Coupling of Markov Chains

Andreas Klappenecker

Texas A&M University

Card Shuffling

Let us consider the following simple procedure to shuffle n cards. Select a card uniformly at random at put it on the top of the deck. Repeat this step.

Observations

This shuffling process is a Markov chain. Any of the $n!$ permutations can be reached from any permutation, so the chain is irreducible. Since with probability $1/n$ the state remains the same, each state is aperiodic, so the Markov chain is aperiodic. Hence the chain has a unique stationary distribution.

Question

What is the stationary distribution of the shuffling Markov chain?

Question

What is the stationary distribution of the shuffling Markov chain?

Answer

The uniform distribution is the stationary distribution on the Markov chain. Indeed, the stationary distribution π satisfies $\pi P = \pi$. More explicitly, if x is a state of the chain and $N(x)$ the set of states that can reach x in the next step, then

$$n = |N(x)|,$$

since the top card in x could have been in n different positions. Thus, we have

$$\pi_x = \frac{1}{n} \sum_{y \in N(x)} \pi_y.$$

Since the uniform distribution satisfies these equations, it must coincide with π.

Question

We know that the stationary distribution is the limiting distribution of the Markov chain. So eventually the states will be uniformly distributed. But we would like to shuffle the cards just a finite number of times.

How many times should we shuffle until the distribution is close to uniform?

Definition

If $p = (p_0, p_1, \ldots, p_{n-1})$ and $q = (q_0, q_1, \ldots, q_{n-1})$ are probability distributions on a finite state space, then

$$d_{TV}(p, q) = \frac{1}{2} \sum_{k=0}^{n-1} |p_k - q_k|$$

is called the **total variation distance** between p and q.

In general, $0 \leqslant d_{TV}(p, q) \leqslant 1$. If $p = q$, then $d_{TV}(p, q) = 0$.

Proposition

Let p_1 and p_2 be discrete probability distributions on a set S. For any subset A of S, we define

$$p_i(A) = \sum_{x \in A} p_i(x).$$

Then

$$d_{TV}(p_1, p_2) = \max_{A \in P(S)} |p_1(A) - p_2(A)|.$$

Proof.

Let $S^{\pm}$ be the set of states such that

$$S^{+} = \{x \in S \mid p_1(x) \geqslant p_2(x)\}$$
$$S^{-} = \{x \in S \mid p_1(x) < p_2(x)\}$$

Then

$$\max_{A \in P(S)} p_1(A) - p_2(A) = p_1(S^{+}) - p_2(S^{+}),$$
$$\max_{A \in P(S)} p_2(A) - p_1(A) = p_2(S^{-}) - p_1(S^{-}).$$

Proof. (Continued)

Since $p_1(S) = p_2(S) = 1$, we have

$$p_1(S^+) + p_1(S^-) = p_2(S^+) + p_2(S^-),$$

hence

$$p_1(S^+) - p_2(S^+) = p_2(S^-) - p_1(S^-).$$

Therefore,

$$\max_{A \in P(S)} |p_1(A) - p_2(A)| = |p_1(S^+) - p_2(S^+)| = |p_1(S^-) - p_2(S^-)|.$$

Proof. (Continued)

Since

$$|p_1(S^+) - p_2(S^+)| + |p_1(S^-) - p_2(S^-)| = \sum_{x \in S} |p_1(x) - p_2(x)|$$
$$= 2d_{TV}(p_1, p_2),$$

we can conclude that

$$\max_{A \in P(S)} |p_1(A) - p_2(A)| = d_{TV}(p_1, p_2).$$

Suppose that we run our shuffling Markov chain until the variation distance between the distribution of the chain and the uniform distribution is less than ϵ.

This is a strong notion of close to uniform, because every permutation of the cards must have probability at most $1/n! + \epsilon$.

The bound on the variation distance gives an even stronger statement: For any subset A of S, the probability that the final permutation is from the set A is at most $\pi(A) + \epsilon$

Example

Suppose someone is trying to make the top card in the deck an ace. If the total variation distance from the distribution p_1 to the uniform distribution p_2 is less than ϵ, then the probability that an ace is the first card of the deck is at most ϵ greater than if we had a perfect shuffle.

Example

As another example, suppose we take a standard 52 card deck and shuffle all the cards, but leave the ace of space on top. In this case, the variation distance between the resulting distribution p_1 and the uniform distribution p_2 could be bounded by considering the set B of states where the ace of space is on the top of the deck.

$$d_{TV}(p_1, p_2) = \max_{A \in P(S)} |p_1(A) - p_2(A)| \geqslant |p_1(B) - p_2(B)|$$
$$= 1 - \frac{1}{52} = \frac{51}{52}.$$

See how easy it is now to obtain a lower bound on the total variation distance?

Notation

Let π be the stationary distribution of a Markov chain with state space S. Let p_x^t denote the distribution of the state of the chain starting at state x after t steps. We define

$$\Delta_x(t) = d_{TV}(p_x^t, \pi).$$

The maximum over all starting states is denoted by

$$\Delta(t) = \max_{x \in S} d_{TV}(p_x^t, \pi).$$

Definition

The **mixing time** $\tau_x(\epsilon)$ of the Markov chain starting in state x is given by

$$\tau_x(\epsilon) = \min\{t : \Delta_x(t) \leqslant \epsilon\}.$$

The **mixing time** $\tau(\epsilon)$ is given by

$$\tau(\epsilon) = \max_{x \in S} \tau_x(\epsilon).$$

A chain is called **rapidly mixing** if and only if $\tau(\epsilon)$ is polynomial in $\log(1/\epsilon)$ and the size of the problem.

Coupling

Motivation

Coupling of Markov chains is a general technique for bounding the mixing time of a Markov chain.

Definition

A **coupling** of a Markov chain M_t with state space S is a Markov chain $Z_t = (X_t, Y_t)$ on the state space $S \times S$ such that

$$\Pr[X_{t+1} = x' \mid Z_t = (x, y)] = \Pr[M_{t+1} = x' \mid M_t = x],$$
$$\Pr[Y_{t+1} = y' \mid Z_t = (x, y)] = \Pr[M_{t+1} = y' \mid M_t = y].$$

In other words, a coupling consists of two copies of the Markov chain M running simultaneously. These two copies are not literal copies; the two chains are not necessarily in same state, nor do they necessarily make the same move. Instead, we mean that each copy behaves exactly like the original Markov chain in terms of its transition probabilities.

We are interested in couplings that

1. bring the two copies of the chain to the same state and then
2. keep them in the same state by having the two chains identical moves once they are in the same state.

When the two copies of the chain reach the same state, they are said to have coupled.

Lemma

Let $Z_t = (X_t, Y_t)$ be a coupling for a Markov chain M on a state space S. Suppose that there exists a T such that for every x, y in S

$$\Pr[X_T \neq Y_T \mid X_0 = x, Y_0 = y] \leqslant \epsilon.$$

Then the mixing time after T steps is at most ϵ, so

$$\tau(\epsilon) \leqslant T.$$

In other words, the total variation distance between the distribution of the chain after T steps and the stationary distribution is at most ϵ.

Proof.

Let X_0 be an arbitrarily chosen value and let Y_0 be chosen according to the stationary distribution. For the given T and ϵ and for any subset A of the set of states S, we have

$$\begin{aligned}
\Pr[X_T \in A] &\geqslant \Pr[(X_T = Y_T) \wedge (Y_T \in A)] \\
&= 1 - \Pr[(X_T \neq Y_T) \vee (Y_T \notin A)] \\
&\geqslant 1 - \Pr[X_T \neq Y_T] - \Pr[Y_T \notin A] \\
&\geqslant \Pr[Y_T \in A] - \epsilon \\
&= \pi(A) - \epsilon.
\end{aligned}$$

The same argument for the set $S - A$ shows that

$$\Pr[X_T \notin A] \geqslant \pi(S - A) - \epsilon,$$

whence

$$\Pr[X_T \in A] \leqslant \pi(A) + \epsilon.$$

Proof. (Continued)

It follows that

$$\max_{x,A} |p_x^T(A) - \pi(A)| \leqslant \epsilon.$$

By the previous proposition, the total variation distance from the stationary distribution is bounded by ϵ. So

$$\tau(\epsilon) \leqslant T.$$

Card Shuffling

Let us analyze how quickly the card shuffling procedure converges to a perfect shuffle.

Recall that in each step, we choose one card uniformly at random and place it on top.

Definition

We will now define a coupling. Choose a position j uniformly at random from 1 to n and then obtain X_{t+1} from X_t by moving the j-th card to the top. Denote the value of this card by C.

To obtain Y_{t+1} from Y_t, move the card with value C to the top.

The coupling is valid, because in both chains the probability a specific card is moved to the top at each step is $1/n$.

Observation

Once a card C is moved to the top, it is always in the same position in both copies of the chain.

Hence, the two copies are sure to become coupled once every card has been moved to the top at least once.

We can bound the number of steps until the chains couple by bounding how many times cards must be chosen uniformly at random before **every card is chosen at least once**.

If the Markov chain runs for at least $n \ln n + cn$ steps, then the probability that a specific card has not been moved to the top at least once is at most

$$\left(1 - \frac{1}{n}\right)^{n \ln n + cn} \leqslant e^{-(\ln n + c)} = \frac{e^{-c}}{n}.$$

By the union bound, the probability that any card has not been moved to the top at least once is at most e^{-c}. Hence, after only

$$n \ln n + n \ln(1/\epsilon) = n \ln(n/\epsilon)$$

steps, the probability that the chains have not coupled is at most ϵ.

The coupling lemma allows us to conclude that the variation distance between the uniform distribution and the distribution of the state of the chain after $n \ln(n/\epsilon)$ steps is bounded above by ϵ.

Random Walk on the Hypercube

Definition

The hypercube has 2^n vertices that are labeled by bit strings of length n.

Two vertices u and v are connected by an edge if and only if their labels differ in exactly one bit.

Markov Chain

At each step, choose a coordinate i uniformly at random from $\{0, \ldots, n-1\}$. The new state x' is obtained from the current state x by keeping all coordinates of x the same, except possibly for x_i. The coordinate x_i is set to 0 with probability $1/2$ and to 1 with probability $1/2$.

Remark

This Markov chain is exactly the random walk on the hypercube, except that with probability $1/2$ the chain stays at the same vertex instead of moving to a new one, so the chain is aperiodic. Evidently, the chain is also irreducible.

Proposition

The stationary distribution of the Markov chain is the uniform distribution.

Indeed, the uniform distribution is reversible for this chain. Since this is an aperiodic irreducible finite Markov chain, the uniform distribution is the unique stationary distribution.

Coupling

We bound the mixing time $\tau(\epsilon)$ of this Markov chain by using the obvious coupling between two copies X_t and Y_t of the Markov chain: at each step, we have both chains make the same move.

With this coupling, the two copies of the chain will surely agree on the i-th coordinate, once the i-th coordinate has been chosen for a move of the Markov chain. Hence the chains will have coupled after all n coordinates have each been chosen at least once.

Mixing Time

The mixing time can therefore be bounded by bounding the number of steps until each coordinate has been chosen at least once by the Markov chain. As in the card shuffling, the probability is less than ϵ that after $n \ln(n/\epsilon)$ steps the chains have not coupled. By the coupling lemma, the mixing time satisfies

$$\tau(\epsilon) \leqslant n \ln(n/\epsilon).$$

This is a rapidly mixing Markov chain.

Convergence to the Stationary Distribution

Proposition

Any finite irreducible aperiodic Markov chain converges to a unique stationary distribution in the limit.

Lemma

For any discrete random variables X and Y, we have

$$d_{TV}(X, Y) \leqslant \Pr[X \neq Y].$$

Proof.

Let A be an event for which $\Pr[X \in A]$ and $\Pr[Y \in A]$ are defined. Then

$$\Pr[X \in A] = \Pr[X \in A \wedge Y \in A] + \Pr[X \in A \wedge Y \notin A]$$
$$\Pr[Y \in A] = \Pr[X \in A \wedge Y \in A] + \Pr[X \notin A \wedge Y \in A].$$

Therefore,

$$\begin{aligned}\Pr[X \in A] - \Pr[Y \in A] &= \Pr[X \in A \wedge Y \notin A] - \Pr[X \notin A \wedge Y \in A]. \\ &\leqslant \Pr[X \in A \wedge Y \notin A] \\ &\leqslant \Pr[X \neq Y].\end{aligned}$$

Thus, we get

$$d_{TV}(X, Y) = \max_A |\Pr[X \in A] - \Pr[Y \in A]| \leqslant \Pr[X \neq Y].$$

□

Proof.

Consider two copies of the chain $\{X_t\}$ and $\{Y_t\}$, where X_0 starts in some arbitrary distribution x and Y_0 starts in a stationary distribution π. Define a coupling between $\{X_t\}$ and $\{Y_t\}$ by the following rule:

1. if $X_t \neq Y_t$, then $\Pr[X_{t+1} = j \wedge Y_{t+1} = j' \mid X_t = i \wedge Y_t = i'] = p_{ij}p_{i'j'}$.
2. if $X_t = Y_t$, then $\Pr[X_{t+1} = Y_{t+1} = j \mid X_t = Y_t = i] = p_{ij}$.

Intuitively, we let both chains run independently until they collide, after which we run them together.

Since each chain individually moves from state i to state j with probability p_{ij} in either case, we have that X_t evolves normally and Y_t remains in the stationary distribution.

Proof. (Continued)

By the second coupling lemma,

$$d_{TV}(p_x^t, \pi) = \max_A |p_x^t(A) - \pi(A)| \leqslant \Pr[X_t \neq Y_t],$$

so it suffices to show that

$$\lim_{t \to \infty} \Pr[X_t \neq Y_t] = 0.$$

Proof. (Continued)

Consider a state i. The first passage time from i to j is the minimum time t such that $p_{ij}^t \neq 0$. Let r be the maximum of all first passage times. Let s be a time such that $p_{ii}^t \neq 0$ for all $t \geqslant s$. Suppose that at time $\ell(r+s)$, we have

$$X_{\ell(r+s)} = j \neq j' = Y_{\ell(r+s)}.$$

Then there are times $\ell(r+s)+u$ and $\ell(r+s)+u'$, where $u, u' \leqslant r$, such that X reaches i at time $\ell(r+s)+u$ and Y reaches i at time $\ell(r+s)+u'$ with nonzero probability.

Proof. (Continued)

Since $(r+s-u) \geqslant s$ and $(r+s-u') \geqslant s$, after having reached i at these times, X and Y both return to i at time $\ell(r+s)+(r+s) = (l+1)(r+s)$ with nonzero probability. Let $\epsilon > 0$ be the product of these nonzero probabilities. Then

$$\Pr[X_{(\ell+1)(r+s)} \neq Y_{(\ell+1)(r+s)}] \leqslant (1-\epsilon)\Pr[X_{\ell(r+s)} \neq Y_{\ell(r+s)}].$$

In general, we have

$$\Pr[X_t \neq Y_t] \leqslant (1-\epsilon)^{\lfloor t/(r+s) \rfloor},$$

whence

$$\lim_{t\to\infty} \Pr[X_t \neq Y_t] = 0. \qquad \square$$