A, repeated random transpositions of n cards in a row can be modeled as repeatedly convolving the following measure:

$$(1) \qquad P(\text{id}) = \frac{1}{n}, \; P(\tau) = \frac{2}{n^2} \text{ for } \tau \text{ a transposition}, P(\pi) = 0 \text{ otherwise }.$$

This section presents a proof of the following theorem

Theorem 5. *Let $k = \frac{1}{2}n \log n + cn$. For $c > 0$,*

$$||P^{*k} - U|| \leq ae^{-2c}$$

for a universal constant a. Conversely, for $c < 0$, as n tends to infinity

$$||P^{*k} - U|| \geq (\frac{1}{e} - e^{-e^{-2c}}) + o(1).$$

The proof introduces some basic results about the representation theory of the symmetric groups. Most all of these will be treated in greater detail in Chapter 7. This problem was first treated by Diaconis and Shahshahani (1981). The present argument is based on simplifications suggested by Leo Flatto, Andrew Odlyzko, and Hansmartin Zeuner. After the proof, several further problems, *to which the same analysis applies*, are described.

Discussion The measure P is constant on conjugacy classes: $P(\eta\pi\eta^{-1}) = P(\pi)$. Thus Schur's lemma implies, for any irreducible representation ρ, $\hat{P}(\rho) =$ constant $\cdot I$. Taking traces, the constant equals $(\frac{1}{n} + \frac{n-1}{n}r(\rho))$ with $r(\rho) = \chi_\rho(\tau)/d_\rho$. Here $\chi_\rho(\tau)$ denotes the character of ρ at any transposition τ and d_ρ denotes the dimension of ρ (see proposition 6 of Chapter 2). Now, the upper bound lemma yields

$$||P^{*k} - U||^2 \leq \frac{1}{4}\sum_\rho^* d_\rho^2(\frac{1}{n} + \frac{n-1}{n}r(\rho))^{2k}.$$

The following heuristic discussion may help understanding. Table 1 gives d_ρ and $\chi_\rho(\tau)$ for $n = 10$. There are 42 irreducible representations of S_{10}. For example, the first entry is $d_\rho = 1$, $\chi_\rho(\tau) = 1$ for the trivial representation. The second entry is $d_\rho = 9$, $\chi_\rho(\tau) = 7$ for the 9-dimensional permutation representation. Except for a few representations at the ends, the ratio $|\chi_\rho(\tau)/d_\rho|$ is small. Suppose it could be shown that $\chi_\rho(\tau)/d_\rho \leq \frac{1}{2}$ for most ρ, then, approximately for n large, $|\frac{1}{n} + \frac{n-1}{n}r(\rho)| \leq \frac{1}{2}$ and the upper bound above would be smaller than

$$\frac{1}{4}(\frac{1}{2})^{2k}\Sigma d_\rho^2 = \frac{1}{4}(\frac{1}{2})^{2k}n! \text{ (using proposition 5 of Chapter 2).}$$

Table 1
Characters of S_{10} (from James and Kerber (1981, pg. 354))

Partition	dim	$\chi_\rho(\tau)$
[10]	1	1
[9,1]	9	7
[8,2]	35	21
[8,1,1]	36	20
[7,3]	75	35
[7,2,1]	160	64
[7,1,1,1]	84	28
[6,4]	90	34
[6,3,1]	315	91
[6,2,2]	225	55
[6,2,1,1]	350	70
[6,1,1,1,1]	126	14
[5,5]	42	14
[5,4,1]	288	64
[5,3,2]	450	70
[5,3,1,1]	567	63
[5,2,2,1]	525	35
[5,2,1,1,1]	448	0
[5,1,1,1,1,1]	126	-14
[4,4,2]	252	28
[4,4,1,1]	300	20
[4,3,3]	210	14
[4,3,2,1]	768	0
[4,3,1,1,1]	525	-35
[4,2,2,2]	300	-20
[4,2,2,1,1]	567	-63
[4,2,1,1,1,1]	350	-70
[4,1,1,1,1,1,1]	84	-28
[3,3,3,1]	210	-14
[3,3,2,2]	252	-28
[3,3,2,1,1]	450	-70
[3,3,1,1,1,1]	225	-55
[3,2,2,2,1]	288	-64
[3,2,2,1,1,1]	315	-91
[3,2,1,1,1,1,1]	160	-64
[3,1,1,1,1,1,1,1]	36	-20
[2,2,2,2,2]	42	-14
[2,2,2,2,1,1]	90	-34
[2,2,2,1,1,1,1]	75	-35
[2,2,1,1,1,1,1,1]	35	-21
[2,1,1,1,1,1,1,1,1]	9	-7
[1,1,1,1,1,1,1,1,1,1]	1	-1

Now using Stirling's formula,

$$(\frac{1}{2})^{2k} n! \doteq e^{-2k \log 2 + n\log n - n + \dots}.$$

It follows that if k is $n \log n$, the upper bound will tend to zero. To complete this heuristic discussion, consider the term arising from the $n-1$ dimensional representation: $d_\rho = n-1$, and $\chi_\rho(\tau) = n-3$. This is easy to see: the trace of the permutation representation for a transposition is $n-2$. The permutation representation is the direct sum of the trivial representation and the $n-1$ dimensional representation so $n-2 = \chi_\rho(\tau)+1$. Here $(\frac{1}{n} + \frac{n-1}{n} r(\rho))^{2k} = (1-\frac{2}{n})^{2k}$. This is a far cry from $(\frac{1}{2})^{2k}$. Persevering, in the upper bound lemma k has to be chosen large enough to kill

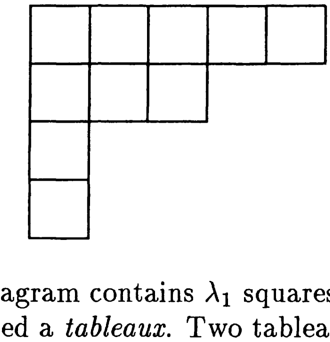
$$(n-1)^2(1-\frac{2}{n})^{2k} = e^{2k \log(1-\frac{2}{n}) + 2\log(n-1)} = e^{-\frac{4k}{n} + 2\log n + 0(\frac{k}{n^2})}.$$

For $k = \frac{1}{2} n \log n + cn$ this is asymptotic to e^{-4c}. Taking square roots gives the e^{-2c} of the theorem.

It will turn out that this is the slowest term, the other terms being geometrically smaller, and most terms being smaller than $(\frac{1}{2})^{2k}$.

This argument is somewhat similar to the bounds for simple random walk on Z_p: terms near the trivial representation needed to be summed up carefully, terms far away were geometrically small and easily dealt with. Putting in the details for Z_p required properties of cosine. For the symmetric group, the representations are usefully thought of as 2-dimensional shapes. Properties of d_ρ and $\chi_\rho(\tau)$ will have to be developed.

To begin a more detailed discussion, consider a partition of n, say $\lambda = (\lambda_1, \dots \lambda_m)$ with $\lambda_1 \geq \lambda_2 \dots \geq \lambda_m$ positive integers with $\lambda_1 + \dots + \lambda_m = n$. There is a one to one correspondence between irreducible representations of S_n and partitions of n. This is carefully described in Chapter 7. For present purposes, the notion of the *diagram* associated to a partition is helpful. An example says things best: the diagram corresponding to (5,3,1,1) is



The first row of the diagram contains λ_1 squares, etc. A diagram containing numbers 1, 2,..., n is called a *tableaux*. Two tableaux are considered equivalent if they have the same row sets:

$$\begin{array}{|c|c|c|c|c|}\hline 5 & 10 & 7 & 8 & 9 \\ \hline 6 & 3 & 4 \\ \cline{1-3} 2 \\ \cline{1-1} 1 \\ \cline{1-1} \end{array} \quad \sim \quad \begin{array}{|c|c|c|c|c|}\hline 9 & 8 & 7 & 10 & 5 \\ \hline 3 & 6 & 4 \\ \cline{1-3} 2 \\ \cline{1-1} 1 \\ \cline{1-1} \end{array}$$

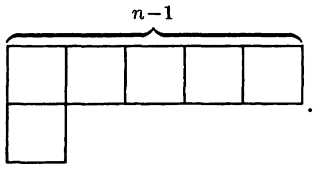
An equivalence class of tableaux is called a called a *tabloid.* There are $n!/\lambda_1!\dots\lambda_m!$ distinct tabloids of a given shape. These are used to build a representation called M^λ as follows. Consider a vector space with basis vectors $\{e_t\}$ where t runs over all tabloids of shape λ. For π a permutation, define $\rho(\pi)$ by defining on basis vectors:

$$\rho(\pi)(e_t) = e_{\pi t}$$

where for example, π applied to the tabloid

$$\begin{array}{lcl} 5\ 10\ 7\ 8\ 9 & & \pi(5)\ \pi(10)\ \pi(7)\ \pi(8)\ \pi(9) \\ 6\ \ 3\ 4 & \text{is the} & \pi(6)\ \pi(3)\ \pi(4) \\ 2 & \text{tabloid} & \pi(2) \\ 1 & & \pi(1) \end{array}$$

Here are some examples: the partition $(n-1,1)$ has $n!/(n-1)! = n$ distinct tabloids, all of shape



These are evidently completely determined by the one number in the second row. Hence the vector space $M^{n-1,1}$ is just the n-dimensional space spanned by the usual basis $e_1,\dots e_n$ with $\rho(\pi)e_i = e_{\pi(i)}$. The partition $n-2,1,1$ gives rise to a vector space $M^{n-2,1,1}$ with basis $\{e_{(i,j)}\}$ and $\rho(\pi)e_{(i,j)} = e_{(\pi(i),\pi(j))}$. The partition $n-2,2$ gives rise to $M^{n-2,2}$ with basis $\{e_{\{i,j\}}\}$ where $\{i,j\}$ runs through all unordered pairs.

The representations M^λ are all reducible except of course for $\lambda = n$. It will be argued that each M^λ contains a well-defined irreducible subspace S^λ, and as λ varies the S^λ range over all the irreducible representations of S_n. The following two facts are all that is needed to prove Theorem 5.

FACT 1. The dimension of the irreducible representation corresponding to partition λ is the number of ways of placing the numbers $1,2,\dots,n$ into the diagram of λ such that the entries in each row and column are increasing.

This fact is Theorem (8.4) in James (1978) who discusses other formulas for the dimension. These are also described in Chapter 7 below. A useful computational corollary is

$(D-1)$ The dimension d_λ of the irreducible representation corresponding to the partition λ satisfies $d_\lambda \le \binom{n}{\lambda_1} d_{\lambda^*}$ with $\lambda^* = (\lambda_2, \lambda_3, \ldots, \lambda_m)$ a partition of $n - \lambda_1$.

Proof. The first row may be chosen in $\binom{n}{\lambda_1}$ ways. For each choice of first row, the number of ways of placing the $n - \lambda_1$ remaining numbers is d_{λ^*}. Of course not all of these combine with the choice of first row to give a monotone total placement. This gives the inequality. □

Fact 2. Let $r(\lambda) = \chi_\lambda(\tau)/d_\lambda$ where $\chi_\lambda(\tau)$ is the character at a transposition and d_λ is the dimension of the irreducible representation corresponding to the partition λ of n. Then

$$(D-2) \qquad r(\lambda) = \frac{1}{n(n-1)} \Sigma[\lambda_j^2 - (2j-1)\lambda_j] = \frac{1}{\binom{n}{2}} \sum_j \binom{\lambda_j}{2} - \binom{\lambda_j'}{2}$$

with λ' the transpose of λ.

This is a special case of a result due to Frobenius who essentially determined formulas for all of the characters. These become unwieldly for complex conjugacy classes. An accessible proof of (D-2) is given in Ingram (1950).

Using Frobenius' formula, Shahshahani and I proved a simple monotonicity result: Call partition λ^1 larger than partition λ^2 if it is possible to get from the diagram of λ^2 to the diagram of λ^1 by moving boxes up to the right. This is a partial order. For example $(5,1) \ge (4,2) \ge (3,3)$, but $(3,3)$ and $(4,1,1)$ are not comparable, though both are larger than $(3,2,1)$. This order is widely used in statistics under the name of majorization (see e.g. Marshall and Olkin (1979)). James (1978, pg. 8) contains further examples.

Lemma 1. *If $\lambda \ge \lambda'$, then $r(\lambda) \ge r(\lambda')$ where $r(\lambda) = \chi_\lambda(\tau)/d_\lambda$ is given by (D-2).*

Proof. It suffices to consider the case where one box is switched from row b to $a (b > a)$, i.e. $\lambda_a = \lambda_a' + 1$, $\lambda_b = \lambda_b' - 1$, $\lambda_c = \lambda_c'$ for $c \ne a, b$. Formula (D-2) shows that

$$\begin{aligned} r(\lambda) - r(\lambda') &= \frac{1}{n(n-1)}\{\lambda_a^2 - (2a-1)\lambda_a - \lambda_a'^2 + (2a-1)\lambda_a' + \\ &\qquad \lambda_b^2 - (2b-1)\lambda_b - \lambda_b'^2 + (2b-1)\lambda_b'\} \\ &= \frac{1}{n(n-1)}\{2\lambda_a' + 1 - (2a-1) - (2\lambda_b' - 1) + (2b-1)\} \\ &= \frac{1}{n(n-1)}\{\lambda_a' - \lambda_b' + b - a + 1\} \ge \frac{4}{n(n-1)} > 0 \end{aligned}$$

since $\lambda_a' \ge \lambda_b'$ and $b - a \ge 1$. This argument is correct even if $\lambda_b = 0$. □

LEMMA 2. *Let λ be a partition of n. Then*

(a) $$r(\lambda) \leq 1 - \frac{2(n-\lambda_1)(\lambda_1+1)}{n(n-1)} \text{ if } \lambda_1 \geq n/2,$$

(b) $$r(\lambda) \leq \frac{\lambda_1 - 1}{n-1}.$$

Proof.
(a) By assumption $\lambda \leq (\lambda_1, n-\lambda_1)$, so it follows from Lemma 1 and (D-2) that

$$\begin{aligned} r(\lambda) &\leq \frac{1}{n(n-1)}\{\lambda_1^2 - \lambda_1 + n^2 - 2n\lambda_1 + \lambda_1^2 - 3n + 3\lambda_1\} \\ &= 1 + \frac{2(\lambda_1^2 + \lambda_1 - n\lambda_1 - n)}{n(n-1)} \\ &= 1 - \frac{2(\lambda_1+1)(n-\lambda_1)}{n(n-1)} \end{aligned}$$

(b) $r(\lambda) = \frac{1}{n(n-1)} \sum_{j=1}^{m} (\lambda_j^2 - (2j-1)\lambda_j) \leq \frac{1}{n(n-1)} \sum_{j=1}^{m} \lambda_j(\lambda_j - 1) \leq \frac{\lambda_1 - 1}{n(n-1)} \sum_{j=1}^{m} \lambda_j = \frac{\lambda_1-1}{n-1}$. □

COROLLARY. *Let λ be such that $r(\lambda) \geq 0$. Then*

$$|\frac{1}{n} + \frac{n-1}{n} r(\lambda)| \leq \begin{cases} 1 - \frac{2(\lambda_1+1)(n-\lambda_1)}{n^2} & \text{if } \lambda_1 \geq n/2 \\ \frac{\lambda_1}{n} & \text{for all } \lambda. \end{cases}$$

Proof of Theorem 5: If λ^t denotes the transpose of λ we certainly have either $r(\lambda) \geq 0$ or $r(\lambda^t) > 0$, because $\chi_\lambda = -\chi_\lambda^t$ (see James 1978, p. 25). Hence

$$\begin{aligned} \Sigma_\lambda^* d_\lambda^2 \left(\frac{1}{n} + \frac{n-1}{n} r(\lambda)\right)^{2k} &\leq \sum_{\lambda: r(\lambda) \geq 0}^{*} d_\lambda^2 \left(\frac{1}{n} + \frac{n-1}{n} r(\lambda)\right)^{2k} \\ &\quad + \sum_{\lambda: r(\lambda) > 0}^{*} d_\lambda^2 \left(\frac{1}{n} - \frac{n-1}{n} r(\lambda^t)\right)^{2k} \\ &\leq 2 \sum_{\lambda: r(\lambda) \geq 0}^{*} d_\lambda^2 |\frac{1}{n} + \frac{n-1}{n} r(\lambda)|^{2k}. \end{aligned}$$

For $\lambda = (n)$, which is contained in the next to last, but not the last sum, this used

$$\begin{aligned} & d_n^2 |\frac{1}{n} - \frac{n-1}{n} r(n)|^{2k} + d_{(n-1,1)}^2 |\frac{1}{n} - \frac{n-1}{n} r(n-1,1)|^{2k} \\ &= (1 - \frac{2}{n})^{2k} + (n-1)^2 (1 - \frac{4}{n})^{2k} \leq (n-1)^2 (1 - \frac{2}{n})^{2k} \\ &= d_{n-1}^2 |1 + \frac{n-1}{n} r(n-1,1)|^{2k}. \end{aligned}$$

In order to bound this sum we split it into two parts according as $\lambda_1 \gtrless (1-\alpha)n$ (where $\alpha \in (0, \frac{1}{4})$ will be chosen below)

$$\sum_{\lambda:r(\lambda)\geq 0}^{*} d_\lambda^2 \left(\frac{1}{n} + \frac{n-1}{n} r(\lambda)\right)^{2k} = \sum_{j=1}^{n} \sum_{\substack{\lambda:r(\lambda)\geq 0 \\ \lambda_1 = n-j}} d_\lambda^2 \left(\frac{1}{n} + \frac{n-1}{n} r(\lambda)\right)^{2k}$$

$(*)$

$$\leq \sum_{j=1}^{\alpha n} \binom{n}{j} \frac{n!}{(n-j)!} \left(1 - \frac{2j(n-j+1)}{n^2}\right)^{2k} + \sum_{j>\alpha n}^{n-1} \binom{n}{j} \frac{n!}{(n-j)!} (1 - \frac{j}{n})^{2k}.$$

To obtain this we used the corollary to (D-2) above and

$$\sum_{\lambda:\lambda_1=\ell} d_\lambda^2 \leq \binom{n}{\ell}^2 \sum_{\lambda::\lambda_1=\ell} d^2_{(\lambda_2,\lambda_3,\ldots,\lambda_m)} = \binom{n}{\ell}^2 \sum_{\lambda'} d^2_{\lambda'} = \binom{n}{\ell} \frac{n!}{\ell!}$$

(where the λ' are the irreducible representations of $S_{n-\ell}$).

In order to give a bound for the first sum shown in $(*)$ above recall that $k = \frac{n}{2}\log n + cn$;

$$\sum_{j=1}^{\alpha n} \binom{n}{j} \frac{n!}{(n-j)!} e^{-4k(\frac{j}{n} - \frac{j^2-j}{n^2})}$$

$$\geq n^2 \cdot e^{-4k/n} \cdot \sum_{j=1}^{\alpha n} \frac{(n-1)!^2}{(n-j)!^2 j!} \cdot e^{-2 \log n \cdot (1-\frac{j}{n})(j-1)},$$

we observe that the factor in front of the sum is exactly e^{-4c} and so all we have to do is to bound $\sum_{j=1}^{\alpha n} \frac{n^{2(j-1)}}{j!} \cdot n^{-2(1-\frac{j}{n})(j-1)} = \sum_{j=1}^{\alpha n} \frac{1}{j!} \cdot n^{\frac{2j(j-1)}{n}}$ for large values of n. The ratio between two consecutive terms in this sum is $\frac{1}{j+1} \cdot n^{4j/n}$, which, as a function of j, is decreasing if $j < \frac{n}{4 \log n}$ and increasing if $j > \frac{n}{4 \log n}$. So if both the first and the last ratio are less than $q < 1$ we may bound the sum by $\frac{1}{1-q}$. But the first ratio is < 1 if $n \geq 17$ and the last one is < 1 if $\frac{1}{\alpha n} n^{4\alpha} < 1$, i.e. $n > (\frac{1}{\alpha})^{1/1-4\alpha}$. This works well if $\alpha < 1/4$.

Now let's consider the second sum

$$\sum_{j>\alpha n} \binom{n}{j} \frac{n!}{(n-j)!} (1 - \frac{j}{n})^{2k} \leq (1-\alpha)^{2cn} \cdot \sum_{j>\alpha n}^{n-1} \binom{n}{j} \frac{n!}{(n-j)!} (1 - \frac{j}{n})^{n \log n}.$$

The factor in front of the sum is $\leq e^{-4c}$ if $n \geq 7$ and α is close enough to $\frac{1}{4}$. Hence it is sufficient to bound the sum for large values of n. The ratio between two consecutive terms is

$$\frac{(n-j)^2}{j+1} (1 - \frac{1}{n-j})^{n \log n},$$

which is decreasing in j. So, if the first of these ratios,

$$\frac{(n-\alpha n)^2}{\alpha n+1}(1-\frac{1}{n-\alpha n})^{n \log n} \leq \frac{(1-\alpha)^2}{\alpha} n^{1-\frac{1}{1-\alpha}}$$

is less than one (this happens if $n \geq (\frac{(1-\alpha)^2}{\alpha})^{\frac{(1-\alpha)}{\alpha}}$) we may bound the sum by n times the first term, that is, by

$$n \cdot \binom{n}{\alpha n} \frac{n!}{(n-\alpha n)!} \cdot (1-\alpha)^{n \log n}.$$

Using Stirling's formula one can show that this tends to 0 (very slowly) if n tends to ∞ and so it must be bounded. This completes the proof of the upper bound part of Theorem 5.

The following argument for the lower bound produces an explicit set A where the variation distance is large. Intuitively, if not enough transpositions have been made, there will be too many cards that occupy their original positions. Let A be the set of all permutations with one or more fixed points. Under U, the chance of one or more fixed points is well known under the name of the matching problem. Feller (1968, Sec. IV.4) implies

$$U(A) = 1 - \frac{1}{e} + 0(\frac{1}{n!}).$$

To bound $P^{*k}(A)$, consider the process for generating P^{*k} using random transpositions $(L_1, R_1), \ldots, (L_k, R_k)$. Let B be the event that the set of labels $\{L_i, R_i\}_{i=1}^{k}$ is strictly smaller than $\{1, \ldots, n\}$. Clearly $A \supset B$. The probability of B is the same as the probability that when $2k$ balls are dropped into n boxes, one or more of the boxes will be empty. Arguments in Feller (1968, Sec. IV.2) imply that the probability of B equals

$$1 - e^{-ne^{-2k/n}} + o(1) \text{ uniformly in } k, \text{ as } n \to \infty.$$

With $k = \frac{1}{2}n \log n + cn, P^{*k}(A) \geq 1 - e^{-e^{-2c}+o(1)}$. Thus

$$\begin{aligned} ||P^{*k} - U|| \geq |P^{*k}(A) - U(A)| &\geq (P^{*k}(A) - U(A)) \\ &\geq (\frac{1}{e} - e^{-e^{-2c}}) + o(1). \end{aligned}$$

□

Remarks.

1) *On Lower Bounds.* The argument for the lower bound is satisfying in that it produces a simple set that "explains" why $\frac{1}{2}n \log n$ steps are needed. On the other hand, the computation involved two classical results which would not generally be available for other random walk problems. It is therefore of

some interest to see how the general approach to lower bounds works out in this case.

The general approach begins with the upper bound argument and chooses the difficult or "slowest" representations to construct a random variable to work with. In the present case, the difficult representation is $S^{n-1,1}$. A reasonable candidate for a random variable is thus the character χ of this representation. Observe that this is exactly the number of fixed points minus one. Under the uniform distribution

$$E_U(\chi(\pi)) = 0, \text{ Var}_U(\chi) = \frac{1}{|G|}\Sigma_\pi \chi^2(\pi) = (\chi|\chi) = 1.$$

Under the convolution measure, with ρ the $n-1$ dimensional representation,

$$E_k(\chi) = \Sigma P^{*k}(\pi) \text{ Tr}\rho(\pi) = \text{ Tr } \Sigma P^{*k}(\pi)\rho(\pi) = \text{ Tr } P^{\hat{*}k}(\rho) = (n-1)(1-\frac{2}{n})^k.$$

Observe that in order to drive this to its correct value zero, k must be $\frac{1}{2}n$ log $n+cn$ for c large. However, this is not enough to lower bound the variation distance since there are random variables with large means which are small with high probability. A second moment is needed. To compute this $E_k(\chi^2)$ is needed. Now χ^2 is the character of the tensor product of χ with itself. It is not difficult to argue that

$$\begin{array}{cccccccccc} & S^{n-1,1} \otimes S^{n-1,1} & = & S^n & \oplus & S^{n-1,1} & \oplus & S^{n-2,2} & \oplus & S^{n-2,1,1} \\ \text{dim} & (n-1)^2 & & 1 & & n-1 & & \frac{n(n-3)}{2} & & \frac{(n-1)(n-2)}{2}. \end{array}$$

An explicit proof of this result can be found on page 97 of James and Kerber (1981).

EXERCISE 13. Using the data above, compute $\text{Var}_k(\chi)$ and use Chebychev's inequality to show that $\frac{1}{2}n$ log $n - cn$ steps are not enough.

2) While limited, the approach developed above gives precise results for some other problems. To begin with, consider random transpositions. The identity is chosen much more frequently than any specific transposition. It is straightforward to carry out the analysis for the probability

$$P_n(\text{id}) = p_n, \; P_n(\tau) = \frac{1-p_n}{\binom{n}{2}}, \; P_n(\pi) = 0 \text{ otherwise }.$$

If $p_n = 1/(1+\binom{n}{2})$, all possible permutations are equally likely. In this case the argument shows that $k = c(n)n^2$ transpositions are needed where $c(n) \to \infty$ with n. This is somewhat surprising; usually, for a given support set, the probability that approaches the uniform most rapidly is uniform on the support set.

Similarly, any simple probability on S_n which is constant on conjugacy classes can be worked with. A key tool is a uniform bound on the characters developed by Vershik and Kerov (1981). A readable account of this is given by Flatto, Odlyzko and Wales (1985). They work out details for probabilities uniform on a fixed conjugacy class c (e.g., all 3 cycles). Their results imply that $\frac{1}{2}n$ log n steps are always sufficient. This is not surprising — choosing a random 3-cycle mixes more cards each time and should result in faster convergence.

A simple problem not covered by available techniques is the rate of convergence for a random involution ($\pi^2 = \text{id}$). There are $\Sigma_\rho d_\rho$ of these, which is asymptotically $(\frac{n}{e})^{n/2} \cdot e^{\sqrt{n}}/\sqrt{2}e^{\frac{1}{4}}$. For this and other properties of involutions see Stanley (1971, pg. 267). Such a measure is constant on conjugacy classes, but the asymptotics haven't been worked out. It is not hard to show that any non trivial conjugacy class generates S_n. See Arad and Herzog (1985). Thus there are many open problems.

Finally, it is straightforward to handle random walks based on measures constant on conjugacy classes of the alternating group A_n. The characters of A_n are available as simple functions of the characters of S_n. James and Kerber (1981) Chapter 2 give these results.

EXERCISE 14. Let n be odd. Let Q be uniform on the set of n cycles in A_n. Show that Q^{*2} is close to uniform for large n. (Hint: See formula 2.3.17 in James and Kerber (1981) or Stanley (1983).)

3) *Connections with Radon Transforms.* The analysis developed in this section has been applied to the study of uniqueness and inversion of the Radon transform by Diaconis and Graham (1985a). Here is a brief description: let G be a group with $d(s,t)$ a bi-invariant metric: $d(rs,rt) = d(sr,tr) = d(s,t)$. Let $f\colon G \to \mathbb{R}$ be a function. Suppose we are told not $f(s)$ but

$$\overline{f}(s) = \sum_{d(s,t)\leq a} f(t) \text{ for all } s \text{ and fixed } a.$$

When do these averages determine f? If $S = \{s\colon d(\text{id},s) \leq a\}$ the Radon transform is $\overline{f}(s) = I_S{*}f(s)$. Taking Fourier transforms, the Radon transform is unique if and only if $\hat{I}_S(\rho)$ is invertible for every irreducible representation ρ.

The study of this problem leads to interesting questions of probability and computational complexity even for groups as simple as Z_2^n. In this case, with d as Hamming distance, when $a = 1, f \to \overline{f}$ is $1-1$ iff n is even; when $a = 2$, iff n is not a perfect square; for $a \geq 4$ iff n is not in a finite set of numbers.

John Morrison (1986) derived exact results for this problem using Gelfand pair tools (Section F below). Jim Fill (1987) gives comprehensive results for Z_n. For applications to data analysis, see Diaconis (1983). For general background, see Bolker (1987).

For $G = S_n$, choose $d(\pi, \eta)$ as the minimum number of transpositions needed to bring π to η. This metric is further discussed in Chapter 6-B. For any bi-invariant metric, I_S is constant on conjugacy classes, so $\hat{I}_S(\rho) = cI$. For $a = 1, c = (1 + \binom{n}{2}\, r(\rho))$. Diaconis and Graham use this result and Frobenius' formula for $r(\rho)$ to argue that $f \to \overline{f}$ is invertible iff $n \in \{1,3,4,5,6,8,10,12\}$.

4) *Perfect codes.* Very similar computations occur in a seemingly different problem treated by Rothaus and Thompson (1966). Let G be a group and T be a subset of G. Say that T *divides* G if there is a set S in G such that each $g \in G$

has a unique representation $st = g$ with s in S and t in T. For example, if T is a subgroup, then T divides G. If $G = S_3$ and $T = \{\mathrm{id}, (12), (13), (23)\}$, then T does not divide S_3.

The construction of codes leads naturally to questions of divisibility: Let G be a group and $d(s,t)$ a G invariant metric on G (i.e., $d(s,t) = d(gs, gt)$). For example, G might be Z_2^n and d might be the Hamming distance, or G might be S_n and $d(s,t)$ might be Cayley's distance: the minimum number of transpositions required to bring s to t (see Chapter 6-B).

A subset $S \subset G$ is called a *code*; *S corrects k errors* if any two code words are at distance more than $2k$ apart; *S is perfect* if G is the disjoint union of balls of fixed radius centered at the elements of S.

Perfect codes are elegant efficient ways of coding data with minimum waste. On Z_2^n the perfect codes have been classified; see MacWilliams and Sloane (1977). The search for codes in other groups is an active area of research.

To see the connection with group divisibility, consider S_n with Cayley's distance. Take T to be a (solid) ball of radius k about the identity. Observe that T divides S_n if and only if there is a perfect code S of this radius — indeed, balls centered at points of S would be disjoint if $TS = S_n$ uniquely.

Rothaus and Thompson considered $k = 1$, i.e. T as the identity together with the set of all transpositions in S_n. To explain their result, observe that a necessary condition for divisibility is $(1 + \binom{n}{2}) | n!$ (after all, disjoint balls of radius 1 have to cover). This rules out $n = 3, 4, 5$ but not 6 for example. They proved that if $(1 + \binom{n}{2})$ is divisible by a prime exceeding $\sqrt{n} + 2$, then T does not divide S_n.

Their argument is *very* similar to the argument for analyzing repeated random transpositions. Interpret the equation $ST = G$ as an equation about the convolution of the indicator functions of the sets S and $T(f_S * f_T = 1$ say). Taking Fourier transforms at an irreducible representation leads to $c(\rho)\hat{f}_S(\rho) = 0$, where $c(\rho) = 1 + \binom{n}{2}\chi_\rho(\tau)/d_\rho$. Now one must study when $c(\rho)$ vanishes (see the previous remark). One really new thing in the Rothaus-Thompson paper is the skillful use of transforms at non-irreducible representations to give checkable divisibility constraints on n. The argument is fairly detailed and will not be given here. Sloane (1982) connects this work with the modern coding literature and gives many further applications. Chihara (1987) extends the results to Chevalley groups.

EXERCISE 15. Rothaus and Thompson report 1, 2, 3, 6, 91, 137, 733, and 907 as the only integers less than 1,000 which fail to satisfy the theorem. The naive criterion does 3, (and S_2 is divisible). Show that S_6 is not divisible.

5) *Varying the measure.* The ideas developed above can be used for some related problem like transpose a random card with the top card, or switch the top k cards with k randomly chosen cards. Here we have a measure on S_n invariant under conjugation by S_k and bi-invariant under S_{nk}. The Fourier transform can be shown to be diagonal with explicitly computable elements. See Diaconis (1986) or Greenhalgh (1988) for further details.

6) *Random reflections.* Similar analyses are possible for other random walks

constant on conjugacy classes. For example, let $G = 0_p$ — the p-dimensional orthogonal group. One practical problem involves algorithms for choosing a random element of G when p is large (e.g. $p = 256$). The usual algorithm begins with p^2 standard normal random variables X_{ij}, forms a matrix $M = \{X_{ij}\}$ and makes M orthogonal using the Gram-Schmidt algorithm. It is easy to show that this results in a random orthogonal matrix uniformly distributed on G. Diaconis and Shahshahani (1987a) discuss this and other algorithms. In carrying out the Gram-Schmidt algorithm, the ith row of M must be modified by subtracting out the inner product of all rows above it. This entails computation of $i - 1$ inner products. Each inner product involves p multiplications and additions. The whole procedure takes order $p \sum_{i=1}^{p} i = 0(p^3)$ operations. This is often too large for practical use.

Sloane (1983) contains a fascinating application to encrypting telephone conversations. Sloane suggested generating a matrix by using random reflections. Geometrically this involves choosing a random point U in the p-sphere and reflecting in the hyperplane orthogonal to U. Algebraically the matrix is $\Gamma = (I - 2UU')$. Observe that the distribution of Γ is constant on conjugacy classes because $\Gamma_1(I - 2UU')\Gamma_1' = (I - 2\Gamma_1 U(\Gamma_1 U)')$. If U is uniform on the p-sphere, $\Gamma_1 U$ is uniform as well. There is a straightforward extension of the upper bound lemma to compact groups. The analysis can be carried out to show that $\frac{1}{2}p \log p + cp$ steps are enough (while $\frac{1}{2}p \log p - cp$ steps are too few). Some details can be found in Diaconis and Shahshahani (1986a).

In this problem, P is singular with respect to the uniform distribution, but P^{*k} has a density for $k \geq p$. Thus variation distance bounds make sense. For random walks on continuous compact groups involving a discrete measure, the distribution is always singular and only bounds in a metic for the weak star topology can be hoped for.

7) *Random walks on linear groups over finite fields.* The problem described above can be carried out over other fields such as C (to generate a random unitary matrix) or F_q - a finite field with $q = p^d$ elements. Here is another problem which should lead to interesting mathematics. Let V be a vector space of dimension d over F_q. Let $SL_d(V)$ be the $d \times d$ invertible matrices with determinant 1. This is a finite group of order $q^{\binom{d}{2}} \Pi_{i=2}^{d}(q^i - 1)$. A *transvection* is a linear transformation in $SL_d(V)$ which is not the identity but fixes all elements of a hyperplane. Suzuki (1982, Sec. 9) shows that if $d \geq 3$, the transvections form a single conjugacy class that generates $SL_d(V)$. Thus, the question "how many random transvections are required to get close to the uniform distribution on $SL_d(V)$?" can be attacked by the method of this section.

E. The Markov chain connection.

1. Introduction.

There is another approach to random walks on groups: treat them as Markov chains with state space G and $|G| \times |G|$ transition matrix $Q(s,t) = Q(ts^{-1})$. In early attempts to understand the problem of random transpositions Joseph Deken did *exact* computations of the second largest eigenvalue for decks of size $n = 2, 3, \dots, 10$. He found it to be $(1 - 2/n)$. This is precisely the constant in the Fourier transform at the "slow" representation (see Theorem 5 of Section D). This striking numerical coincidence suggested that (a) the $(1-2/n)$ result must hold for all n, and (b) there is a close connection between the Markov chain and group representation approach. Some of this was worked out by Diaconis and Shahshahani (1981), who showed that the eigenvalues of the transition matrix are precisely the eigenvalues of $\hat{Q}(\rho)$, each appearing with multiplicity d_ρ.

The following discussion uses work of Matthews (1985). It results in a sort of diagonalization of the transition matrix and an exact determination of eigenvalues and eigenvectors where these are available. This allows us to use results from classical Markov chain theory.

2. A spectral decomposition of the transition matrix.

Let G be a finite group with elements $\{s_1, \dots, s_N\}, N = |G|$. For a probability Q on G, construct $Q(i,j) = Q(s_j s_i^{-1})$ — the chance that the walk goes from s_i to s_j in one step. Suppose that the irreducible representations are numbered $\rho_1, \dots, \rho_K$. Define

(1)
$$M_k = \begin{pmatrix} \hat{Q}(\rho_k) & & 0 \\ & \ddots & \\ 0 & & \hat{Q}(\rho_k) \end{pmatrix}$$

a $d_k^2 \times d_k^2$ block matrix with $\hat{Q}(\rho_k)$ the Fourier transform of Q at ρ_k.

(2) Let M be the $N \times N$ block diagonal matrix $\begin{pmatrix} M_1 & & 0 \\ & \ddots & \\ 0 & & M_K \end{pmatrix}$.

Suppose that a basis has been chosen so that each irreducible representation is given by a unitary matrix. Define

(3)
$$\psi_k(s) = \sqrt{\frac{d_k}{N}}(\rho_k(s)_{11}, \rho_k(s)_{21}, \dots, \rho_k(s)_{d_k 1}, \rho_k(s)_{12}, \dots, \rho_k(s)_{d_k d_k})^T,$$

a column vector of length d_k^2. Let $\phi(s) = (\psi_1(s)^T, \psi_2(s)^T, \dots, \psi_K(s)^T)^T$ be a column vector of length N obtained by concatenating the $\psi_k(s)$ vectors.

(4) Let ϕ be the $N \times N$ matrix $(\phi(s_1), \dots, \phi(s_N))$ and ϕ^* its conjugate transpose .

Theorem 6. *The transition matrix $Q(i,j)$ can be written*

$$Q = \phi^* M^* \phi \tag{5}$$

Remarks. The Schur orthogonality relations show that ϕ is a unitary matrix. So (5) is a decomposition similar to the traditional eigenvalue, eigenvector decomposition. It implies that each eigenvalue of $\hat{Q}(\rho)$ is an eigenvalue of $Q(i,j)$ with multiplicity d_ρ. Together these are all the eigenvalues of $Q(i,j)$. If M is diagonal (e.g. Q constant on conjugacy classes or bi-invariant on a Gelfand pair (Section F below)), then (5) is the spectral decomposition of Q with respect to an orthonormal basis of eigenvectors.

Proof of Theorem 6: The Fourier inversion theorem gives

$$Q(i,j) = \frac{1}{N}\sum_{k=1}^{K} d_k \mathrm{Tr}[\hat{Q}(\rho_k)\rho_k(s_i)\rho_k(s_j^{-1})] = \frac{1}{N}\sum_{k=1}^{K} d_k \mathrm{Tr}[\rho_k(s_j^{-1})\hat{Q}(\rho_k)\rho_k(s_i)].$$

Expanding the trace, this equals

$$\sum_{k=1}^{K} \psi_k(s_j)^* M_k \psi_k(s_i).$$

□

3. The first hit distribution.

Let G be a finite group and Q a probability on G. For $s,t \in G$, define F_{st}^n = the probability that t is first hit at time n starting at s. For $|z| < 1$ let $F_{st}(z) = \sum_{n=1}^{\infty} F_{st}^n z^n$.

Theorem 7. *For $|z| < 1, (I - z\hat{Q}(\rho))$ is invertible and*

$$F_{st}(z) = \frac{\Sigma_\rho d_\rho \mathrm{Tr}[\rho(st^{-1})(I - z\hat{Q}(\rho))^{-1}]}{\Sigma_\rho d_\rho \mathrm{Tr}[I - z\hat{Q}(\rho)]^{-1}}.$$

Proof. Using the notation of Section 2, $Q(z) = \sum_{n=1}^{\infty} z^n Q^n = \{\phi^*(I - zM^*)^{-1}\phi\}$. Kemperman (1961, pg. 18–19) gives the standard result

$$F_{st}(z) = Q_{st}(z)/Q_t(z).$$

The result follows from this and (5) above. It is given a direct independent proof in Section H. It is mentioned here to underscore the availability of the Markov chain machine in situations where all the eigenvalues and eigenvectors of the transition matrix are known.

4. On generalized circulants.

The technique we have developed for analyzing random walks gives rise to a class of "patterned matrices" for which we can explicitly determine all the eigenvalues and eigenvectors. Let G be a finite group of order g. Let $s_1, \ldots, s_g$ be an enumeration of the elements of G. Let P be a probability measure on G. *The transition matrix associated with* P is the $g \times g$ matrix with i, j entry $P(s_j s_i^{-1})$. If a random walk on G is thought of as a Markov chain with G as state space, the i, j entry is the probability of a transition from state s_i to state s_j. We have been working with measures which are constant on conjugacy classes. Generalizing this somewhat define a *G-circulant* as a $g \times g$ matrix with i, j entry $f(s_j s_i^{-1})$ with f constant on conjugacy classes.

Examples. If G is Abelian, then the equivalence classes consist of single elements. If G is cyclic, then a G circulant is an ordinary circulant: a $g \times g$ matrix in which each row is a cyclic shift of the first row. For $G = S_3$ the equivalence classes are $\{\text{id}\}, \{(1\ 2), (1\ 3), (2\ 3)\}\{(1\ 2\ 3), (1\ 3\ 2)\}$. If $f(\text{id}) = a, f(1\ 2) = b, f(1\ 2\ 3) = c$ and the group is labelled in order (using $\left(\begin{smallmatrix} 1 & 2 & 3 \\ \alpha & \beta & \gamma \end{smallmatrix}\right)$ notation) (1 2 3)(1 3 2)(2 1 3)(2 3 1)(3 1 2)(3 2 1), we get

$$\begin{pmatrix} a & b & b & c & c & b \\ b & a & c & b & b & c \\ b & c & a & b & b & c \\ c & b & b & a & c & b \\ c & b & b & c & a & b \\ b & c & c & b & b & a \end{pmatrix}$$

value	typical vector	dim
$a + 3b + 2c$	(1 1 1 1 1 1)	1
$a - 3b + 2c$	(1 −1 −1 1 1 −1)	1
$a - c$	(2 0 0 −1 −1 0)	4

Let G be the 8 element quarternion group $G = \{\pm 1, \pm i, \pm j, \pm k\}$ with multiplication given by
$$\begin{tikzcd} & i \arrow[dr] & \\ k \arrow[ur] & & j \arrow[ll] \end{tikzcd}$$
. Thus $ij = k, kj = -i$, etc. There are five conjugacy classes: $\{+1\}, \{-1\}, \{\pm i\}, \{\pm k\}$. Let them have weight a, b, c, d, e. Label the group 1, -1, i, -i, j, -j, k, -k. We get

$$\begin{pmatrix} a & b & c & c & d & d & e & e \\ b & a & c & c & d & d & e & e \\ c & c & a & b & e & e & d & d \\ c & c & b & a & e & e & d & d \\ d & d & e & e & a & b & c & c \\ d & d & e & e & b & a & c & c \\ e & e & d & d & c & c & a & b \\ e & e & d & d & c & c & b & a \end{pmatrix}$$

value	typical vector	dim
$a + b + 2c + 2d + 2e$	(1 1 1 1 1 1 1 1)	1
$a + b + 2c - 2d - 2e$	(1 1 1 1 −1 −1 −1 −1)	1
$a + b + 2d - 2c - 2e$	(1 1 −1 −1 1 1 −1 −1)	1
$a + b + 2e - 2c - 2d$	(1 1 −1 −1 −1 −1 1 1)	1
$a - b$	(1 −1 0 0 0 0 0 0)	2

Theorem 8. *Let M be a G-circulant. Then M has an eigenvalue λ_ρ for each irreducible representation ρ of G,*

$$\lambda_\rho = \frac{1}{d_\rho} \Sigma f(g) \chi_\rho(g),$$

the eigenvalue λ_ρ occurs with multiplicity d_ρ^2.

Proof. The spectral decomposition of Section 2 above proves a stronger result: it gives the eigenvectors as an explicit arrangement of the matrix entries of the irreducible representations.

Remarks.

1. There is a lovely book called *Circulant Matrices* by Phillip Davis (1979). It seems like a nice project to go through the book and generalize all the results to G-circulants.
2. Note that the character vector $(\chi_\rho(s_1)\ldots\chi_\rho(s_g))$ is always an eigenvector for λ_ρ.
3. The argument generalizes easily to a transitive action of G on a finite set X. If P is a probability on G, then P induces a Markov chain on X. The transition matrix of this chain has the same eigenvalues as the matrices $\hat{P}(\rho)$, where ρ runs over the irreducible representations of G that appear in the permutation representation of G on X. This is developed in Section F which follows.
4. Example 3 of Section C suggests some further extensions. This begins with the Markov chain $X_n = 2X_{n-1} + \in_n \pmod{p}$ with $\in_i$ i.i.d. taking values $0, \pm 1$ with probability $\frac{1}{3}$. The transition matrix M of this chain is not a circulant, but the argument shows that its ath power is a circulant, where a is the order of 2 (mod p). Thus one knows, up to an ath root of unity, all the eigenvalues of M. Remark 2 of the example suggests many further situations where a similar analysis is possible.

F. Random walks on homogeneous spaces and Gelfand pairs.

1. Homogeneous spaces.

There is an extension of the basic set up which is useful. It involves the Markov chain induced by a random walk under the action of a group. This arises in some of the introductory examples: for instance, in considering the recurrence $X_n = a_nX_{n-1} + b_nX_{n-2} \pmod{p}$, a random walk on 2×2 matrices was considered. The matrices act on pairs (X_n, X_{n-1}). To understand the behavior of X_n it is not necessary to bound the rate of convergence on the group of matrices, but only on the set of non-zero pairs. Similarly, the grand tour example in section A4 only involved the action of the orthogonal group on lines or planes.

Definition. Let G be a finite group and X be a finite set. An *action* of G on X is a mapping from $G \times X \to X$ which we will denote $(s, x) \to s \cdot x$ or simply sx. It must satisfy: id $\cdot x = x$ and $s \cdot (t \cdot x) = (st) \cdot x$. Define an equivalence on X by $x \sim y$ if for some $s \in G, sx = y$. The equivalence classes are called *orbits. G operates transitively on X* if there is only one orbit. A set with a group acting transitively is called a *homogeneous space.*

When G operates transitively, the following canonical representation of X is useful. Fix $x_0 \in X$. Let N — the *isotropy subgroup* of x_0 - be the set of $s \in G$ with $sx_0 = x_0$. The group G acts on the coset space G/N. There is an isomorphisim between X and G/N respecting the action of G. We will identify

X with $x_0, x_1, \ldots, x_n$, a set of coset representatives for N in G. It will always be assumed that $x_0 = \text{id}$.

Example 1. The symmetric group S_n acts on $\{1, 2, \ldots n\}$ transitively. The isotropy subgroup is isomorphic to S_{n-1} — as all permutations fixing 1. Coset representatives can be chosen as the identity and the transpositions $(12), \ldots, (1n)$.

A probability P on G induces a probability $\tilde{P}$ on $X = G/N$ by $\tilde{P}(x_i) = P(x_i N)$. Similarly, if P^{*k} denotes the convolution of P with itself k times, P^{*k} induces a probability on X. We can think of a process with values in G, say id, s_1, $s_2 s_1$, $s_3 s_2 s_1, \ldots$. The induced process in X is x_0, $s_1 x_0$, $s_2 s_1 x_0, \ldots$.

EXERCISE 16. Let the finite group G act on the finite set X, partitioning X into orbits θ_i. If P and Q are probabilities on X which are G-invariant, then

$$||P - Q|| = \frac{1}{2}\Sigma|P(\theta_i) - Q(\theta_i)|.$$

Thus, the variation distance between P and Q equals the distance between their restrictions to the set of orbits. This is a special case of the following result: if P and Q are probabilities on a σ-algebra $\mathcal{F}$ and if a sub-σ-algebra $B \subset \mathcal{F}$ is sufficient for P and Q, then $||P - Q||_{\mathcal{F}} = ||P - Q||_{\mathcal{B}}$. See Diaconis and Zabell (1982) for details and applications.

LEMMA 3. *Let G act transitively on X. Let P be a probability on G. The induced process is a doubly stochastic Markov chain on X with transition matrix $P_x(y) = P(yNx^{-1})$.*

Proof. For the induced processes, the chance of going from x to y in one step is $P_x(y)$ defined as $P\{s: sx = y\} = P\{yNx^{-1}\}$. For a Markov chain, the chance of going from x to y in two steps is of course

$$P_x^2(y) = \sum_z P_x(z) P_z(y).$$

The chance that the chain in question is at y in two steps is

$$P * P(yN) = \sum_s P(yNs^{-1})P(s).$$

Let $s = xn$, we get

$$= \sum_{x,n} P(yNn^{-1}x^{-1}) \cdot P(xn) = \sum_x P(yNx^{-1}) \cdot P(xN) = P_{x_0}^2(y).$$

The last computation is essentially the inductive step of a proof that the measure induced by P^{*k} on X equals $P_{x_0}^k(y)$. □

To state the next result, introduce $L(X)$ — the set of all functions from X into the complex numbers. The action of G on X induces an action of G on $L(X)$

by $sf(x) = f(s^{-1}x)$. This is a $1-1$ linear mapping of $L(X)$, and so yields a representation of G. The representation splits into a direct sum of irreducible representations ρ.

LEMMA 4. *(Upper bound lemma). Let G operate transitively on the finite set X. Let N be the isotropy subgroup. Let P be a right N invariant probability on $G, \tilde{P}$ the induced probability on X, and U the uniform distribution on X. Then*

$$||P - U||^2 \le \frac{1}{4}\Sigma^* d_\rho Tr\{\hat{\tilde{P}}(\rho)\hat{\tilde{P}}(\rho)^*\}$$

where the sum is over all nontrivial irreducible representations that occur in $L(X)$.

Proof. Let $\tilde{U}$ be the uniform probability on G.

$$\begin{aligned}(\Sigma_x|P(x) - U(x)|)^2 \le |X|\Sigma_x|P(x) - U(x)|^2 &= |X|\,|N|\Sigma_s|\tilde{P}(s) - \tilde{U}(s)|^2 \\ &= \Sigma^*_\rho d_\rho \mathrm{Tr}(\hat{\tilde{P}}(\rho)\hat{\tilde{P}}(\rho)^*).\end{aligned}$$

In the last step, the Plancherel theorem was used together with the facts that a right N invariant function has zero Fourier transform if ρ does not occur in $L(X)$. This follows from the following lemma and remark. □

LEMMA 5. *Let ρ, V be an irreducible representation of the finite group G. Let $N \subset G$ be a subgroup and $X = G/N$ the associated homogeneous space. The number of times that ρ appears in $L(X)$ equals the dimension of the space of N fixed vectors in $\rho, V(=$ $dim\{v \in V\colon \rho(n)v = v$ for all $n \in N\})$.*

Proof. Let $\{\delta_x(\cdot)\}$ be a basis for $L(X)$. The character χ for the representation of G on $L(X)$ is

$$\begin{aligned}\chi(s) &= |\{x\colon \delta_x(s^{-1}y) = \delta_x(y)\}| = |x\colon sx = x| \\ &= |x\colon x^{-1}sx \in N|.\end{aligned}$$

Now, the number of times ρ appears in this representation is

$$\begin{aligned}(\chi_\rho|\chi) &= \frac{1}{|G|}\sum_s \chi_\rho(s)\chi(s) = \frac{1}{|G|}\sum_s \chi_\rho(s) \sum_{\substack{x,n \\ x^{-1}sx=n}} 1 \\ &= \frac{1}{|G|}\sum_n \chi_\rho(n) \sum_{\substack{s,x \\ x^{-1}sx=n}}\end{aligned}$$

But, for any fixed n,

$$\sum_{\substack{s,x \\ x^{-1}sx=n}} 1 = |X|.$$

To see this observe that for fixed $n, x, s = xnx^{-1}$ is determined. Further, if $x^{-1}sx = n$, then $(tx)^{-1}tst^{-1}(tx) = n$ for all $t \in G$. Since G operates transitively on X, for every $y \in X$ there is a unique s^* such that $y^{-1}s^*y = n$.

Since $|G|/|X| = |N|$,

$$(\chi_\rho|\chi) = \frac{1}{N}\Sigma_n\chi_\rho(n) = (\chi_\rho|1)_N.$$

The right side is the number of times the trivial representation appears in χ_ρ restricted to N. This is just the dimension of the space of N fixed-vectors. □

Remarks. Lemma 5 is a special case of Frobenius' reciprocity formula. The representation $L(X)$ is the trivial representation of N induced up to G. Frobenius' formula says the number of times a representation ρ of G appears in the induction of λ (a representation of N) to G equals the multiplicity of λ in ρ restricted to N. Chapters 3 and 7 of Serre (1977) give further development. The general result is proved by essentially the same combinatorial argument. For present purposes, Lemma 5 is all that is needed.

Using Lemma 5, if ρ does not occur in $L(X)$, the trivial representation does not occur in ρ restricted to N. Now, the orthogonality relations (Corollary 2 of Schurs lemma in Chapter 2) yield $\Sigma_n\rho(n) = 0$. For a right N invariant function f on G,

$$\hat{f}(\rho) = \Sigma_x\ f(x)\ \rho(x)\Sigma_n\ \rho(n) = 0.$$

This completes the proof of the upper bound Lemma 4.

The next section discusses a collection of examples where a huge simplification occurs.

2. *Gelfand pairs*

This is a class of examples where the Fourier analysis becomes simple. Consider, as above, a group G acting transitively on a finite set X with isotropy subgroup N. A function $f: G \to C$ is called *N-bi-invariant* if $f(n_1sn_2) = f(s)$ for all $s \in G, n_1, n_2 \in N$.

Definition. G, N is called a *Gelfand pair* if the convolution of N bi-invariant functions is commutative.

One value of this definition comes from a long list of examples. Some of these are discussed later in this section. Letac (1981, 1982) or Bougerol (1983) present very readable surveys of available results. The following theorem is basic:

Theorem 9. *The following three conditions are equivalent*

(1) *G, N is a Gelfand pair.*

(2) *The decomposition of $L(X)$ is multiplicity free.*

(3) *For every irreducible representation (ρ, V) there is a basis of V such that $\hat{f}(\rho) = \begin{pmatrix} * & 0 \\ 0 & 0 \end{pmatrix}$ (a matrix with zero entries except perhaps in the $(1,1)$ position) for all N-bi-invariant functions f.*

Proof. Assume (2), so $L(X) = V_1 \oplus V_2 \oplus \ldots \oplus V_m$ say. This is multiplicity free, so by Lemma 5 above, each V_i has a non-trivial one-dimensional space of N invariant functions. Choose so called spherical functions $s_i \in V_i$ to be left N-invariant and normalized so $s_i(\mathrm{id}) = 1$. Complete s_i to a basis for V_i chosen so $\rho_i(n) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & * \end{smallmatrix}\right)$ for all $n \in N$ (the top block is 1×1, the bottom block is $(d_i - 1) \times (d_i - 1)$).

For f an N-bi-invariant function,

$$\hat{f}(\rho_i) = \sum_s f(t)\rho_i(t) = \sum_{x,n} f(xn)\rho_i(xn) = \sum_x \rho_i(x) f(x) \sum_n \rho_i(n).$$

But ρ_i restricted to N has a one-dimensional space of fixed vectors. By the orthogonality relations for the matrix entries, the (a, b) entry satisfies

$$\sum_n \rho_i^{ab}(n) = \begin{cases} |N| & \text{if } a = b = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Thus $\hat{f}(\rho_i)$ has the form $\Sigma f(x)\left(\begin{smallmatrix} * \\ \vdots \\ * \end{smallmatrix}\ 0\right) = \left(\begin{smallmatrix} * \\ \vdots \\ * \end{smallmatrix}\ 0\right)$. This argument works for any right invariant function f. For left invariant f, a similar argument shows that $\hat{f}(\rho_i)$ has form $\left(\begin{smallmatrix} * * \cdots * \\ 0 \end{smallmatrix}\right)$. From Lemma 5, if ρ does not appear in $L(X)$, $\hat{f}(\rho) = 0$. This shows (2) implies (3).

Clearly (3) implies (1) by taking Fourier transforms. To finish off, suppose (1) but some ρ_i has multiplicity $j > 1$ in $L(X)$. Pick a basis of V_i with first j coordinates spanning the N-invariant space. Take M_1, M_2 any two non-commuting $j \times j$ matrices. Define f_1, f_2 on G by

$$\begin{gathered} \hat{f}_1(\rho) = \hat{f}_2(\rho) = 0 \text{ if } \rho \neq \rho_i \\ \hat{f}_1(\rho_i) = \begin{pmatrix} M_1 & 0 \\ 0 & 0 \end{pmatrix}, \ \hat{f}_2(\rho_i) = \begin{pmatrix} M_2 & 0 \\ 0 & 0 \end{pmatrix}. \end{gathered}$$

By Fourier inversion, these are non-zero, N-bi-invariant functions and $f_1 * f_2 \neq f_2 * f_1$. □

COROLLARY. *Let (G, N) be a Gelfand pair with $L(X) = V_1 \oplus \ldots \oplus V_m$. Each V_i contains a unique N-invariant function s_i with $s_i(id) = 1$. If the Fourier transform of an N invariant probability P on X is defined by*

$$\hat{P}(i) = \Sigma_x s_i(x) P(x),$$

then, for U the uniform distribution on X

$$||P_{x_0}^k - U||^2 \leq \frac{1}{4} \sum_{i=1}^m d_i |\hat{P}(i)|^{2k}.$$

Remarks. The corollary follows from the theorem above and the upper bound lemma of the last section. The s_i are called *spherical functions.* They have been explicitly computed for many groups. From part (3) of the theorem $s_i(x) = \rho_i(x)_{11}$. This sometimes serves as a definition, or as a way of computing spherical functions: take (ρ_i, V_i), a unitary representation that appers in $L(X)$. By the theorem, V_i contains a one-dimensional space of N fixed vectors. Let u be a unit N-fixed vector. Then $s_i(x) =< \rho_i(x)u, u >$. The s_i are left N invariant functions on X. They are also N-bi-invariant functions on G. Turning things around, if the spherical functions are known, the * in Theorem 9-3 can be computed as $\sum\limits_{t \in G} f(t)s(t)$.

EXERCISE 17. Let χ_i be the character of a representation ρ_i that appears in $L(X)$. Show

$$s_i(x) = \frac{1}{|N|} \sum_{n \in N} \chi_i(xn).$$

Thus the spherical functions are determined by characters.

3. *Example: The Bernoulli-Laplace model of diffusion.*

As a specific example of the techniques discussed above consider the following model of diffusion suggested originally by Daniel Bernoulli and Laplace. Feller (1968, p. 378) contains the history. There are two urns, the first containing n red balls, the second containing n white balls. At each stage, a ball is picked at random from each urn and the two are switched. Evidently, many switches mix things up and it is not difficult to show that once things reach equilibrium they evolve (approximately) as an Ornstein-Uhlenbeck process (at least for large n). The problem is, how many switches are required to reach equilibrium? In what follows, we show that $\frac{n}{4} \log n + cn$ switches suffice.

It is just as simple to solve the same problem with r red balls in the first urn and b black balls in the second urn. Let $n = r + b$. A convenient mathematical model for this has $X = S_n/S_r \times S_b$; thus X can be thought of as the set of r element subsets of a set with n elements. For $x, y \in X$ define the distance $d(x,y) = r - |x \cap y|$. This is a metric (see Chapter 6-D), and the random walk problem becomes the following: start at $x_0 = \{1, 2, \ldots, r\}$. Choose an element inside x_0 and an element outside x_0 and switch them. This chooses a set x at distance one from x_0 at random. The following result is proved by Diaconis and Shahshahani (1987b).

Theorem 10. *For nearest neighbor random walk on the r sets of an n set, if* $k = \frac{r}{2}(1 - \frac{r}{n}) \log n + cr$ *then*

$$||P_{x_0}^k - U||^2 \leq ae^{-dc}$$

for positive universal constants a and d.

Proof. Without loss, take $r \leq n/2$. The space decomposes as $L(X) = V_0 \oplus V_1 \oplus \ldots \oplus V_r$ where V_i is the irreducible representation of the symmetric group S_n corresponding to the partition $(n - i, i)$. James (1978) gives this result as well as $\dim(V_i) = \binom{n}{i} - \binom{n}{i-1}$. In particular, the pair $(S_n, S_r \times S_b)$ is a Gelfand pair.

The spherical functions have been determined by Karlin and McGregor (1961) in studying an equivalent formulation in a genetics example (Moran's model). Stanton (1984) contains this result in modern language. The spherical functions turn out to be classically studied orthogonal functions called the Dual Hahn or Eberlein polynomials. The function $s_i(x)$ only depends on the distance $d(x, x_0)$ and is a polynomial in d given by

$$s_i(d) = \sum_{m=0}^{i} \frac{(-i)_m(i-n-1)_m(-d)_m}{(r-n)_m(-r)_m m!} \quad 0 \le i \le r,$$

where $(j)_m = j(j+1)\dots(j+m-1)$. Thus,

$$s_0(d) = 1, \; s_1(d) = 1 - \frac{nd}{r(n-r)},$$
$$s_2(d) = 1 - \frac{2(n-1)d}{r(n-r)} + \frac{(n-1)(n-2)d(d-1)}{(n-r)(n-r-1)r(r-1)}.$$

The basic probability P for this problem is supported on the $r(n-r)$ sets of distance one from the set $\{1, \dots, r\}$. Thus the Fourier transform of P at the ith spherical function is

$$\hat{P}(i) = s_i(1) = 1 - \frac{i(n-i+1)}{r(n-r)} \quad 0 \le i \le r. \tag{3}$$

Now the corollary to Theorem 9

$$||P^k - U||^2 \le \frac{1}{4}\sum_{i=1}^{r}\{\binom{n}{i} - \binom{n}{i-1}\}(1 - \frac{i(n-i+1)}{r(n-r)})^{2k}.$$

To bound this sum, consider first the term for $i = 1$,

$$(n-1)(1 - \frac{n}{r(n-r)})^{2k}.$$

This is essentially

$$e^{-\frac{2kn}{r(n-r)} + \log n}.$$

Thus k must be $\frac{r}{2}(1 - \frac{r}{n}) \log n$ at least to kill this term. If $r = n/2$, this becomes $\frac{r}{4} \log r$. If $r = o(n)$, this becomes $\frac{r}{2} \log n$.

Next consider the final term

$$\left(\binom{n}{r} - \binom{n}{r-1}\right)(\frac{1}{n-r})^{2k}.$$

This is certainly bounded above by

$$\frac{n^r}{r!}\frac{1}{(\frac{n}{2})^{2k}} = e^{-2k \log \frac{n}{2} + r \log n - \log r!}.$$

In any case, if k is of order $\frac{r}{2}(1 - \frac{r}{n}) \log n$, this tends to zero exponentially fast. The intermediate terms are always geometrically smaller than the extreme terms, just as with the argument for random transpositions. Further details are in Diaconis and Shahshahani (1987b). □

Remark 1. As described in Section E, the analysis gives a precise formula for the eigenvectors and eigenvalues of the transition matrix of this problem treated as a Markov chain. Karlin and McGregor (1961) essentially derived this result without using group theory. Their application was to a similar problem arising as a genetics model due to Moran. A clear discussion of Moran's model can be found in Ewens (1979, Sec. 3.3). Diaconis and Shahshahani give applications to a learning problem discussed by Piaget.

Remark 2. As usual with approximation, some precision has been lost to get a clean statement. The basic result is the bound of the corollary to Theorem 9. When $r = 1$ for example there is only one term: $(n-1)(\frac{1}{n-1})^{2k}$. For $k = 1$ taking square roots gives $\frac{1}{2}\frac{1}{\sqrt{n-1}}$ as an upper bound for the variation distance. Elementary considerations show that the exact distance in this case is $1/n$. Here, when n is large, use of the upper bound lemma gives the right answer for the number of steps required (namely 1) to make the distance small but an overestimate for the distance itself.

EXERCISE 18. Consider two urns, the left containing n red balls, the right containing n black balls. At each stage "a" balls are chosen at random from each urn and the two sets are switched. Show that this is bi-invariant. Show that for fixed a, as $n \to \infty$, this speeds things up by a factor of a (so $\frac{1}{4a}n \log n$ moves suffice).

Remark 3. A reasonable number of other problems allow very similar analysis. Stanton (1984) contains a list of finite homogeneous spaces arising from Chevalley groups where (a) the associated $L(X)$ is a Gelfand pair, and (b) the spherical functions are explicitly known orthogonal polynomials. One case of particular interest is a walk in the set of r-dimensional subspaces of an s dimensional vector space over a finite field. See Greenhalgh (1988) for details. In all cases, there is a natural metric so that nearest neighbor walks on X allow a 1-dimensional analysis. For the example of r-dimensional subspaces the distance is $d(x,y) = r - \dim(x \cap y)$.

A special case of this analysis is nearest neighbor walk on the cube $X = Z_2^n$. Here G can be represented as the semi-direct product of Z_2^n with S_n. This is the group of pairs (x, π) for $x \in Z_2^n$, $\pi \in S^n$. It acts on $y \in Z_2^n$ by $(x,\pi)(y) = \pi y + x$. Multiplication in G is composition of repeated transformations. Choosing $x_0 = 0$, the isotropy subgroup is $N = \{(0,\pi) : \pi \in S_n\} \cong S_n$. It is easy to verify that $L(X) = V_0 \oplus V_1 \oplus \ldots \oplus V_n$ where V_j is the subspace spanned by the functions $\{f_y(x)\}_{|y|=j}$ and $f_y(x) = (-1)^{x \cdot y}$. Thus G, N is a Gelfand pair and $\dim V_j = \binom{n}{j}$. The spherical functions $s_j(x)$ again only depend on $d(x,0)$ (with $d(x,y) = \#\text{places}(x_i \neq y_i)$) and are polynomials in d called Krawtchouk polynomials:

$$s_j(d) = \frac{1}{\binom{n}{j}} \sum_{m=0}^{j} (-1)^m \binom{d}{m}\binom{n-d}{j-m}.$$

The upper bound found by treating this problem as a Gelfand pair is the same as the upper bound by treating the problem on the group Z_2^n (Example 2 of Section C).

Remark 4. The theory of Gelfand pairs can be developed without using group theory. One advantage of the present program is that it offers a route to follow for problems where the representation is not multiplicity free. For example, consider the Bernoulli-Laplace urn model with 3 urns; the first containing n red, the second containing n white, the third containing n blue balls. At each stage, a pair of urns is chosen at random, then a randomly picked pair of balls is switched. Analysis of the contents of even the first urn is complicated by the fact that the associated representation of S_{3n} on $L(X)$, with $X = S_{3n}/S_n \times S_n \times S_n$, has multiplicity. (See Young's rule in Chapter 7.) This is an open problem.

There is a useful sufficient condition for showing that (G, N) is Gelfand without explicitly determining the decomposition of $L(X)$.

LEMMA 6. *(Gelfand's lemma). Let τ be $1 - 1$ homomorphism $\tau{:}\, G \to G$ with the property $s^{-1} \in N\tau(s)N$ for all $s \in G$. Then (G, N) is a Gelfand pair.*

Proof. Note first that for bi-invariant functions $f(s^{-1}) = f(\tau(s))$ and $\tau(N) \subset N$. If f is bi-invariant, define $\check{f}(s) = f(s^{-1})$, $f^\tau(s) = f(\tau(s))$. Thus $\check{f} = f^\tau$. Now $f * g(t) = \Sigma_s f(ts^{-1})g(s)$, so

$$f \check{*} g(t) = \Sigma_s f(t^{-1}s^{-1})g(s) = \Sigma_z f(z^{-1})g(zt^{-1}) = \Sigma_z \check{f}(z)\check{g}(tz^{-1}) = \check{g} * \check{f}(t),$$
$$(f * g)^\tau(t) = \Sigma_s f(\tau(t)s^{-1})g(s) = \Sigma f(\tau(t)\tau(s)^{-1})g(\tau(s)) = f^\tau * g^\tau(t).$$

It thus follows that for all bi-invariant f, g

$$\check{f} * \check{g} = g \check{*} f = (g * f)^\tau = g^\tau * f^\tau = \check{g} * \check{f},$$

so $f * g = g * f$. □

Example 1. Let N be any group, A an Abelian group and suppose N acts on A. Form the semi-direct product $G = N \times_s A$ as the set of pairs (n, a) with $(n_2, a_2)(n_1, a_1) = (n_2n_1, n_2a_1 + a_2); (n, a)^{-1} = (n^{-1}, -n^{-1}a)$. These are all Gelfand pairs as one sees by considering the $1 - 1$ homomorphisms $\tau(n, a) = (n, -a)$. This satisfies $(n, a)^{-1} = (n^{-1}, 0)(n, -a)(n^{-1}, 0)$.

As examples we have the dihedral groups, the group of the cube $(S_n \times_s Z_2^n)$, the affine group $Z_m^* \times_s Z_m$. The Euclidean group $SO^d \times_s \mathbb{R}^d$ is also a Gelfand pair.

Example 2. (Groups of isometries). Let (X, d) be a finite metric space on which G acts. Suppose d is G invariant. Say G acts 2 *point homogeniously* if for all (x_1, y_1), (x_2, y_2) with $d(x_1, y_1) = d(x_2, y_2)$ there is an s such that $sx_1 = x_2$, $sy_1 = y_2$. Observe that G operates transitively (take $x_1 = y_1$, $x_2 = y_2$, then $sx_1 = x_2$ for some s). Pick $x_0 \in X$, let $N = \{s \in G{:}\, sx_0 = x_0\}$. Then (G, N) is Gelfand with $\tau(s) = s$. To see this observe $d(x_0, sx_0) = d(x_0, s^{-1}x_0)$. Thus there is an n so $nsx_0 = s^{-1}x_0$. This implies $sns \in N$, so $s^{-1} \in NsN$.

There are *many* special cases for this construction – most notably graphs whose automorphism groups act 2 point homogeniously. Biggs (1984) gives an extensive list, and a detailed recipe for determining the associated spherical functions. As a special case, consider X as the k sets of an n set with distance $d(x,y) = k - |x \cap y|$. The symmetric group S_n operates 2 point homogeniously. The isotropy subgroup is $S_k \times S_{n-k}$, and we have recaptured the Bernoulli-Laplace model. A continuous example has $X = S^n$ (the n-sphere), $G = SO(n)$.

It is interesting to know when (G, N) can be shown Gelfand by the existence of a homomorphism τ. Diaconis and Garsia (1988) show that $\tau(s) = s$ works if and only if the representation of G in the space of real functions on X is multiplicity free. They also present counter examples (a Gelfand pair that doesn't admit an automorphisim) and discussion.

We have seen that Fourier analysis of bi-invariant functions on a Gelfand pair offers a rich theory and collection of examples. The commutativity, which makes life so easy here, is also present in the analysis of functions that are constant on conjugacy classes. It is not surprising that one can be regarded as a special case of the other.

EXERCISE 19. Let G be a finite group. Let $G \times G$ act on G by $(s,t)x = s^{-1}xt$. The isotropy subgroup N in $G \times G$ is isomorphic to G as is the quotient space X. Show that $(G \times G, N)$ is a Gelfand pair. The decomposition of $L(X)$ is into $\rho \oplus \tilde{\rho}$ with $\tilde{\rho}(s) = \rho^*(s^{-1})$ as ρ varies over irreducible representations of G. These are all distinct irreducible representations of $G \times G$. Find the spherical functions in terms of the characters and show how Fourier analysis of N invariant functions on X via Gelfand pair techniques is the same as Fourier analysis of functions constant on conjugacy classes as developed in section D.

There are two generalizations of Gelfand pairs worth mentioning here: association schemes and Hypergroups.

An *association scheme* is a finite set X with a collection of relations $R_0, R_1, \ldots, R_d$. Take R_i as a zero-one matrix indexed by X with a 1 in position (x,y) if x and y are related in the ith relation. The R_i's satisfy axioms: (1) $R_0 = \text{Id}$, (2) $\Sigma R_i = J$ (matrix of all ones), (3) for every i there is an i' such that $R_i^t = R_{i'}$, (4) $R_i R_j = \Sigma p_{ij}^k R_k$ for non-negative integers p_{ij}^k. If $R_i R_j = R_j R_i$, the association scheme is called commutative.

Commutative association schemes have an interesting elementary theory. MacWilliams and Sloane (1981) give an efficient development. Bannai and Ito (1986, 1987) give a very well done encyclopedic treatment.

Because of (4) the set of all linear combinations of the R_i form a_n algebra. For commutative association schemes the R_i can be simultaneously diagonalized. For many examples, this diagonalization is very explicit.

As one example, take G a group, H a subgroup with $X = G/H$ and (G, H) a Gelfand pair. Then G acts on $X \times X$ by $g(x,y) = (gx, gy)$. Take the orbits of this action as relations on X. These relations form a commutative association scheme with algebra isomorphic to the convolution algebra $L(X)$.

In the other direction, consider a commutative association scheme X. The axioms imply that every row and column of R_i have the same number k_i of ones in each row and column. Thus R_i/k_i is a doubly stochastic matrix. If $w_i \geq 0$ sum to 1,

$$M = \sum_{i=0}^{d} \frac{w_i}{k_i} R_i$$

is doubly stochastic and so defines a Markov chain on X. The point is, for hundreds of examples, this Markov chain is explicitly diagonalizable using available information. Classical Markov chain techniques can then be used to derive answers to the usual questions. Diaconis and Smith (1987) derive an appropriate upper bound lemma and carry through some examples that don't arise from groups.

Association schemes were originally developed by statisticians for analysis of variance problems. Speed (1987) shows how they have come to life recently for new statistical applications. Coding theorists, combinatorialists, and finite group theorists have been the principal developers of assocation schemes in recent years. Bannai and Ito (1986, 1987) survey these developments and examples.

A *Hypergroup* begins with a set X and introduces a product on probabilities on X — so the product of two pointmasses is a probability (which is not usually a point mass). For example, a product can be introduced on the conjugacy classes of a group: e.g. in the symmetric group, the product of two transpositions can be the identity, a 3-cycle on the product of two 2-cycles. These occur with certain mass. As a second example, the set of irreducible representations on a compact group can be made into a Hypergroup using the Tensor product and its associated weights.

A reasonable amount of theory and examples have been developed. There has started to be a payoff to more classical areas. For example, Bochner's theorem for Gelfand pairs or class functions follows from Hypergroup Theorems of Johanson (1981). It is still unknown in general cases. Gallardo (1987) presents a nice example of Fourier analysis for a class of birth and death processes that is available by interpreting the decomposition of Tensor products on $SU(2)$ as rules for births and deaths. Zeuner (1987) gives a unified treatment of central limit problmes on Hypergroups and pointers to related literature.

Hypergroups offer a continuous generalization of association schemes. They appear to offer an extension worth keeping track of.

There are many further topics to discuss relating to Gelfand pairs. The interested reader is referred to the annotated bibliography in Section G.

G. Some references on Gelfand pairs.

The literature on Gelfand pairs is already sizeable. I hope the following annotated bibliography will help. The articles by Bougerol and Stanton are very clear and give details. The articles by Sloane and Heyer have extensive bibliographies.

Bailey, R. and Rowley, C. A. (1987). General balance and treatment permutations. Techical Report, Statistics Department, Rothamsted Experimental Station. Harpenden, Herts, AL5 2JQ, United Kingdom.

This paper is important in offering a bridge betwen the mathematics of Gelfand pairs and an important component of designed experiments — gener-

alized balance. Many experimental designs are constructed using group theory. The paper shows that many such designs automatically have nice statistical properties.

Biggs, N. (1974). *Algebraic Graph Theory.* Cambridge University Press, London.

Chapters 20, 21 discuss "distance transitive graphs." These are what we called two-point homogeneous. Graph theorists have derived lots of facts about the eigenvalue, eigenvectors of these groups redeveloping the tools of Gelfand pairs.

Bougerol, P. (1983). *Un Mini-Cours Sur Les Couples de Guelfand.* Pub. du Laboratoire de Statistique et Probabilities, Universite Paul Sabatier, Toulouse.

A terrific set of lectures with complete proofs and no "baloney," many examples.

Bougerol, P. (1981). Theoreme Central Limite Local Sur Certains Groupes de Lie. *Ann. Scient. Ec. Norm. Sup. 4th Ser. 1*, **14**, 403–432.

A serious application in probability, showing how general results (*not* restricted to bi-invariant functions) can be derived using the machinery of Gelfand pairs.

Cartier, P. (1972). Fonctions Harmoniques Sur Un Arbe. *Symposia Math.* **9**, 203–270.

An elegant combinatorial derivation of all properties of this Gelfand pair. See Sawyer (1978) for an application.

Diaconis, P. and Graham, R. L. (1985). The Radon Transform Z_2^k. *Pacific Jour.* **118**, 323–345.

This can all be carried over to bi-invariant neighborhoods on Gelfand pairs.

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

A longer version of Section F-3 above.

Dieudonne, J. (1978). *Treatise on Analysis VI.* Academic Press, New York.

A reasonably self-contained single source. Weighted toward the analyst, but it's possible to read.

Farrell, R. (1976). *Techniques of Multivariate Calculation.* Lecture Notes in Math, No. 520. Springer-Verlag, Berlin.

The only attempt at a beginning to end treatment of the mathematics of multivariate analysis that really does zonal polynomials.

Gangolli, R. (1972). Spherical functions on semi-simple Lie groups. In *Symmetric Spaces*, W. Boothby and G. Weiss. Marcel Dekker, New York.

Gangolli's article is a well written introduction to computations on continuous groups involving the Laplacian and its generalizations. The whole book consists of survey articles, roughly on the same topic.

Guillemin, V. and Sternberg, S. (1984). Multiplicity free spaces. *J. Differential Geometry* **19**, 31–56.

This is included to show this area is still under active development.

Helgason, S. (1978). *Differential Geometry Lie Groups and Symmetric Spaces.* Academic Press, New York.

Helgason, S. (1984). *Groups and Geometric Analysis: Integral Geometry Invariant Differential Operators, and Spherical Functions.* Academic Press, New York.

These two books give a comprehensive modern treatment of continuous Gelfand Pairs.

Helgason, S. (1973). Functions on symmetric spaces, pp. 101–146 in Harmonic Analysis on Homogeneous Spaces. *Proc. Symposia Math.* **24**, American Mathematical Society. Providence.

This entire volume shows how "grown-ups" use Gelfand pairs to do general representation theory.

Heyer, H. (1983). Convolution semigroups of probability measures on Gelfand pairs. *Expo. Math.* **1**, 3–45.

Contains a 62 item bibliography (mainly analytic, but useful).

James, A. T. (1975). Special functions of matrix and single argument in statistics. *Theory and Application of Special Functions*, R. Askey ed.

This is a summary of years of work on the example GL_n/O_n. This is a central example in the piece of math statistics known as multivariate analysis. The spherical functions, known as zonal polynomials, are used to derive distributions of things like the largest eigenvector in the covariance matrix of a normal sample.

Karlin, S. and McGregor, J. (1961). The Hahn polynomials, formulas and an application. *Scripta Math.* **23**, 33–46.

One of the earliest derivations of the special functions of $S_n/S_k \times S_{n-k}$. The applications are to a genetics model for random mating in a population with two alleles due to Moran. Many useful properties of the spherical functions are derived without mention of group theory.

Kramer, M. (1979). Sphärische Untergruppen in Kompacten Zusammenhangenden Lie Gruppen. *Composito Math.* **38**, 129–153.

He classifies, for G compact, simple, connected, Lie, all subgroups K such that (G, K) is Gelfand.

Letac, G. (1981). Problemes classiques de probabilite sur un couple de Gelfand. In *Lecture Notes in Math.* **861** (Springer-Verlag).

A very clear, elementary survey explaining a dozen applications in probability. Highly recommended.

Letac, G. (1982). Les fonctions spheriques d'un couple de Gelfand symetrique et les chaines de Markov. *Advances Appl. Prob.* **14**, 272–294.

A very clear, useful survey, explaining in particular a method of computing the spherical functions in "small" cases, so one can hope to guess at the answer.

Saxl, J. (1981). On multiplicity — free permutation representations. In *Finite Geometries and Designs*, London Math. Soc. Lecture notes, Series 48, Cambridge University Press, 337–353.

This classifies all subgroups of S_n which yield a Gelfand pair. Aside from $S_k \times S_{n-k}$ and small twists like $A_k \times A_{n-k}$ (A_k the alternating groups) the only "interesting" example is $S_2 Wr S_n$ which gives the Zonal Polynomials. See Diaconis (1987).

Saw, J. G (1977). Zonal polynomials: an alternative approach. *Jour. Multivariate Analysis* **7**, 461–467.

Derives properties of the spherical functions of GL_n/O_n without any group theory (but lots of "standard" properties of the Wishart distribution).

Sawyer, S. (1978). Isotropic random walks in a tree. *Zeit. Wahr.* **42**, 279–292.

A fascinating application of Gelfand pairs and *p*-adic numbers to salmon fishing!

Sloane, N. J. A. (1975). An introduction to association schemes and coding theory. *Theory and Applications of Special Functions*, R. Askey, ed.

Long, friendly introduction to the use of the tools of interest to coding theory.

Sloane, N. J. A. (1982). Recent bounds for codes, sphere packings and related problems obtained by linear programming and other methods. *Contemp. Math* **9**, 153–185.

Great, friendly article on the use of Gelfand pairs. Bibliography of 163 items.

Soto-Andrade, J. (1985). En torna a las funciones esfericas (caso finito). *Notas de la Sociedad de Matematica de Chile* IV, 71–94.

There is an active group working in Chile on Gelfand pairs. There are several other papers in this volume on this subject. A valuable thesis: *Caracteres de Espacios de Gelfand Finitos* by S. Garcia Zambrano (1984), also contains much of interest, in particular a careful discussion of spherical functions for the "orthogonal group" over a finite field.

Stanton, D. (1984). Orthogonal polynomials and Chevalley groups. In R. Askey et al (eds.) *Special Functions: Group Theoretical Aspects and Applications*, 87–92.

An important, clear, friendly survey of a dozen explicit calculations of spherical functions for *finite* spaces. Highly recommended.

Takemura, A. (1984). *Zonal Polynomials.* Institute of Mathematical Statistics. Hayward.

The best introduction to zonal polynomials for statisticians. No group theory, but lots of Wishart distributions.

H. First hitting times

Fourier analysis has been used to bound rates of convergence through the upper bound lemma. In this section a different application is presented. This permits sharp approximation of first passage probabilities and first return times for random walk. As an application, the classical gambler's ruin is given a new presentation. The arguments lean heavily on Good (1951).

Let Q be a probability on a finite group G. The random walk determined by Q starting at x is denoted Q_x^{*k}. Thus $Q_x^{*0}(y) = \delta_x(y)$, $Q_x^{*1}(y) = Q(yx^{-1})$, $Q_x^{*2}(y) = \Sigma_z Q(y(zx)^{-1})Q(z)$. In general, $Q_x^{*k} = Q_{id}^{*k} * \delta_x$.

Let $S \subset G$ be a set of elements called "sinks." We consider the random walk, starting at x and absorbed the first time it hits S. To rule out trivialities, assume $x \notin S$.

Let $a_k(t)$ be defined as the probabililty of *arriving* at the group element t at time k. If $t \notin S$, this is the chance of the walk being at t at time k, without having hit any sites in S. If $t \in S$, this is the chance of first being absorbed at t at time k.

Let $b_k(t)$ be defined as the probability of arriving at t, at time k, in the unrestricted random walk ($S = \phi$). The a's and b's are related via

LEMMA 7.

$$b_k(t) = a_k(t)\delta_{S^c}(t) + \sum_{s \in S}\sum_{j=0}^{k} a_j(s)b_{k-j}(ts^{-1}x).$$

Proof. Divide the set of paths of length k from x to t into $1 + (k+1)|S|$ classes. The first consists of paths that avoid all sinks. A typical path in one of the other classes hits a sink s for the first time at j (Probability $a_j(s)$) and then goes from s to t in the next $k - j$ steps (Probability $b_{k-j}(ts^{-1}x)$). By finite additivity, $b_k(t)$ is the sum of the probabilities of the classes. □

The convolution suggests generating functions (Fourier analysis on Z). Let $A(t,z), B(t,z)$ be the generating functions

$$\sum_{k=0}^{\infty} a_k(t)z^k, \ \sum_{k=0}^{\infty} b_k(t)z^k.$$

COROLLARY.

$$\begin{aligned} B(t,z) = \Sigma b_k(t)z^k &= \Sigma\Big(a_k(t)\delta_{S^c}(t) + \sum_{s \in S}\sum_{j=0}^{k} a_j(s)b_{k-j}(ts^{-1}x)\Big)z^k \\ &= \delta_{S^c}(t)a(t,z) + \sum_{s \in S} A(s,z)B(ts^{-1}x,z). \end{aligned}$$

Remark. Here is the way this formulation works: usually, we have a closed form expression for $B(t,z)$ for all t. Letting t run through S, the corollary gives $|S|$ equations in the $|S|$ unknowns $A(s,z)$. These can be solved (if $|S|$ is not too large, or is "symmetric") and then the corollary gives an expression for $A(t,z)$ for all t. The group theory enters as follows:

LEMMA 8. *With notation as above, $B(t,z)$ equals*

$$\frac{1}{|G|}\Sigma_\rho d_\rho Tr(\rho(xt^{-1}) \cdot (I - z\hat{Q}(\rho))^{-1}).$$

The inverse exists, at least for $|z| < 1$. The sum is over all irreducible representations of G.

Proof. This is just the Fourier inversion theorem applied to the definition of $B(t, z)$. □

Classical Gambler's Ruin: Peter and Paul flip a fair coin, Peter wins \$1 if the coin comes up heads; Paul wins \$1 if the coin comes up tails. Peter starts with $\$\ell$, Paul starts with $\$m$. The game stops at time T when one of them has no money left.

This can be analyzed as simple random walk on Z_n, where $n = \ell + m$. The particle starts at ℓ, and the game ends the first time zero is hit. For example, suppose Peter has 1 and Paul has 4

0

4 ①

3 2

A walk starting at 1 stops after 1 step to the left (Peter wiped out) or 4 steps to the right (Paul wiped out), etc.

Here there is one sink, namely, zero. From Lemmas 7, 8

$$A(0,z) = \frac{B(0,z)}{B(\ell,z)} = \frac{\sum_{j=0}^{n-1} \frac{e^{2\pi i j\ell/n}}{1-z\cos(\frac{2\pi j}{n})}}{\sum_{j=0}^{n-1} \frac{1}{1-z\cos(\frac{2\pi j}{n})}}.$$

Note that the numerator and denominator both have simple poles at $z = 1$. It follows that the left side is analytic in $|z| \le 1$. Writing

$$\frac{1}{1-z} + \sum_{j=1}^{n-1} \frac{e^{2\pi i j\ell/n}}{1 - z\cos(\frac{2\pi j}{n})} = \left\{\frac{1}{1-z} + \sum_{j=1}^{n-1} \frac{1}{1-z\cos(\frac{2\pi j}{n})}\right\}\{A(0,1) \\ - (1-z)A'(0,1) + \dots\}$$

Here A' denotes differentiation with respect to the second argument. Comparing coefficient as $z \to 1$ (set $1 - z = \varepsilon$) gives

Result 1. $A(0,1) = 1$ *so absorption is certain.*

Result 2.
$E(T) = A'(0,1) = \sum_{j=1}^{n-1} \frac{1-e^{2\pi i j\ell/n}}{1-\cos(\frac{2\pi j}{n})} = \sum_{j=1}^{n-1}(1 - cos(\frac{2\pi j\ell}{n}))\ (1 - cos(\frac{2\pi j}{n}))^{-1}$. Here is a curious consequence. By an elementary argument (Feller (1968, Chapter 14) $E(T) = \ell(n-\ell)$. This gives a curious trigonometric identity. Pass to the limit, with ℓ fixed, $n \to \infty$, the following emerges:

$$\int_0^1 \frac{1-\cos(2\pi\ell t)}{1-\cos(2\pi t)} = \ell.$$

It is also straightforward to pass to the limit in the original generating function:

Result 3.

Case 1. Let ℓ be fixed and let $n \to \infty$:

$$A(0,z) \to \frac{\int_0^1 \frac{\cos(2\pi t\ell)}{1-z\cos(2\pi t)}}{\int_0^1 \frac{1}{1-z\cos(2\pi t)}} = \left(\frac{1-(1-z^2)^{\frac{1}{2}}}{z}\right)^{\ell}.$$

The second identity is derived as follows: by expanding both sides, verify $\int_0^{2\pi} \frac{1}{1-z\cos t} = 2\pi(1-z^2)^{-\frac{1}{2}}$. Then, for $\ell = 1$, $A(0,z) = \frac{N(z)}{D(z)}$ with $D - zN = 1$. This gives the right side when $\ell = 1$. The general case follows from the convolution interpretation of the left side.

Case 2. Let $\ell = \theta n$ for $0 < \theta < 1$ fixed. Make the change of variables $z = e^{-\lambda/n^2}$. Then as n tends to $\infty, E\{e^{\lambda T/n^2}\}$ tends to

$$\frac{\sum_{j=0}^{\infty} \frac{\cos\,(2\pi\theta j)}{\lambda+(2\pi j)^2/2}}{\sum_{j=0}^{\infty} \frac{1}{\lambda+(2\pi j)^2/2}}.$$

This last function is the Laplace transform of a probability measure on $\mathbb{R}^+$.

Exercise 20. Consider nearest neighbor walk on the cube Z_2^n as described in Example 2 of Section C. Let T be the first return time to zero. Prove that $E(T) = 2^n$, and show that $T/2^n$ has a limiting exponential distribution.

Remarks. Flatto, Odlyzko, and Wales (1985) carried out similar computations for random transpositions. All of these computations use only 1 sink. Using 2 sinks, one can derive the chance that Peter wins in gambler's ruin, or the law of the maximum of random walk, given that its first return is at time $2k$; see Smith and Diaconis (1988) for references to the literature. Similar results on the cube would give results for fluctuations and excursions of the Ornstein-Uhlenbeck process, or any of the birth and death chains described in Section F.

An elegant application of first hitting distributions for simple random walk on the n-cube to the analysis of algorithms appears in Aldous (1983b). Consider finding the minimum of a function $f: Z_2^n \to \mathbb{R}$. Clearly general functions take order 2^n steps for any algorithm on average. People in operations research hoped that "local-global" functions with the property that if $f(x)$ is not a minimum, then $f(y) < f(x)$ for some neighbor y of x, would be a useful special class.

The obvious algorithm is: start someplace, and take your smallest neighbor as your next place, etc. Craig Tovey showed there were some exponentially bad examples, but naturally created functions seemed to work in order n^2 steps.

Aldous treated the problem as a two-person game: nature picks a function, we pick an algorithm. We pay nature the number of steps it takes our algorithm to find the minimum. Because both sets of possibilities are finite (things only depend on the relative values) the game has a value v. Aldous showed that the value was approximately $2^{n/2}$.

The two strategies are easy to describe: your (randomized) strategy is to pick vertices $x_1, x_2 \ldots, x_J$ at random ($J \doteq 2^{n/2}$) and then use the obvious algorithm starting at $\min(f(x_i))$.

Nature's strategy involves choosing a random local-global function as follows. Start simple random walk at a random point x_0. Let $f(x)$ be the number of steps until the walk first hits x. Thus $f(x_0) = 0$, and for any other x there is a neighbor y of x which was visited first. Thus f is local-global. By careful analysis of random walk, Aldous is able to show it takes order $2^{n/2}$ steps to find the minimum with any algorithm.