2. More on Markov chains, Examples and Applications

2.1 Branching Processes

The branching process model we will study was formulated in 1873 by Sir Francis Galton,[*] who was interested in the survival and extinction of family names. Suppose children inherit their fathers' names, so we need only keep track of fathers and sons. Consider a male who is the only member of his generation to have a given family name, so that the responsibility of keeping the family name alive falls upon him—if his line of male descendants terminates, so does the family name. Suppose for simplicity that each male has probability $f(0)$ of producing no sons, $f(1)$ of producing one son, and so on. Here is a question: What is the probability that the family name eventually becomes extinct?

Galton brought the problem to his mathematician friend, Rev. H. W. Watson, who devised the method of analysis using probability generating functions that is still used today. However, a minor mathematical slip caused Galton and Watson to get the answer to the main question wrong. They believed that the extinction probability is 1 — all names are doomed to eventual extinction. We will see below that this is false: if the expected number of sons is greater than 1, the branching process model produces lines of descent that have positive probability of going on forever.

Let us begin with a more formal description of the branching process. Thinking of G_t as the number of males in generation t, start with $G_0 = 1$. If $G_t = i$ then write $G_{t+1} = X_{t1} + X_{t2} + \cdots + X_{ti}$; here X_{tj} denotes the number of sons fathered by the jth man in generation t. Assume the random variables $\{X_{tj} : t \geq 0, j \geq 1\}$ are *iid* with probability mass function f, so that $\mathbb{P}\{X_{tj} = k\} = f(k)$ for $k = 0, 1, \ldots$. To avoid trivial cases we

[*] See Jagers (1975) and Guttorp (1991) for more on the history.

assume that $f(0) > 0$ and $f(0) + f(1) < 1$. ⟦Why are these trivial?⟧ We are interested in the extinction probability $\rho = \mathbb{P}_1\{G_t = 0 \text{ for some } t\}$.

It is clear from the verbal description of the process that $\{G_t : t \geq 0\}$ is a Markov chain. We can say a few interesting things about the process directly from general results of the previous chapter. Clearly state 0 is absorbing. Therefore, for each $i > 0$, since $\mathbb{P}_i\{G_1 = 0\} = (f(0))^i > 0$, the state i must be transient—this follows from Theorem (1.24). Consequently, we know that with probability 1, each state $i > 0$ is visited only a finite number of times. From this, a bit of thought shows that, with probability 1, the chain must either get absorbed at 0 eventually or approach ∞. ⟦Exercise: *Think a bit and show this.*⟧

We can obtain an equation for ρ by the idea we have used before a number of times—e.g. see exercise ([1.3])—namely, conditioning on what happens at the first step of the chain. This gives

$$\rho = \sum_{k=0}^{\infty} \mathbb{P}\{G_1 = k \mid G_0 = 1\}\mathbb{P}\{\text{eventual extinction} \mid G_1 = k\}.$$

Evidently, since the males all have sons independently (in the terminology above, the random variables X_{tj} are independent), we have $\mathbb{P}\{\text{eventual extinction} \mid G_1 = k\} = \rho^k$. This holds because the event of eventual extinction, given $G_1 = k$, requires each of the k male lines starting at time 1 to reach eventual extinction; this is the intersection of k independent events, each of which has probability ρ. Thus, ρ satisfies

$$\rho = \sum_{k=0}^{\infty} f(k)\rho^k =: \psi(\rho). \tag{2.1}$$

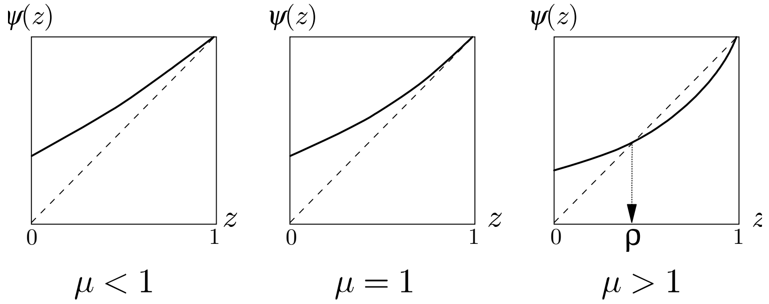
The last sum is a function of ρ; for each distribution f there is a corresponding function of ρ, which we have denoted by ψ. So ρ satisfies $\psi(\rho) = \rho$: the extinction probability ρ is a fixed point of ψ.

So the function ψ, which is called the *probability generating function* of the probability mass function f, arises in a natural and interesting way in this problem. Let us pause to collect a few of its properties. The first two derivatives are given by

$$\psi'(z) = \sum_{k=1}^{\infty} k f(k) z^{k-1}, \quad \psi''(z) = \sum_{k=2}^{\infty} k(k-1) f(k) z^{k-2}$$

for $z \in (0,1)$. Since these are positive, the function ψ is strictly increasing and convex on (0,1). Also, clearly $\psi(0) = f(0)$ and $\psi(1) = 1$. Finally, notice that $\psi'(1) = \sum k f(k) = \mu$, where μ denotes $E(X)$, the expected number of sons for each male.

These properties imply that the graph of ψ over [0,1] must look like one of the three following pictures, depending on the value of $\mu = \psi'(1)$.



So you can see what happens. Since $\psi(1) = 1$, the equation $\psi(\rho) = \rho$ always has a trivial solution at $\rho = 1$. When $\mu \le 1$, this trivial solution is the only solution, so that, since the probability ρ of eventual extinction satisfies $\psi(\rho) = \rho$, it must be the case that $\rho = 1$. When $\mu > 1$, there is one additional solution, indicated by the arrow in the picture. This solution was missed by Watson and Galton (1875), leading them to believe that the probability of extinction would be 1 in this case as well. We will show that this was incorrect, and that the probability of extinction is the smaller solution of the equation $\psi(\rho) = \rho$.

Thus, assuming $\mu > 1$ and defining r to be the smaller solution of $\psi(r) = r$, we want to show that $\rho = r$. Since $\psi(\rho) = \rho$, we know that ρ must be either r or 1. Defining $p_t = \mathbb{P}_1\{G_t = 0\}$, observe that as $t \to \infty$,

$$p_t \uparrow \mathbb{P}_1\Big[\bigcup_{n=1}^{\infty}\{G_n = 0\}\Big] = \rho.$$

Therefore, to rule out the possibility that $\rho = 1$, it is sufficient to prove the following statement:

$$p_t \le r \text{ for all } t.$$

To prove this by induction, observe that $p_0 = 0$, so that the statement holds for $t = 0$. Next observe that

$$p_{t+1} = \mathbb{P}_1\{G_{t+1} = 0\} = \sum_{i=0}^{\infty} \mathbb{P}_1\{G_1 = i\}\mathbb{P}_1\{G_{t+1} = 0 \mid G_1 = i\} = \sum_{i=0}^{\infty} f(i)(p_t)^i,$$

that is, $p_{t+1} = \psi(p_t)$. Thus, using the induction hypothesis $p_t \le r$ and the fact that the function ψ is increasing, we obtain $p_{t+1} \le \psi(r) = r$, which completes the proof.

(2.2) Example. Suppose each man has 3 children, with each child having probability 1/2 of being male, and different children being independent. What is the probability that a particular man's line of male descendants will eventually become extinct? Here the distribution f is the binomial distribution Bin(3,1/2), so that $\mu = 3/2 > 1$. Thus, we know that the probability ρ of extinction is less than 1. Here $f(0) = 1/8$, $f(1) = 3/8$, $f(2) = 3/8$, and $f(3) = 1/8$, so that the equation $\psi(r) = r$ becomes

$$\frac{1}{8} + \frac{3}{8}r + \frac{3}{8}r^2 + \frac{1}{8}r^3 = r,$$

or $r^3 + 3r^2 - 5r + 1 = 0$. Fortunately, $r = 1$ is a solution (as it must be!), so we can factor it out, getting the equation $(r - 1)(r^2 + 4r - 1) = 0$. Solving the quadratic equation gives $\rho = \sqrt{5} - 2 = 0.2361$. The man can rest assured that with probability $1 - \rho = 0.7639$ his glorious family name will live on forever. □

2.2 Time Reversibility

Let $X_0, X_1, \ldots$ be a Markov chain having probability transition matrix $P = P(i, j)$. Imagine that I recorded a movie of the sequence of states $(X_0, \ldots, X_n)$, and I am showing you the movie on my fancy machine that can play the tape forward or backward equally well. Can you tell by watching the sequence of transitions on the movie whether I am showing it forward or backward?

Of course, we are assuming that you know the transition matrix P; otherwise, this would be an unreasonable request. There are cases in which distinguishing the direction of time is very easy. For example, if the state space is $\{1, 2, 3\}$ and $P(1, 2) = P(2, 3) = P(3, 1) = 1$ ⟦one of our standard periodic examples⟧, observing just one transition of the chain is enough to tell you for sure the direction of time; for example, a "movie" in which we observe 3 followed by 2 must be running backward.

That one was easy. Let's consider another example: do you think a stationary Ehrenfest chain is time-reversible? Here the state space is $\{0, 1, \ldots, d\}$, say, and $X_0 \sim \text{Bin}(d, 1/2)$, the stationary distribution of the chain. It is clear in this case that you will not be able to tell *for sure* from observing any finite movie $(X_0, \ldots, X_n)$ which direction the movie is being shown—a sequence has positive probability if and only if its reversal also has positive probability. But we are asking whether or not you can get *any sort of probabilistic hint* about the direction in which the movie is being shown, and I am willing to show you as long a segment as you would like to request. So you can have plenty of data to look at. One might suspect that it should be possible to make this sort of distinction. For example, we know that the Ehrefest chain has a "restoring force" that pulls it toward the level $d/2$, where half the balls are in each of the two urns. So, for instance, if we observe a long sequence that moves from $(3/4)d$ down toward $d/2$, we might favor the explanation that the movie is being shown forward, since otherwise we are observing a long sequence moving against the restoring force.

Did you buy that? I hope not, because in fact we will see that the Ehrenfest chain is time-reversible: no movie, no matter how long, will give you any probabilistic information that is useful in distinguishing the direction of time. ⟦And the argument suggested above really didn't make much sense — what comes down must have gone up.⟧

Here is a definition that captures the concept.

(2.3) Definition. *We say that a Markov chain $\{X_n\}$ is time-reversible if, for each n,*

$$(X_0, X_1, \ldots, X_n) \overset{\mathcal{D}}{=} (X_n, X_{n-1}, \ldots, X_0)$$

that is, the joint distribution of $(X_0, X_1, \ldots, X_n)$ is the same as the joint distribution of $(X_n, X_{n-1}, \ldots, X_0)$.

Suppose $\{X_n\}$ is time-reversible. As a particular consequence of the definition, we see that $(X_0, X_1) \overset{\mathcal{D}}{=} (X_1, X_0)$. This, in turn, implies that $X_1 \overset{\mathcal{D}}{=} X_0$, that is, $\pi_1 = \pi_0$. Thus, in view of the relation $\pi_1 = \pi_0 P$, we obtain $\pi_0 = \pi_0 P$, so that the initial distribution π_0 is stationary. Not surprisingly, we have found that a time-reversible chain must be stationary.

We will write π for the distribution π_0 to emphasize that it is stationary. So $X_n \sim \pi$ for all n. The condition $(X_0, X_1) \overset{\mathcal{D}}{=} (X_1, X_0)$ says that $\mathbb{P}\{X_0 = i, X_1 = j\} = \mathbb{P}\{X_1 = i, X_0 = j\}$ for all i, j; that is,

$$\pi(i)P(i,j) = \pi(j)P(j,i) \text{ for all } i, j. \tag{2.4}$$

We have shown that the condition (2.4) together with $X_0 \sim \pi$ is necessary for a chain to be reversible. In fact, these two conditions are also sufficient for reversibility.

(2.5) Proposition. *The Markov chain $\{X_n\}$ is time-reversible if and only if the distribution π of X_0 satisfies $\pi P = \pi$ and the condition (2.4) holds.*
To see this, observe that (2.4) gives, for example,

$$\begin{aligned}
\mathbb{P}\{X_0 = i, X_1 = j, X_2 = k\} &= [\pi(i)P(i,j)]P(j,k) \\
&= [P(j,i)\pi(j)]P(j,k) \\
&= P(j,i)[\pi(j)P(j,k)] \\
&= P(j,i)[P(k,j)\pi(k)] \\
&= \pi(k)P(k,j)P(j,i) \\
&= \mathbb{P}\{X_0 = k, X_1 = j, X_2 = i\} \\
&= \mathbb{P}\{X_2 = i, X_1 = j, X_0 = k\},
\end{aligned}$$

that is, $(X_0, X_1, X_2) \overset{\mathcal{D}}{=} (X_2, X_1, X_0)$. Notice how (2.4) allowed the π factor to propagate through the product from the left end to the right, reversing the direction of all of the transitions along the way. The same trick allows us to deduce the general equality required in the definition (2.3).

The equalities in (2.4) have a nice interpretation in terms of probability flux. Recall ⟦as discussed in one of your homework problems⟧ that the flux from i to j is defined as $\pi(i)P(i,j)$. So (2.4) says that the flux from i to j is the same as the flux from j to i—flux balances between each pair of states. These are called the "detailed balance" (or "local balance") equations; they are more detailed than the "global balance equations" $\pi(j) = \sum_i \pi(i)P(i,j)$ that characterize stationarity. Global balance, which can be rewritten as $\sum_i \pi(j)P(j,i) = \sum_i \pi(i)P(i,j)$ says that the total flux leading out of state j is the same as the total flux into state j. If we think of a system of containers of fluid connected by tubes, one container for each state, and we think of probability flux as fluid flowing around the system, global balance says that the flow out of container j is balanced by the flow into j, so that the fluid level in container j stays constant, neither rising nor falling. This is a less stringent requirement than detailed balance, which requires flow to balance between each pair of containers.

A more probabilistic interpretation is this: think of $\pi(i)P(i,j)$ as the limiting long run fraction of transitions made by the Markov chain that go from state i to state j. Time reversibility requires that the long run fraction of i-to-j transitions is the same as that of the j-to-i transitions, for all i and j. This is a more stringent requirement than stationarity, which equates the long run fraction of transitions that go out of state i to the long run fraction of transitions that go into state i.

The mathematical formulation of this relationship is simple.

(2.6) PROPOSITION. *If the local balance equations (2.4) hold, then the distribution π is stationary.*

PROOF: Summing the local balance equations (2.4) over i gives the global balance equations

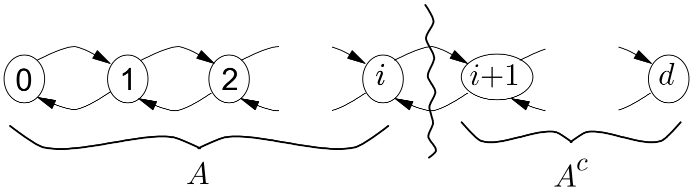
$$\sum_i \pi(i)P(i,j) = \sum_i \pi(j)P(j,i) = \pi(j).$$

□

So why is the Ehrenfest chain time-reversible? The Ehrenfest chain is an example of a *birth and death chain*, which is defined to be a Markov chain whose states consist of nonnegative integers and whose transitions increase or decrease the state by at most 1. That is, interpreting the current state of the chain as the population count of living individuals, the population can change by at most 1 in a transition, which might represent a birth, a death, or no change. The time reversibility of the Ehrenfest chain is an example of a more general fact.

(2.7) CLAIM. *All stationary birth and death chains are reversible.*

To show this, consider a stationary birth and death chain on the state space $\mathcal{S} = \{0, 1, \ldots, d\}$.



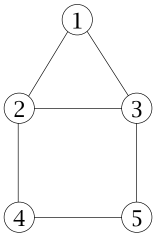
We ask: does $\pi(i)P(i,j) = \pi(j)P(j,i)$ for all i,j? Since a birth and death chain has $P(i,j) = 0$ if $|i-j| > 1$, we need only consider the case where $j = i+1$. Recall from

Exercise [1.11] (the exercise on probability flux) that for any subset $A \subset \mathcal{S}$, the flux from A to A^c must equal the flux from A^c to A. Taking $A = \{0, \ldots, i\}$ as in the picture gives just what we want: $\pi(i)P(i, i+1) = \pi(i+1)P(i+1, i)$. This establishes the claim. □

(2.8) Example. Another important class of examples of time-reversible Markov chains is the ***random walk on a graph***. Defining $d(i)$ to be the degree of node i, the random walk moves according to the matrix $P(i, j) = 1/(d(i))$ for each neighbor j of node i, and $P(i, j) = 0$ otherwise. Then it is easy to check that the distribution

$$\pi(i) = \frac{d(i)}{\sum_{j \in \mathcal{S}} d(j)}$$

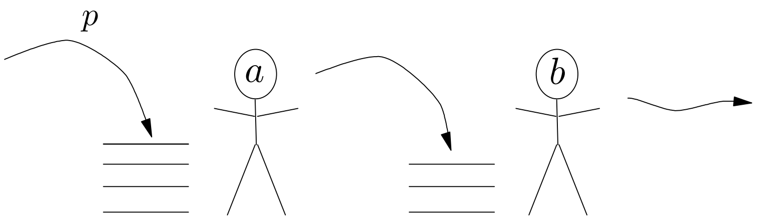
satisfies the detailed balance equations. Thus, the random walk is time-reversible, and π is its stationary distribution.



For example, consider a random walk on the house graph above. The degrees are $(d_1, d_2, d_3, d_4, d_5) = (2, 3, 3, 2, 2)$. So the stationary distribution is $(\pi_1, \pi_2, \pi_3, \pi_4, \pi_5) = (2/12, 3/12, 3/12, 2/12, 2/12)$. □

2.3 More on Time Reversibility: A Tandem Queue Model

Consider two office workers Andrew and Bertha who have a lot of paper work to do. When a piece of paper arrives at the office, it goes first to Andrew for processing. When he completes his task, he puts the paper on Bertha's pile. When she completes her processing of the paper, it is sent out of the office.

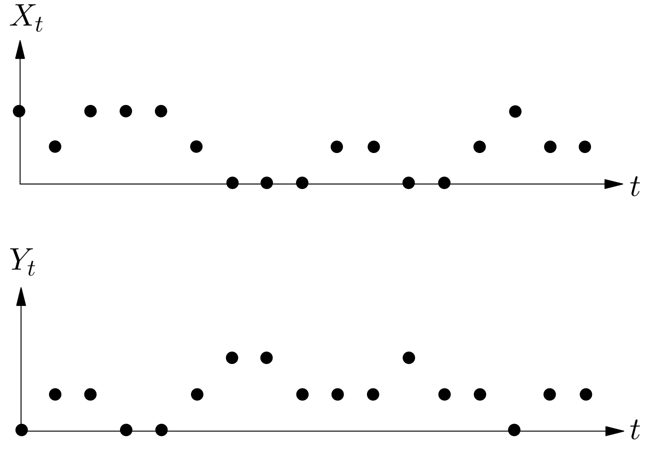


Let's specify a model of this in more detail. Consider a Markov chain whose state at time n is (X_n, Y_n), where X_n is the number of papers in Andrew's work pile and Y_n is the number of papers in Bertha's pile. Suppose that $X_n = i \geq 0$. Then with probability p, a new piece of work enters the office, resulting in $X_{n+1} = i + 1$. For definiteness, it is helpful to paint a detailed picture, so let's suppose that any new arrival is placed onto the *top* of Andrew's current pile of papers, preempting any other task he might have been working on at the time. Thus, if a new paper arrives, Andrew cannot complete his processing of any previously received papers that period. If Andrew receives no new arrival in a period and $i > 0$, then with probability a Andrew completes the processing of the paper on top of his pile, resulting in $X_{n+1} = i - 1$. Thus, in summary, Andrew's pile evolves as follows. Given $X_n = i > 0$,

$$X_{n+1} = \begin{cases} i+1 & \text{with probability } p \\ i & \text{with probability } (1-p)(1-a) \\ i-1 & \text{with probability } (1-p)a, \end{cases}$$

and if $X_n = 0$, then $X_{n+1} = 1$ with probability p and $X_{n+1} = 0$ with probability $1 - p$.

Let us assume that Bertha's pile changes in the same way as Andrew's, except that she gets her new work from Andrew's completed papers rather than from the outside, and her service-completion probability is b rather than a. A sample path of the $\{(X_n, Y_n)\}$ process is shown in the picture. Notice that in each period in which the X process decreases, the Y process increases: work completed by Andrew is sent to Bertha.



The $\{X_n\}$ process is a birth and death chain in its own right. Letting q_a denote the downward transition probability $q_a = (1-p)a$, a stationary distribution π_a for Andrew's process X exists if $p < q_a$, in which case familiar probability flux reasoning gives $\pi_a(i)p = \pi_a(i+1)q_a$, or $\pi_a(i+1)/\pi_a(i) = p/q_a$, so that

$$\pi_a(i) = \left(\frac{p}{q_a}\right)^i \left(1 - \frac{p}{q_a}\right) \quad \text{for } i = 0, 1, \ldots.$$

Here is where time reversibility allows us to make an interesting and useful observation. Assume $X_0 \sim \pi_a$. Then we know that $\{X_n\}$, being a stationary birth and death process, is time reversible. Define A_n to be 1 if Andrew has an "arrival" at time n and 0 otherwise, so that $A_n = 1$ occurs when $X_n = X_{n-1} + 1$. Define another indicator random variable D_n to be 1 if Andrew has a "departure" at time n, that is, when $X_n = X_{n-1} - 1$. Considering k to be the present time, clearly the present queue size X_k is independent of the future arrivals $A_{k+1}, A_{k+2}, \ldots$. This obvious fact, when applied to the reversed process, gives something interesting. In the reversed process, if k again represents the "present," then "future arrivals" correspond to the departures $D_k, D_{k-1}, \ldots$. Therefore, we can say that the departures $(D_1, D_2, \ldots, D_k)$ are independent of the queue size X_k. This is quite surprising, isn't it? Also, since reversibility implies that arrivals in the reversed process have the same probabilistic behavior as arrivals in the forward process, we see that the departures $D_1, D_2, \ldots$ are *iid* Bernoulli(p). Thus, the output process of Andrew's queue is the same probabilistically as the input process of his queue. Isn't that interesting? For example, we have found that the departure process does not depend on the service completion probability a.

Bertha's queue size $\{Y_n\}$ is also a birth and death chain, with a similar structure as Andrew's. We have just shown that Bertha's input consists of *iid* Bernoulli(p) random variables, just as Andrew's input does. Defining the downward transition probability $q_b =$

$(1-p)b$ for $\{Y_n\}$, if $p < q_b$ the stationary distribution π_b is given by

$$\pi_b(i) = \left(\frac{p}{q_b}\right)^i \left(1 - \frac{p}{q_b}\right) \quad \text{for } i = 0, 1, \ldots.$$

Now we are ready to show a surprising property of the stationary distribution of (X_k, Y_k): the two queue sizes X_k and Y_k are independent! That is, we claim that the stationary distribution π of $\{(X_n, Y_n)\}$ takes the product form

$$\pi(i, j) = \pi_a(i)\pi_b(j).$$

I'll try to say the idea loosely in words first, then more carefully. It is helpful to imagine that at each time, Andrew flips a coin with probability a of heads, and if he gets heads he completes a piece of work, if that is possible—i.e. if there is work in his pile and if he has not received a new arrival. Bertha does the same, only with her coin having probability b. Back to the question: Supposing that X_0 and Y_0 are independent with distributions π_a and π_b, we want to see that X_n and Y_n are also independent with distributions π_a and π_b. We know the marginal distributions of X_n and Y_n are π_a and π_b; independence is the real question. The key is the observation made above that X_n is independent of Andrew's departures up to time n, which are the same as Bertha's arrivals up to time n. So since Y_n is determined Y_0, Bertha's arrivals up to time n, and Bertha's service coin flips, all of which are independent of X_n, we should have Y_n independent of X_n.

To establish this more formally, assuming that $(X_0, Y_0) \sim \pi$, we want to show that $(X_1, Y_1) \sim \pi$. Since π_a and π_b are stationary for $\{X_n\}$ and $\{Y_n\}$, we know that $X_1 \sim \pi_a$ and $Y_1 \sim \pi_b$, so our task is to show that X_1 and Y_1 are independent. Let A_k^X denote the indicator of an arrival to Andrew's desk at time k. Let $S_k^X = 1$ if at time k Andrew's "service completion coin flip" as described in the previous paragraph comes up heads, and $S_k^X = 0$ otherwise. Define S_k^Y analogously for Bertha. We are assuming that the random variables X_0, Y_0, A_1^X, S_1^X, and S_1^Y are independent. But we can write X_1 and A_1^Y as functions of (X_0, A_1^X, S_1^X). ⟦The precise functional forms are not important, but just for fun,

$$X_1 = \begin{cases} X_0 + 1 & \text{if } A_1^X = 1 \\ X_0 & \text{if } A_1^X = 0 \text{ and } S_1^X = 0 \\ X_0 - 1 & \text{if } A_1^X = 0 \text{ and } S_1^X = 1 \text{ and } X_0 > 0 \\ 0 & \text{if } A_1^X = 0 \text{ and } X_0 = 0 \end{cases}$$

and

$$A_1^Y = \begin{cases} 1 & \text{if } X_0 > 0 \text{ and } A_1^X = 0 \text{ and } S_1^X = 1 \\ 0 & \text{otherwise} \end{cases}$$

is one way to write them.⟧ So Y_0 is independent of (X_1, A_1^Y). But we know that X_1 is independent of whether there is a departure from Andrew's queue at time 1, which is just the indicator A_1^Y. Therefore, the 3 random variables Y_0, X_1, and A_1^Y are independent. Finally, observe that S_1^Y is independent of (Y_0, X_1, A_1^Y), so that the 4 random variables Y_0, X_1, A_1^Y, and S_1^Y are all independent. Thus, since Y_1 is a function of (Y_0, A_1^Y, S_1^Y), it follows that X_1 and Y_1 are independent.

2.4 The Metropolis method

This is a very useful general method for using Markov chains for simulation. The idea is a famous one due to Metropolis et al. (1953), and is known as the *Metropolis method*. Suppose we want to simulate a random draw from some distribution π on a finite set $\mathcal{S}$. By the Basic Limit Theorem above, one way to do this (approximately) is to find an irreducible, aperiodic probability transition matrix P satisfying $\pi P = \pi$, and then run a Markov chain according to P for a sufficiently long time.

Suppose we have chosen some connected graph structure $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ on our set $\mathcal{S}$. That is, we think of each element of $\mathcal{S}$ as a node, and we imagine a collection $\mathcal{E}$ of edges, where each edge joins some pair of nodes. If nodes i and j are joined by an edge, we say i and j are *neighbors*. Let $\mathcal{N}(i)$ denote the set of neighbors of node i. ⟦We'll assume that $i \notin \mathcal{N}(i)$ for all i.⟧ Just to make sure the situation is clear, I want to emphasize that this graph structure is not imposed on us, and there is not a single, unique, magic choice that will work. The set of edges $\mathcal{E}$ is ours to choose; we have a great deal of freedom here, and different choices will lead to different matrices P satisfying $\pi P = \pi$.

As a preliminary observation, recall from Example (2.8) that a random walk on $\mathcal{G}$, which has probability transition matrix

$$P_{\mathrm{rw}}(i,j) = \begin{cases} \frac{1}{d(i)} & \text{if } j \in \mathcal{N}(i) \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

has stationary distribution

$$\pi_{\mathrm{rw}}(i) = \frac{d(i)}{\sum_{j \in \mathcal{S}} d(j)},$$

where $d(i)$ is the degree of node i. To reduce typographical and conceptual clutter, let us write this as

$$\pi_{\mathrm{rw}}(i) \propto d(i),$$

by omitting the denominator, which is simply a normalization constant ⟦constant in that it does not depend on i⟧ that makes the probabilities add up to 1. The Basic Limit Theorem tells us that (assuming aperiodicity holds) if we run the random walk for sufficiently long, then we get arbitrarily close to achieving the distribution π_{rw}.

Thus, simply running a random walk on $\mathcal{G}$ would solve our simulation problem if we happened to want to simulate from π_{rw}. In general, however, we will want to simulate from some different, arbitrary distribution π, which we will write in the form

$$\pi(i) \propto d(i) f(i). \tag{2.10}$$

That is, we are interested in modifying the relative probabilities of the natural random walk stationary distribution by some multiplicative function f. Our goal here is a simple way to modify the random walk transition probabilities in such a way that the modified probability transition matrix has stationary distribution π. The Metropolis method solves this problem by defining the probability transition matrix

$$P(i,j) = \begin{cases} \frac{1}{d(i)}\min\{1, \frac{f(j)}{f(i)}\} & \text{if } j \in \mathcal{N}(i) \\ 1 - \sum_{k\in\mathcal{N}(i)} P(i,k) & \text{if } j = i \\ 0 & \text{otherwise.} \end{cases} \tag{2.11}$$

The verification that this method works is simple.

(2.12) Claim. *For π defined by (2.10) and $(P(i,j))$ defined by (2.11), we have $\pi P = \pi$.*

Proof: For $j \in \mathcal{N}(i)$,

$$\pi(i)P(i,j) \propto f(i)\min\{1, \frac{f(j)}{f(i)}\} = \min\{f(i), f(j)\}$$

is symmetric in i and j, so we have

$$\pi(i)P(i,j) = \pi(j)P(j,i). \tag{2.13}$$

In fact, (2.13) holds for all i and j, since it is trivial to verify (2.13) in the cases when $i = j$ or $j \notin N(i)$. Summing (2.13) over i gives

$$\sum_i \pi(i)P(i,j) = \pi(j)\sum_i P(j,i) = \pi(j),$$

so that $\pi P = \pi$. □

Notice that the last proof showed that the "detailed balance" equations $\pi(i)P(i,j) = \pi(j)P(j,i)$ that characterize time-reversible Markov chains hold.

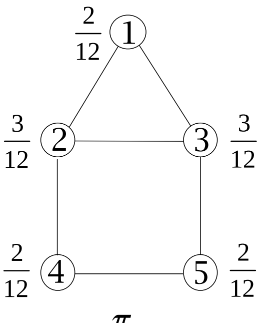
▷ *As we know from our discussion of the Basic Limit Theorem, the fact that a Markov chain has stationary distribution π does not in itself guarantee that the Markov chain will converge in distribution to π. Exercise [2.10] gives conditions under which this convergence holds.*

Running the Metropolis chain (using the P from (2.11)) is actually a simple modification of performing a random walk (using P_{rw} from (2.9)). To run the random walk, at each time, we choose a random neighbor and go there. We can think of running the Metropolis chain as follows. Suppose we are currently at state i and we are about to generate our next random transition. We start out, in the same way as the random walk, by choosing a random neighbor of i; let's call our choice j. The difference between the Metropolis chain and the random walk is that in the Metropolis chain, we might move to j, or we might stay at i. So let's think of j as our "candidate" state, and we next make a probabilistic decision about whether to "accept the candidate" and move to j, or "reject the candidate" and stay at i. The probability that we accept the candidate is the extra factor

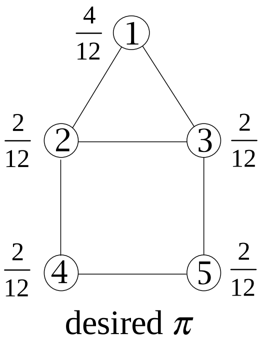
$$\min\{1, \frac{f(j)}{f(i)}\} \tag{2.14}$$

that appears in (2.11). Thus, having nominated a candidate j, we look at the values $f(j)$ and $f(i)$. If $f(j) \geq f(i)$, the minimum in (2.14) is 1, and we definitely move to j. If $f(j) < f(i)$, the minimum in (2.14) is $f(j)/f(i)$, and we move to j with probability $f(j)/f(i)$. This makes qualitative sense: for example, if $f(j)$ is much smaller than $f(i)$, this means that, relative to the random walk stationary distribution π_{rw}, our desired distribution π places much less probability on j than on i, so that we should make a transition from i to j much less frequently than the random walk does. This is accomplished in the Metropolis chain by usually rejecting the candidate j.

(2.15) Example. To illustrate the Metropolis method in a simple way, we'll discuss an artificial toy example where we don't really need the method. Suppose the distribution π on $\mathcal{S} = \{1,2,3,4,5\}$ is $(4/12, 2/12, 2/12, 2/12, 2/12)$, and suppose the graph structure we choose on $\mathcal{S}$ is the "house" graph from Example (2.8). Thus, we want to be on the roof of the house (state 1) with probability 4/12, and at each of the other states with equal probability, 2/12.

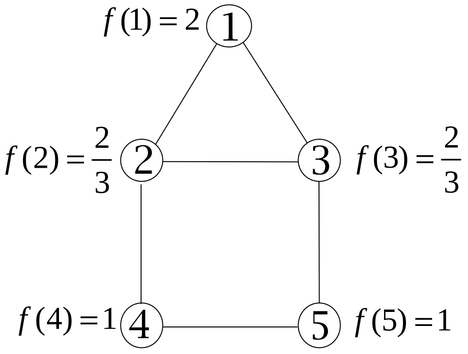


π_{rw}

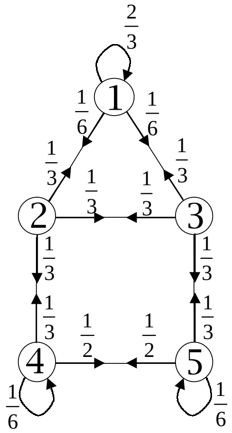


desired π

Comparing with the distribution π_{rw} that we found in Example (2.8) ⟦and is reproduced in the figure above⟧ we can calculate the f ratios to be $f(1) = (4/12)/(2/12) = 2$, and, similarly, $f(2) = 2/3 = f(3)$ and $f(4) = 1 = f(5)$.



And now the Metropolis formula for P shows that to modify the random walk to have the desired stationary distribution, we run the process depicted in the figure below.



These transition probabilities were calculated as follows. For example,

$$\begin{aligned}
P(1,2) &= \frac{1}{2}\min\left\{1, \frac{2/3}{2}\right\} = \frac{1}{6} = P(1,3), \\
P(1,1) &= 1 - \frac{1}{6} - \frac{1}{6} = \frac{2}{3}, \\
P(2,1) &= \frac{1}{3}\min\left\{1, \frac{2}{2/3}\right\} = \frac{1}{3}, \\
P(2,3) &= \frac{1}{3}\min\left\{1, \frac{2/3}{2/3}\right\} = \frac{1}{3}, \\
P(2,4) &= \frac{1}{3}\min\left\{1, \frac{1}{2/3}\right\} = \frac{1}{3}, \\
P(2,2) &= 1 - \frac{1}{3} - \frac{1}{3} - \frac{1}{3} = 0,
\end{aligned}$$

and so on. □

The Metropolis method has a nice property: actually we do not even quite have to be able to write down or compute the probabilities $\pi(i)$ in order to be able to use the method to simulate from π! That is, as is clear from (2.11), to run the Metropolis chain, all we need to know are ratios of the form $f(j)/f(i)$, or, equivalently, ratios of the form $\pi(j)/\pi(i)$ $\llbracket$these are equivalent by (2.10), and because we assume we know the degrees $d(j)$ and $d(i)$$\rrbracket$. That is, we do not have to know any of the $\pi(i)$'s explicitly to use the method; all we have to know are the *ratios* $\pi(j)/\pi(i)$. Now this may not seem like a big deal, but there are cases in which we would like to simulate from a distribution π for which the ratios $\pi(j)/\pi(i)$ are easy to compute, while the individual probabilities $\pi(j)$ and $\pi(i)$ are extremely difficult to compute. This happens when π is known only up to a complicated multiplicative normalization constant. One simple example of this you have already seen: in our problem of simulating a uniformly distributed 4×4 table with given row and column sums, the desired probability of any given table is the reciprocal of the number of tables satisfying the given restrictions—a number that we do not know! (Remember, in fact, a bonus of being able to generate a nearly uniformly distributed table is that it leads to a method for approximating the number of such tables.) So in this problem, we do not know the individual probabilities of the form $\pi(i)$. But the ratio $\pi(j)/\pi(i)$ is simply 1 for a uniform distribution! Now, simulating from a uniform distribution is admittedly a special case, and a symmetric probability transition matrix will do the trick. For a more general class of examples, in the Bayesian approach to statistics, suppose the unknown parameter of interest is $\theta \in \Theta$, where Θ is a finite parameter space. Suppose our prior distribution ⟦probability mass function⟧ for θ is $\mu(\theta)$ and the likelihood is $P(x \mid \theta)$. Both $\mu(\theta)$ and $P(x \mid \theta)$ are known to us because we specify them as part of our probability model. The posterior distribution for θ given x is

$$P(\theta \mid x) = \frac{\mu(\theta)P(x \mid \theta)}{G},$$

where

$$G = \sum_{\theta' \in \Theta} \mu(\theta')P(x \mid \theta').$$

The sum G may be very difficult to compute; in statistical mechanics it is the infamous "partition function." However, for given x, if we want to simulate from the posterior distribution $P(\cdot \mid x)$, we can do so using the Metropolis method; although the distribution itself may be hard to compute because G is hard to compute, the ratios

$$\frac{P(\theta_1 \mid x)}{P(\theta_2 \mid x)} = \frac{\mu(\theta_1)P(x \mid \theta_1)}{\mu(\theta_2)P(x \mid \theta_2)}$$

are easy to compute.

2.5 Simulated annealing

Simulated annealing is a recent and powerful technique for addressing large, complicated optimization problems. Although the idea is so simple it may sound naive, the simulated annealing method has enabled people in some cases to find better answers to bigger problems than any previously known method.

Suppose we would like to minimize some "cost function" $c(\cdot)$ defined on a set $\mathcal{S}$. For example, $c(\cdot)$ might be a function of d variables defined on the simplest interesting domain $\mathcal{S}$, namely, the domain $\mathcal{S} = \{0,1\}^d$, in which each of the d variables may take on only the two values 0 and 1. That is, this $\mathcal{S}$ is the set of 2^d d-tuples $i = (i_1, \ldots, i_d)$; we could think of these as the vertices of the d-dimensional unit cube. So for $d = 10$ variables, $\mathcal{S}$ contains $2^{10} \approx 1000$ points. If we want to solve a problem with $d = 20$, 30, or 40 variables, the number of points in $\mathcal{S}$ rises to about one million, one billion, and one trillion, respectively. Have our computers gotten fast enough that we can just about handle a trillion points now? Well, if we then just add 20 more variables to the problem, all of a sudden our computers are a million times too slow again. So even though computers are getting faster all the time, clearly our appetite for solving larger and more complex problems grows much, much faster. "But come now, who really deals with functions of 60 or 100 variables?" you may be wondering. Well, consider, for example, an image processing problem, in which we want to calculate an optimal guess at an image, given a noisy, corrupted version. If we are dealing with a black and white image, so that each pixel can be encoded as a 0 or a 1, then our state space is exactly of the form $\{0,1\}^d$ that we have been discussing. How big are typical values of d? Very big: d is the number of pixels in the image, so if we have a (quite crude!) 200×200 image, then $d = 40,000$. This is *much* bigger than 60 or 100! I hope this gives you some inkling of the inconceivable vastness of many quite ordinary combinatorial optimization problems, and a feeling for the utter hopelessness of *ever* solving such problems by exhaustive enumerative search of the state space.

Stuart and Donald Geman used simulated annealing in their approach to image restoration. Another famous example of a difficult optimization problem is the traveling salesman problem in which a salesman is given a set of cities he must visit, and he wants to decide which city to visit first, second, and so on, in such a way that his total travel distance is minimized. For us, in addition to its practical importance, simulated annealing provides a nice illustration of some of the Markov chain ideas we have discussed so far, as well as an excuse to learn something about time-inhomogeneous Markov chains.

2.5.1 Description of the method

The method is a combination of the familiar idea of using Markov chains for simulation and a new idea ("annealing") that provides the connection to optimization. We have already discussed the very general Metropolis method that allows us to simulate approximately from any desired distribution on $\mathcal{S}$. But what does simulation have to do with optimization or "annealing" or whatever?

Our goal is to find an $i \in \mathcal{S}$ minimizing $c(i)$, where c is a given cost function defined on the set of nodes $\mathcal{S}$ of a graph. As discussed above, an exact solution of this problem may be an unattainable ideal in practice, but we would like to come as close as possible. For each $T > 0$, define a probability distribution $\alpha_T = \{\alpha_T(i) : i \in \mathcal{S}\}$ on $\mathcal{S}$ by

$$\alpha_T(i) = \frac{d(i)e^{-c(i)/T}}{G(T)}, \tag{2.16}$$

where again $d(i)$ is the degree of node i and of course

$$G(T) = \sum_{i \in \mathcal{S}} d(i)e^{-c(i)/T}$$

is just the normalizing constant that makes (2.16) a probability distribution. The letter "T" stands for "temperature."

We have defined a family of probability distributions on $\mathcal{S}$; corresponding to each positive T there is a distribution α_T. These distributions have an important property that explains why we are interested in them. To state this property, let $\mathcal{S}^*$ denote the set of global minimizers of $c(\cdot)$, that is,

$$\mathcal{S}^* = \{i^* \in \mathcal{S} : c(i^*) = \min_{i \in \mathcal{S}} c(i)\}.$$

Our goal is to find an element of $\mathcal{S}^*$. Define a distribution α^* on $\mathcal{S}^*$ by

$$\alpha^*(i) = \frac{d(i)}{\sum_{j\in\mathcal{S}^*} d(j)} \tag{2.17}$$

if $i \in \mathcal{S}^*$, and $\alpha^*(i) = 0$ otherwise. The important thing to keep in mind about the distribution α^* is that it puts positive probability only on globally optimal solutions of our optimization problem.

(2.18) Fact. *As $T \downarrow 0$, we have $\alpha_T \xrightarrow{\mathcal{D}} \alpha^*$, that is, $\alpha_T(i) \to \alpha^*(i)$ for all $i \in \mathcal{S}$.*

The symbol "$\xrightarrow{\mathcal{D}}$" stands for convergence in distribution.

▷ *Exercise [2.16] asks you to prove (2.18).*

Thus, we have found that, as $T \downarrow 0$, the distributions α_T converge to the special distribution α^*. If we could somehow simulate from α^*, our optimization problem would be solved: We would just push the $\boxed{\text{simulate from } \alpha^*}$ button, and out would pop a random element of $\mathcal{S}^*$, which would make us most happy. Of course, we cannot do that, since we are assuming that we do not already have the answer to the minimization problem that we are trying to solve! However, we *can* do something that seems as if it should be nearly as good: simulate from α_T. If we do this for a value of T that is pretty close to 0, then since α_T is pretty close to α^* for that T, presumably we would be doing something pretty good.

So, fix a $T > 0$. How do we simulate from α_T? We can use the Metropolis idea to create a probability transition matrix $A_T = (A_T(i,j))$ such that $\alpha_T A_T = \alpha_T$, and then run a Markov chain according to A_T.

⟦A note on notation: I hope you aren't bothered by the use here of α and A for stationary distributions and probability transition matrices related to simulated annealing. The usual letters π and P have been so overworked that using different notation for the special example of simulated annealing should be clearer ultimately. Although they don't look much like π and P, the letters α and A might be remembered as mnemonic for "simulαted Annealing" at least.⟧

A glance at the definition of α_T in (2.16) shows that we are in the situation of the Metropolis method as described in (2.10) with the choice $f(i) = e^{-c(i)/T}$. So as prescribed by (2.11), for $j \in \mathcal{N}(i)$ we take

$$\begin{aligned} A_T(i,j) &= \frac{1}{d(i)} \min\left\{1, \frac{e^{-c(j)/T}}{e^{-c(i)/T}}\right\} \\ &= \frac{1}{d(i)} \begin{cases} 1 & \text{if } c(j) \le c(i) \\ e^{-(c(j)-c(i))/T} & \text{if } c(j) > c(i). \end{cases} \end{aligned} \tag{2.19}$$

The specification of A_T is completed by the obvious requirements that $A_T(i,i) = 1 - \sum_{j\in\mathcal{N}(i)} A_T(i,j)$ and $A_T(i,j) = 0$ if $j \notin \mathcal{N}(i)$.

For any fixed temperature $T_1 > 0$, if we run a Markov chain $\{X_n\}$ having probability transition matrix A_{T_1} for "a long time," then the distribution of $\{X_n\}$ will be very close to

α_{T_1}. If our goal is to get the chain close to the distribution α^*, then continuing to run this chain will not do much good, since the distribution of X_n will get closer and closer to α_{T_1}, not α^*. So the only way we can continue to "make progress" toward the distribution α^* is to decrease the temperature to T_2, say, where $T_2 < T_1$, and continue to run the chain, but now using the probability transition matrix A_{T_2} rather than A_{T_1}. Then the distribution of X_n will approach the distribution α_{T_2}. Again, after the distribution gets very close to α_{T_2}, continuing to run the chain at temperature T_2 will not be an effective way to get closer to the desired distribution α^*, so it makes sense to lower the temperature again.

It should be quite clear intuitively that, if we lower the temperature slowly enough, we can get the chain to converge in distribution to α^*. For example, consider "piecewise constant" schedules as discussed in the last paragraph. Given a decreasing sequence of positive temperatures $T_1 > T_2 > \ldots$ such that $T_n \downarrow 0$, we could start by running a chain for a time n_1 long enough so that the distribution of the chain at time n_1 is within a distance of 1/2 ⟦in total variation distance, say⟧ from α_{T_1}. Then we could continue to run at temperature T_2 until time n_2, at which the distribution of the chain is within 1/3 of α_{T_2}. Then we could run at temperature T_3 until we are within 1/4 of of α_{T_3}. And so on. Thus, as long as we run long enough at each temperature, the chain should converge in distribution to α^*. We must lower the temperature slowly enough so that the chain can always "catch up" and remain close to the stationary distribution for the current temperature.

Once we have had this idea, we need not lower the temperature in such a piecewise constant manner, keeping the temperature constant for many iterations and only then changing it. Instead, let us allow the temperature to change at each at each step of the Markov chain. Thus, each time n may have its own associated temperature T_n, and hence its own probability transition matrix A_{T_n}. The main theoretical result that has been obtained for simulated annealing says that for any problem there is a cooling schedule of the form

$$T_n = \frac{a}{\log(n+b)} \tag{2.20}$$

⟦where a and b are constants⟧ such that, starting from any state $i \in \mathcal{S}$, the ⟦time-inhomogeneous⟧ Markov chain $\{X_n\}$ will converge in distribution to α^*, the uniform distribution on the set of global minima.

Accordingly, a simulated annealing procedure may be specified as follows. Choose a "cooling schedule" $T_0, T_1, \ldots$; the schedules we will discuss later will have the property that $T_n \downarrow 0$ as $n \to \infty$. Choose the inital state X_0 according to a distribution ν_0. Let the succession of states $X_0, X_1, X_2, \ldots$ form a time-inhomogeneous Markov chain with probability transition matrices $A_{T_0}, A_{T_1}, A_{T_2}, \ldots$, so that

$$\mathbb{P}\{X_{n+1} = j \mid X_n = i\} = A_{T_n}(i,j)$$

and

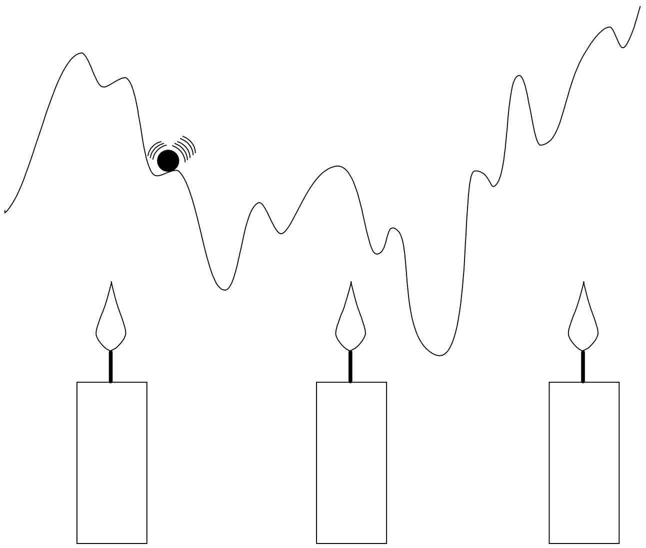
$$X_n \sim \nu_n = \nu_0 A_{T_0} A_{T_1} \cdots A_{T_{n-1}}. \tag{2.21}$$

By the way, what is all this talk about "temperature," "cooling schedule," "annealing," and stuff like that? I recommend you consult the article by Kirkpatrick et al., but I'll try to say a few words here. Let's see, I'll start by looking up the word "anneal" in my dictionary. It gives the definition

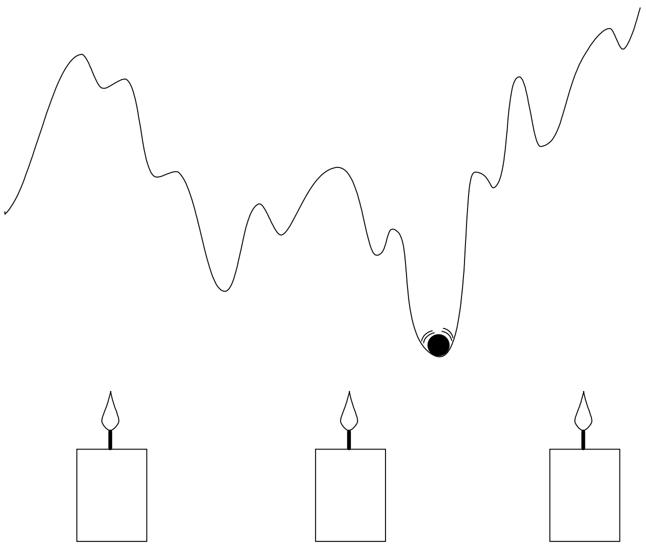
> To free (glass, metals, etc.) from internal stress by heating and gradually cooling.

The idea, as I understand it, is this. Suppose we want to have a metal that is as "stress-free" as possible. That corresponds to the metal being "relaxed," in a low energy state. The metal is happy when it has low energy, and it wants to get to such a state, just as a ball will roll downhill to decrease its potential energy. In the liquid state, the atoms of a metal are all sliding around in all sorts of random orientations. If we quickly freeze the metal into a solid by cooling it quickly, its atoms will freeze into a configuration that has all sorts of haphazard orientations that give the metal unnecessary excess potential energy. In contrast, if we start with the metal as a liquid and then cool it extremely slowly, then the atoms have plenty of time to explore around, find, and work themselves into a nice, happy, ordered, low-energy configuration. Thus, nature can successfully address an optimization problem—minimization of the energy of the metal—by slow cooling. I hope that gives the flavor of the idea at least.

For another image that captures certain aspects of simulated annealing, imagine a bumpy frying pan containing a marble made of popcorn. At high temperature, the marble is jumping around quite wildly, ignoring the ups and downs of the pan.



As we lower the temperature, the popcorn begins to settle down into the lower parts of the pan.

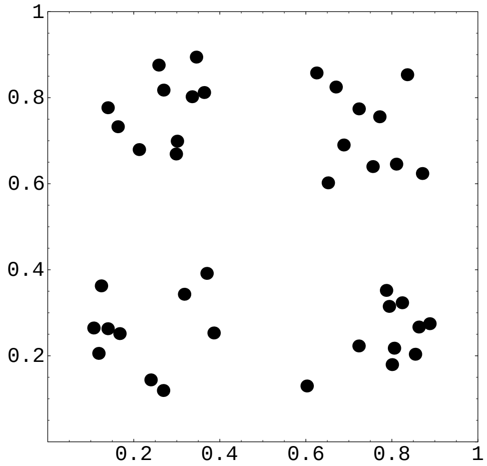


I'd like to make sure it is clear to you how to run the method, so that you will be able to implement it yourself and experiment with it if you'd like. So let us pause for a moment to imagine what it is like to move around according to a Markov chain having probability transition matrix A_T specified by (2.19). Here is one way to describe it. Suppose that we are now at time n sitting at node i of the graph, and we are trying to decide where to go next. First we choose at random one of the d neighbors of i, each neighbor being chosen with the same probability $1/d(i)$. Say we choose node j. Then j becomes out "candidate" for where to go at time $n+1$; in fact, at time $n+1$ we will either move to j (accept the candidate) or stay at i. To decide between these two alternatives, we must look at the values of $c(i)$ and $c(j)$. If $c(j) \leq c(i)$, we definitely move to j. If $c(j) > c(i)$, we might or might not move to j; in fact we move to j with probability $e^{-(c(j)-c(i))/T}$ and stay at i with the complementary probability $1 - e^{-(c(j)-c(i))/T}$. At high temperature, even when $c(j) > c(i)$, the probability $e^{-(c(j)-c(i))/T}$ is close to 1, so that we accept all candidates with high probability. Thus, at high temperature, the process behaves nearly like a random walk on $\mathcal{S}$, choosing candidates as a random walk would, and almost always accepting them. At lower temperatures, the process still always accepts "downhill moves" $[\![$those with $c(j) \leq c(i)]\!]$, but has a lower probability of accepting uphill moves. At temperatures very close to zero, the process very seldom moves uphill.

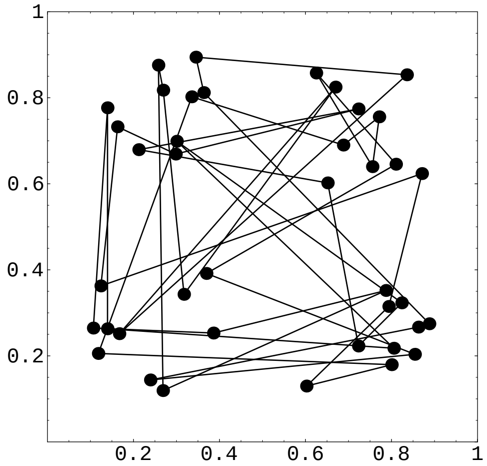
Note that when $c(j) \leq c(i)$, moving to j "makes sense"; since we are trying to minimize $c(\cdot)$, decreasing c looks like progress. However, simulated annealing is not just another "descent method," since we allow ourselves positive probability of taking steps that *increase* the value of c. This feature of the procedure prevents it from getting stuck in local minima.

You may have asked yourself the question, "If we want to bring the temperature down to 0, then why don't we just run the chain at temperature 0?" But now you can see that at temperature $T = 0$, the process will never make an uphill move. Thus, running at temperature 0 is a descent method, which will get stuck in local minima, and therefore will not approach approach global minima. At temperature 0, the chain all of a sudden loses the nice properties it has at positive temperatures—it is no longer irreducible, and so the basic limit theorem doesn't apply.

(2.22) Example [Traveling salesman problem]. The figure below shows the location of 40 cities. A traveling salesman who lives in one of the cities wants to plan a tour in such a way that he visits all of the cities and then returns home while traveling the shortest possible total distance.



To solve this problem by simulated annealing, we start with some legal tour. We'll start in a very dumb way, with a tour that is simply a random permutation of the 40 cities.

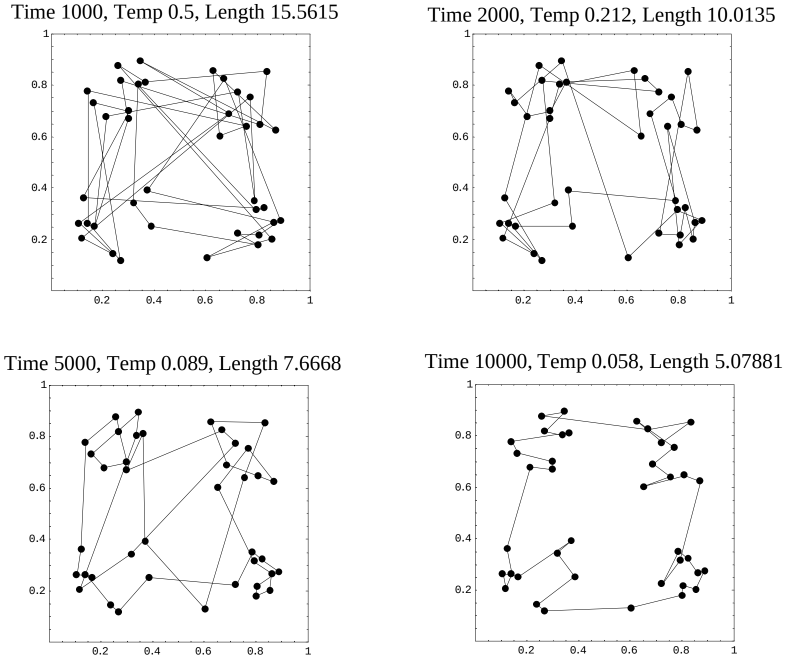


As we know, simulated annealing gradually modifies and, we hope, improves the tour by proposing random modifications of the tour, and accepting or rejecting those modifications randomly based on the current value of the temperature parameter and the amount of improvement or deterioration produced by the proposed modification.

Here our state space $\mathcal{S}$ is the set of all possible traveling salesman tours. Each such tour can be specified by a permutation of the cities; for example, if there are 8 cities, the tour $(1, 3, 5, 7, 2, 4, 6, 8)$ means that we start at city 1, then go to 3, 5, 7, 2, 4, 6, and 8 in that order, and finally return from 8 back to 1. ⟦Note that since the tours are closed circuits because the salesman returns to his starting point, we could consider, for example, the tour $(3, 5, 7, 2, 4, 6, 8, 1)$ to be the same tour as $(1, 3, 5, 7, 2, 4, 6, 8)$. So in this way different permutations can represent the same tour. We could say these two equivalent permutations are “rotations” of each other.⟧

Recall that the Metropolis method, and simulated annealing, require us to choose a neighborhood structure on $\mathcal{S}$. That is, for each tour i in $\mathcal{S}$, we must say which other tours are the neighbors of i, and then the simulated annealing method chooses a random candidate neighbor at each iteration. There are many possible ways to do this. One way that seems to move around $\mathcal{S}$ nicely and is also easy to work with is to choose two cities on the tour randomly, and then reverse the portion of the tour that lies between them. For example, if we are currently at the tour $(1, 2, 3, 4, 5, 6, 7, 8)$, we might choose the two cities 4 and 6, and change from the tour $(1, 2, 3, \underbrace{4, 5, 6}, 7, 8)$ to the neighboring tour $(1, 2, 3, \underbrace{6, 5, 4}, 7, 8)$. Another convenient feature of this definition of neighbors is that it is easy to calculate the difference in total lengths between neighboring tours; no matter how many cities we are considering, the *difference* in tour lengths is calculated using only four intercity distances — the edges that are broken and created as we break out part of the tour and then reattach it in the reversed direction.

Using the type of moves just described, it was quite easy to write a computer program, and fun to watch it run. Here is how it went:



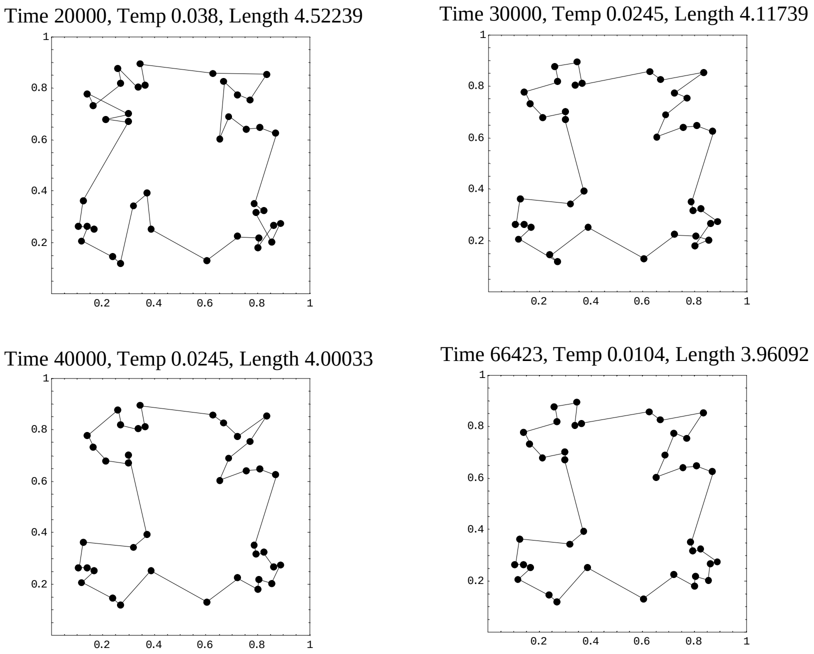
Time 1000, Temp 0.5, Length 15.5615

Time 2000, Temp 0.212, Length 10.0135

Time 5000, Temp 0.089, Length 7.6668

Time 10000, Temp 0.058, Length 5.07881

If we were watching a movie of this, we would see the path slithering around and gradually untangling itself. The slithering is violent at high temperatures. It starts to get the gross features right, removing those silly long crossings from one corner of the picture to another. And now let’s watch it continue.


Time 20000, Temp 0.038, Length 4.52239

Time 30000, Temp 0.0245, Length 4.11739

Time 40000, Temp 0.0245, Length 4.00033

Time 66423, Temp 0.0104, Length 3.96092

Our salesman ends up with a very nice tour. I checked to make sure this tour is a local optimum — no neighboring tour is better. Maybe it's even the globally optimal tour. □

Doesn't this method of random searching seem extraordinarily dumb? No particular knowledge or strategic ideas about the particular problem are built in. The Markov property means that there is no memory; any successes or failures from the past are forgotten. But then again, evolution is "dumb" in the same way, accumulating random changes, gently guided by natural selection, which makes favorable changes more likely to persist. And look at all of the marvelous solutions evolution has produced.

2.5.2 The Main Theorem

First some notation. Define the matrix $A^{(n)}$ by

$$A^{(n)} = \prod_{k=0}^{n-1} A_{T_k}.$$

Recalling that ν_n denotes the distribution of X_n, (2.21) says that $\nu_n = \nu_0 A^{(n)}$. Let $A^{(n)}(i, \cdot)$ denote the ith row of the matrix $A^{(n)}$. This row may be interpreted as the distribution of X_n given that $X_0 = i$. We will continue using the notation $\|\mu - \nu\|$ for the total variation distance

$$\|\mu - \nu\| = \sup_{S \subset \mathcal{S}} |\mu(S) - \nu(S)| = \frac{1}{2} \sum_{i \in \mathcal{S}} |\mu_i - \nu_i|$$

between the distributions μ and ν.

We also need some definitions related to the structure of the graph $\mathcal{G}$ and the cost function $c(\cdot)$. For two nodes $i, j \in \mathcal{G}$, define the *distance* $\rho(i,j)$ to be the length ⟦i.e., number of edges⟧ of a path from i to j of minimal length. Define the *radius* r of $\mathcal{G}$ by

$$r = \min_{i \in \mathcal{S}_c} \max_{j \in \mathcal{S}} \rho(i,j),$$

where

$$\mathcal{S}_c = \{i \in \mathcal{S} : c(j) > c(i) \text{ for some } j \in \mathcal{N}(i)\},$$

that is, $\mathcal{S}_c$ is the set of all nodes that are not local maxima of the function c. ⟦Note that r actually depends on the function c in addition to the graph $\mathcal{G}$, but we won't indicate this in the notation.⟧ Define the number L to be the largest "local fluctuation" of $c(\cdot)$, that is,

$$L = \max_{i \in \mathcal{S}} \max_{j \in \mathcal{N}(i)} |c(j) - c(i)|.$$

Finally, recall the definition of α^* from (2.17).

After only 38 million definitions (counting multiplicity), we are ready to state the main theorem.

(2.23) Theorem [Main Theorem]. *For any cooling schedule $T_0, T_1, \ldots$ satisfying*

(i) $T_{n+1} < T_n$ *for all* $n \ge 0$

(ii) $T_n \to 0$ *as* $n \to \infty$

(iii) $\sum_k \exp(-rL/T_{kr-1}) = \infty$,

we have

$$\|A^{(n)}(i,\cdot) - \alpha^*\| \to 0$$

as $n \to \infty$, for all $i \in \mathcal{S}$.

The proof of this theorem is a rather easy application of some theory of time-inhomogeneous Markov chains. We'll develop this theory in the next section, and then come back to simulated annealing to prove Theorem (2.23).

Here is a specific family of cooling schedules that satisfy the conditions of the theorem: taking $\gamma \ge rL$, let

$$T_n = \frac{\gamma}{\log(n)}$$

for $n > 1$. ⟦And let T_0 and T_1 be arbitrary, subject to the monotonicity condition (i).⟧ It is easy to check that the conditions of the theorem hold; (iii) boils down to the statement that $\sum_k k^{-p} = \infty$ if $p \le 1$.

Note that the "convergence in distribution" sort of conclusion of the theorem is weaker than almost sure ("a.s.") convergence ⟦i.e., convergence with probability 1⟧. There is good reason for this: a.s. convergence indeed does not hold in general for simulated annealing. Intuitively, we should not expect to be able to get a.s. convergence—if our procedure has the nice property of always escaping eventually from local minima, then we must live with the fact that it will also eventually "escape" from global minima. That is, the process will not get "stuck" in ⟦that is, converge to⟧ a global minimum. It is true that, along a typical sample path, the process will spend a larger and larger fraction of its time at or near global minima as n increases. However, it will also make infinitely many (although increasingly rare) excursions away from global minima.

▷ *I would say that the theorem is certainly elegant and instructive. But if you come away thinking it is only a beginning and is not yet answering the "real" question, I would consider that a good sign. Exercise [2.15] asks for your thoughts on the matter.*

2.6 Ergodicity Concepts for Time-Inhomogeneous Markov chains

In this section we will work in the general setting of a time-inhomogeneous Markov chain on a countably infinite state space $\mathcal{S}$; we will revert to finite $\mathcal{S}$ when we return to the specific problem of simulated annealing.

(2.24) Notation. *For a time-inhomogeneous Markov chain $\{X_n\}$, let P_n denote the probability transition matrix governing the transition from X_n to X_{n+1}, that is, $P_n(i,j) = \mathbb{P}\{X_{n+1} = j \mid X_n = i\}$. Also, for $m < n$ define $P^{(m,n)} = \prod_{k=m}^{n-1} P_k$, so that $P^{(m,n)}(i,j) = \mathbb{P}\{X_n = j \mid X_m = i\}$.*

(2.25) Definition. $\{X_n\}$ *is* strongly ergodic *if there exists a probability distribution π^* on $\mathcal{S}$ such that*

$$\lim_{n\to\infty} \sup_{i\in\mathcal{S}} \|P^{(m,n)}(i,\cdot) - \pi^*\| = 0 \quad \textit{for all } m.$$

For example, we will prove Theorem (2.23) by showing that under the stated conditions, the simulated annealing chain $\{X_n\}$ is strongly ergodic, with limiting distribution α^* as given in (2.17).

The reason for the modifier "strongly" is to distinguish the last concept from the following weaker one.

(2.26) Definition. $\{X_n\}$ *is* weakly ergodic *if*

$$\lim_{n\to\infty} \sup_{i,j\in\mathcal{S}} \|P^{(m,n)}(i,\cdot) - P^{(m,n)}(j,\cdot)\| = 0 \quad \textit{for all } m.$$

To interpret these definitions a bit, notice that weak ergodicity is a sort of "loss of memory" concept. It says that at a large enough time n, the chain has nearly "forgotten" its state at time m, in the sense that the distribution at time n would be nearly the same no matter what the state was at time m. However, there is no requirement that the distribution be converging to anything as $n \to \infty$. The concept that incorporates convergence in addition to loss of memory is strong ergodicity.

What is the role of the "for all m" requirement? Why not just use $\lim_{n\to\infty}\sup_{i\in\mathcal{S}}\|P^{(0,n)}(i,\cdot)-\pi^*\|=0$ for strong ergodicity and $\lim_{n\to\infty}\sup_{i,j\in\mathcal{S}}\|P^{(0,n)}(i,\cdot)-P^{(0,n)}(j,\cdot)\|=0$ for weak ergodicity? Here are a couple of examples to show that these would not be desirable definitions. Let $\mathcal{S}=\{1,2\}$ and $P_0=\begin{pmatrix}1/4 & 3/4\\ 1/4 & 3/4\end{pmatrix}$. Then with these definitions, $\{X_n\}$ would be strongly ergodic even if $P_k=I$ for all $k\ge 1$ and $\{X_n\}$ would be weakly ergodic for *any* sequence of probability transition matrices $P_1,P_2,\ldots$. This seems silly. We want more "robust" concepts that cannot be determined just by one matrix P_0, but rather depend on the whole sequence of transition matrices.

Incidentally, since our goal with respect to simulated annealing is the main theorem (2.23) above, we are really interested in proving strong ergodicity. However, we will find weak ergodicity to be a useful stepping stone on the way toward that goal.

2.6.1 The Ergodic Coefficient

This will be a useful quantity in formulating sufficient conditions for weak and strong ergodicity. For a probability transition matrix $P=(P(i,j))$, the *ergodic coefficient* $\delta(P)$ of P is defined to be the maximum total variation distance between pairs of rows of P, that is,

$$\begin{aligned}\delta(P)&=\sup_{i,j\in\mathcal{S}}\|P(i,\cdot)-P(j,\cdot)\|\\&=\frac{1}{2}\sup_{i,j\in\mathcal{S}}\sum_{k\in\mathcal{S}}|P(i,k)-P(j,k)|\\&=\sup_{i,j\in\mathcal{S}}\sum_{k\in\mathcal{S}}(P(i,k)-P(j,k))^+.\end{aligned}$$

The basic idea here is that $\delta(P)$ being small is "good" for ergodicity. For example, the extreme case is $\delta(P)=0$, in which case all of the rows of P are identical, and so P would cause a Markov chain to lose its memory completely in just one step: $\nu_1=\nu_0P$ does not depend on ν_0.

Here is a useful lemma.

(2.27) Lemma. *$\delta(PQ)\le\delta(P)\delta(Q)$ for probability transition matrices P, Q.*

Proof: By definition, $\delta(PQ)=\sup_{i,j\in\mathcal{S}}\sum_{k\in\mathcal{S}}[(PQ)_{ik}-(PQ)_{jk}]^+$, where here and throughout this proof, for readability let us use subscripts to denote matrix entries. Fix a pair of states i, j, and let $A = \{k \in \mathcal{S} : (PQ)_{ik} > (PQ)_{jk}\}$. Then

$$\begin{aligned}
\sum_{k\in\mathcal{S}}[(PQ)_{ik} - (PQ)_{jk}]^+ &= \sum_{k\in A}[(PQ)_{ik} - (PQ)_{jk}] \\
&= \sum_{k\in A}\sum_{l\in\mathcal{S}}[P_{il}Q_{lk} - P_{jl}Q_{lk}] \\
&= \sum_{l\in\mathcal{S}}[P_{il} - P_{jl}]\sum_{k\in A}Q_{lk} \\
&\le \sum_{l\in\mathcal{S}}(P_{il} - P_{jl})^+\Big[\sup_l\sum_{k\in A}Q_{lk}\Big] - \sum_{l\in\mathcal{S}}(P_{il} - P_{jl})^-\Big[\inf_l\sum_{k\in A}Q_{lk}\Big] \\
&= \sum_{l\in\mathcal{S}}(P_{il} - P_{jl})^+\Big[\sup_l\sum_{k\in A}Q_{lk} - \inf_l\sum_{k\in A}Q_{lk}\Big] \\
&\le \Big[\sum_{l\in\mathcal{S}}(P_{il} - P_{jl})^+\Big]\delta(Q) \\
&\le \delta(P)\delta(Q),
\end{aligned}$$

where the last equality uses the fact that $\sum_l(P_{il} - P_{jl})^+ = \sum_l(P_{il} - P_{jl})^-$, which holds because P is a probability transition matrix. Since i and j were arbitrary in $\mathcal{S}$, we are done.
□

▷ *Exercises [2.17] and [2.18] lead to an alternative proof of Lemma (2.27) using coupling.*

2.6.2 Sufficient Conditions for Weak and Strong Ergodicity

Sufficient conditions are given in the next two propositions.

(2.28) PROPOSITION. *If there exist $n_0 < n_1 < n_2 < \cdots$ such that $\sum_k[1 - \delta(P^{(n_k,n_{k+1})})] = \infty$, then $\{X_n\}$ is weakly ergodic.*

(2.29) PROPOSITION. *If $\{X_n\}$ is weakly ergodic and if there exist $\pi_0, \pi_1, \dots$ such that π_n is a stationary distribution for P_n for all n and $\sum_n \|\pi_n - \pi_{n+1}\| < \infty$, then $\{X_n\}$ is strongly ergodic. In that case, the distribution π^* in the definition (2.25) is given by $\pi^* = \lim_{n\to\infty}\pi_n$.*

Recall that strong ergodicity is like weak ergodicity ⟦loss of memory⟧ together with convergence. The extra condition $\sum_n \|\pi_n - \pi_{n+1}\| < \infty$ is giving this convergence in (2.29).

PROOF OF PROPOSITION (2.28). It follows directly from the definitions we have given that weak ergodicity is equivalent to the condition that $\lim_{n\to\infty}\delta(P^{(m,n)}) = 0$ for all m. By assumption, $\sum_{k\ge K}[1 - \delta(P^{(n_k,n_{k+1})})] = \infty$ for all K. We will use the following little fact about real numbers: if $0 \le a_n \le 1$ for all n and $\sum_k a_k = \infty$, then $\prod_k(1 - a_k) = 0$. ⟦Proof: under the assumed conditions, $0 \le \prod(1 - a_k) \le \prod e^{-a_k} = e^{-\sum a_k} = 0$.⟧ From this we obtain $\prod_{k\geq K}\delta(P^{(n_k,n_{k+1})}) = 0$ for all K. That is, $\lim_{L\to\infty}\prod_{k=K}^{L-1}\delta(P^{(n_k,n_{k+1})}) = 0$ for all K. However, from Lemma (2.27),

$$\delta(P^{(n_K,n_L)}) \leq \prod_{k=K}^{L-1}\delta(P^{(n_k,n_{k+1})}).$$

Therefore, $\lim_{L\to\infty}\delta(P^{(n_K,n_L)}) = 0$ for all K. Clearly this implies that $\lim_{n\to\infty}\delta(P^{(m,n)}) = 0$ for all m. □

Proof of Proposition (2.29). Suppose that $\{X_n\}$ is weakly ergodic and $\sum\|\pi_n - \pi_{n+1}\| < \infty$, where each π_n is stationary for P_n. Let $\pi^* = \lim \pi_n$; clearly the limit exists by the assumption $\sum\|\pi_n - \pi_{n+1}\| < \infty$. Let k be fixed. Then for any $l > k$ and $m > l$ we have

$$\begin{aligned}\|P^{(k,m)}(i,\cdot) - \pi^*\| &\leq \|P^{(k,m)}(i,\cdot) - \pi_l P^{(l,m)}\| \\ &\quad + \|\pi_l P^{(l,m)} - \pi_m\| + \|\pi_m - \pi^*\|.\end{aligned} \tag{2.30}$$

Let $\epsilon > 0$. We will show that the right-hand side can be made less than ϵ if m is large enough; we'll do this by making a judicious choice of l. The last term is the simplest; clearly there exists M_3 such that $\|\pi_m - \pi^*\| \leq \epsilon/3$ for all $m \geq M_3$. For the second term, note that

$$\begin{aligned}\pi_l P^{(l,m)} &= \pi_l P_l P^{(l+1,m)} = \pi_l P^{(l+1,m)} \\ &= \pi_{l+1}P^{(l+1,m)} + [\pi_l - \pi_{l+1}]P^{(l+1,m)},\end{aligned}$$

so that

$$\pi_l P^{(l,m)} - \pi_m = [\pi_{l+1}P^{(l+1,m)} - \pi_m] + [\pi_l - \pi_{l+1}]P^{(l+1,m)}.$$

Applying this relation recursively gives

$$\pi_l P^{(l,m)} - \pi_m = \sum_{n=l}^{m-1}[\pi_n - \pi_{n+1}]P^{(n+1,m)}$$

So $\|\pi_l P^{(l,m)} - \pi_m\| \leq \sum_{n=l}^{m-1}\|\pi_n - \pi_{n+1}\|$. ⟦Why? **Exercise.**⟧ Therefore, since $\sum\|\pi_n - \pi_{n+1}\| < \infty$, we can make $\|\pi_l P^{(l,m)} - \pi_m\| \leq \epsilon/3$ by taking m and l large enough, say, $m \geq l \geq L_2$.

Finally, for the first term on the right-hand side of (2.30), note that

$$\|P^{(k,m)}(i,\cdot) - \pi_l P^{(l,m)}\| = \|[P^{(k,l)}(i,\cdot) - \pi_l]P^{(l,m)}\|.$$

However, for any given l, we can make the last expression less than $\epsilon/3$ by taking m large enough—by weak ergodicity, at a large enough time m, the chain doesn't "remember" whether its distribution at time l was $P^{(k,l)}(i,\cdot)$ or π_l! So, for all l, there is an $M_1(l)$ such that if $m \geq M_1(l)$ then $\|P^{(k,m)}(i,\cdot) - \pi_l P^{(l,m)}\| < \epsilon/3$. So we are done: if $m \geq \max\{M_3, M_1(L_2)\}$, then $\sup_i\|P^{(k,m)}(i,\cdot) - \pi^*\| \leq \epsilon$. □

Notice how the hypotheses of the last result were used to get the terms on the right-hand side of (2.30) small: weak ergodicity took care of the first term, and $\sum\|\pi_n - \pi_{n+1}\| < \infty$ took care of the other two.

2.7 Proof of the Main Theorem of Simulated Annealing

We will show that if conditions (i)–(iii) of Theorem (2.23) hold, then the time-inhomogeneous Markov chain $\{X_n\}$ for simulated annealing satisfies the sufficient conditions for strong ergodicity.

We have used two sets of parallel notation in the last two sections; one (involving α's and A's) that we introduced specifically to discuss simulated annealing and one (involving π's and P's) used in the general theory of time-inhomogeneous Markov chains. Let's relate them and make sure there is no confusion. The general notation P_n refers to the probability transition matrix used in going from X_n to X_{n+1}. In the case of simulated annealing, since the chain is operating at temperature T_n during that transition, we have $P_n = A_{T_n}$. Similarly, π_n, the stationary distribution associated with P_n, is α_{T_n} for simulated annealing. We'll also continue to use the notation $P^{(m,n)} = \prod_{k=m}^{n-1} P_k = \prod_{k=m}^{n-1} A_{T_k}$.

To establish weak ergodicity, we want to show that the sufficient condition $\sum_k [1 - \delta(P^{(n_k,n_{k+1})})] = \infty$ holds for some sequence $\{n_k\}$. In fact, we will show that the condition holds for the sequence $n_k = kr$, that is,

$$\sum_k [1 - \delta(P^{(kr-r,kr)})] = \infty. \tag{2.31}$$

We will do this by finding an upper bound for $\delta(P^{(kr-r,kr)})$ that will guarantee that $\delta(P^{(kr-r,kr)})$ does not get too close to 1.

Recall the definition of the radius $r = \min_{i \in \mathcal{S}_c} \max_{j \in \mathcal{S}} \rho(i,j)$. Let i_0 denote a *center* of the graph, that is, a node at which the minimum in the definition of r is assumed. Also recall the definition $L = \max_{i \in \mathcal{S}} \max_{j \in \mathcal{N}(i)} |c(j) - c(i)|$.

(2.32) Claim. *For all sufficiently large m we have*

$$P^{(m-r,m)}(i, i_0) \geq D^{-r} \exp(-rL/T_{m-1}) \quad \textit{for all } i \in \mathcal{S},$$

where $D = \max_j d(j)$.

Proof: It follows immediately from the definitions of $P_{ij}(n)$, D, and L that for all n, $i \in \mathcal{S}$, and $j \in \mathcal{N}(i)$, we have

$$P_{ij}(n) = \frac{1}{d(i)} \min\{1, e^{-(c(j)-c(i))/T_n}\} \geq D^{-1} e^{-L/T_n}.$$

Since we did not allow i_0 to be a local maximum of $c(\cdot)$, we must have $\{j \in \mathcal{N}(i_0) : c(j) > c(i_0)\}$ nonempty, so that as $n \to \infty$,

$$P_{i_0,i_0}(n) = 1 - \sum_{j \in \mathcal{N}(i_0)} P_{i_0 j}(n) \to 1 - \sum_{\substack{j \in \mathcal{N}(i_0) \\ c(j) \leq c(i_0)}} \frac{1}{d(i)} = \sum_{\substack{j \in \mathcal{N}(i_0) \\ c(j) > c(i_0)}} \frac{1}{d(i)} > 0.$$

Therefore, $P_{i_0,i_0}(n) \geq D^{-1} \exp(-L/T_n)$ clearly holds for large enough n.

Let $i \in \mathcal{S}$, and consider $P^{(m-r,m)}(i, i_0) = \mathbb{P}\{X_m = i_0 \mid X_{m-r} = i\}$. Clearly this probability is at least the conditional probability that, starting from i at time $m - r$, the chain takes a specified shortest path from i to i_0 ⟦which is of length at most r, by the definition of r⟧, and then holds at i_0 for the remaining time until m. However, by the previous paragraph, if m is large enough, this probability is in turn bounded below by

$$\prod_{n=m-r}^{m-1} D^{-1}e^{-L/T_n} \geq D^{-r}e^{-rL/T_{m-1}},$$

where we have used the assumption that T_n is decreasing in n here. Thus, $P^{(m-r,m)}(i,i_0) \geq D^{-r}e^{-rL/T_{m-1}}$ for large enough m. □

Taking $m = kr$ in the last claim, we get for all sufficiently large k that

$$P^{(kr-r,kr)}(i,i_0) \geq D^{-r}e^{-rL/T_{kr-1}} \quad \text{for all } i.$$

Thus, $P^{(kr-r,kr)}$ is a probability transition matrix having a column ⟦namely, column i_0⟧ all of whose entries are at least $D^{-r}\exp(-rL/T_{kr-1})$. Next we use the general observation that if a probability transition matrix Q has a column all of whose entries are at least a, then $\delta(Q) \leq 1-a$. ⟦Exercise [2.19] asks you to prove this.⟧ Therefore, for large enough k, we have $1-\delta(P^{(kr-r,kr)}) \geq D^{-r}\exp(-rL/T_{kr-1})$, which, by assumption (iii) of the main theorem, gives (2.31). This completes the proof of weak ergodicity.

Finally, we turn to the proof of strong ergodicity. Recall that the stationary distribution π_n for P_n is given by $\pi_n(i) = (G(n))^{-1}d(i)\exp[-c(i)/T_n]$. By our sufficient conditions for strong ergodicity, we will be done if we can show that $\sum \|\pi_{n+1} - \pi_n\| < \infty$. This will be easy to see from the following monotonicity properties of the stationary distributions π_n.

(2.33) Lemma. *If $i \in \mathcal{S}^*$ then $\pi_{n+1}(i) > \pi_n(i)$ for all n. If $i \notin \mathcal{S}^*$ then there exists $\tilde{n}_i$ such that $\pi_{n+1}(i) < \pi_n(i)$ for all $n \geq \tilde{n}_i$.*

Thus, as the temperature decreases, the stationary probabilities of the optimal states increase. Also, for each nonoptimal state, as the temperature decreases to 0, the stationary probability of that state decreases eventually, that is, when the temperature is low enough. The proof is calculus; just differentiate away. I'll leave this as an exercise.

From this nice, monotonic behavior, the desired result follows easily. Letting $\tilde{n}$ denote $\max\{\tilde{n}_i : i \notin \mathcal{S}^*\}$,

$$\begin{aligned}
\sum_{n=\tilde{n}}^{\infty} \|\pi_{n+1} - \pi_n\| &= \sum_{n=\tilde{n}}^{\infty}\sum_{i\in\mathcal{S}} (\pi_{n+1} - \pi_n)^+ \\
&= \sum_{n=\tilde{n}}^{\infty}\sum_{i\in\mathcal{S}^*} (\pi_{n+1} - \pi_n) \\
&= \sum_{i\in\mathcal{S}^*}\sum_{n=\tilde{n}}^{\infty} (\pi_{n+1} - \pi_n) \\
&= \sum_{i\in\mathcal{S}^*} (\pi^*(i) - \pi_{\tilde{n}}(i)) \;\leq\; \sum_{i\in\mathcal{S}^*} \pi^*(i) \;=\; 1,
\end{aligned}$$

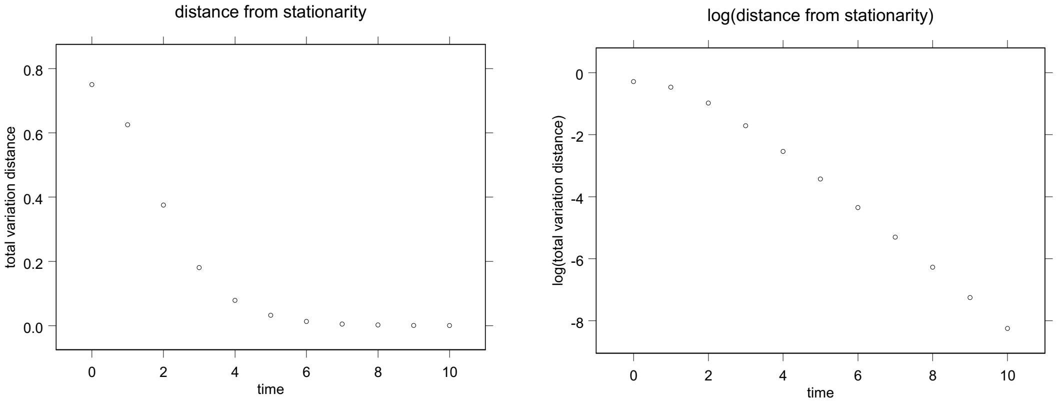
so that clearly $\sum_{n=\tilde{n}}^{\infty} \|\pi_{n+1} - \pi_n\| < \infty$. This completes the proof of strong ergodicity.

REFERENCES: This proof came from a paper of Mitra, Romeo, and Sangiovanni-Vincentelli called "Convergence and finite-time behavior of simulated annealing" [*Advances in Applied Probability*, **18**, 747–771 (1986)]. The material on time-inhomogeneous Markov chains is covered very nicely in the book *Markov chains: Theory and Applications* by Isaacson and Madsen (1976).

2.8 Card Shuffling: Speed of Convergence to Stationarity

We have seen that for an irreducible, aperiodic Markov chain $\{X_n\}$ having stationary distribution π, the distribution π_n of X_n converges to π in the total variation distance. For example, this was used in Example (1.22), which showed how to generate a nearly uniformly distributed 4×4 table having given row and column sums by simulating a certain Markov chain for a long enough time. The inevitable question is: How long is "long enough"? We could ask how close (in total variation, say) is the Markov chain to being uniformly distributed after 100 steps? How about 1000? 50 billion?

In certain simple Markov chain examples we have discussed, it is easy to figure out the rate of convergence of π_n to π. For instance, for the Markov frog example (1.2), starting in the initial distribution $\pi_0 = (1, 0, 0)$, say, we can compute the distributions $\pi_1, \pi_2, \ldots$ by matrix multiplication and compare them to the stationary distribution $\pi = (1/4, 3/8, 3/8)$, getting the results shown in Figure (2.34).



(2.34) FIGURE. *Speed of convergence to stationarity in Markov frog example from Chapter 1.*

Notice the smooth geometric rate of decrease: the log distance decreases linearly, which means that the distance decreases to 0 geometrically. Coupling is a technique that can sometimes shed some light on questions of this sort. For example, in Exercise ([1.27]) we showed that $\|\pi_n - \pi\| \le \frac{2}{3}\left(\frac{11}{16}\right)^n$ for all n, ⟦and, in fact, $\|\pi_n - \pi\| \le \frac{2}{3}\left(\frac{1}{4}\right)^n$⟧ which gives much more information than just saying that $\|\pi_n - \pi\| \to 0$.

In this section we will concentrate on a simple shuffling example considered by Aldous and Diaconis in their article "Shuffling cards and stopping times." Again, the basic question is: How close is the deck to being "random" (i.e. uniformly distributed over the 52! possible permutations) after n shuffles? Or, put another way, how many shuffles does it take to shuffle well? Of course, the answer depends on the details of the method of shuffling; Aldous and Diaconis say that for the riffle shuffle model, which is rather realistic, the answer is "about 7." In this sort of problem, in contrast to the smooth and steady sort of decrease depicted in Figure (2.34), we will see that an interesting *threshold phenomenon* arises.

2.8.1 "Top-in-at-random" Shuffle

This shuffle, which is the model of shuffling we will analyze in detail, is simpler than the riffle shuffle for which Aldous and Diaconis gave the answer "about 7 is enough." The analysis we will discuss here also comes from their paper.

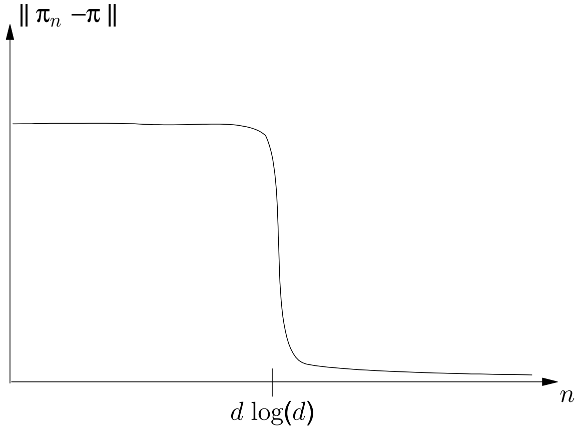
The top-in-at-random method would be a really silly way to shuffle a deck of cards in practice, but it is an appealingly simple model for a first example to study. One such shuffle consists of taking the top card off the deck and then inserting it back into the deck in a random position. "Random" here means that the top card could end up back on top (in which case the shuffle didn't change the deck at all), or it could end up second from top, third from top, ..., or at the bottom of the deck: altogether 52 equally likely positions.

Repeated performances of this shuffle on a deck of cards produces a sequence of "states" of the deck. This sequence of states forms a Markov chain $\{X_n\}$ having state space $\mathcal{S}_{52}$, the group of permutations of the cards. This Markov chain is irreducible, aperiodic, and has stationary distribution $\pi =$ Uniform on $\mathcal{S}_{52}$ (i.e. probability $1/(52!)$ for each permutation); therefore, by the Basic Limit Theorem, we may conclude that $\|\pi_n - \pi\| \to 0$ as $n \to \infty$.

▷ *Explaining each assertion in the previous sentence is Exercise [2.20].*

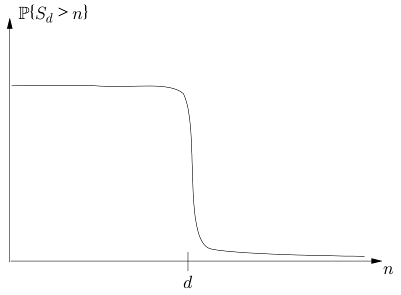
2.8.2 Threshold Phenomenon

Suppose we are working with a fresh deck of d cards, which we know are in the original order: card 1 on top, card 2 just below card 1, ..., and card d on the bottom. Then $\|\pi_0 - \pi\| = 1 - (1/d!)$. We also know that $\|\pi_n - \pi\| \to 0$ as $n \to \infty$, by the Basic Limit Theorem. It is natural to presume that the distance from stationarity $\|\pi_n - \pi\|$ decreases to 0 in some smooth, uneventful manner. However, the fact of the matter is that $\|\pi_n - \pi\|$ stays close to 1 for a while, then it undergoes a rather abrupt decrease from nearly 1 to nearly 0, and this abrupt change happens in a relatively small neighborhood of the value $n = d \log d$. That is, for large d the graph of $\|\pi_n - \pi\|$ versus n looks rather like the next picture.



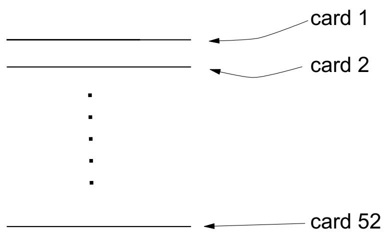
The larger the value of the deck size d, the sharper (relative to $d \log d$) the drop in $\|\pi_n - \pi\|$ near $n = d \log d$.

Well, this is hardly a gradual or uneventful decrease! This interesting behavior of $\|\pi_n - \pi\|$ has been called the *threshold phenomenon*. The phenomenon is not limited to this particular shuffle or even to shuffling in general, but rather seems to occur in a wide variety of interesting Markov chains. Aldous and Diaconis give a number of examples in which threshold phenomena can be shown to occur, but they point out that the phenomenon is not yet very well understood, in the sense that general conditions under which the phenomenon does or does not occur are not known.



This seems weird, doesn't it? As a partial antidote to the uneasy feelings of mystery that tend to accompany a first exposure to the threshold phenomenon idea, let's think about a simple problem that is familiar to all of us, in which a threshold of sorts also occurs. Suppose $X_1, X_2, \ldots$ are *iid* random variables with mean 1 and variance 1. In fact, for simplicity, let's suppose that they have the Normal distribution $N(1,1)$. Thinking of d as a large number, let $S_d = X_1 + \cdots + X_d$, and consider the probability $\mathbb{P}\{S_d > n\}$ as a function of n. Since $S_d \sim N(d,d)$, it is easy to see that the graph of $\mathbb{P}\{S_d > n\}$ has a "threshold" near $n = d$ if d is large. In fact, the length of a neighborhood about $n = d$ in which $\mathbb{P}\{S_d > n\}$ decreases from nearly 1 to nearly 0 is of order $\sqrt{d}$. However, if d is large, then $\sqrt{d}$ is vanishingly small compared with d. Thus, for large d, the graph of $\mathbb{P}\{S_d > n\}$ versus n (plotted on a scale in which $n = d$ is at some moderate, fixed distance from $n = 0$) will indeed appear as a sharp dropoff near $n = d$. In particular, note that the existence of a threshold for large d does *not* say that the dropoff from near 1 to near 0 takes place over a shorter and shorter time interval as d increases; it is just that the length (here $O(\sqrt{d})$) of that dropoff interval is smaller and smaller in comparison with the location (here around d) of that interval.

2.8.3 A random time to exact stationarity



Let's give a name to each card in the deck: say "card 1" is 2♡, "card 2" is 3♡, ..., "card 51" is K♠, "card 52" is A♠. Suppose we start with the deck in the pristine order shown to the right. Wouldn't it be nice if we could say, "After 1000 shuffles the deck will be exactly random," or maybe "After 1,000,000 shuffles the deck will be exactly random"? Well, sorry. We can't. We know π_n gets closer and closer to the uniform distribution as n increases, but unfortunately π_n will *never* become *exactly* random, even if n is 53 bezillion.

However, it *is* possible to find a *random* time T at which the deck becomes exactly uniformly distributed, that is, $X_T \sim \text{Unif}(\mathcal{S}_{52})$! Here is an example of such a random time. To describe it, let's agree that "card i" always refers to the same card ⟦e.g. card 52 = A♠⟧, while terms like "top card," "card in position 2," and so on just refer to whatever card happens to be on top, in position 2, and so on at the time under consideration. Also note that we may describe a sequence of shuffles simply by a sequence of *iid* random variables $U_1, U_2, \ldots$ uniformly distributed on $\{1, 2, \ldots, 52\}$: just say that the ith shuffle moves the top card to position U_i. Define the following random times:

$$\begin{aligned}
T_1 &= \inf\{n : U_n = 52\} = \text{1st time a top card goes below card 52},\\
T_2 &= \inf\{n > T_1 : U_n \geq 51\} = \text{2nd time a top card goes below card 52},\\
T_3 &= \inf\{n > T_2 : U_n \geq 50\} = \text{3rd time a top card goes below card 52},\\
&\vdots\\
T_{51} &= \inf\{n > T_{50} : U_n \geq 2\} = \text{51st time a top card goes below card 52},
\end{aligned}$$

and

$$T = T_{52} = T_{51} + 1. \tag{2.35}$$

It is not hard to see that T has the desired property and that X_T is uniformly distributed. To understand this, start with T_1. At time T_1, we know that some card is below card 52; we don't know which card, but that will not matter. After time T_1 we continue to shuffle until T_2, at which time another card goes below card 52. At time T_2, there are 2 cards below card 52. Again, we do not know which cards they are, but *conditional on which 2 cards are below card 52, each of the two possible orderings of those 2 cards is equally likely.* Similarly, we continue to shuffle until time T_3, at which time there are some 3 cards below card 52, and, whatever those 3 cards are, each of their (3!) possible relative positions in the deck is equally likely. And so on. At time T_{51}, card 52 has risen all the way up to become the top card, and the other 51 cards are below card 52 (now we *do* know which cards they are), and those 51 cards are in random positions (i.e. uniform over 51! possibilities). Now all we have to do is shuffle one more time to get card 52 in random position, so that at time $T = T_{52} = T_{51} + 1$, the whole deck is random.

Let us find ET. Clearly by the definitions above we have $T_1 \sim \text{Geom}(1/52)$, $(T_2 - T_1) \sim \text{Geom}(2/52)$, ..., $(T_{51} - T_{50}) \sim \text{Geom}(51/52)$, and $(T_{52} - T_{51}) \sim \text{Geom}(52/52) = 1$. Therefore,

$$\begin{aligned} ET &= E(T_1) + E(T_2 - T_1) + \cdots + (T_{51} - T_{50}) + E(T_{52} - T_{51}) \\ &= 52 + (52/2) + (52/3) + \cdots + (52/51) + (52/52) \\ &= 52\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{51} + \frac{1}{52}\right) \approx 52 \log 52. \end{aligned}$$

Analogously, if the deck had d cards rather than 52, we would have obtained $ET \sim d \log d$ (for large d), where T is now a random time at which the whole deck of d cards becomes uniformly distributed on $\mathcal{S}_d$.

2.8.4 Strong Stationary Times

The main new idea in the analysis of the shuffling example is that of a strong stationary time. As we have observed, the random variable T that we just constructed has the property that $X_T \sim \pi$. T also has two other important properties. First, X_T is independent of T. Second, T is a ***stopping time***, that is, for each n, one can determine whether or not $T = n$ just by looking at the values of $X_0, \ldots, X_n$. In particular, to determine whether or not $T = n$ it is not necessary to know any "future" values $X_{n+1}, X_{n+2}, \ldots$.

A random variable having the three properties just enumerated is called a strong stationary time.

(2.36) Definition. *A random variable T satisfying*

(i) *T is a stopping time,*

(ii) *X_T is distributed as π, and*

(iii) *X_T is independent of T*

*is called a **strong stationary time**.*

So what's so good about strong stationary times? For us, the answer is contained in the following theorem, which says that strong stationary times satisfy the same inequality that we derived for coupling times in (1.37).

(2.37) Lemma. *If T is a strong stationary time for the Markov chain $\{X_n\}$, then $\|\pi_n - \pi\| \leq \mathbb{P}\{T > n\}$ for all n.*

Proof: Letting $A \subseteq \mathcal{S}$, we will begin by showing that

$$\mathbb{P}\{T \leq n, X_n \in A\} = \mathbb{P}\{T \leq n\}\pi(A). \tag{2.38}$$

To see this, let $k \leq n$ and write $\mathbb{P}\{T = k, X_n \in A\} = \sum_i \mathbb{P}\{T = k, X_k = i, X_n \in A\} = \sum_i \mathbb{P}\{T = k, X_k = i\}\mathbb{P}\{X_n \in A \mid T = k, X_k = i\}$. But $\mathbb{P}\{X_n \in A \mid T = k, X_k = i\} = \mathbb{P}\{X_n \in A \mid X_k = i\} =: P^{n-k}(i, A)$ by the Markov property and the assumption that T

is a stopping time. Also, $\mathbb{P}\{T = k, X_k = i\} = \mathbb{P}\{T = k, X_T = i\} = \mathbb{P}\{T = k\}\mathbb{P}\{X_T = i\} = \mathbb{P}\{T = k\}\pi(i)$ by properties (ii) and (iii) of the definition of strong stationary time. Substituting gives $\mathbb{P}\{T = k, X_n \in A\} = \mathbb{P}\{T = k\} \sum_i \pi(i)P^{n-k}(i, A) = \mathbb{P}\{T = k\}\pi(A)$ by the stationarity of π. Summing this over $k \le n$ gives (2.38).

Next, similarly to the proof of the coupling inequality, we decompose according to whether $T \le n$ or $T > n$ to show that for $A \subseteq \mathcal{S}$

$$\begin{aligned}\pi_n(A) - \pi(A) = \mathbb{P}\{X_n \in A\} - \pi(A) &= \mathbb{P}\{X_n \in A, T \le n\} + \mathbb{P}\{X_n \in A, T > n\} - \pi(A) \\ &= \pi(A)\mathbb{P}\{T \le n\} + \mathbb{P}\{X_n \in A, T > n\} - \pi(A) \quad \text{by (2.38)} \\ &= \mathbb{P}\{X_n \in A, T > n\} - \pi(A)\mathbb{P}\{T > n\}.\end{aligned}$$

Since each of the last two quantities lies between 0 and $\mathbb{P}\{T > n\}$, we conclude that $|\pi_n(A) - \pi(A)| \le \mathbb{P}\{T > n\}$, so that $\|\pi_n - \pi\| \le \mathbb{P}\{T > n\}$. □

2.8.5 Proof of threshold phenomenon in shuffling

Let $\Delta(n)$ denote $\|\pi_n - \pi\|$. The proof that the threshold phenomenon occurs in the top-in-at-random shuffle consists of two parts: Roughly speaking, the first part shows that $\Delta(n)$ is close to 0 for n slightly larger than $d \log d$, and the second part shows that $\Delta(n)$ is close to 1 for n slightly smaller than $d \log d$, where in both cases the meaning of "slightly" is "small relative to $d \log d$."

The first part is addressed by the next result.

(2.39) THEOREM. *For T as defined in (2.35), we have*

$$\Delta(d \log d + cd) \le \mathbb{P}\{T > d \log d + cd\} \le e^{-c} \quad \textit{for all } c \ge 0.$$

Note that for each fixed c, cd is small relative to $d \log d$ if d is large enough.

PROOF: The first inequality in (2.39) is just Lemma (2.37), so our task is to prove the second inequality. Recall that, as discussed above, $T_1 \sim \text{Geom}(1/d)$, $T_2 - T_1 \sim \text{Geom}(2/d)$, $T_3 - T_2 \sim \text{Geom}(3/d)$, ..., and $T - T_{d-1} = T_d - T_{d-1} \sim \text{Geom}(d/d) = 1$. It is also clear that $T_1, T_2 - T_1, T_3 - T_2, \ldots,$ and $T - T_{d-1}$ are independent. Thus,

$$T \sim \text{Geom}\left(\frac{1}{d}\right) \oplus \text{Geom}\left(\frac{2}{d}\right) \oplus \cdots \oplus \text{Geom}\left(\frac{d-1}{d}\right) \oplus 1,$$

where the symbol "$\oplus$" indicates a sum of independent random variables. However, observe that the distribution $1 \oplus \text{Geom}[(d-1)/d] \oplus \cdots \oplus \text{Geom}[1/d]$ is also the distribution that arises in the famous *coupon collector's problem.* ⟦To review the coupon collector's problem: Suppose that each box of Raisin Bran cereal contains one of d possible coupons numbered $\{1, \ldots, d\}$, with the coupons in different boxes being independent and uniformly distributed on $\{1, \ldots, d\}$. The number of cereal boxes a collector must buy in order to obtain a complete set of d coupons has the distribution $1 \oplus \text{Geom}[(d-1)/d] \oplus \cdots \oplus \text{Geom}[1/d]$.⟧

To find a bound on $\mathbb{P}\{T > n\}$, let us adopt this coupon collecting interpretation of T. For each $i = 1, \ldots, d$ define an event

$$B_i = \{\text{coupon } i \text{ does not appear among the first } n \text{ cereal boxes}\}.$$

Then the event $T > n$ is just the union $\bigcup_{i=1}^d B_i$, so that

$$\mathbb{P}\{T > n\} \le \sum_{i=1}^d \mathbb{P}(B_i) = \sum_{i=1}^d \left(\frac{d-1}{d}\right)^n = d\left(1 - \frac{1}{d}\right)^n \le de^{-n/d},$$

where the last inequality uses the fact that $1 - x \le e^{-x}$ for all numbers x. Setting $n = d\log d + cd$ gives

$$\mathbb{P}\{T > n\} \le de^{-(\log d + c)} = e^{-c},$$

as desired. □

For the second part of the proof, let us temporarily be a bit more fastidious about notation: Instead of just writing π_n and π, let us write $\pi_n^{(d)}$ and $\pi^{(d)}$ to indicate explicitly the dependence of the various distributions on the deck size d.

(2.40) Theorem. *Let $k(d) = d\log d - c_d d$, where $\{c_d\}$ is a sequence of numbers that approaches infinity as $d \to \infty$. Then*

$$\|\pi_{k(d)}^{(d)} - \pi^{(d)}\| \to 1 \quad \textit{as } d \to \infty.$$

Note: The case of interest for establishing the threshold at $d\log d$ is when $c_d = o(\log d)$, since in that case $k(d) \sim d\log d$.

Proof: Let's start with some fumbling around, intended to provide some glimmer of hope that we might have been able to think of this proof ourselves. The proof proceeds by bounding $\|\pi_{k(d)}^{(d)} - \pi^{(d)}\|$ below by something that is close to 1 for large d. By the definition of the total variation distance $\|\cdot\|$, this may be done by finding events A_d such that $\|\pi_{k(d)}^{(d)}(A_d) - \pi^{(d)}(A_d)\|$ is close to 1. OK, now let's drop those pesky d's from the notation, and say that for large d, *we want to find events A such that $\pi_k(A)$ is large (close to 1) while $\pi(A)$ is small (close to 0).* Fumbling time...

- How about $A = \{\text{card } d \text{ is still on the bottom}\}$?
  - Is $\pi(A)$ small? Yes: $\pi(A) = 1/d$.
  - Is $\pi_k(A)$ large? No: since $k \gg d$, clearly $\mathbb{P}\{T_1 > k\} = \mathbb{P}\{\text{Geom}(1/d) > k\}$ is not large, so that $\mathbb{P}\{\text{card } d \text{ is still on the bottom at time } k\}$ is not large.
- How about $A = \{\text{cards } d-1 \text{ and } d \text{ are still in their original positions}\}$?
  - Is $\pi(A)$ small? Yes: $\pi(A) = 1/[d(d-1)]$.
  - Is $\pi_k(A)$ large? No: in fact, it is smaller than the previous $\pi_k(A)$ we just considered. You should be ashamed of yourself for that suggestion!

- How about just requiring cards $d-1$ and d still be in their original *order*, that is, $A = \{\text{card } d-1 \text{ still above card } d \text{ in the deck}\}$?
  - Is $\pi_k(A)$ large? Maybe; this doesn't seem very obvious.
  - Is $\pi(A)$ small? No: $\pi(A) = 1/2$.
- Well, that may look discouraging. But with a little more thought we can at least extend the previous idea to get $\pi(A)$ small while keeping a "maybe" for $\pi_k(A)$ being large, as follows. How about

  $$A = A_{d,a} = \{\text{cards } d-a+1, d-a+2, \ldots, d \text{ still in their original order}\}?$$

  - Is $\pi_k(A)$ large? Still maybe.
  - Is $\pi(A)$ small? $\pi(A) = 1/(a!)$, so ***yes*** if a increases with d.

Let's review what we are doing. We are given a sequence of numbers $\{c_d\}$ such that $c_d \to \infty$ as $d \to \infty$. [[As noted, we are interested in the case where we also have $c_d = o(\log d)$, but the proof will not require this.]] For each d, we have also defined a number $k = k(d) = d \log d - c_d d$. For each d and each $a \le d$ we have identified an event $A = A_{d,a}$. What we want to show is that there is a sequence $\{a_d\}$ of values of a such that, as $d \to \infty$, we have $\pi_k(A) \to 1$ and $\pi(A) \to 0$. [[Actually we should write these statements as $\pi^{(d)}_{k(d)}(A_{d,a_d}) \to 1$ and $\pi^{(d)}(A_{d,a_d}) \to 0$, but I doubt any of us really wants that.]]

As for getting the second statement, since $\pi(A) = 1/(a!)$, any sequence $\{a_d\}$ that tends to infinity as $d \to \infty$ will suffice.

To get the first statement to hold we need a little more analysis. Suppose that in k shuffles, card $d-a+1$ has not yet "risen to the top of the deck." In this case, clearly the event A occurs. Letting U denote the number of shuffles required for card $d-a+1$ to rise to the top of the deck, we thus have

$$\pi_k(A) \ge \mathbb{P}\{U > k\}.$$

Note that

$$U \sim \text{Geom}\left(\frac{a}{d}\right) \oplus \text{Geom}\left(\frac{a+1}{d}\right) \oplus \cdots \oplus \text{Geom}\left(\frac{d-1}{d}\right).$$

The plan now is to use Chebyshev's inequality to show that we can cause $\mathbb{P}\{U > k\} \to 1$ to hold by choosing a_d appropriately. This will show that $\pi_k(A) \to 1$, and hence prove the theorem.

The ingredients needed to use Chebyshev are $\mathbb{E}(U)$ and $\text{Var}(U)$. Since $\mathbb{E}[\text{Geom}(p)] = 1/p$, we have

$$\mathbb{E}(U) = d\left(\frac{1}{a} + \frac{1}{a+1} + \cdots + \frac{1}{d-1}\right) = d\big(\log d - \log a + o(1)\big)$$

where the second equality assumes only that $a_d \to \infty$. Next, since $\text{Var}[\text{Geom}(p)] = (1-p)/p^2 \le 1/p^2$, using independence gives

$$\text{Var}(U) \le d^2\left(\frac{1}{a^2} + \frac{1}{(a+1)^2} + \cdots\right) =: \epsilon(a) d^2,$$

where $\epsilon(a) \to 0$ as $a \to \infty$ (or $d \to \infty$).

So here it is in a nutshell. Since $\text{Var}(U) = o(d^2)$, so that U has standard deviation $\text{SD}(U) = o(d)$, all we have to do is choose a_d so that the difference

$$\mathbb{E}(U) - k = d(\log d - \log a_d + o(1)) - d(\log d - c_d) \sim d(c_d - \log a_d)$$

is large compared with $\text{SD}(U)$, that is, at least the order of magnitude of d. But that's easy; for example, if we choose $a_d = e^{c_d/2}$, then $\mathbb{E}(U) - k \sim d(c_d/2)$, whose order of magnitude is larger than d.

To say this in a bit more detail, choose $a_d = e^{c_d/2}$, say. ⟦Many other choices would also do.⟧ Then of course $a_d \to \infty$. So we have

$$\begin{aligned}\mathbb{P}\{U > k(d)\} &= \mathbb{P}\{U > d(\log d - c_d)\}\\ &= \mathbb{P}\{U - \mathbb{E}(U) > d(\log d - c_d) - d(\log d - \log a_d + o(1))\}\\ &= \mathbb{P}\{U - \mathbb{E}(U) > -d(c_d - \log a_d + o(1))\}.\end{aligned}$$

Substituting $a_d = e^{c_d/2}$, this becomes

$$\begin{aligned}\mathbb{P}\{U > k(d)\} &= \mathbb{P}\{U - \mathbb{E}(U) > -d(c_d/2 + o(1))\}\\ &\ge \mathbb{P}\{|U - \mathbb{E}(U)| < d(c_d/2 + o(1))\}\\ &\ge 1 - \frac{\text{Var}(U)}{d^2(c_d/2 + o(1))^2}\\ &\ge 1 - \frac{\epsilon(a_d)}{(c_d/2 + o(1))^2}.\end{aligned}$$

Since the last expression approaches 1 as $d \to \infty$, we are done. □

This completes our analysis of the top-in-at-random shuffle. There are lots of other interesting things to look at in the Aldous and Diaconis paper as well as some of the other references. For example, a book of P. Diaconis applies group representation theory to this sort of problem. The paper of Brad Mann is a readable treatment of the riffle shuffle.

2.9 Exercises

[2.1] For a branching process $\{G_t\}$ with $G_0 = 1$, define the probability generating function of G_t to be ψ_t, that is,

$$\psi_t(z) = \mathbb{E}(z^{G_t}) = \sum_{k=0}^{\infty} z^k \mathbb{P}\{G_t = k\}.$$

With ψ defined as in (2.1), show that $\psi_1(z) = \psi(z)$, $\psi_2(z) = \psi(\psi(z))$, $\psi_3(z) = \psi(\psi(\psi(z)))$, and so on.

[2.2] With ψ_t defined as in Exercise [2.1], show that $\mathbb{P}\{G_t = 1\} = \psi_t'(0)$.

[2.3] Consider a branching process with offspring distribution Poisson(2), that is, Poisson with mean 2. Calculate the extinction probability ρ to four decimal places.

[2.4] As in the previous exercise, consider again a branching process with offspring distribution Poisson(2). We know that the process will either go extinct or diverge to infinity, and the probability that it is any fixed finite value should converge to 0 as $t \to \infty$. In this exercise you will investigate how fast such probabilities converge to 0. In particular, consider the probability $\mathbb{P}\{G_t = 1\}$, and find the limiting ratio

$$\lim_{t\to\infty} \frac{\mathbb{P}\{G_{t+1} = 1\}}{\mathbb{P}\{G_t = 1\}}.$$

This may be interpreted as a rate of geometric decrease of $\mathbb{P}\{G_t = 1\}$.

⟦Hint: use the result of Exercise [2.2].⟧

[2.5] Consider a branching process $\{G_t\}$ with $G_0 = 1$ and offspring distribution $f(k) = q^k p$ for $k = 0, 1, \ldots$, where $q = 1 - p$. So f is the probability mass function of $X - 1$, where $X \sim \text{Geom}(p)$.

(a) Show that

$$\frac{\psi(z) - (p/q)}{\psi(z) - 1} = \frac{p}{q}\left(\frac{z - (p/q)}{z - 1}\right).$$

(b) Derive the expressions

$$\psi_t(z) = \begin{cases} \frac{p[(q^t - p^t) - qz(q^{t-1} - p^{t-1})]}{q^{t+1} - p^{t+1} - qz(q^t - p^t)} & \text{if } p \neq 1/2 \\ \frac{t - (t-1)z}{t + 1 - tz} & \text{if } p = 1/2. \end{cases}$$

⟦Hint: The first part of the problem makes this part quite easy. If you are finding yourself in a depressing, messy calculation, you are missing the easy way. For $p \neq 1/2$, consider the fraction $[\psi_t(z) - (p/q)]/[\psi_t(z) - 1]$.⟧

(c) What is the probability of ultimate extinction, as a function of p?
⟦Hint: Observe that $\mathbb{P}\{G_t = 0\} = \psi_t(0)$.⟧

[2.6] Let $\{G_t\}$ be a supercritical (i.e. $\mu = \mathbb{E}(X) > 1$) branching process with extinction probability $\rho \in (0, 1)$. Let $B = \bigcup\{G_t = 0\}$ denote the event of eventual extinction.

(a) Show that $\mathbb{E}(z^{G_t} \mid B) = (1/\rho)\psi_t(\rho z)$.

(b) Consider again the example of Exercise [2.5], with $p < 1/2$. Let $\{\tilde{G}_t\}$ be a branching process of the same form as $\{G_t\}$, except with the probabilities p and q interchanged. So $\{\tilde{G}_t\}$ is subcritical, and goes extinct with probability 1. Show that the G process, conditional on the event B, behaves like the $\tilde{G}$ process, in the sense that $\mathbb{E}(z^{G_t} \mid B) = E(z^{\tilde{G}_t})$.

(c) Isn't that interesting?

[2.7] Consider an irreducible, time-reversible Markov chain $\{X_t\}$ with $X_t \sim \pi$, where the distribution π is stationary. Let A be a subset of the state space. Let $0 < \alpha < 1$, and define on the same state space a Markov chain $\{Y_t\}$ having probability transition matrix Q satisfying, for $i \neq j$,

$$Q(i,j) = \begin{cases} \alpha P(i,j) & \text{if } i \in A \text{ and } j \notin A \\ P(i,j) & \text{otherwise.} \end{cases}$$

Define the diagonal elements $Q(i,i)$ so that the rows of Q sum to 1.

(a) What is the stationary distribution of $\{Y_t\}$, in terms of π and α?

(b) Show that the chain $\{Y_t\}$ is also time-reversible.

(c) Show by example that the simple relationship of part (1) need not hold if we drop the assumption that X is reversible.

[2.8] Let $\{X_t\}$ have probability transition matrix P and initial distribution π_0. Imagine observing the process until time n, seeing $X_0, X_1, \ldots, X_{n-1}, X_n$. The time reversal of this sequence of random variables is $X_n, X_{n-1}, \ldots, X_1, X_0$, which we can think of as another random process $\tilde{X}$. That is, given the Markov chain X, define the reversed process $\{\tilde{X}_t\}$ by $\tilde{X}_t = X_{n-t}$.

(a) Show that

$$\mathbb{P}\{X_t = j \mid X_{t+1} = i, X_{t+2} = x_{t+2}, \ldots, X_{t+n} = x_{t+n}\} = \frac{\pi_t(j)P(j,i)}{\pi_{t+1}(i)}$$

(b) Use part (a) to show that the process $\{\tilde{X}_t\}$ is a Markov chain, although it is not time homogeneous in general.

(c) Suppose $\{X_t\}$ has stationary distribution π, and suppose X_0 is distributed according to π. Show that the reversed process $\{\tilde{X}_t\}$ is a time-homogeneous Markov chain.

[2.9] Let $p = (p_1, \ldots, p_d)$ be a probability mass function on $\{1, \ldots, d\}$. Consider the residual lifetime chain, discussed in Exercise [1.14], which has probability transition matrix

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ d-1 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & \cdots & d-2 & d-1 \end{array} \\ \begin{pmatrix} p_1 & p_2 & p_3 & \cdots & p_{d-1} & p_d \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ & & & & & \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \end{array}$$

and stationary distribution $\pi(i) = \mathbb{P}\{X > i\}/\mathbb{E}(X)$, where X denotes a random variable distributed according to p.

(a) Find $\tilde{P}$, the probability transition matrix for the reversed chain.

(b) In renewal theory, the time since the most recent renewal is called the *age*, and the process $\{A_t\}$, whose state A_t is the age at time t, is called the *age process*. Show that the matrix $\widetilde{P}$ that you have just found is the probability transition matrix of the age process. That is, the time-reversed residual lifetime chain is the age chain.

[2.10] Let's think about the irreducibility and aperiodicity conditions of the Basic Limit Theorem as applied to the Metropolis method. Suppose that the graph structure on $\mathcal{S}$ is a connected graph. Let π be any distribution other than π_{rw}, the stationary distribution of a random walk on the graph. Show that the Basic Limit Theorem implies that the Metropolis chain converges in distribution to π.

[2.11] Why was the condition $\pi \neq \pi_{\text{rw}}$ needed in Exercise [2.10]?

[2.12] ⟦Metropolis-Hastings method⟧ For simplicity, let us assume that π is positive, so that we won't have to worry about dividing by 0. Choose any probability transition matrix $Q = (Q(i,j))$ ⟦again, suppose it is positive⟧, and define $P(i,j)$ for $i \neq j$ by

$$P(i,j) = Q(i,j) \min\left(1, \frac{\pi(j)Q(j,i)}{\pi(i)Q(i,j)}\right),$$

and of course define $P(i,i) = 1 - \sum_{j \neq i} P(i,j)$. Show that the probability transition matrix P has stationary distribution π. Show how the Metropolis method we have discussed is a special case of this Metropolis-Hastings method.

[2.13] ⟦Computing project: traveling salesman problem⟧ Make up an example of the traveling salesman problem; it could look similar to the first figure in Example (2.22) if you'd like. Write a program to implement simulated annealing and produce a sequence of figures showing various improving traveling salesman tours. You could even produce a slithering snake movie if you are so inspired.

[2.14] For simulated annealing, temperature schedules of the form (2.20) decrease excruciatingly slowly. It is reasonable to ask whether we could decrease the temperature faster and still retain a guarantee of convergence in distribution to global optima. Let c be a positive number, and consider performing simulated annealing with the cooling schedule $T_n = bn^{-c}$. Of course, this schedule decreases faster than (2.20), no matter how small c is. Can you give an example that shows that such a schedule decreases too fast, in the sense that the process has positive probability of getting stuck in a local minimum forever? Thus, even $T_n = n^{-.0001}$ cools "too fast"!

[2.15] ⟦A creative writing, essay-type question⟧ Do you care about convergence in distribution to a global minimum? Does this property of simulated annealing make you happy?

[2.16] Prove (2.18).

[2.17] Here is yet another interpretation of total variation distance. Let μ and ν be distributions on a finite set $\mathcal{S}$. Show that

$$\|\mu - \nu\| = \min \mathbb{P}\{X \neq Y\},$$

where the minimum is taken over all $\mathbb{P}$, X, and Y such that X has distribution μ and Y has distribution ν.

[2.18] Prove Lemma (2.27) using coupling.

Hint: Defining $R = PQ$, we want to show that for all i and j,

$$\|R_{i\cdot} - R_{j\cdot}\| \leq \sup_{k,l} \|P_{k\cdot} - P_{l\cdot}\| \sup_{k,l} \|Q_{k\cdot} - Q_{l\cdot}\|.$$

Construct Markov chains $X_0 \xrightarrow{P} X_1 \xrightarrow{Q} X_2$ and $Y_0 \xrightarrow{P} Y_1 \xrightarrow{Q} Y_2$ with $X_0 = i$ and $Y_0 = j$. Take (X_1, Y_1) to be a coupling achieving the total variation distance $\|P_{i\cdot} - P_{j\cdot}\|$. Then, conditional on (X_1, Y_1), take X_2 and Y_2 to achieve the total variation distance $\|Q_{X_1\cdot} - Q_{Y_1\cdot}\|$.

[2.19] Show that if a probability transition matrix Q has a column all of whose entries are at least a, then $\delta(Q) \leq 1 - a$.

[2.20] Repeated performances of the top-in-at-random shuffle on a deck of cards produces a Markov chain $\{X_n\}$ having state space $\mathcal{S}_{52}$, the group of permutations of the cards. Show that this Markov chain is irreducible, aperiodic, and has stationary distribution $\pi =$ Uniform on $\mathcal{S}_{52}$ (i.e. probability $1/(52!)$ for each permutation), so that, by the Basic Limit Theorem, we may conclude that $\|\pi_n - \pi\| \to 0$ as $n \to \infty$.

[2.21] Why do we require the "strong" in strong stationary times? That is, in Definition (2.36), although I'm not so inclined to question the requirement $X_T \sim \pi$, why do we require X_T to be independent of T? It is easy to see where this is used in the proof of the fundamental inequality $\|\pi_n - \pi\| \leq \mathbb{P}\{T > n\}$, but that is only a partial answer. The real question is whether the fundamental inequality could fail to hold if we do not require X_T to be independent of T. Can you find an example?

Things to do

- Add a section introducing optimal stopping, dynamic programming, and Markov decision problems.