The Markov Chain Monte Carlo Method

Idea: define an ergodic Markov chain whose stationary distribution is the desired probability distribution.

Let $X_0, X_1, X_2, \ldots, X_n$ be the run of the chain.

The Markov chain converges to its stationary distribution from any starting state X_0 so after some sufficiently large number r of steps, the distribution at of the state X_r will be close to the stationary distribution π of the Markov chain.

Now, repeating with X_r as the starting point we can use X_{2r} as a sample etc.

So $X_r, X_{2r}, X_{3r}, \ldots$ can be used as almost independent samples from π.

The Markov Chain Monte Carlo Method

Consider a Markov chain whose states are independent sets in a graph $G = (V, E)$:

1. X_0 is an arbitrary independent set in G.
2. To compute X_{i+1}:
   1. Choose a vertex v uniformly at random from V.
   2. If $v \in X_i$ then $X_{i+1} = X_i \setminus \{v\}$;
   3. if $v \notin X_i$, and adding v to X_i still gives an independent set, then $X_{i+1} = X_i \cup \{v\}$;
   4. otherwise, $X_{i+1} = X_i$.

- The chan is irreducible
- The chain is aperiodic
- For $y \neq x$, $P_{x,y} = 1/|V|$ or 0.

$N(x)-$ set of neighbors of x. Let $M \geq \max_{x \in \Omega} |N(x)|$.

Lemma

Consider a Markov chain where for all x and y with $y \neq x$, $P_{x,y} = \frac{1}{M}$ if $y \in N(x)$, and $P_{x,y} = 0$ otherwise. Also, $P_{x,x} = 1 - \frac{|N(x)|}{M}$. If this chain is irreducible and aperiodic, then the stationary distribution is the uniform distribution.

Proof.

We show that the chain is time-reversible, and apply Theorem 7.10. For any $x \neq y$, if $\pi_x = \pi_y$, then

$$\pi_x P_{x,y} = \pi_y P_{y,x},$$

since $P_{x,y} = P_{y,x} = 1/M$. It follows that the uniform distribution $\pi_x = 1/|\Omega|$ is the stationary distribution. □

The Metropolis Algorithm

Assuming that we want to sample with non-uniform distribution. For example, we want the probability of an independent set of size i to be proportional to λ^i.

Consider a Markov chain on independent sets in $G = (V, E)$:

1. X_0 is an arbitrary independent set in G.
2. To compute X_{i+1}:
    1. Choose a vertex v uniformly at random from V.
    2. If $v \in X_i$ then set $X_{i+1} = X_i \setminus \{v\}$ with probability $\min(1, 1/\lambda)$;
    3. if $v \notin X_i$, and adding v to X_i still gives an independent set, then set $X_{i+1} = X_i \cup \{v\}$ with probability $\min(1, \lambda)$;
    4. otherwise, set $X_{i+1} = X_i$.

Lemma

For a finite state space Ω*, let* $M \geq \max_{x\in\Omega} |N(x)|$*. For all* $x \in \Omega$*, let* $\pi_x > 0$ *be the desired probability of state* x *in the stationary distribution. Consider a Markov chain where for all* x *and* y *with* $y \neq x$*,*

$$P_{x,y} = \frac{1}{M} \min\left(1, \frac{\pi_y}{\pi_x}\right)$$

if $y \in N(x)$*, and* $P_{x,y} = 0$ *otherwise. Further,* $P_{x,x} = 1 - \sum_{y\neq x} P_{x,y}$*. Then if this chain is irreducible and aperiodic, the stationary distribution is given by the probabilities* π_x*.*

Proof.

We show the chain is time-reversible. For any $x \neq y$, if $\pi_x \leq \pi_y$, then $P_{x,y} = \frac{1}{M}$ and $P_{y,x} = \frac{1}{M}\frac{\pi_x}{\pi_y}$. It follows that $\pi_x P_{x,y} = \pi_y P_{y,x}$. Similarly, if $\pi_x > \pi_y$, then $P_{x,y} = \frac{1}{M}\frac{\pi_y}{\pi_x}$ and $P_{y,x} = \frac{1}{M}$, and it follows that $\pi_x P_{x,y} = \pi_y P_{y,x}$. □

Note that the Metropolis Algorithm only needs the ratios π_x/π_y's. In our construction, the probability of an independent set of size i is λ^i/B for $B = \sum_x \lambda^{size(x)}$ although we may not know B.

Coupling and MC Convergence

- An Ergodic Markov Chain converges to its stationary distribution.
- How long do we need to run the chain until we sample a state in almost the stationary distribution?
- How do we measure distance between distributions?
- How do we analyze speed of convergence?

Variation Distance

Definition

The *variation distance* between two distributions D_1 and D_2 on a countably finite state space S is given by

$$||D_1 - D_2|| = \frac{1}{2} \sum_{x \in S} |D_1(x) - D_2(x)|.$$

See Figure 11.1 in the book:
The total area shaded by upward diagonal lines must equal the total areas shaded by downward diagonal lines, and the variation distance equals one of these two areas.

Lemma

For any $A \subseteq S$, *let* $D_i(A) = \sum_{x \in A} D_i(x)$, *for* $i = 1, 2$. *Then,*

$$||D_1 - D_2|| = \max_{A \subseteq S} |D_1(A) - D_2(A)|.$$

Let $S^+ \subseteq S$ be the set of states such that $D_1(x) \geq D_2(x)$, and $S^- \subseteq S$ be the set of states such that $D_2(x) > D_1(x)$.
Clearly

$$\max_{A \subseteq S} D_1(A) - D_2(A) = D_1(S^+) - D_2(S^+),$$

and

$$\max_{A \subseteq S} D_2(A) - D_1(A) = D_2(S^-) - D_1(S^-).$$

But since $D_1(S) = D_2(S) = 1$, we have

$$D_1(S^+) + D_1(S^-) = D_2(S^+) + D_2(S^-) = 1,$$

which implies that

$$D_1(S^+) - D_2(S^+) = D_2(S^-) - D_1(S^-).$$

$$\max_{A\subseteq S}|D_1(A)-D_2(A)| = |D_1(S^+)-D_2(S^+)| = |D_1(S^-)-D_2(S^-)|.$$

and

$$|D_1(S^+)-D_2(S^+)| + |D_1(S^-)-D_2(S^-)| = \sum_{x\in S}|D_1(x)-D_2(x)|$$

$$= 2||D_1-D_2||,$$

we have

$$\max_{A\subseteq S}|D_1(A)-D_2(A)| = ||D_1-D_2||,$$

Rate of Convergence

Definition

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

$$\Delta_x(t) = ||p_x^t - \pi|| \,; \qquad \Delta(t) = \max_{x \in S} \Delta_x(t).$$

That is, $\Delta_x(t)$ is the variation distance between the stationary distribution and p_x^t, and $\Delta(t)$ is the maximum of these values over all states x.
We also define

$$\tau_x(\epsilon) = \min\{t : \Delta_x(t) \leq \epsilon\} \,; \qquad \tau(\epsilon) = \max_{x \in S} \tau_x(\epsilon).$$

That is, $\tau_x(\epsilon)$ is the first step t at which the variation distance between p_x^t and the stationary distribution is less than ϵ, and $\tau(\epsilon)$ is the maximum of these values over all states x.

Coupling

Definition

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

$$\begin{aligned} \Pr(X_{t+1} = x' | Z_t = (x, y)) &= \Pr(X_{t+1} = x' | X_t = x); \\ \Pr(Y_{t+1} = y' | Z_t = (x, y)) &= \Pr(Y_{t+1} = y' | Y_t = y). \end{aligned}$$

The Coupling Lemma

Lemma (Coupling 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 *so that for every* $x, y \in S$,

$$\Pr(X_T \neq Y_T \mid X_0 = x, Y_0 = y) \leq \epsilon.$$

Then

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

That is, for any initial state, the variation distance between the distribution of the state of the chain after T *steps and the stationary distribution is at most* ϵ.

Proof:

Consider the coupling when Y_0 is chosen according to the stationary distribution and X_0 takes on any arbitrary value. For the given T and ϵ, and for any $A \subseteq S$

$$\begin{aligned}
\Pr(X_T \in A) &\geq \Pr((X_T = Y_T) \cap (Y_T \in A)) \\
&= 1 - \Pr((X_T \neq Y_T) \cup (Y_T \notin A)) \\
&\geq (1 - \Pr(Y_T \notin A)) - \Pr(X_T \neq Y_T) \\
&\geq \Pr(Y_T \in A) - \epsilon \\
&= \pi(A) - \epsilon.
\end{aligned}$$

Here we used that when Y_0 is chosen according to the stationary distribution, then, by the definition of the stationary distribution, $Y_1, Y_2, \ldots, Y_T$ are also distributed according to the stationary distribution.

Similarly,

$$\Pr(X_T \notin A) \geq \pi(S \setminus A) - \epsilon$$

or

$$\Pr(X_T \in A) \leq \pi(A) + \epsilon$$

It follows that

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

.

Example: Shuffling Cards

- Markov chain:
  - States: orders of the deck of n cards. There are $n!$ states.
  - Transitions: at each step choose one card, uniformly at random, and move to the top.
- The chain is irreducible: we can go from any permutation to any other using only moves to the top (at most n moves).
- The chain is aperiodic: it has loops as top card is chosen with probability $\frac{1}{n}$.
- Hence, by Theorem 7.10, the chain has a stationary distribution π.

- A given state x of the chain has $|N(x)| = n$: the new top card can be anyone of the n cards.
- Let π_y be the probability of being in state y under π, then for any state x:

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

- $\pi = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$ is a solution and hence the stationary distribution is the uniform stationary distribution

- Given two such chains: X_t and Y_t we define the coupling:
  - The first chain chooses a card uniformly at random and moves it to the top.
  - The second chain moves the same card (it may be in a different location) to the top.
  - Once a card is on the top in both chains at the same time it will remain in the same position in both chains!
  - Hence we are sure the chains will be equal once every card has been picked at least once.
  - So we can use the coupon collector argument:
  - after running the chain for at least $n \ln n + cn$ steps the probability that a specific card (e.g. ace of spades) has not been moved to the top yet is at most

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

- Hence the probability that there is some card which was not chosen by the first chain in $n \ln n + cn$ steps is at most e^{-c}.
- After $n \ln n + n \ln (1/\epsilon) = n \ln (n/\epsilon)$ steps the variation distance between our chain and the uniform distribution is bounded by ϵ, implying that

$$\tau(\epsilon) \leq n \ln(n\epsilon^{-1}).$$

Example: Random Walks on the Hypercube

- Consider n-cube, with $N = 2^n$ nodes., Let $\bar{x} = (x_1, \ldots, x_n)$ be the binary representation of x. Nodes x and y are connected by an edge iff $\bar{x}$ and $\bar{y}$ differ in exactly one bit.
- Markov chain on the n-cube: at each step, choose a coordinate i uniformly at random from $[1, n]$, and set x_i to 0 with probability $1/2$ and 1 with probability $1/2$.
- The chain is irreducible, finite and and aperiodic so it has a unique stationary distribution π.

- A given state x of the chain has $|N(x)| = n + 1$: the
- Let π_y be the probability of being in state y under π, then for any state x:

$$\begin{aligned}\pi_x &= \sum_{y \in N(x)} \pi_y P_{y,x} \\ &= \frac{1}{2}\pi_x + \frac{1}{2n} \sum_{y \in N(x) \setminus x} \pi_y\end{aligned}$$

- $\pi = (\frac{1}{2^n}, \frac{1}{2^n}, \ldots, \frac{1}{2^n})$ solves this and hence it is the stationary distribution.

- Coupling: both chains choose the same bit and give it the same value.
- The chains couple when all bits have been chosen.
- By the Coupling Lemma the mixing time satisfies

$$\tau(\epsilon) \leq n \ln(n\epsilon^{-1}).$$

Example: Sampling Independent Sets of a Given Size

Consider a Markov chain whose states are independent sets of size k in a graph $G = (V, E)$:

1. X_0 is an arbitrary independent set of size k in G.
2. To compute X_{t+1}:
   (a) Choose uniformly at random $v \in X_t$ and $w \in V$.
   (b) if $w \notin X_t$, and $(X_t - \{v\}) \cup \{w\}$ is an independent set, then $X_{t+1} = (X_t - \{v\}) \cup \{w\}$
   (c) otherwise, $X_{t+1} = X_t$.

- Assume $k \leq n/(3\Delta + 3)$, where Δ is the maximum degree.
- The chain is irreducible as we can convert an independent set X of size k into any other independent set Y of size k using the operation above (exercise 11.11).
- The chain is aperiodic as there are loops.
- For $y \neq x$, $P_{x,y} = 1/|V|$ (if they differ in exactly one vertex) or 0.
- By Lemma 10.7 the stationary distribution is the uniform distribution.

Convergence Time

Theorem

Let G be a graph on n vertices with maximum degree $\leq \Delta$. For $k \leq n/(3\Delta + 3)$,

$$\tau(\epsilon) \leq O(kn \ln \epsilon^{-1}).$$

Coupling:

1. X_0 and Y_0 are arbitrary independent sets of size k in G.
2. To compute X_{t+1} and Y_{t+1}:
    1. Choose uniformly at random $v \in X_t$ and $w \in V$.
    2. if $w \notin X_t$, and $(X_t - \{v\}) \cup \{w\}$ is an independent set, then $X_{t+1} = (X_t - \{v\}) \cup \{w\}$, otherwise, $X_{t+1} = X_t$.
    3. If $v \notin Y_t$ choose v' uniformly at random from $Y_t - X_t$, else $v' = v$.
    4. if $w \notin Y_t$, and $(Y_t - \{v'\}) \cup \{w\}$ is an independent set, then $Y_{t+1} = (Y_t - \{v'\}) \cup \{w\}$, otherwise, $Y_{t+1} = Y_t$.

Let $d_t = |X_t - Y_t|$,

- $|d_{t+1} - d_t| \leq 1$.
- $d_{t+1} = d_t + 1$: must be $v \in X_t \cap Y_t$ and there is move in only one chain. Either w or some neighbor of w must be in $(X_t - Y_t) \cup (Y_t - X_t)$

$$\Pr(d_{t+1} = d_t + 1) \leq \frac{k - d_t}{k} \frac{2d_t(\Delta + 1)}{n}.$$

- $d_{t+1} = d_t - 1$: sufficient $v \notin Y_t$ and w and its neighbors are not in $X_t \cup Y_t - \{v, v'\}$. $|X_t \cup Y_t| = k + d_t$

$$\Pr(d_{t+1} = d_t - 1) \geq \frac{d_t}{k} \frac{n - (k + d_t - 2)(\Delta + 1)}{n}.$$

Conditional expectation:
There are only 3 possible valuse for d_{t+1} given the value of $d_t > 0$, namely $d_t - 1, d_t, d_t + 1$. Hence, using the formula for conditional expectation we have

$$\begin{aligned}
\mathbf{E}[d_{t+1} \mid d_t] &= (d_t + 1)\Pr(d_{t+1} = d_t + 1) + d_t \Pr(d_{t+1} = d_t) + (d_t - 1)\Pr(d_{t+1} = d_t - 1) \\
&= d_t(\Pr(d_{t+1} = d_t - 1) + \Pr(d_{t+1} = d_t) + \Pr(d_{t+1} = d_t + 1)) \\
&+ \Pr(d_{t+1} = d_t + 1) - \Pr(d_{t+1} = d_t - 1) \\
&= d_t + \Pr(d_{t+1} = d_t + 1) - \Pr(d_{t+1} = d_t - 1)
\end{aligned}$$

Now we have for $d_t > 0$,

$$\begin{aligned}
\mathbf{E}[d_{t+1} \mid d_t] &= d_t + \Pr(d_{t+1} = d_t + 1) - \Pr(d_{t+1} = d_t - 1) \\
&\leq d_t + \frac{k - d_t}{k} \frac{2d_t(\Delta + 1)}{n} - \frac{d_t}{k} \frac{n - (k + d_t - 2)(\Delta + 1)}{n} \\
&= d_t \left(1 - \frac{n - (3k - d_t - 2)(\Delta + 1)}{kn}\right) \\
&\leq d_t \left(1 - \frac{n - (3k - 3)(\Delta + 1)}{kn}\right).
\end{aligned}$$

Once $d_t = 0$, the two chains follow the same path, thus
$\mathbf{E}[d_{t+1} \mid d_t = 0] = 0$.

$$\mathbf{E}[d_{t+1}] = \mathbf{E}[\mathbf{E}[d_{t+1} \mid d_t]] \leq \mathbf{E}[d_t] \left(1 - \frac{(n - 3k + 3)(\Delta + 1)}{kn}\right).$$

$$\mathbf{E}[d_t] \leq d_0 \left(1 - \frac{n - (3k + 3)(\Delta + 1)}{kn}\right)^t.$$

$$\mathbf{E}[d_t] \leq d_0 \left(1 - \frac{n - (3k+3)(\Delta+1)}{kn}\right)^t.$$

Since $d_0 \leq k$, and d_t is a non-negative integer,

$$\Pr(d_t \geq 1) \leq \mathbf{E}[d_t] \leq k \left(1 - \frac{n - (3k-3)(\Delta+1)}{kn}\right)^t \leq k\mathrm{e}^{-t\frac{n-(3k-3)(\Delta+1)}{kn}}.$$

For $k \leq n/(3\Delta + 3)$ the variantion distance converges to zero and

$$\tau(\epsilon) \leq \frac{kn \ln(k\epsilon^{-1})}{n - (3k-3)(\Delta+1)}.$$

In particular, when k and Δ are constants, $\tau(\epsilon) = O(\ln \epsilon^{-1})$.

Theorem

Given two distributions σ_X *and* σ_Y *on a state space* S*, Let* $Z = (X, Y)$ *be a random variable on* $S \times S$*, where* X *is distributed according to* σ_X *and* Y *is distributed according to* σ_Y*. Then*

$$\Pr(X \neq Y) \geq \|\sigma_X - \sigma_Y\|.$$

Moreover, there exists a joint distribution $Z = (X, Y)$*, where* X *is distributed according to* σ_X *and* Y *is distributed according to* σ_Y*, for which equality holds.*

Variation distance is nonincreasing

Recall that $\delta(t) = \max_x \Delta_x(t)$, where $\Delta_x(t)$ is the variational distance bewteen the stationary distribution and the distribution of the state of the Markov chain after t steps when it starts at state x.

Theorem

For any ergodic Markov chain M_t, $\Delta(t+1) \leq \Delta(t)$.