CONVERGENCE RATES OF MARKOV CHAINS

STEVEN P. LALLEY

CONTENTS

1. ORIENTATION

Finite-state Markov chains have stationary distributions, and irreducible, aperiodic, finite-state Markov chains have *unique* stationary distributions. Furthermore, for any such chain the $n-$step transition probabilities converge to the stationary distribution. In various applications – especially in Markov chain Monte Carlo, where one runs a Markov chain on a computer to simulate a random draw from the stationary distribution – it is desirable to know *how many* steps it takes for the $n-$step transition probabilities to become close to the stationary probabilities. These notes will introduce several of the most basic and important techniques for studying this problem, *coupling* and *spectral analysis*. We will focus on two Markov chains for which the convergence rate is of particular interest: (1) the *random-to-top shuffling model* and (2) the *Ehrenfest urn model*. Along the way we will encounter a number of fundamental concepts and techniques, notably *reversibility, total variation distance*, and *Poissonization*.

1.1. **Example 1: Random-to-Top Card Shuffling.** "Random-to-top" is not an especially efficient way to randomize a deck of cards, but it is one of the easiest of all shuffling schemes to analyze mathematically. It works as follows. Assume that the deck consists of N cards, labelled $1, 2, 3, \ldots, N$, and assume that initially the cards are in order. At each time $t = 1, 2, \ldots$, select a card at random, independently of all past selections, and move it to the top of the deck.

Problem: How many repetitions are needed to "randomize" the deck?

Before we try to solve this problem, we must settle on an appropriate notion of randomness. A poker player would consider the deck to be in a perfectly random state if each of the $N!$ different possible orderings of the deck were equally likely, and so this will be our operational definition of "perfectly random".

Proposition 1. *Let X_t be the ordering of the deck after t repetitions of random-to-top. The sequence $\{X_t\}_{t=0,1,2,\ldots}$ is an aperiodic, irreducible Markov chain on the set Γ_N of all $N!$ orderings of the deck, and its (unique) stationary distribution is the uniform distribution on Γ_N.*

Proof: It is not difficult to see that the Markov chain is aperiodic and irreducible. That the stationary distribution is uniform will be shown in section 1.5 below. □

Of course, we won't insist that the deck be in a perfectly random state when we begin dealing the cards, as this would require the dealer to shuffle infinitely many times. Thus, we will eventually need to make precise what we mean by "close" to perfectly random.

1.2. **Example 2: The Ehrenfest Urn Model of Diffusion.** An important part of the day-to-day business of physics is to understand how bulk properties of matter arise from statistical regularities in the behavior of its constituent atoms and molecules. In principle, the time evolution of any physical system is governed by differential equations; however, explicit solution of these equations is rarely possible, even for small systems, and even qualitative properties of the solutions are usually difficult or impossible to deduce. Therefore, statistical physicists invent and study mathematical models in which the interactions between molecules are merely caracatured, in the hope that these models will shed light on the behavior of real physical systems.

The *Ehrenfest urn model*, proposed in 1906 as a simple model of diffusion, is typical in certain respects of processes studied by statistical physicists. The physical system consists of N identical (but distinguishable) molecules, each of which may at any time be in either the right or the left half of a cubical box. In the real physical system, the motion of these molecules is rather complicated. The Ehrenfests decided to study a simpler model for how the molecules move from one side of the box to the other: they postulated that at each time $t = 1, 2, \ldots$ one of the molecules would be chosen at random and then moved to the other side of the box.

Here is a simplified (and in certain important respects, a more natural) version of the model: Each of N balls resides at any given time in one of two urns, labelled 0 and 1. The system evolves in time as follows: At each time $t = 1, 2, \ldots$ one ball is selected at random from among the N balls in the system and then moved to a *random* urn (that is, a fair coin is tossed, and the ball is placed in urn A if the toss is a head and in urn B otherwise).

A natural choice for the state variable of the system is the number X_t of balls in 1 at time t. It is not difficult to see that the sequence $\{X_t\}$ is an irreducible Markov chain with finite state space $\{0,1,2,\dots,N\}$, and transition probabilities

$$
\begin{aligned}
(1)\qquad p(k,k+1) &= (1-k/N)/2;\\
p(k,k-1) &= k/2N;\\
p(k,k) &= 1/2.
\end{aligned}
$$

This Markov chain has a unique equilibrium distribution, which we will determine shortly.

Problem: How many steps are necessary for the system to "reach equilibrium" if all of the balls begin in urn 0?

The Ehrenfest random walk on the hypercube The random variable X_t that records the number of balls in urn 1 at time t does not carry all of the information about the system, as it does not record *which* balls are in which urns. Another natural choice of state variable is the N-vector S_t whose i^{th} coordinate is 1 if the i^{th} ball is in 1 and 0 otherwise. Henceforth, to distinguish between the processes S_t and X_t, we shall refer to the sequence S_t as the *Ehrenfest random walk*, and the sequence X_t as the *Ehrenfest urn chain*. Observe that each of the processes $\{X_t\}$ and $\{S_t\}$ is a Markov chain, but their state spaces are different: the state space of the the Ehrenfest urn chain is the set of integers $\{0,1,2,\dots,N\}$, while the state space of the Ehrenfest random walk is the $N-$*dimensional hypercube* $\{0,1\}^N$, which consists of all 2^N sequences of length N with entries 0 or 1. Notice also that X_t is a function of S_t: the variable X_t may be recovered from S_t by adding the entries of S_t.

The hypercube can be made into an abelian group: $N-$vectors of 0s and 1s are added component-wise modulo 2. The group is denoted by $\mathbb{Z}_2^N$, because it is gotten by pasting together N copies of the two-element group $\mathbb{Z}_2$. The group structure is useful in the study of the Ehrenfest process, because S_{t+1} is obtained from S_t by adding (mod 2) a Bernoulli-$(1/2)$ random variable to a randomly chosen coordinate of S_t: thus,

$$(2)\qquad S_t = S_0 + \sum_{j=1}^{t} \xi_j,$$

where the random variables $\xi_1,\xi_2,\dots$ are independent and identically distributed with distribution

$$
\begin{aligned}
(3)\qquad P\{\xi_j = 0\} &= 1/2 && \text{and}\\
P\{\xi_j = e_k\} &= (2N)^{-1} && \text{for each } k=1,2,\dots,N.
\end{aligned}
$$

Here 0 is the 0 vector in $\mathbb{Z}_2^N$, and e_k is the kth unit vector that is, the vector with a 1 in the kth coordinate and zeros elsewhere.

Stationary Distributions. It is easy to guess the equilibrium distribution for the Ehrenfest random walk $\{S_t\}$. In equilibrium, each of the balls is equally likely to be in either of the two urns, and for any two balls i,j the events

$$\{i \text{ in urn } A\} \qquad \text{and} \qquad \{j \text{ in urn } A\}$$

should be independent. Consequently, the stationary distribution of the Markov chain S_t should be the uniform distribution on the set $\mathbb{Z}_2^N$. In section 1.4 below, we will prove that this guess is correct.

Once we know the equilibrium distribution of the random walk S_t, it is easy to recover that of the Ehrenfest urn process X_t. Recall that X_t can be recovered from S_t be adding the coordinates. Consequently, one may obtain a random variable X whose distribution is the equilibrium distribution of the Markov chain X_t by choosing a random element S of the hypercube $\mathbb{Z}_2^N$ with the uniform distribution and adding its coordinates. (EXERCISE: Explain why.) But the coordinates of a uniformly distributed random element S of the hypercube are independent, identically distributed Bernoulli-$(1/2)$ random variables! Thus, we should expect (and will prove shortly) that the stationary distribution of the Ehrenfest chain X_t will be the Binomial distribution with parameters N and $p = 1/2$.

1.3. **Example 3: Metropolis-Hastings Algorithm.** The Metropolis-Hastings algorithm is a basic strategy for simulation. The problem is this: we are given some complicated probability distribution π on a large state space $\mathcal{X}$ (often much to large to enumerate) and we would like to estimate some feature of this distribution (e.g., the mean of some function on $\mathcal{X}$) by simulation. What often makes the problem doubly difficult is that the distribution π might have a normalizing constant that is unknown and impossible to calculate directly. (This is generally the case in Bayesian statistics, where the posterior distribution involves a denominator given in the form of an integral or sum over the parameter space.) In such cases, π is given in the form

$$\pi(x) = Cw(x) \quad \text{where} \quad C^{-1} = \sum_{y \in \mathcal{X}} w(y)$$

for a weight function w that is easy to compute. The Metropolis-Hastings strategy is to run a Markov chain whose transition probability matrix on $\mathcal{X}$ satisfies the *detailed balance* equations

$$w(x)p(x,y) = w(y)p(y,x). \tag{4}$$

Clearly, if the transition probabilities satisfy these equations, then they satisfy the same equations with w replaced by π. We will see in section1.4 below that if the transition probabilities satisfy these equations then π is a stationary distribution for the transition probability matrix. If the Markov chain with this transition probability matrix is irreducible then the ergodic theorem (i.e., the strong law of large numbers) guarantees that if the Markov chain runs long enough then the empirical distribution of states visited will closely approximate the stationary distribution. The obvious question of interest for the user: How long must I run the Markov chain so as to be pretty sure that the approximation is good?

There is an easy and flexible way to find solutions to the detailed balance equations (4) due to Metropolis. The idea is this: Suppose that $\pi(x) = Cw(x)$ is a probability distribution on a finite set $\mathcal{X}$ such that $\pi(x) > 0$ for every x, and suppose that $\mathcal{X}$ is the vertex set of a connected graph $\mathcal{G}$ with edge set $\mathcal{E}$. Choose d large enough that no vertex is an endpoint of more than d edges. For each pair $x, y \in \mathcal{X}$ of distinct vertices, define

$$\begin{aligned} p(x,y) &= \left(\frac{w(y)}{w(x)} \wedge 1 \right) /d \quad \text{if } x, y \text{ share an edge;} \\ p(x,y) &= 0 \quad \text{if } x, y \text{ do not share an edge; and} \\ p(x,x) &= 1 - \sum_{y \neq x} p(x,y). \end{aligned} \tag{5}$$

Observe that $p(x,x) \geq 0$, because $p(x,y) \leq 1/d$ for each pair x, y of neighboring vertices, and no vertex x has more than d neighbors. Thus, equations (5) defines a system of transition probabilities on $\mathcal{X}$. It is easy to check (EXERCISE: Do it!) that the detailed balance equations hold for this set of transition probabilities, and so π is a stationary distribution (see Lemma 2 in section 1.4 below). The clever thing about the set (5) of transition probabilities is that you don't need to know the normalizing constant C to compute $p(x,y)$, but nevertheless the stationary distribution is π.

Since the graph $\mathcal{G}$ is connected, the transition probability matrix is irreducible, and so the stationary distribution is unique; but the Markov chain may be periodic (EXERCISE: Find an example!). However, by choosing d strictly larger than the degree (:= number of incident edges) of some vertex, one may obtain an aperiodic, irreducible transition probability matrix by (5).

Note: You might wonder why we don't just use the complete graph on the vertex set $\mathcal{X}$. (The complete graph on a set $\mathcal{X}$ is the graph that has an edge between every pair $x, y \in \mathcal{X}$.) The reason is this: In many problems, the weight function $w(x)$ is computable, but not easily so. In such problems it is often advantageous to restrict possible transitions $x \to y$ to those pairs for which the *ratio* $w(y)/w(x)$ is easy to compute.

1.4. **Reversible Markov Chains.** Here is a simple but extremely useful sufficient condition for checking that a candidate distribution is in fact a stationary distribution for a given Markov chain:

Lemma 2. *If $\{p(x,y)\}_{x,y\in\mathcal{X}}$ are the one-step transition probabilities of an irreducible Markov chain on the state space $\mathcal{X}$, and if $\{\pi(x)\}_{x\in\mathcal{X}}$ is a probability distribution on $\mathcal{X}$ such that for any two states $x, y \in \mathcal{X}$,*

$$\pi(x)p(x,y) = \pi(y)p(y,x), \tag{6}$$

then π is the unique equilibrium distribution for the Markov chain.

Proof: This is an easy calculation. It will follow immediately from Proposition 3 below. □

The equations (6) are called the *detailed balance* equations. It is quite easy to see that they hold for the transition probabilities of the Ehrenfest random walk, with $\pi =$ the uniform distribution on the hypercube. It is also quite routinee to check that the detailed balance equations hold for the Ehrenfest urn chain transition probabilities (1) and the Binomial-$(N, 1/2)$ distribution (EXERCISE: CHECK THIS!).

Note that the detailed balance equations are not a *necessary* condition for stationarity: For example, the uniform distribution on the $N!$ possible orderings of a deck of N cards is stationary for the random-to-top shuffle, as will be shown in section 1.5 below, but the detailed balance equations don't hold. In fact, the Markov chains that satisfy detailed balance equations are somewhat atypical of Markov chains in general. They are, however, of great importance:

Definition 1. A Markov chain on a finite or countable state space $\mathcal{X}$ with transition probabilities $p(x,y)$ is said to be *reversible* if there is a positive function $w : \mathcal{X} \to (0,\infty)$ such that for any two states $x, y \in \mathcal{X}$,

$$w(x)p(x,y) = w(y)p(y,x). \tag{7}$$

The weight function w is not required to be a probability distribution, and there are important examples, such as the simple random walk on the integers, where the detailed balance equations (7) hold *only* for a weight function w that satistfies $\sum_x w(x) = \infty$.

Proposition 3. *Suppose that $p(x, y)$ are the transition probabilities of an irreducible random walk on a finite or countable state space $\mathcal{X}$, and suppose that they satisfy the detailed balance equations with respect to some weight function w. Then*

(A) For any integer $n \geq 1$, the $n-$step transition probabilities $p_n(x, y)$ also satisfy the detailed balance equations with respect to the weight function w.

(B) If the transition probabilities $p(x, y)$ also satisfy the detailed balance equations with respect to another weight function v, then the weight functions w and v must be proportional, that is, there exists a positive scalar α such that $v(x) = \alpha w(x)$ for all $x \in \mathcal{X}$.

(C) If the Markov chain with transition probabilities $p(x, y)$ is positive recurrent, and if the transition probabilities satisfy the detailed balance equations with respect to some weight function w, then w must be proportional to the stationary distribution.

Proof: The proofs of (A) and (B) are left to the reader as exercises (HINT: For (B), you must make use of the irreducibility of the Markov chain!). To prove (C), we make use of (A), which asserts that the $n-$step transition probabilities satisfy the detailed balance equations relative to w:

$$w(x)p_n(x, y) = w(y)p_n(y, x). \tag{8}$$

If the Markov chain is aperiodic (we have already assumed that it is irreducible), then the $n-$step transition probabilities converge to the stationary distribution, and so taking the limit in (8) shows that

$$w(x)\pi(y) = w(y)\pi(x). \tag{9}$$

This implies that the weight function w and the stationary distribution π are proportional. If the Markov chain is not aperiodic, then this argument does not apply directly, because the $n-$step transition probabilities do not converge for Markov chain with periodicity. However, if the period is (say) d, then the matrix

$$\mathbb{Q} = \frac{1}{d}\sum_{j=1}^{d} \mathbb{P}^j \tag{10}$$

is aperiodic and irreducible (Exercise!), has stationary distribution π, and by (8) satisfies the detailed balance equations

$$w(x)q(x, y) = w(y)q(y, x) \tag{11}$$

with the same weight function w as does p. Since $\mathbb{Q}$ is aperiodic, it follows by the argument above that w is proportional to π. □

Proposition 4. *(Kolmogorov's Criterion) An irreducible Markov chain on a finite or countable state space $\mathcal{X}$ with transition probabilities $p(x, y)$ is reversible if and only if for every finite* cycle *$x_0, x_1, x_2, \dots, x_{n-1}, x_n = x_0$ of states,*

$$\prod_{i=0}^{n-1} p(x_i, x_{i+1}) = \prod_{i=0}^{n-1} p(x_{i+1}, x_i). \tag{12}$$

A *cycle* is just a finite sequence of states whose first and last elements are the same. In words, Kolmogorov's Criterion asserts that a Markov chain is reversible if and only if for every cycle, the probability of traversing the cycle is the same as the probability of traversing it in reverse order.

Proof: It is easy to prove that if the Markov chain is reversible then the equation (12) must hold. (EXERCISE: Do it!) The interesting direction is the converse, that if the equation (12) holds for every cycle then the detailed balance equations must hold for some weight function. For this, we will construct a suitable weight function.

First, let $x, y \in \mathcal{X}$ be any two states such that $p(x, y) > 0$. I claim that it must also be the case that $p(y, x) > 0$. Here is why: Since the the Markov chain is irreducible, there must be a cycle $x_0, x_1, x_2, \ldots, x_n$ with $x_0 = x_n = x$ and $x_1 = y$ such that all steps through the cycle have positive transition probabilities, that is, $p(x_i, x_{i+1}) > 0$. But the Kolmogorov equations (12) imply that all the reverse steps also have positive transition probabilities $p(x_{i+1}, x_i) > 0$. Since the first step of the cycle is from x to y, it follows that $p(y, x) > 0$.

Now fix a state $x_* \in \mathcal{X}$ to serve as a reference point: call this state the *origin*. Observe that if the detailed balance equations hold for some weight function, then they will hold for any scalar multiple of it, and so we may as well restrict our search to potential weight functions that satisfy $w(x_*) = 1$. Now if the detailed balance equations are to hold, it must be that for any state y with $p(x, y) > 0$,

$$w(y) = w(x_*) \frac{p(x_*, y)}{p(y, x_*)} = \frac{p(x_*, y)}{p(y, x_*)}.$$

Note that the denominator is positive, by the preceding paragraph. Thus, the weight function is determined for any state y that is accessible from the origin in one step. To extend the definition of w to all states, we use again the hypothesis that the Markov chain is irreducible: If $y \in \mathcal{X}$ is any state, there must be a finite sequence $x_0, x_1, x_2, \ldots, x_n$ beginning at $x_0 = x_*$ and ending at $x_n = y$ such that $p(x_i, x_{i+1}) > 0$ for every i. By the result of the previous paragraph, it must also be the case that $p(x_{i+1}, x_i) > 0$. So let's try to define $w(y)$ as follows:

$$w(y) = \prod_{i=0}^{n-1} \frac{p(x_i, x_{i+1})}{p(x_{i+1}, x_i)}. \tag{13}$$

The obvious problem, of course, is that there may be many different positive-probability paths from the origin to y, so how do we choose among them in defining $w(y)$? It turns out that it doen't matter: any two paths will give the same value for $w(y)$. To see this, suppose that there are two positive-probability paths, say β, γ, from x_* to y. Following β to y and then γ in reverse back to x_* gives a cycle, for which equation (12) must hold. This implies (why?) that either path will give the same value $w(y)$ in (13). Thus, w is well-defined.

It remains to show that the detailed balance equations will hold for *all* pairs of states x, y. Let $x, y \in \mathcal{X}$ be any pair of states such that $p(x, y) > 0$, and let $\beta = (x_0, x_1, x_2, \ldots, x_m)$ and $\gamma = (y_0, y_1, y_2, \ldots, y_n)$ be positive-probability paths from $x_0 = y_0 = x_*$ to $x_m = x$ and $y_n = y$. By definition of w,

$$w(x) = \prod_{i=0}^{m-1} \frac{p(x_i, x_{i+1})}{p(x_{i+1}, x_i)} \quad \text{and}$$

$$w(y) = \prod_{i=0}^{n-1} \frac{p(y_i, y_{i+1})}{p(y_{i+1}, y_i)}.$$

To see that

$$w(x)p(x, y) = w(y)p(y, x),$$

just apply the Kolmogorov equation (12) to the cycle gotten by following β to x, then jumping to y, and then following γ in reverse back to x_*. □

Remark. Those of you who haven't yet thrown away your copy of ROSS will see that he gives a shorter proof of Kolmogorov's Criterion (Th. 4.7.2). Unfortunately, (a) his proof requires that the Markov chain be positive recurrent, and (b) it is not constructive. The proof given above *is* constructive: it shows you how to compute the weight function for a Markov chain that satisfies Kolmogorov's equations.

1.5. Exercises: Reversiblity, Symmetries and Stationary Distributions.

Problem 1. Birth-Death Chains: A *birth-death* chain is an irreducible Markov chain whose state space $\mathcal{Y}$ is an interval of the integers, that is,

$$\begin{aligned} &\mathcal{Y} = [0, M] := \{0, 1, 2, \dots, M\} &&\text{or} \\ &\mathcal{Y} = \mathbb{Z}_+ := \{0, 1, 2, \dots\} &&\text{or} \\ &\mathcal{Y} = \mathbb{Z} := \{\text{all integers}\}, \end{aligned}$$

and for which only transitions to nearest neighbors are allowed. In this problem only the state space $\mathbb{Z}_+$ will be considered. Assume that the nonzero transition probabilities are

$$\begin{aligned} p(x, x+1) &:= \beta_x > 0 &&\text{for } x \geq 0 \\ p(x, x-1) &:= \alpha_x > 0 &&\text{for } x \geq 1 \\ p(0, 0) &:= \alpha_0 > 0. \end{aligned}$$

(A) Find a weight function $w(x)$ such that the detailed balance equations (6) hold with respect to w.
(B) Prove that your weight can be normalized to give a stationary probability distribution if and only if

$$\sum_{x=1}^{\infty} \prod_{j=0}^{x-1} \frac{\beta_j}{\alpha_{j+1}} < \infty.$$

(C) Conclude that a birth-death chain on the nonnegative integers is positive recurrent if and only if the inequality in (B) holds.

Problem 2. Prove assertions (A) and (B) of Proposition 3.

Problem 3. Let $\mathbb{P} = (p(i, j))$ be an irreducible transition probability matrix on a finite state space $\mathcal{Y}$. An *automorphism* (or *symmetry*) of the transition kernel $\mathbb{P}$ (or, more informally, of the the Markov chain with this transition kernel) is a one-to-one mapping $T : \mathcal{Y} \to \mathcal{Y}$ such that for every pair $i, j \in \mathcal{Y}$,

$$p(i, j) = p(T(i), T(j)).$$

Let π be the unique stationary probability distribution for the transition probability matrix $\mathbb{P}$. (Recall that the stationary distribution is unique if $\mathbb{P}$ is irreducible.) Suppose that $T : \mathcal{Y} \to \mathcal{Y}$ is an automorphism of $\mathbb{P}$. Show that for every $i \in \mathbb{Y}$,

$$\pi(i) = \pi(T(i)).$$

Problem 4. Consider the random-to-top shuffle described in section 1.1. At time t, the state of the deck is the list

$$X_t = (X_t(1), X_t(2), \dots, X_t(N))$$

in which the ith entry $X_t(i)$ records the identity of the ith card from the top of the deck. Thus, the sequence $X_0, X_1, \dots$ is a Markov chain on the state space $\mathcal{S}_N$, the set of all permutations of the set

$$[N] := \{1, 2, 3, \dots, N\}.$$

(a) Explain briefly why the Markov chain X_t is aperiodic and irreducible. (b) Show that for any two permutations π, σ there is an automorphism T of the transition kernel such that $T(\pi) = \sigma$. (c) Conclude that the unique stationary distribution of the Markov chain is the uniform distribution on $\mathcal{S}_N$.

Problem 5. Consider the Ehrenfest random walk S_t on the hypercube $G = \mathbb{Z}_2^N$. (a) Show that for any two states $x, y \in G$ there is an automorphism T of the transition probability kernel such that $T(x) = y$. (b) Conclude that the uniform distribution on G is the unique stationary distribution.

2. Coupling

The term *coupling* is used loosely by probabilists to describe constructions in which two (or sometimes more) random objects (random variables, or random vectors, or sequences of random variables) are built simultaneously on the same probability space in such a way that information about one provides information about the other. In such constructions the two random objects are dependent (otherwise neither would give any information about the other); the trick in using a coupling argument is to arrange things so that the random objects interact in just the right way. This is a bit vague, but the examples to follow will illustrate how the strategy often works.

2.1. **Coupling and Total Variation Distance.** How close is close? Since our goal is to be able to say something about the number of steps a Markov chain must run to get "close" to equilibrium, we must have a notion of what it means for the distribution of X_t to be "close" to the stationary distribution. The most natural mathematical definition of the "closeness" of two probability distributions on a finite set uses the so-called *total variation* distance.

Definition 2. Let μ and ν be two probability distributions on a finite or denumerable set $\mathcal{X}$. The *total variation* distance between μ and ν is defined to be

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_{x \in X} |\mu\{x\} - \nu\{x\}| \tag{14}$$

$$= \max_{A \subseteq \mathcal{X}} (\mu(A) - \nu(A)). \tag{15}$$

Notice that the total variation distance between μ and ν is 0 if and only if $\mu = \nu$, and that the total variation distance is 1 (the maximum possible value) if and only if the measures μ and ν are supported by disjoint subsets of $\mathcal{X}$. Also, if μ_n is a sequence of probability distributions, then

$$\lim_{n \to \infty} \|\mu_n - \mu\|_{TV} = 0 \qquad \text{iff} \qquad \lim_{n \to \infty} \mu_n\{x\} = \mu\{x\} \; \forall \, x \in \mathcal{X}. \tag{16}$$

There is an equivalent characterization of total variation distance that uses the notion of *coupling*. Define a *coupling* of the probability measures μ and ν to be a probability distribution λ on the Cartesian product $\mathcal{X} \times \mathcal{X}$ whose marginal distributions are μ and ν, respectively, that is,

$$\mu\{x\} = \sum_{y \in \mathcal{X}} \lambda\{x, y\} \quad \forall x \in \mathcal{X} \qquad \text{and} \tag{17}$$

$$\nu\{y\} = \sum_{x \in \mathcal{X}} \lambda\{x, y\} \quad \forall y \in \mathcal{X}. \tag{18}$$

Notice that there are in general many couplings of two given probability measures μ, ν. Define a *maximal coupling* to be a coupling that assigns the largest possible probability to the *diagonal*

$$(\mathcal{X} \times \mathcal{X})_{\text{diagonal}} := \{(x, y) \in \mathcal{X} \times \mathcal{X} \,:\, x = y\}.$$

Proposition 5. *For any pair μ, ν of probability distributions on a finite set $\mathcal{X}$ there is a maximal coupling λ, and*

$$\lambda(\mathcal{X} \times \mathcal{X}) - \lambda(\mathcal{X} \times \mathcal{X})_{\textit{diagonal}} = \|\mu - \nu\|_{TV}. \tag{19}$$

Proof: Homework. HINT: First try to guess what the diagonal entries $\lambda\{x, x\}$ have to be. □

2.2. **Coupling Constructions and Convergence of Markov Chains.** In studying convergence rates of Markov chains, what one needs is not just a simultaneous construction of two random variables but a simultaneous construction of two (or more) *infinite sequences* of random variables. Probabilists somewhat loosely refer to any such construction as a *coupling*. The two (or more) sequences, say $\{X_t\}_{t \in \mathbb{N}}$ and $\{Y_t\}_{t \in \mathbb{N}}$, are often both Markov chains with the same transition probability matrix $\mathbb{P}$, but with different initial conditions. It will then be the case that for any $t \in \mathbb{N}$ the joint distribution of (X_t, Y_t) is a coupling of the marginal distributions μ_t^X and μ_t^Y of X_t and Y_t, respectively. In particular, it will follow from Proposition 5 above (why?) that

$$\|\mu_t^X - \mu_t^Y\|_{TV} \leq P\{X_t \neq Y_t\}. \tag{20}$$

Doeblin Coupling: This is the most basic coupling of two Markov chains, used first by Doeblin. Let $p(x, y)$ be the transition probability kernel of Markov chain on a finite or countable state space $\mathcal{X}$. Define a transition probability kernel on the Cartesian product $\mathcal{X} \times \mathcal{X}$ by

$$\begin{aligned} q((x, y), (x', y')) &= p(x, x')p(y, y') && \text{if } x \neq y \\ &= p(x, x') && \text{if } x = y \text{ and } x' = y' \\ &= 0 && \text{otherwise.} \end{aligned} \tag{21}$$

It is easy to check that this is defines a transition probability matrix, that is, that the row sums are all 1. Thus, there is a Markov chain (X_t, Y_t) on $\mathcal{X} \times \mathcal{X}$ with transition probabilities $q(\cdot, \cdot)$. This Markov chain behaves according to the following rules: if $X_t \neq Y_t$ then the two components make independent jumps according to the transition probability kernel p, and if $X_t = Y_t$ then they make the *same* jump. Thus, the diagonal is an absorbing set.

Proposition 6. *Let (X_t, Y_t) be a Markov chain on $\mathcal{X} \times \mathcal{X}$ with transition probability kernel q. Then each of the component processes X_t and Y_t is, by itself, a Markov chain on $\mathcal{X}$ with transition probabilities p. Moreover, if the transition probability kernel p is ergodic (irreducible and aperiodic) and the state space $\mathcal{X}$ is finite, then with probability one (X_t, Y_t) will eventually be absorbed in the diagonal, that is, eventually $X_t = Y_t$.*

Proof: It is easy to see that marginally the sequence X_t is a Markov chain with transition probabilities $p(x, x')$ by summing over the variable y' in (21) above. It is also obvious that the diagonal is an absorbing set for the Markov chain (X_t, Y_t). Thus, what remains is to prove that the diagonal is *accessible* from every state $(x, y) \in \mathcal{X} \times \mathcal{X}$. For this, it is enough to show that there is a positive-probability path from (x, y) to the diagonal. Here we use the hypothesis that the transition kernel p is aperiodic and irreducible: this assures that for some integer $n \geq 1$ the $n-$step transition probability matrix $\mathbb{P}^n$ has all entries positive. In particular, there exists $z \in \mathcal{X}$ such that both $p_n(x, z) > 0$ and $p_n(y, z) > 0$. But this implies that there are sequences of states

$$x = x_1, x_2, \dots, x_n = z \quad \text{and}$$
$$y = y_1, y_2, \dots, y_n = z$$

such that $p(x_i, x_{i+1}) > 0$ and $p(y_i, y_{i+1}) > 0$ for each i. Then

$$q((x_i, y_i), (x_{i+1}, y_{i+1})) > 0$$

for each $i < k$, where k is the smallest integer such that $x_k = y_k$, and so $(x_i, y_i)_{1 \leq i \leq k}$ is a positive-probability path from (x, y) to the diagonal point (z, z). □

Proposition 6 has important ramifications for the Markov chain X_t. Suppose that the initial state $X_0 = x$ is nonrandom, but that the initial state Y_0 is random, with distribution π, where π is the stationary distribution for the transition probability kernel p. Then for every time $t \geq 0$, the distribution of Y_t is π. On the other hand, the distribution of X_t is not (in general) π for any finite time t. By Proposition 6, the pair (X_t, Y_t) will eventually reach the diagonal for a first time

$$\tau = \min\{t \,:\, X_t = Y_t\}, \tag{22}$$

after which it will be absorbed. Thus, by inequality (20), with μ_t denoting the distribution of X_t when the Markov chain is started in state $X_0 = x$,

$$\boxed{\|\mu_t - \pi\|_{TV} \leq P\{\tau > t\}.} \tag{23}$$

It should be noted that this conclusion is true not only for the Doeblin coupling — the particular coupling determined by the kernel q defined by (21) above — but for *any* coupling. That is, if X_t and Y_t are both Markov chains with transition probability matrix p, with Y_0 having the stationary distribution π and $X_{0=x}$, and if $X_t = Y_t$ for all times $t \geq \tau$, then inequality (23) holds. In applications of the coupling method, the hard work is usually in finding a clever way to build such a pair X_t, Y_t so that $P\{\tau > t\}$ can be efficiently approximated. Later we will build couplings for random-to-top shuffling and for the Ehrenfest chain.

Note also that the inequality (23) is, in a sense, a "quantitative" version of Kolmogorov's convergence theorem for ergodic Markov chains: By Proposition 6, eventual absorption in the diagonal is certain, so $\tau < \infty$ with probability one, and so $P\{\tau > t\} \to 0$ as $t \to \infty$.

Thus, (23) implies that the total variation distance between μ_t and π converges to 0 as $t \to \infty$. This in turn implies that for each state $x \in \mathcal{X}$,

$$\lim_{t\to\infty} \mu_t\{x\} = \pi\{x\}.$$

Proposition 7. *Let X_t be an ergodic Markov chain on a finite state space $\mathcal{X}$ with transition probability matrix $p(x, y)$ and with initial distribution $\mu = \mu_0$. Let π be the unique stationary distribution for the transition probability matrix $p(x, y)$, and let μ_t be the distribution of X_t. Then the total variation distance $\|\mu_t - \pi\|_{TV}$ converges* monotonically *to zero as $t \to \infty$.*

Proof: We have already noted that the total variation distance converges to 0 as $t \to \infty$. We must show that the total variation distance is (weakly) monotonically decreasing in t. Since the initial distribution $\mu = \mu_0$ is arbitrary, it suffices to show that the total variation distance decreases in the first step, that is, that

$$\|\mu_1 - \pi\|_{TV} \leq \|\mu_0 - \pi\| \tag{24}$$

For this we use the characterization of the total variation norm by maximal coupling given in Proposition 5. In particular, there is a maximal coupling $\lambda = \lambda_0$ of the probability distributions $\mu = \mu_0$ and π, and for this maximal coupling,

$$\|\mu_0 - \pi\|_{TV} = \lambda\{(u, v) \,:\, u \neq v\}.$$

Consider the Markov chain (X_t, Y_t) with transition probabilities q as defined by equations (21) above, with initial distribution λ_0, and let λ_1 be the distribution of the pair (X_1, Y_1). Then λ_1 is a coupling of the distributions μ_1 and π, because (a) each of the processes X_t and Y_t is, by itself, a Markov chain with transition probabilities $p(x, y)$; and (b) consequently, the marginal distributions of X_t and Y_t are μ_t and π. Furthermore, the distribution λ_1 attaches at least as much probability to the diagonal as does λ_0, because the diagonal is an absorbing set for the transition probability kernel q. Therefore, by Proposition 5, the inequality (24) holds. □

2.3. **Couplings for the Ehrenfest Urn and Random-to-Top Shuffling.**

Both random-to-top shuffling and the Ehrenfest urn are systems with a large number of component subsystems that interact weakly: For card-shuffling, the components are the individual cards; and for the Ehrenfest urn, the components are the individual balls. For such multi-component systems it is often possible to construct couplings by forcing the coupled chains to manipulate the same component at each step.

Example: Random-to-top shuffling The coupling will have two decks of cards each undergoing random-to-top shuffling. The states of the two decks at any time t will be denoted by X_t and Y_t, respectively; each of these is a random ordering of the integers $1, 2, \dots, N$. The initial state X_0 of the first deck will be some particular nonrandom ordering, for instance, the ordering

$$1234 \cdots N$$

The initial state Y_0 of the second deck will be uniformly distributed on the set of all $N!$ possible orderings. The pair (X_t, Y_t) evolves as follows: at each time t, a random integer J_t between 1 and N is chosen, independently of all past choices, and the card labelled J_t in each of the two decks is moved to the top. Observe that the evolution of each of the two decks follows the random-to-top rule, but that the evolutions of the two decks are *not* independent (as they were in the Doeblin coupling defined by equation (21)). Define τ to be the first time t at which *all* of the cards have been moved at least once. Observe that at time τ the states of the two decks coincide; moreover, once they coincide, they coincide forever after. Therefore, by inequality (23) above, the total variation distance between the distribution of X_t and the uniform distribution is bounded above by $P\{\tau > t\}$.

Example: Random walk on the hypercube The idea is similar to that used in random-to-top shuffling. The coupling will involve two distinct sets of N balls (say "red" and "blue") labelled $1, 2, \dots, N$, each evolving as in the Ehrenfest urn model. The states of the two system will be denoted by R_t and B_t, respectively; each of these is a random $N-$ vector of 0s and 1s. The initial state R_0 is a particular nonrandom vector, perhaps the vector 0 of all zeros; and the initial state B_0 is uniformly distributed on the set $\mathbb{Z}_2^N$ of all possible configurations. The pair (R_t, B_t) evolves as follows: at each time, a random coordinate $J = J_t$ and a Bernoulli-$(1/2)$ random variable ξ_t are chosen, independent of all previous choices. The pair (R_{t+1}, B_{t+1}) is obtained from (R_t, B_t) by changing the Jth coordinate in each of R_t and B_t to ξ_t (that is, both red and blue balls labelled J_t are moved to urn ξ_t). Once again, each of the sequences R_t and B_t behaves individually as a random walk on the hypercube. Define τ to be the first time at which all of the balls have been touched at least once; then $R_t = B_t$ for all $t \geq \tau$. Therefore, as for random-to-top shuffling, the total variation distance between the distribution of X_t and the uniform distribution is bounded above by $P\{\tau > t\}$.

2.4. **The Coupon Collector's Problem.** The couplings described in section 2.3 have an important connection with a famous problem of elementary probability known as the *coupon collector's* problem. Roughly stated, the problem is this: If random draws are made repeatedly and independently from a population of N objects (e.g., "balls" or "cards"), how many draws are needed to make it reasonably likely that every member of the population will be selected at least once? In both of the couplings described in section 2.3, the coalescence time τ is the first time that all N objects have been sampled at least once, and the total variation distance from the equilibrium distribution at time t is bounded above by the tail probability $P\{\tau > t\}$.

The problem of approximating the tail probability $P\{\tau > n\}$ may be reformulated as a question about the distribution of the number U_n of balls that are *not* seen in the first n draws: in particular,

$$P\{\tau > n\} = P\{U_n \geq 1\}. \tag{25}$$

To study the distribution of U_n, we define for each index $j \in [N] := \{1, 2, \dots, N\}$ and each integer $n \geq 1$ indicator random variables

$$Y_j = Y_{n,j} = \begin{cases} 0 & \text{if } j \text{ is chosen in the first } n \text{ draws} \\ 1 & \text{if } j \text{ is } \textit{not} \text{ chosen in the first } n \text{ draws;} \end{cases} \tag{26}$$

then

$$U_n = \sum_{j=1}^{N} Y_{n,j}. \tag{27}$$

Lemma 8.

$$EY_{n,j} = (1 - 1/N)^n \tag{28}$$

$$EY_{n,j}Y_{n,k} = (1 - 2/N)^n. \tag{29}$$

Proof: This is an easy calculation. □

Note that the random variables Y_j and Y_k are negatively correlated. (In fact, one can see this without doing any calculations by observing that if ball j is not chosen then it is *more* likely that ball k will be chosen.) Equations (28) and (29) imply that

$$\begin{aligned}\mathrm{cov}(Y_j, Y_k) &= EY_jY_k - EY_jEY_k \\ &= (1-2/N)^n - (1-1/N)^{2n} < 0\end{aligned}$$

Therefore, the variance of U_n is *smaller* than the variance of the sum of N independent, identically distributed Bernoulli random variables with mean (28):

$$\begin{aligned}\text{(30)}\qquad \mathrm{var}(U_n) &= NEY_1(1-EY_1) + \sum\sum\nolimits_{k\neq j}\mathrm{cov}(Y_j, Y_k) \\ &\leq NEY_1(1-EY_1) \\ &\leq NEY_1 = EU_n \qquad \text{and} \\ \text{(31)}\qquad EU_n &= N(1-1/N)^n.\end{aligned}$$

Proposition 9. *Suppose that $N \to \infty$ and $n \to \infty$ in such a way that $n/N \to C$ for some constant $0 < C < \infty$. Then for every $\epsilon > 0$,*

$$\lim_{n\to\infty} P\{|U_n/N - \exp\{-C\}| \geq \epsilon\} = 0. \tag{32}$$

Proof: If $n/N \to C$ as $N \to \infty$ then $(1-1/N)^n \to \exp\{-C\}$, and so by equation (31), EU_n/N converges to $\exp\{-C\}$. Hence, for sufficiently large N, the difference $EU_n/N - \exp\{-C\}$ will be less than $\epsilon/2$ in magnitude. By Chebyshev's inequality and inequality (30),

$$\begin{aligned}P\{|U_n - EU_n| \geq N\epsilon/2\} &\leq 4\mathrm{var}(U_n)/N^2\epsilon^2 \\ &\leq 4EU_n/N^2\epsilon^2 \\ &\longrightarrow 0\end{aligned}$$

as $N \to \infty$. □

Proposition 9 shows that it is not enough to draw CN times to have any reasonable chance of seeing all N balls in the population, because with CN draws there is a very high probability that about Ne^{-C} balls (a positive fraction of the population) will not be seen. Thus, if we want to have a good chance of seeing all the balls we must draw $n \gg N$ times. How much larger than N should the sample size n be? The answer is hidden in formula (31) for the expected number of balls that will not be seen in n draws. Keeping in mind that $(1-1/N)^N \approx e^{-1}$ for large N, we see that the number n of draws that will get $EU_n \approx 1$ is about $N \log N$. In fact, the following is true.

Lemma 10. *Suppose that $N \to \infty$ and $n \to \infty$ in such a way that $n = \alpha N \log N + \beta N + O(1)$. Then*

$$\left(1 - \frac{1}{N}\right)^n \sim \frac{e^\beta}{N^\alpha}. \tag{33}$$

Note: The notation $O(1)$ indicates that the remainder is bounded as $n, N \to \infty$. The symbol $\sim$ indicates that the ratio of the two sides converges to 1.

Proof: This is based on the behavior of the natural logarithm function near the argument 1: $\log(1-x) = -x + O(x^2)$ as $x \to 0$. Consequently,

$$\begin{aligned}(1-1/N)^n &= \exp\{(\alpha N \log N + \beta N + O(1)) \log(1-1/N)\}\\ &= \exp\{(\alpha N \log N + \beta N + O(1))(-1/N + O(1)/N^2)\}\\ &= N^{-\alpha} e^{\beta} \exp\{O(1)(\log N)/N\}\\ &\sim N^{-\alpha} e^{\beta}.\end{aligned}$$

□

Proposition 11. *Suppose that $N \to \infty$ and $n \to \infty$ in such a way that $n \sim \alpha N \log N$ for some constant $\alpha > 0$. Then*

$$\lim_{n\to\infty} P\{U_n = 0\} = 1 \qquad \text{if } \alpha > 1; \quad \text{and} \tag{34}$$

$$\lim_{n\to\infty} P\{U_n = 0\} = 0 \qquad \text{if } \alpha < 1. \tag{35}$$

In fact, if $\alpha < 1$ then for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P\{U_n \leq N^{1-\alpha-\varepsilon}\} = 0. \tag{36}$$

Proof: This is another simple exercise in the use of the Chebyshev inequality. □

2.5. Exercises.

Problem 6. Poissonization: There is another useful technique for dealing with problems like the coupon collector problem that avoids the need to deal with correlated Bernoulli random variables. It is based on the following simple device: instead of drawing a sample of *fixed* size n, draw a sample of a *random* size κ, where κ is independent of the draws, and has a Poisson distribution with parameter λ. Let $Z_j = Z_{\kappa,j}$ be the number of times ball j is drawn in the first κ draws, and let Y_j be the indicator of the event $Z_j = 0$.

(a) Prove that the random variables $Z_1, Z_2, \dots, Z_N$ are i.i.d., each with Poisson distribution λ/N. (b) Conclude that the random variables Y_j are i.i.d. Bernoulli, with success parameter $\exp\{-\lambda/N\}$. (c) Use the results (a) and (b) to give another proof of the first two relations in Proposition 11. HINT: Use the Chebyshev inequality to show that the event $|\kappa - \lambda| > \lambda^{1/2+\varepsilon}$ has vanishingly small probability as $\lambda \to \infty$.

Problem 7. The Coin Collector: A collector makes random draws repeatedly from a set of N coins, each weighted so that the probability of falling Heads when tossed is p. After each draw, he tosses the coin drawn and observes whether it falls Heads or Tails. Let τ be the first time at which the collector has observed a Head on every one of the N coins. Show that there is a constant $C = C_p$ such that for every $\varepsilon > 0$, as $N \to \infty$,

$$P\{\tau \leq (C-\varepsilon) N \log N\} \longrightarrow 0 \qquad \text{and} \tag{37}$$

$$P\{\tau \leq (C+\varepsilon) N \log N\} \longrightarrow 1. \tag{38}$$

2.6. **Convergence Rates for the Ehrenfest Urn and Random-to-Top.** Our results concerning the coupon collector's problem give us some immediate information about the number of steps necessary to reach equilibrium in the Ehrenfest urn and in random-to-top shuffling. Recall that in each of these Markov chains, the total variation distance between the distribution μ_t of the state variable after t steps and the equilibrium distribution π is bounded above by $P\{\tau > t\}$, where τ is the first time that all of the cards or balls in the urn have been touched at least once. Furthermore, $P\{\tau > n\} = P\{U_n \geq 1\}$. By Proposition 11, if N is large and $n > \alpha N \log N$ then $P\{U_n \geq 1\} \approx 0$. This implies the following corollary.

Corollary 12. *Let μ_t denote the distribution of the state variable X_t or S_t for either random-to-top shuffling of an $N-$card deck or for the Ehrenfest random walk on the $N-$dimensional hypercube $\mathbb{Z}_2^N$. Suppose that $N \to \infty$ and $t \to \infty$ in such a way that $t \sim \alpha N \log N$ for some constant $\alpha > 1$. Then*

$$\|\mu_t - \pi\|_{TV} \longrightarrow 0. \tag{39}$$

Thus, a little more than $N \log N$ steps suffice to reach equilibrium. Are $N \log N$ steps actually needed? For random-to-top shuffling, the answer isn't yet known (as far as I am aware, despite my claim in a previous version of these notes that it is), but you will show in the homework that it takes at least $(1/2)N \log N$ steps. For the Ehrenfest urn scheme, however, the answer is known to be No: in fact, as we will see, just a bit more than $(1/2)N \log N$ steps are needed to reach equilibrium. In section 3 we will develop some additional technical machinery that will give us very precise information about the total variation distance to uniformity. However, even without this machinery we can use what we know about coupon collecting to show that it takes at least $(1/2)N \log N$ steps to reach equilibrium.

Proposition 13. *Let μ_t denote the distribution of the state variable S_t for the Ehrenfest random walk on the $N-$dimensional hypercube $\mathbb{Z}_2^N$, where the initial state is the vector 0 (all balls start in urn 0). Suppose that $N \to \infty$ and $t \to \infty$ in such a way that for some $\delta > 0$,*

$$t \leq ((1/2) - \delta) N \log N. \tag{40}$$

Then

$$\|\mu_t - \pi\|_{TV} = 1. \tag{41}$$

Proof: Recall that the equilibrium distribution π is the uniform distribution on $\mathbb{Z}_2^N$. Under this distribution, the coordinate random variables are i.i.d. Bernoulli-$(1/2)$. Thus, if A is the event that the sum of the coordinates is less than $N/2 - N^{1/2+\varepsilon}$, then

$$\lim_{N\to\infty} \pi(A) = 0, \tag{42}$$

by Chebyshev's inequality (or by the Central Limit Theorem). We will show that if $t \leq ((1/2) - \delta) N \log N$, where $\delta > \varepsilon$, then

$$\lim_{N\to\infty} \mu_t(A) = 1. \tag{43}$$

The relation (41) will then follow from the definition (15) of the total variation distance.

Assume, then, that $t \leq ((1/2) - \delta) N \log N$. By relation (36), for large N it is nearly certain that $U_t \geq N^{(1+\delta)/2}$, where U_t is the number of balls not touched in the first t steps. Each of these balls must therefore remain in urn 0 at time t. On the other hand, each ball that *is* touched in the first t steps has a $50/50$ chance of being in urn 0 at time t.

By the Central Limit Theorem (or Chebyshev's inequality), the chance that fewer than $(N-U_t)/2 - N^{1/2}\log N$ of these are in urn 0 is nearly 0 for large N. Consequently, with probability approaching 1, the number of balls in urn 0 at time t is at least

$$\begin{aligned} U_t + [(N-U_t)/2 - N^{1/2}\log N] &= (N+U_t)/2 - N^{1/2}\log N \\ &\geq N/2 + N^{(1+\delta)/2} - N^{1/2}\log N \end{aligned}$$

If $\varepsilon < \delta/2$ then $N^{(1+\delta)/2} - N^{1/2}\log N > N^{1/2+\varepsilon}$ for all large N, and so (43) must hold. □

We have now established that after $(1.00001)N\log N$ steps the Ehrenfest urn is close to equilibrium, but after $(.49999)N\log N$ steps it is still very far from equilibrium. All of this suggests that the "right" answer is $CN\log N$ for some $C \in [.5, 1]$. What is the right value of C? To answer this, we will next discuss the technique known as *spectral analysis*. But first, some problems.

2.7. **Exercises.**

Problem 8. Top-to-Random Shuffling. "Top- to-random" shuffling works as you might guess: at each step, the top card of the deck is moved to a randomly chosen position in the deck. (More precisely, with probability $1/N$ the top card remains in place, and the state of the deck remains unchanged; and for each $k = 2, 3, \dots, N$, with probability $1/N$ the top card is inserted just below the kth card.)

(a) Verify that the stationary distribution is uniform on the $N!$ permutations of the cards.

The object of the rest of the problem is to determine the number of steps necessary to randomize a deck of N cards. For simplicity, assume that the initial state X_0 of the deck is $1, 2, \dots, N)$, that is, that the cards are initially arranged in order.

(b) Let τ be the first time that card N (which started on the bottom) is moved from the top to a random position. Show that τ is a *strong stationary time*, that is, show that the state of the deck X_τ at time τ has the uniform distribution *and* is independent of τ. HINT: Think about what happens to those cards that are inserted *below* card N before time τ.

(c) Conclude that the total variation distance to the uniform distribution after n steps is no more than $2P\{\tau > n\}$.

(d) Let $\sigma_1, \sigma_2, \dots$ be the successive times at which the top card is moved to a position below card N, and let $\xi_k = \sigma_k - \sigma_{k-1}$(where $\sigma_0 = 0$). Show that (i) the random variables $\xi_1, \xi_2, \dots$ are independent; and (ii) each ξ_k has a geometric distribution with parameter (???) (you figure it out).

(e) Finally, use what you know about sums of independent geometric random variables to show that for any $\varepsilon > 0$

$$\lim_{N\to\infty} P\{\sigma_N > (1+\varepsilon)N\log N\} = 0.$$

Problem 9. Mixing Rate for Random-to-Top Shuffling. The object of this problem is to show that it takes *at least* $(\frac{1}{2} - \varepsilon)N\log N$ steps to randomize a deck of N cards by random-to-top shuffling.

(a) The first step is to learn something about the structure of a random permutation. Let $Y = (Y_1, Y_2, \dots, Y_N)$ be a randomly chosen permutation of the integers $[N] = \{1, 2, \dots, N\}$.

Fix $k_1 < k_2 < \cdots < k_M$. What is the probability that

$$Y_{k_1} < Y_{k_2} < \cdots < Y_{k_M}?$$

(b) A *rising sequence* in a permutation $y = (y_1, y_2, \dots, y_N)$ is a subsequence $k_1 < k_2 < \cdots < k_M$ such that

$$y_{k_1} < y_{k_2} < \cdots < y_{k_M}.$$

Prove that for any $\varepsilon > 0$,

$$\lim_{N \to \infty} P\{\text{there is a rising sequence of length } N^{\frac{1}{2}+\varepsilon} inY\} = 0. \tag{44}$$

HINT: How many increasing sequences $k_1 < k_2 < \cdots < k_M$ of length $M = N^{\frac{1}{2}+\varepsilon}$ are there? Combine this information with the answer to part (a).

(c) By part (b) you know that if you examine a deck of cards and find a rising sequence of length $N^{\frac{1}{2}+\varepsilon}$ then *the deck wasn't random* (or if it was then you hit on a very unlikely event). Now suppose that you start with an ordered deck

$$1, 2, \dots, N$$

and do random-to-top k times. Show that if $k < (\frac{1}{2} - 2\varepsilon) N \log N$ then with probability approaching one (as $N \to \infty$) there will be a rising sequence of length $M = N^{\frac{1}{2}+\varepsilon}$. HINT: What happens to the cards you don't touch in the k shuffles?

3. Spectral Analysis

3.1. Transition Kernel of a Reversible Markov Chain.

Reversible Markov chains are in many ways simpler than Markov chains generally, and not only (or even primarily) because the detailed balance equations provide a simple way to find the stationary distribution. Far more important is the fact that the transition kernel of a reversible Markov chain is *self-adjoint*. The ramifications of this go far beyond what we will discuss in this course; here we will discuss only the case where the state space $\mathcal{X}$ is finite, and where the stationary distribution π is the uniform distribution on $\mathcal{X}$. In this case, the detailed balance equations read

$$p(x, y) = p(y, x); \tag{45}$$

thus, the transition probability matrix $\mathbb{P}$ is *symmetric*. Here is what should immediately come to mind in connection with symmetric matrices:

Spectral Theorem for Symmetric Matrices. *Let A be a symmetric, real, $m \times m$ matrix. Then all eigenvalues of A are real, and there is a complete orthonormal basis (ONB) of real eigenvectors.*

Thus, if you use the right basis – namely, the orthonormal basis of normalized eigenvectors – then the matrix will be diagonal, and the diagonal entries will be the eigenvalues. Here is another way to express this:

$$A = \sum_{j=1}^{m} \lambda_j u_j u_j^T, \tag{46}$$

where u_j is an orthonormal basis of eigenvectors (written as column vectors), u_j^T are the transposes of the vectors u_j, and λ_j are the eigenvalues corresponding to the eigenvectors u_j, that is,

$$Au_j = \lambda_j u_j. \tag{47}$$

An easy consequence of the spectral representation (46) is that the *matrix norm* of A is the maximum of $|\lambda_j|$ over all eigenvalues. What this means is that for any vector v of norm 1, the norm of Av is at most $\max |\lambda_j|$. (EXERCISE: Check this.)

The orthogonality of the eigenvectors is especially important: it guarantees, among other things, that when A is multiplied by itself repeatedly, the result is

$$A^n = \sum_{j=1}^{m} \lambda_j^n u_j u_j^T. \tag{48}$$

(Exercise: Prove this.) Recall from your first course in probability theory that the *characteristic function* of a sum of n i.i.d. random variables is just the nth power of the characteristic function of the first term. Formula (48), when applied to a symmetric transition probability matrix $A = \mathbb{P}$, indicates that, in a certain sense, the *spectrum* of $\mathbb{P}$ (the set of eigenvalues) is the natural analogue for reversible Markov chains of the characteristic function for sums of i.i.d. random variables.

Proposition 14. *Let $\mathbb{P}$ be a symmetric, aperiodic, irreducible transition probability matrix. Then $\lambda_1 = 1$ is a simple eigenvalue, with eigenvector $\mathbf{1}$ (the column vector with all entries 1), and all other eigenvalues λ_j satisfy*

$$-1 < \lambda_j < 1. \tag{49}$$

Remark. This result further indicates that the spectrum of the transition probability matrix acts as an analogue of the characteristic function for sums of i.i.d. r.v.s. In particular, *reversibility* is the natural analogue for Markov chains of *symmetry* for probability distributions on the real line (recall that the characteristic function of a symmetric probability distribution on $\mathbb{R}$ is real-valued).

Proof: The fact that $\mathbb{P}\mathbf{1} = \mathbf{1}$ is merely the assertion that the row sums are all 1, which follows because $\mathbb{P}$ is a transition probability matrix. Thus $\lambda_1 = 1$ is an eigenvalue, with eigenvector $\mathbf{1}$ (normalized, it is $u_1 = \mathbf{1}/\sqrt{m}$). What remains to be shown is that all other eigenvalues are smaller than 1 in modulus.

Let u_j be any nonzero eigenvector orthogonal to $\mathbf{1}$: then $\mathbb{P}u_j = \lambda_j u_j$ for some eigenvalue λ_j. We don't yet know that $|\lambda_j| < 1$, but we do know that $\mathbf{1}^T u_j = 0$. Becasue $\mathbb{P}$ is symmetric, its stationary distribution must be the uniform distribution $\mathbf{1}/m$, and so the $n-$step transition probability matrix $\mathbb{P}^n$ must converge as $n \to \infty$ to the matrix with all entries equal to $1/m$. Consequently,

$$\lim_{n\to\infty} \mathbb{P}^n u_j = \mathbf{1}^T u_j / m = 0.$$

But because $\mathbb{P}u_j = \lambda_j u_j$,

$$\mathbb{P}^n u_j = \lambda_j^n u_j.$$

Thus, it must be the case that $\lambda_j^n \to 0$. This implies that (i) the eigenvalue $\lambda_1 = 1$ is simple; and (ii) all other eigenvalues are strictly less than one in modulus. □

Let's take stock of what we have learned: If $\mathbb{P}$ is a symmetric, aperiodic, irreducible transition probability matrix, then $\lambda_1 = 1$ is a simple eigenvalue with eigenvector $\mathbf{1}$, and all other eigenvalues λ_j are less than one in absolute value. Furthermore, the nth power of $\mathbb{P}$ is given by

$$\mathbb{P}^n = \sum_{j=1}^{m} \lambda_j^n u_j u_j^T, \tag{50}$$

with $u_1 = \mathbf{1}/\sqrt{m}$, so that $u_1 u_1^T$ is the constant matrix with all entries $1/m$. Consequently, the difference

$$\mathbb{P}^n(x,y) - \pi(y) = \mathbb{P}^n(x,y) - 1/m \tag{51}$$

is just the x, y entry of the matrix

$$A^n := \mathbb{P}^n - u_1 u_1^T = \sum_{j=2}^{m} \lambda_j^n u_j u_j^T. \tag{52}$$

Consequently, the rate of convergence of $\mathbb{P}^n$ as $n \to \infty$ is controlled by the *nontrivial* eigenvalues ($\lambda_j \neq 1$) and their eigenvectors. Following are two estimates on the total variation norm $\|\mathbb{P}^n(x,\cdot) - \text{uniform}\|_{TV}$ in terms of the nontrivial eigenvalues.

Proposition 15. *Let $\mathbb{P}$ be a symmetric, aperiodic, irreducible $m \times m$ transition probability matrix with nontrivial eigenvalues $\lambda_2, \lambda_3, \dots, \lambda_m$, written according to multiplicity and in decreasing order of absolute value. Then*

$$\|\mathbb{P}^n(x,\cdot) - \textit{uniform}\|_{TV} \leq \sqrt{m}|\lambda_2|^n. \tag{53}$$

Furthermore, if $\mathbb{P}^{2n}(x,x)$ has the same value for all states x, then

$$\|\mathbb{P}^n(x,\cdot) - \textit{uniform}\|_{TV} \leq \left\{\sum_{j=2}^{m} \lambda_j^{2n}\right\}^{1/2}. \tag{54}$$

Remark. The hypothesis that $\mathbb{P}^{2n}(x,x)$ has the same values for all states x is often satisfied by reversible Markov chains whose stationary distribution is uniform, because the uniformity of the stationary distribution often arises from symmetry. The hypothesis holds for the Ehrenfest random walk, in particular. The reason for wanting this is that the estimate (54) is often sharper than (53).

Proof: The total variation distance in each of the inequalities is just the sum of the absolute values of the entries in the x row of the matrix $A^n = \mathbb{P}^n - \mathbf{1}\mathbf{1}^{\mathbf{T}}/m$; by the Cauchy-Schwartz inequality,

$$
\begin{aligned}
(55) \qquad \left(\sum_y |p_n(x,y) - m^{-1}|\right)^2 &\leq \left(\sum_y m|p_n(x,y) - m^{-1}|^2\right)\left(\sum_y m^{-1}\right) \\
&= \sum_y m|p_n(x,y) - m^{-1}|^2 \\
&= \sum_y m(p_n(x,y)^2 - 2m^{-1}p_n(x,y) + m^{-2}) \\
&= \sum_y mp_n(x,y)^2 - 2 + 1 \\
&= mp_{2n}(x,x) - 1 \\
&= mA^{2n}(x,x)
\end{aligned}
$$

the last by the Chapman-Kolmogorov equations and the fact that $p(x,y) = p(y,x)$. Observe that this string of inequalities implies that $p_{2n}(x,x) \geq 1/m$ for any finite-state Markov chain with a symmetric transition probability matrix.

Thus, the square of the total variation distance is bounded above by m times the diagonal entry $A^{2n}(x,x)$, where A is the matrix defined by (52). This entry equals $e_x^T A^{2n} e_x$, where e_x is the unit vector with a 1 in entry x and all other entries 0. But

$$
\begin{aligned}
e_x^T A^{2n} e_x &= e_x^T A^n A^n e_x \\
&= \|A^n e_x\|^2 \\
&\leq \lambda_2^{2n}.
\end{aligned}
$$

(Recall that the matrix norm of a symmetric matrix is the magnitude of the largest eigenvalue in absolute value.) Combining this estimate with the string of inequalities (55) in the previous paragraph, we obtain the first assertion (53) of the proposition.

Finally, if $\mathbb{P}^{2n}(x,x)$ is the same for all states x then $A^{2n}(x,x)$ is also the same for all states x, and so $mA^{2n}(x,x) = \sum_y A^{2n}(y,y)$, which is just the trace of A^{2n}. Now the trace of a symmetric matrix is the sum of its eigenvalues, in this case $\sum_{j=2}^m \lambda_j^{2n}$ (note that one of the eigenvalues is 0: see (52)). This proves the second assertion (54). □

3.2. **Spectrum of the Ehrenfest random walk.** It is possible to give simple, explicit formulas for the eigenvalues and eigenvectors of the transition probability matrix $\mathbb{P}$ of the Ehrenfest random walk. The reason for this has to do with the fact that the Ehrenfest random walk actually has a representation as a sum of i.i.d. random variables valued in the group $\mathbb{Z}_2^N$; in fact, the eigenvectors of the transition probability matrix will turn out to be the *group characters* of $\mathbb{Z}_2^N$, which are the analogues of the complex exponentials $e^{i\theta x}$. It would take us too far afield to study Fourier analysis on $\mathbb{Z}_2^N$ in general, so I will only write down the eigenvectors and check that they actually are the eigenvectors.

Recall that the state space of the Ehrenfest random walk is the hypercube $\mathbb{Z}_2^N = \{0,1\}^N$, and so the transition probability matrix $\mathbb{P}$ is a $2^N \times 2^N$ matrix. Thus, there will be 2^N vectors in the orthonormal basis of eigenvectors. Because there is no canonical linear ordering of the elements of the state space, it is more natural to think of vectors as functions $u : \mathbb{Z}_2^N \to \mathbb{R}$ than as column vectors of length 2^N: The entries of the vector u are just the values $u(x)$, where $x \in \mathbb{Z}_2^N$, and the dot product of two vectors (functions) u, v is just

$$u^T v = \sum_{x \in \mathbb{Z}_2^N} u(x)v(x). \tag{56}$$

Similarly, the x entry of of $\mathbb{P}u$ is given by

$$\mathbb{P}u(x) = \sum_{y \in \mathbb{Z}_2^N} p(x,y)u(y). \tag{57}$$

We are looking for solutions to the eigenvector equation $\mathbb{P}u = \lambda u$.

Definition 3. The *group characters* are the functions u_J defined by

$$u_J(x) = u_J(x_1, x_2, \dots, x_N) = \prod_{j \in J} (-1)^{x_j}. \tag{58}$$

Here J is an arbitrary subset of $[N] := \{1, 2, \dots, N\}$.

Observe that there are precisely 2^N group characters, one for each of the 2^N different subsets of $[N]$. Each character is a function taking values in the set $\{-1, 1\}$; the character $u_\emptyset$ is the function that is identically one.

Proposition 16. *The group characters, when divided by $2^{N/2}$, form an orthonormal basis, and each character is an eigenvector of the transition probability matrix $\mathbb{P}$ of the Ehrenfest random walk:*

$$\mathbb{P}u_J = \left(1 - \frac{|J|}{N}\right) u_J. \tag{59}$$

Thus, the eigenvalues of $\mathbb{P}$ are $\beta_k := (1 - k/N)$ for $k = 0, 1, 2, \dots, N$, and β_k has multiplicity $\binom{N}{k}$.

Proof: Let J, J' be subsets of $[N]$, and let $K = J \Delta J'$ be their symmetric difference, that is, $K = (J \setminus J') \cup (J' \setminus J)$. Observe that $u_J u_{J'} = u_K$, because every index j that is in either both or neither of J, J' will be written twice in the product $\prod_J \prod_{J'}$. This is relevant because in taking the dot product of u_J and $u_{J'}$ you multiply their entries together and then sum: thus,

$$u_J^T u_{J'} = \sum_{x \in \mathbb{Z}_2^N} u_K(x). \tag{60}$$

I claim that this sum is zero if $K \neq \emptyset$ and 2^N if $J = \emptyset$. The latter is obvious, because if $K = \emptyset$ then $u_K(x) = 1$ for every $x \in \mathbb{Z}_2^N$. That the sum is 0 when $K \neq \emptyset$ is almost as obvious: let i be the smallest element of K, and for each state x let x' be the state that agrees with x in all coordinates except the ith, where it differs. Then $u_K(x) = -u_K(x')$; hence, the terms of the sum (60) cancel in pairs. This proves that the characters, when normalized, are orthonormal. The normalizing constant is $1/\sqrt{N}$, because the sum (60) has the value 2^N when $K = \emptyset$. Since any 2^N nonzero orthogonal vectors in an inner product space of dimension 2^N must constitute a basis, it follows that the normalized characters form an orthonormal basis.

It remains to verify that the eigenvalue λ_J corresponding to the eigenvector u_J is $(1 - 2|J|/N)$. Fix an element $x \in \mathbb{Z}_2^N$, and consider the sum

$$\mathbb{P}u_J(x) = \sum_{y \in \mathbb{Z}_2^N} p(x, y) u_J(y).$$

that defines $\mathbb{P}u_J(x)$: this is just the expected value of $u_J(S_1)$ given that $S_0 = x$, where S_n are the successive states of the Ehrenfest random walk. Recall how the state S_1 is gotten from the initial state S_0: one of the coordinates $j \in [N]$ is chosen at random, and this entry of $S_0 = x$ is changed randomly (to a Bernoulli-1/2). What effect will this have on the value of u_J? If the coordinate j that is chosen is *not* in J, then there will be no change: $u_J(S_1) = u_J(S_0)$. If the coordinate j *is* in J, however, then $u_J(S_1) = \varepsilon u_J(S_0)$ where $\varepsilon = \pm 1$ with probability $1/2$.. Consequently,

$$E(u_J(S_1) \,|\, S_0 = x) = u_J(x)(1 - |J|/N) - u_J(x)|J|/2N + u_J(x)|J|/2N,$$

and so u_J is indeed an eigenvector, with eigenvalue $\lambda_j = 1 - |J|/N$. □

3.3. **Rate of convergence of the Ehrenfest random walk.** We are now prepared to determine the mixing rate of the Ehrenfest random walk. By Proposition 16, we know that the eigenvalues of the transition probability matrix are the numbers $\beta_k = 1 - k/N$, where $k = 0, 1, 2, \ldots, N$, and that eigenvalue β_k occurs with multiplicity $\binom{N}{k}$. Proposition 15 provides bounds on the total variation distance from the stationary distribution in terms of the eigenvalues. So let's see what becomes of inequality (54) when

$$n = \frac{1}{2} N \log N + CN.$$

The right side of (54) involves a sum over all of the eigenvalues λ_j *except* the trivial eigenvalue $\lambda_1 = 1$. Hence

$$
\begin{aligned}
(61)\qquad (\|\mathbb{P}^n(x,\cdot) - \text{uniform}\|_{TV})^2 &\le \sum_{j=2}^{2^N} \lambda_j^{2n} \\
&= \sum_{k=1}^{N} \binom{N}{k} \left(1 - \frac{k}{N}\right)^{2n} \\
&\le \sum_{k=1}^{N} \frac{N^k}{k!} \exp\{2n \log(1 - k/N)\} \\
&\le \sum_{k=1}^{N} \frac{N^k}{k!} \exp\{-2nk/N\} \\
&\le \sum_{k=1}^{N} \frac{N^k}{k!} \exp\{-k \log N - 2Ck\} \\
&= \sum_{k=1}^{N} \frac{e^{-2Ck}}{k!} \\
&\le \sum_{k=1}^{\infty} \frac{e^{-2Ck}}{k!} \\
&= \exp\{\exp\{-2C\}\} - 1.
\end{aligned}
$$

Note that this final bound is small when C is large. Thus, a little more than $(1/2)N \log N$ steps suffice for the Ehrenfest random walk to reach equilibrium. Recall that, by the arguments of section 2.6, at least $((N/2) - \delta N) \log N$ steps are necessary to reach equilibrium. Therefore, we have now established the following.

Theorem 17. *Let μ_n denote the distribution of the state variable S_n for the Ehrenfest random walk on the $N-$dimensional hypercube $\mathbb{Z}_2^N$, where the initial state is the vector 0 (that is, all balls in urn 0). Then for every $\delta > 0$, as $N \to \infty$,*

$$
(62)\qquad \min_{n \le (N/2 - N\delta) \log N} \|\mu_n - \pi\|_{TV} \longrightarrow 1 \qquad \textit{and}
$$

$$
(63)\qquad \max_{n \ge (N/2 + N\delta) \log N} \|\mu_n - \pi\|_{TV} \longrightarrow 0.
$$

Problem 10. Consider a variant of the Ehrenfest random walk in which, at each step, the ball that is selected is returned to the urn it came from with probability $1 - p$ and moved to the other urn with probability p. (The case studied in the notes is $p = 1/2$.)

(a) Calculate the spectrum of the transition probability matrix. (b) Show that for some constant $C > 0$, about $CN \log N$ steps are needed to reach equilibrium, and identify C in terms of p.