Markov Chains and Card Shuffling

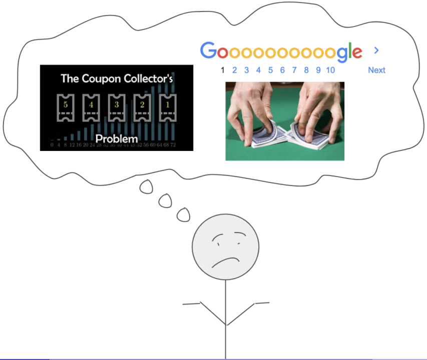
Christina Li, Yuxin Xie, William Yue

November 29, 2020

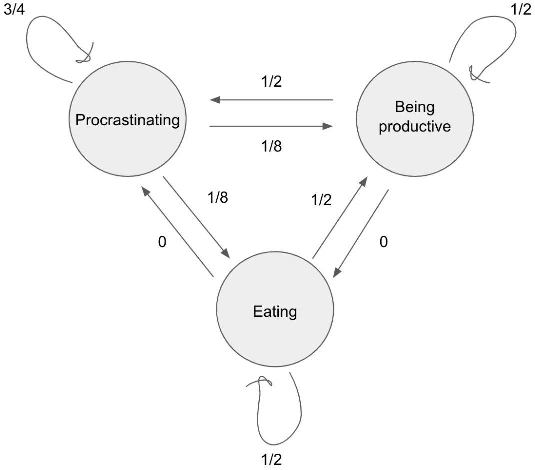
Results Covered

Have you ever wondered about:

- **Coupon-collecting problem:** if there are *n* types of coupons, and the probability of acquiring each of the types is the same, then what's the expected number of coupons that will have been collected before having collected all *n* types of coupons?
- **Google's PageRank**: how does Google order their search results?
- **Card-shuffling**: how many times you need to shuffle a deck of cards before the deck is sufficiently "random"?



Markov Chain Example: Student on School Night



Initial probabilities:

- Procrastinating: 7/8
- Being productive: 1/8
- Eating: 0

Notice: probability of transitioning to a particular activity only depends on current activity

Preliminaries

States are the nodes of the Markov Chain.

Definition (State-space)

A *state-space* Ω is a countable set $\{i, j, k, \dots\}$ where each $i \in \Omega$ is a *state*.

The **initial distribution** describes the probabilities of starting the Markov Chain from a particular state.

Definition (Initial distribution)

An *initial distribution* over Ω is a distribution $\lambda = (\lambda_i : i \in \Omega)$ such that $0 \leq \lambda_i \leq 1$ and $\sum_i \lambda_i = 1$.

A **transition matrix** stores the probabilities p_{ij} of moving from state i to j in a Markov Chain.

Definition (Transition matrix)

A *transition matrix* P is $P = (p_{ij} : i, j \in I)$ with $p_{ij} \geq 0$ for all i, j and $\sum_{j \in I} p_{ij} = 1$.

The distribution after taking a step is λP. To represent a transition after n steps, the *n*-**step transition matrix** is P^n.

Definition of Markov Chain

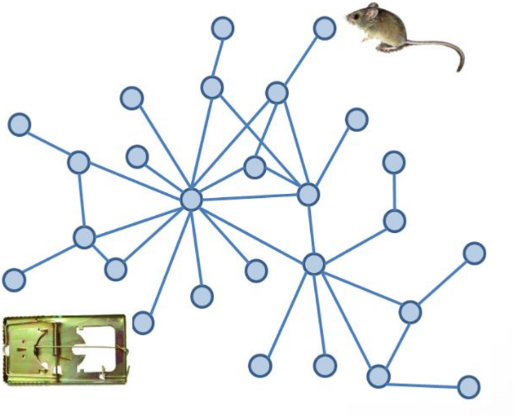
Definition (Markov Chain)

$(X_n)_{n\geq 0}$ is a *Markov chain* Markov(λ, P) with initial distribution λ and transition matrix P if for all $n \geq 0$ and $i_0, \ldots, i_{n+1} \in \Omega$:

1. $\mathbb{P}(X_0 = i_0) = \lambda_{i_0}$;
2. $\mathbb{P}(X_{n+1} = i_{n+1} \mid X_0 = i_0, \ldots, X_n = i_n) = \mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n) = p_{i_n i_{n+1}}$

When To Stop, When To Catch

We need to know when to catch the mouse!



Stopping Time

Definition (Stopping Time)

A stopping time is a random variable $T : \Omega \longrightarrow \mathbb{N} \cup \{\infty\}$ such that the event at $T = n$ only depends on information already known, which are $X_0, X_1, X_2, \cdots, X_n$.

To determine whether or not a random variable is a stopping time, we look for a stopping rule, a mechanism that tells us whether to continue or stop based on the present and past events (information already known)

Example

1. H_i, the time when the mouse hits one of the traps, is a stopping time.
2. $T = H_i + 1$, 1 step after the hitting time, is a stopping time.
3. $T = H_i - 1$, 1 step before the hitting time, is not a stopping time.

Example: Coupon-Collecting

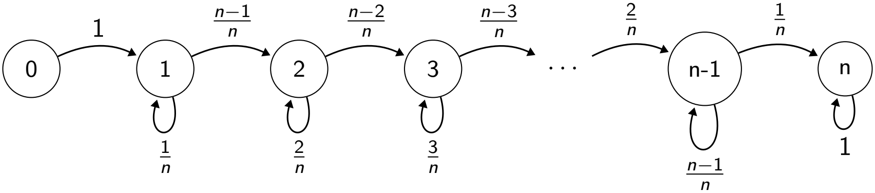
Example

Assume that there are n types of different coupons in total, and the possibility of acquiring each one of the types is the same.

- τ is the total number of coupons in hand when the collection set includes all types of coupons for the first time.
- τ_i is the total number of coupons in hand when the collection set includes i types of coupons for the first time. τ_i is the stopping time here

Why is this a Markov Chain

Each state in the chain is the number of distinct coupon types in hand.



Example: Coupon-Collecting

Proposition

Assume that each time, a new coupon is chosen randomly and uniformly. Then the expected value of τ is:

$$\mathbb{E}(\tau) = n \sum_{l=1}^{n} \frac{1}{l}$$

Proof.

$$\tau_n = \tau_1 + (\tau_2 - \tau_1) + \cdots + (\tau_n - \tau_{n-1})$$

The success probability for the random variable $\tau_i - \tau_{i-1}$ (between two stopping times) is $\frac{n-i+1}{n}$. Therefore,

$$\mathbb{E}(\tau) = \sum_{i=1}^{n} \mathbb{E}(\tau_i - \tau_{i-1}) = n \sum_{i=1}^{n} \frac{1}{n-i+1} = n \sum_{l=1}^{n} \frac{1}{l},\ (l = n - i + 1)$$

□

Example: Coupon-Collecting

Proposition

For any $c > 0$,

$$\mathbb{P}(\tau > \lceil n \log n + cn \rceil) \leq e^{-c}$$

Proof.

Let $\mathbb{P}(A_i)$ denote the probability that the i-th type of coupon has not yet occurred among the $\lceil n \log n + cn \rceil$ coupons already collected. Then,

$$\mathbb{P}(\tau > \lceil n \log n + cn \rceil) = \mathbb{P}(\cup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} \mathbb{P}(A_i)$$

For each coupon among the $\lceil n \log n + cn \rceil$ coupons, the probability that it is not the i-th type is $1 - \frac{1}{n}$.

$$\sum_{i=1}^{n} \left(1 - \frac{1}{n}\right)^{\lceil n \log n + cn \rceil} = n \left(1 - \frac{1}{n}\right)^{\lceil n \log n + cn \rceil} \leq n e^{\left(-\frac{n \log n + cn}{n}\right)} = e^{-c}$$

□

Invariant Distribution

Definition (Invariant distribution)

An invariant distribution λ is a row vector $(\lambda_i : i \in I)$ where $\sum_i \lambda_i = 1$ and

$$\lambda P = \lambda.$$

We also call these **stationary distributions**, as shown by the following theorem.

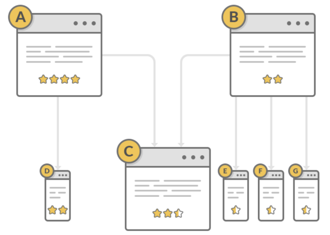
Theorem

Let $(X_n)_{n \geq 0}$ be Markov(λ, P) *and suppose that λ is invariant for P. Then $(X_{m+n})_{n \geq 0}$ is also* Markov(λ, P)*.*

Example: Google's PageRank

Example

The PageRank algorithm orders webpages according to its *relevance*. If webpage i with a relevance of R_i links to k webpages, then i contributes $\frac{R_i}{k}$ relevance to each of the k webpages. The relevance of a webpage is the sum of the relevances attributed to it by the websites that link to it.



Example: Google's Pagerank

Probabilistic view: rank by the probability that a *random surfer* on the Internet will arrive at that page.

Definition (Random Surfer's Movement)

Let n be the number of webpages on the Internet and $L(i)$ be the number of hyperlinks on webpage i. The transition matrix P describing the movement of the random surfer is

$$p_{ij} = \begin{cases} \frac{1}{L(i)} & \text{if } L(i) > 0 \text{ and } (i, j) \in E, \\ \frac{1}{n} & \text{if } L(i) = 0. \end{cases}$$

After adding some small changes, we know that this scenario is a stationary distribution, so we solve $\pi P = \pi$. If $\pi_i > \pi_j$, then webpage i is more relevant than webpage j and is consequently ranked higher.

Properties of Markov Chains

Definition (Irreducibility)

A Markov Chain in which any two states can be reached from each other is irreducible.

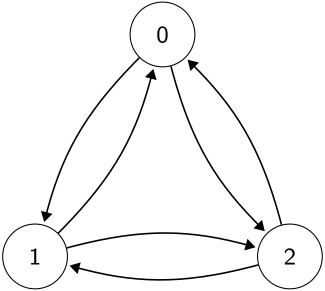
Definition (Periodic and Aperiodic State)

State i is periodic if the number of steps required to return to i have a greatest common divisor, d, larger than 1 (such as 3, 6, 9, ...). Thus, for periodic states, when $d \nmid n$, $p_{ii}^{(n)} = 0$, and the period is d. In contrast, states are aperiodic if $\exists n_1, \cdots, n_k \geq 1, k \geq 2$ with no common divisor and $p_{ii}^{(n_j)} > 0$ for all values of j from $1 \rightarrow k$.

Ergodic Chain

Definition (Ergodic Chain)

A finite ergodic chain is a Markov chain on a finite state space that is aperiodic and irreducible.



Convergence to Equilibrium

Theorem

Assume that P is the transition matrix for an ergodic Markov chain with invariant distribution π. For any initial distribution, $\mathbb{P}(X_n = j) \to \pi_j$ as $n \to \infty$

What if, there's a pattern in everything!

Convergence to Equilibrium (cont.)

Proof.

Assume $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$ are independent, and the two Markov chains are Markov(λ, P) and Markov(π, P). Let T be $T =$inf $\{n \geq 1 : X_n = Y_n = b\}$
Define a new chain, $W_n = (X_n, Y_n)$ on $I \times I$.

$$\tilde{p}_{(i,k)(j,l)} = p_{ij}p_{kl} \quad \mu_{(i,k)} = \lambda_i\pi_k \quad \tilde{\pi}_{(i,k)} = \pi_i\pi_k$$

Using the aperiodicity of the transition matrix P, we know that when n becomes sufficiently large, for any state i,j,k,l,

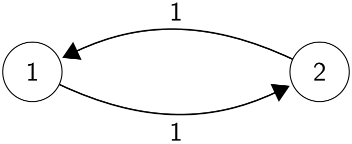
$$\tilde{p}^{(n)}_{(i,k)(j,l)} = p^{(n)}_{ij}p^{(n)}_{kl} > 0 \quad \mathbb{P}(T < \infty) = 1$$

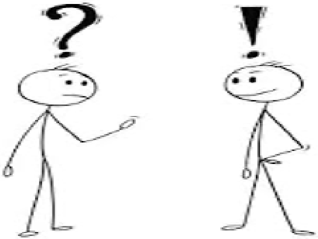
$$\mathbb{P}(X_n = i) = \mathbb{P}(X_n = i, n) + \mathbb{P}(X_n = i, n < T) = \mathbb{P}(Y_n = i, n \geq T) + \mathbb{P}(X_n = i, n < T)$$
$$= \pi_i - \mathbb{P}(Y_n = i, n < T) + \mathbb{P}(X_n = i, n < T)$$

Obviously, $\mathbb{P}(Y_n = i, n < T) \leq \mathbb{P}(n < T)$ and $\mathbb{P}(n < T) \to \mathbb{P}(T = \infty$, which is 0 as $n \to \infty$ (because $\mathbb{P}(T < \infty) = 1$). Therefore, $\mathbb{P}(Y_n = i, n < T) \to 0$. Similarly, $\mathbb{P}(X_n = i, n < T) \to 0$. Therefore, $\mathbb{P}(X_n = i) \to \pi_i$. □

* Convergence to Equilibrium (cont.)

When you look back at the conditions...... they make sense!





Total Variation

Definition (Total Variation Distance)

Assume that there are two distributions, μ and ν on the event space, ω, total variation distance is defined as follows:

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|$$

Coin	$P(H)$	$P(T)$	$P(L)$
Coin A	$\frac{1}{2}$	$\frac{1}{2}$	0
Coin B	$\frac{2}{3}$	$\frac{1}{4}$	$\frac{1}{12}$

The difference between the two distributions for the three events (H, T, and L) are $\frac{1}{6}$, $\frac{1}{4}$, and $\frac{1}{12}$ respectively. The total variation is thus $\frac{1}{2}(\frac{1}{6} + \frac{1}{4} + \frac{1}{12}) = \frac{1}{4}$.

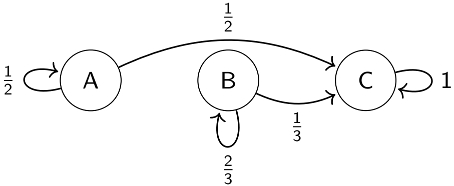
Mixing Time

Definition (Mixing Time)

Mixing time, t_{mix}, is defined as follows:

$$t_{\text{mix}}(\varepsilon) = \min\{t : d(t) \leq \varepsilon\}$$

where $d(t) = \max_{x \in \Omega} \|P^t(x) - \pi\|_{\text{TV}}$. Here $P^t(x)$ is the distribution at time t starting from initial state x.



$$\pi = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad P^2 \begin{pmatrix} a \\ b \\ 1-a-b \end{pmatrix} = \begin{pmatrix} \frac{a}{4} \\ \frac{4b}{9} \\ 1 - \frac{a}{4} - \frac{4b}{9} \end{pmatrix} \quad d(2) = \frac{4}{9}$$

Shuffling Cards

Two methods of shuffling cards:

- The top-to-random shuffle: ~ 300 repetitions
- The classic riffle shuffle: ~ 8 repetitions

The Top-to-Random Shuffle

Example (The Top-to-Random Shuffle)

Take the top card and insert it randomly into the deck. Repeat.

We can view this algorithm as a random walk on the group of permutations S_n, which has an underlying Markov chain.

This is aperiodic, irreducible, and finite, so it's ergodic. Therefore it converges to the uniform stationary distribution π.

Strong Stationary Times

Definition

A *strong stationary time* τ for (X_t) is a randomized stopping time such that the distribution of X_τ is π and is independent of τ:

$$\mathbb{P}_x\{\tau = t, X_\tau = y\} = \mathbb{P}_x\{\tau = t\}\pi(y).$$

In other words, it's a stopping time where you can be certain that you've reached a stationary distribution.

Example (Top-to-Random Shuffle)

Let τ_{top} be the randomized stopping time of when the card that was originally at the bottom of the deck is first inserted randomly into the deck. Then, τ_{top} is a strong stationary time.

Bounding Mixing Times

We are interested primarily in

$$t_{\text{mix}} := t_{\text{mix}}\left(\frac{1}{4}\right) = \min\left\{t : d(t) \le \frac{1}{4}\right\},$$

where, recall, $d(t) = \max_{x\in\Omega} \|P^t(x) - \pi\|_{\text{TV}}$.

Proposition

If τ is a strong stationary time, then the maximal total variation distance to the stationary distribution π at time t is bounded:

$$d(t) \le \max_{x\in\Omega} \mathbb{P}_x\{\tau > t\}.$$

Intuition: "if it's unlikely that τ is large, then $d(t)$ should be small."

Bounding Mixing Times

Proposition

$$\max_{x\in\Omega}\|P^t(x)-\pi\|_{TV} =: d(t) \leq \max_{x\in\Omega}\mathbb{P}_x\{\tau > t\}.$$

Proof (which we will glance over).

Fix $x \in \Omega$. Then,

$$\|P^t(x)-\pi\|_{\mathrm{TV}} = \sum_{\substack{y\in\Omega \\ P^t(x,y)<\pi(y)}} \pi(y)\left[\frac{\pi(y)-P^t(x,y)}{\pi(y)}\right]$$

$$\leq \max_{y\in\Omega}\left[\frac{\pi(y)-P^t(x,y)}{\pi(y)}\right] \leq \max_{y\in\Omega}\left[1-\frac{\mathbb{P}_x\{X_t=y,\tau\leq t\}}{\pi(y)}\right]$$

$$= 1-\frac{\mathbb{P}_x\{\tau\leq t\}\pi(y)}{\pi(y)} = \mathbb{P}_x\{\tau > t\},$$

as desired. □

Top-to-Random Shuffle

It suffices for

$$d(t) \le \max_{x\in\Omega} \mathbb{P}_x\{\tau_{\text{top}} > t\} \le \frac{1}{4}.$$

τ_{top} actually behaves the same as **coupon collector problem**, so recall that

$$\mathbb{P}(\tau_{\text{top}} > \lceil n\log n + cn\rceil) \le e^{-c},$$

Therefore, suffices for $c = \log 4$. Hence,

$$t_{\text{mix}} \le n\log n + \log(4)n.$$

For $n = 52$, this gives $\boxed{\text{278 shuffles}}$.

Riffle Shuffle

Now it's time for the real deal: riffle shuffles.

Example (Riffle and Inverse Riffle Shuffles)

- Split deck into top M and bottom $n-M$ with binomial$(n, 1/2)$ distribution. At any point in time, drop bottom card of top pile with probability $\frac{a}{a+b}$ and of bottom pile with probability $\frac{b}{a+b}$.
- Label all the cards either 0 or 1 randomly. Place all cards labeled 0 at the top of the deck.

Riffle Shuffle

Example (Riffle and Inverse Riffle Shuffles)

- Split deck into top M and bottom $n - M$ with binomial$(n, 1/2)$ distribution. At any point in time, drop bottom card of top pile with probability $\frac{a}{a+b}$ and of bottom pile with probability $\frac{b}{a+b}$.
- Label all the cards either 0 or 1 randomly. Place all cards labeled 0 at the top of the deck.

Proposition

The first algorithm generates the distribution Q on S_n, namely

$$Q(\sigma) = \begin{cases} (n+1)/2^n & \textit{if } \sigma = \mathrm{id}, \\ 1/2^n & \textit{if } \sigma \textit{ has exactly two rising sequences}, \\ 0 & \textit{otherwise.} \end{cases}$$

while the second algorithm generates the inverse distribution $\hat{Q}$.

Riffle Shuffle

Luckily,

$$\|P^t(\mathrm{id}) - \pi\|_{\mathrm{TV}} = \|\hat{P}^t(\mathrm{id}) - \pi\|_{\mathrm{TV}}.$$

Only consider the second algorithm.

Proposition

For each card in the deck, keep track of all its bits, writing new bits in a string to the left. Let τ be the number of inverse riffle shuffles at the time when all cards have different binary labels. Then, τ is a strong stationary time.

Proof.

Note that cards with different binary labels will be sorted by size as a binary number. Every binary string is equally likely. □

Riffle Shuffle

Proposition

For the riffle shuffle on an n-card deck, $t_{mix} \leq 2\log_2(4n/3)$ *for sufficiently large n.*

Proof.

$$\mathbb{P}(\tau \leq t) = \prod_{k=0}^{n-1} \left(1 - \frac{k}{2^t}\right),$$

Let $t = 2\log_2(n/c)$ for some constant c. Take logs:

$$\log \prod_{k=0}^{n-1} \left(1 - \frac{k}{2^t}\right) = \sum_{k=0}^{n-1} \log\left(1 - \frac{c^2 k}{n^2}\right) = -\sum_{k=0}^{n-1} \left(\frac{c^2 k}{n^2} + \mathcal{O}\left(\frac{k^2}{n^4}\right)\right)$$

by the Taylor Series expansion $\log(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \cdots$. Now,

$$\log \mathbb{P}(\tau \leq t) = -\frac{c^2 n(n-1)}{2n^2} + \mathcal{O}\left(\frac{n^3}{n^4}\right) = -\frac{c^2}{2} + \mathcal{O}\left(\frac{1}{n}\right).$$

□

Riffle Shuffle

Proposition

For the riffle shuffle on an n-card deck, $t_{mix} \leq 2\log_2(4n/3)$ *for sufficiently large n.*

Proof.

$$\log \mathbb{P}(\tau \leq t) = -\frac{c^2 n(n-1)}{2n^2} + \mathcal{O}\left(\frac{n^3}{n^4}\right) = -\frac{c^2}{2} + \mathcal{O}\left(\frac{1}{n}\right).$$

Therefore,

$$\lim_{n\to\infty} \log \mathbb{P}(\tau \leq t) = -\frac{c^2}{2} \implies \lim_{n\to\infty} \frac{\mathbb{P}(\tau \leq t)}{e^{-c^2/2}} = 1.$$

Now, take $c < \sqrt{2\log(4/3)} \approx 0.759$, so when $n \to \infty$ we have $\mathbb{P}(\tau \leq t) \to \frac{3}{4}$, so $\mathbb{P}(\tau > t) \to \frac{1}{4}$. Therefore, for large n, we can just take $c = \frac{3}{4}$ to get $t_{\text{mix}} \leq 2\log_2(n/(3/4))$, which is the bound in the proposition. □

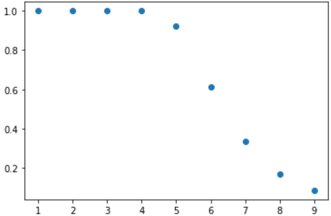
For $n = 52$, this gives about $\boxed{\text{12 shuffles}}$.

Explicit Analysis of Riffle Shuffle

To get a precise estimate, we can explicitly compute the total variation distance

$$\|P^t - \pi\|_{\mathrm{TV}} = \frac{1}{2}\sum_{\sigma\in S_n} |P^t(\sigma) - \pi(\sigma)| = \frac{1}{2}\sum_{\sigma\in S_n} \left|P^t(\sigma) - \frac{1}{52!}\right|.$$

This is actually a nontrivial computation given the $52! \approx 10^{68}$ terms. With tricks:



We see a sharp drop-off at $\boxed{\text{7-8 shuffles}}$.

Conclusion

We would like to thank Slava Gerovitch, Pavel Etingof, Tanya Khovanova, and the MIT PRIMES program, as well as our mentor, Chun Hong Lo. In addition, we would like to thank Zoom for allowing us to meet virtually every week.