[AI2613 Lecture 4] Metropolis Algorithm, Countable Infinite Markov Chain

March 23, 2023

1 Reversible Markov Chains

A Markov chain P over state space $[n]$ is *(time) reversible* if there exists some distribution π satisfying

$$\forall i, j \in [n],\ \pi(i)P(i,j) = \pi(j)P(j,i).$$

This family of identities is called *detailed balance conditions.* Moreover, the distribution π must be a stationary distribution of P. To see this, note that

$$\pi^{\mathsf{T}}P(j) = \sum_{i\in[n]} \pi(i)P(i,j) = \sum_{i\in[n]} \pi(j)P(j,i) = \pi(j).$$

The name *reversible* comes from the fact that for any sequence of variables $X_0, X_1, \ldots, X_t$ following the chain which start from the stationary distribution, the distribution of $(X_0, X_1, \ldots, X_{t-1}, X_t)$ is identical to the distribution of $(X_t, X_{t-1}, \ldots, X_1, X_0)$, namely for all $x_0, x_2, \ldots x_t \in [n]$,

$$\begin{aligned}
&\mathbf{Pr}_{X_0\sim\pi}\left[X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t\right] \\
&= \pi(x_0)P(x_0,x_1)\cdots P(x_{t-1},x_t) \\
&= \pi(x_t)P(x_t,x_{t-1})\cdots P(x_1,x_0) \\
&= \mathbf{Pr}_{X_0\sim\pi}\left[X_0 = x_t, X_1 = x_{t-1}, \ldots, X_t = x_0\right]
\end{aligned}$$

We will study reversible chains since their transition matrices are essentially *symmetric* in some sense, so many powerful tools in linear algebra apply. We will also see that reversible chains are general enough for most of our (algorithmic) applications. You can verify that the the random walks on the hypercube is reversible Markov chains with respect to uniform distribution.

Recall the two conditions of FTMC: irreducibility and aperiodicity. Since the transition graph is undirected if we only consider the connectivity, irreducibility is equivalent to the connectivity of the transition graph. Aperiodicity, on the other hand, is equivalent to that the graph is *not* bipartite.

2 The Metropolis Algorithm

Given a distribution π over a state space Ω, how can we design a Markov chain P so that π is the stationary distribution of P? The *Metropolis algorithm* provides a way to achieve the goal as long as the transition graph G is connected and undirected.

Let Δ be the maximum degree of the transition graph except selfloop (that is $\Delta \triangleq \max_{u\in[n]} \sum_{v\neq u\in[n]} \mathbb{1}[(u,v) \in E]$). We describe the following process to construct a transition matrix P: Choose $k \in [\Delta + 1]$ uniformly at random. For any $i \in [n]$, let $\{j_1, j_2, \dots, j_d\}$ be the d neighbours of i. We consider the transition at state i:

- If $d + 1 \le k \le \Delta + 1$, do nothing.
- If $k \le d$,
  - propose to move from i to j_k.
  - accept the proposal with probability $\min\left\{\frac{\pi(j_k)}{\pi(i)}, 1\right\}$.

Then the transition matrix is, for $i, j \in [n]$,

$$P(i,j) = \begin{cases} \frac{1}{\Delta+1}\min\left\{\frac{\pi(j)}{\pi(i)}, 1\right\}, & \text{if } i \neq j; \\ 1 - \sum_{k \neq i} P(i,k), & \text{if } i = j. \end{cases}$$

We can verify that P is reversible with respect to π:

$$\forall i, j \in \Omega :$$
$$\pi(i)P(i,j) = \pi(i) \cdot \frac{1}{\Delta+1}\min\left\{\frac{\pi(j)}{\pi(i)}, 1\right\} = \frac{\min\{\pi(i), \pi(j)\}}{\Delta+1} = \pi(j)P(j,i).$$

The advantage of the Metropolis algorithm is that we do not need to know π in order to implement the algorithm. We only need to know the quantity $\frac{\pi(j)}{\pi(i)}$, which is much easier to compute in many applications.

Example 1 *We give a toy example to show how the algorithm works. Consider a graph with* 3 *vertices* $\{a, b, c\}$. *There are undirected edges between* (a, b), (b, c) *and* (a, c) *and selfloops for each vertex. In this situation,* $\Delta = 2$. *If we want to design a transition matrix* P *with stationary distribution* $(\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$, *by Metropolis algorithm we have*

$$P(a,b) = \frac{1}{2+1} \cdot \frac{2}{3} = \frac{2}{9},$$
$$P(a,c) = \frac{1}{2+1} \cdot \frac{1}{3} = \frac{1}{9},$$
$$P(a,a) = 1 - \frac{1}{9} - \frac{2}{9} = \frac{2}{3}.$$

3 *Sample Proper Coloring*

Let's consider the problem of sampling proper colorings. Given a graph $G = (V, E)$, we want to color the vertices using q colors under the condition that no two adjacent vertices share the same color. More formally, a coloring of G is a mapping $c : V \mapsto [q]$, and we call it *proper* iff $\forall \{u, v\} \in E, c(u) \neq c(v)$. The proper coloring problem is **NP**-hard in general. However, for $q > \Delta$ there always exists a proper coloring that can be easily obtained by a greedy algorithm, where Δ is the maximum degree of the graph.

If we want to count the number of proper colorings, then the problem becomes harder. It is known that for every $q \ge \Delta$, the problem is #**P**-hard. On the other hand, we can use a uniform sampler to obtain an algorithm to approximately counting the number of proper coloring, at an arbitrarily low cost in the precision.

In fact, it is known that an approximate counting algorithm is equivalent to an uniform sampler in many cases (for example, sampling proper coloring). We only show one direction here: a sampler implies an algorithm for approximate counting. Given a graph $G = (V, E)$ with $V = [n]$, let C be the set of proper colorings and $Z = |C|$. Suppose we have an oracle that can uniformly generate a proper coloring from C. Fix a proper coloring σ. We have

$$\begin{aligned}\frac{1}{Z} &= \mathbf{Pr}_{x\sim C}\left[x=\sigma\right]\\ &= \mathbf{Pr}_{x\sim C}\left[x(1)=\sigma(1)\wedge x(2)=\sigma(2)\wedge\dots\right]\\ &= \prod_{i=1}^{n}\mathbf{Pr}\left[x(i)=\sigma(i)\;\middle|\;\bigcap_{j<i}x(j)=\sigma(j)\right].\end{aligned}$$

The above probability can be estimated by taking a number of samples from the oracle, and computing the ratio between colorings such that $x(j) = \sigma(j)$ for $j \leq i$ and ones that $x(j) = \sigma(j)$ for $j < i$. Moreover, the ratio we just estimated is bounded below by an inverse polynomial and therefore polynomial number of sample suffices to estimate ratio accurately. The strategy works even if the sampler is an approximate one. Hence one can approximately compute Z. See [JVV86] for more details.

Now we use MCMC to do sampling. Consider the following Markov chain to sample proper colorings:

- Pick $v \in V$ and $c \in [q]$ uniformly at random.
- Recolor v with c if possible.

The chain is aperiodic since self-loops exist in the walk. For $q \geq \Delta + 2$, the chain is irreducible. The bound $q \geq \Delta + 2$ is tight for irreducibility since when $q = \Delta + 1$, each proper coloring of complete graph is frozen. It is still an open problem if the mixing time of the chain is polynomial in the size of the graph under the condition $q \geq \Delta + 2$. The best bound so far requires that $q \geq (\frac{11}{6} - \varepsilon)\Delta$ for a certain constant $\varepsilon > 0$. Here, we shall give a rapid mixing proof when $q > 4\Delta$ using the method of coupling.

It is indeed a Metropolis algorithm. Let

$$\sigma^{v\leftarrow c}(u) = \begin{cases}\sigma(u) & u \neq v\\ c & u = v.\end{cases}$$

$\sigma^{v\leftarrow c}$ is a neighbor of σ on the transition graph, and we accept it if $\sigma^{v\leftarrow c}$ is a proper coloring, i.e. $\frac{\pi(\sigma^{v\leftarrow c})}{\pi(\sigma)} = 1$.

The coupling we used is simple: Both players pick same v and c to move. However, we are not able to reduce the analyze the coupling to *coupon collector* as we did before. We introduce a more general method to analyze couplings. We define a certain distance $d(x, y)$ for any two configurations $x, y \in \Omega$. We can assume without loss of generality that if $x \neq y$ then $d(x, y) \geq 1$ since Ω is finite. Consider a coupling ω_t of μ_t, ν_v. Then for every $t \geq 0$ and $(X_t, Y_t) \sim \omega_t$, we try to establish

$$\mathbf{E}\left[d(X_{t+1}, Y_{t+1}) \mid (X_t, Y_t)\right] \leq (1-\alpha)d(X_t, Y_t)$$

for some $\alpha \in (0, 1]$. In other words, $\{d(X_t, Y_t)\}_{t\geq 0}$ is a supermartingale. This implies that for every $t \geq 1$,

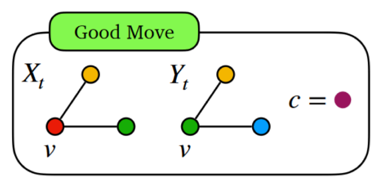
$$\mathbf{E}\left[d(X_t, Y_t)\right] \leq (1-\alpha)\mathbf{E}\left[d(X_{t-1}, Y_{t-1})\right] \leq (1-\alpha)^t d(X_0, Y_0).$$

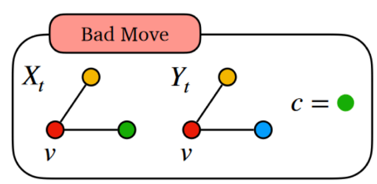
If we have a universal upper bound for $d(X_0, Y_0)$, say n, then by coupling lemma

$$\begin{aligned} D_{\mathrm{TV}}(\mu_t, \nu_t) &\leq \mathbf{Pr}_{(X_t,Y_t)\sim\omega_t}\left[X_t \neq Y_t\right] \\ &= \mathbf{Pr}\left[d(X_t, Y_t) \geq 1\right] \\ &\leq \mathbf{E}\left[d(X_t, Y_t)\right] \\ &\leq (1-\alpha)^t \cdot n. \end{aligned}$$

Now come back to our problem of sampling proper colorings. Suppose X_t, Y_t are two proper colorings. We define the distance $d(X_t, Y_t)$ as their Hamming distance, i.e. the number of vertices colored differently in two colorings. Our coupling of two chains is that we always choose the same v, c in each step. The distance between two colorings can change at most 1 since only v is affected. The possible changes can be divided into two kinds:

- Good move: $X_t(v) \neq Y_t(v)$, and both change into c successfully. It will decrease distance by 1.
- Bad move: $X_t(v) = Y_t(v)$, one succeeds and one fails in the changing. It will increase distance by 1.





Consider the probabilities of two types of moves. For good moves, w.p. $\frac{d(X_t,Y_t)}{n}$, $X_t(v) \neq Y_t(v)$, and there are at least $q - 2\Delta$ choices of c to make it a good move. So

$$\begin{aligned} \mathbf{Pr}\left[d(X_{t+1}, Y_{t+1}) = d(X_t, Y_t) - 1\right] &= \mathbf{Pr}_{(v,c)\in V\times[q]}\left[(v, c) \text{ is a good move}\right] \\ &\geq \frac{d(X_t, Y_t)}{n} \cdot \frac{q-2\Delta}{q}. \end{aligned}$$

For bad moves, there exists a neighbor w of v such that its color is different in two colorings, and in one coloring w is of color c. By a counting argument, we have

$$\mathbf{Pr}\left[d(X_{t+1}, Y_{t+1}) = d(X_t, Y_t) + 1\right] = \mathbf{Pr}_{(v,c)\in V\times[q]}\left[(v, c) \text{ is a bad move}\right] \leq \frac{\Delta d(X_t, Y_t)}{n} \cdot \frac{2}{q}.$$

Therefore,

$$\begin{aligned} \mathbf{E}\left[d(X_{t+1}, Y_{t+1}) | (X_t, Y_t)\right] &= d(X_t, Y_t) + \mathbf{Pr}\left[d(X_{t+1}, Y_{t+1}) = d(X_t, Y_t) + 1\right] - \mathbf{Pr}\left[d(X_{t+1}, Y_{t+1}) = d(X_t, Y_t) - 1\right] \\ &\leq d(X_t, Y_t) + \frac{\Delta d(X_t, Y_t)}{n} \cdot \frac{2}{q} - \frac{d(X_t, Y_t)}{n} \cdot \frac{q-2\Delta}{q} \\ &\leq d(X_t, Y_t)\left(1 - \frac{q-4\Delta}{nq}\right). \end{aligned}$$

In the case $q > 4\Delta$, if we want

$$D_{\mathrm{TV}} \leq \left(1 - \frac{1}{nq}\right)^t n \leq \varepsilon,$$

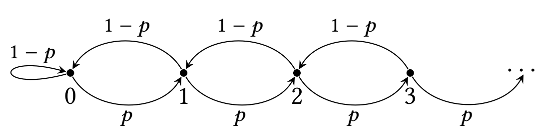
we have the mixing time is bounded by

$$\tau_{\mathrm{mix}}(\varepsilon) \leq nq \log \frac{n}{\varepsilon}.$$

4 *Countably Infinite Markov Chains*

We have proved that finite Markov chain must have a stationary distribution using Perron-Frobenius Theorem. However, when the Markov chain has infinite states, even it's countable infinite, there is something going wrong.

Consider the following random walk on $\mathbb{N}$. The state space is $\mathbb{N}$ and at each state i, go to $i+1$ w.p. p and go to $i-1$ w.p $1-p$ (if $i = 0$, w.p. $1-p$ stay put).



Let π be the stationary distribution of this Markov chain (if there exists a stationary distribution). We have that

$$\pi(0) = \pi(0)(1-p) + \pi(1)(1-p) \qquad \Longrightarrow \pi(1) = \frac{p}{1-p}\pi(0),$$

$$\pi(1) = \pi(0)p + \pi(2)(1-p) \qquad \Longrightarrow \pi(2) = \frac{p}{1-p}\pi(1),$$

$$\cdots$$

$$\pi(i) = \pi(i-1)p + \pi(i+1)(1-p) \qquad \Longrightarrow \pi(i+1) = \frac{p}{1-p}\pi(i).$$

$$\cdots$$

Note that π is a distribution, so $\sum_{i=0}^{\infty} \pi(i) = 1$. Then, we have

- If $p < \frac{1}{2}$, that is, $\frac{p}{1-p} < 1$, then $\sum_{i=0}^{\infty} \left(\frac{p}{1-p}\right)^i \pi(0) = 1$. By direct calculation we have $\pi(0) = \frac{1-2p}{1-p}$ and $\pi(i) = \left(\frac{p}{1-p}\right)^i \frac{1-2p}{1-p}$ for $i \in \mathbb{N}$.
- If $p > \frac{1}{2}$, then $\frac{p}{1-p} > 1$. When $i \to \infty$, if $\pi(0) \neq 0$, $\pi(i) \to \infty$. This yields that $\pi(0) = \pi(1) = \cdots = \pi(i) = \cdots = 0$. The Markov chain doesn't have a stationary distribution in this case.
- If $p = \frac{1}{2}$, $\frac{p}{1-p} = 1$. Then $\pi(0) = \pi(1) = \cdots = \pi(i) = \cdots$ and $\sum_{i=0}^{\infty} \pi(0) = 1$. This yields that $\pi(0) = 0$ and there is no stationary distribution in this case.

4.1 Recurrence

Definition 1 *For $i \in \Omega$, let $T_i > 0$ be the first hitting time of state i. Let $\mathbf{P}_i = \mathbf{Pr}[\cdot|X_0 = i]$. We say a state i is* recurrent *if $\mathbf{P}_i[T_i < \infty] = 1$, o.w. we say the state is* transient.

$T_i \triangleq \min\{t > 0|X_t = i\}$.

Let $N_i \triangleq \sum_{t=0}^{\infty} \mathbb{1}[X_t = i]$, then we have the following propositions.

Recall the probability space of a stochastic process. One can view the outcomes of the probability space is the set of infinite sequence of real numbers between $[0, 1]$, namely $\hat{\Omega} = [0, 1]^{\mathbb{N}}$. The sigma-algebra can be defined in a way similar to the problem 1 in our first homework. Therefore, the random variable T_i is therefore a function $\hat{\Omega} \to \mathbb{R}$.

Proposition 2 *If i is recurrent, then $\mathbf{P}_i[N_i = \infty] = 1$.*

Proof. Assume that $\mathbf{P}_i[N_i = \infty] < 1$. Then there exists $\Omega' \subseteq \hat{\Omega}$ such that $N_i < \infty$ on Ω' and $\mathbf{P}_i[\Omega'] > 0$. This means that with probability larger than 0, we will never reach state i after the last time we visit it. This is in contradiction with the fact that i is recurrent. □

Proposition 3 *If i is recurrent and there exists a finite path from i to j, then*

- $\mathbf{P}_i[T_j < \infty] = 1$.
- $\mathbf{P}_j[T_i < \infty] = 1$.
- *j is recurrent.*

Proof.

- Let $q \triangleq \mathbf{P}_i[\text{reach } j \text{ before returning to } i]$. Since there is a finite path from i to j, we have $q > 0$ and $\mathbf{P}_i[\text{visit } i\ n \text{ times before reaching } j] = (1-q)^n$.

  Assume that $\mathbf{P}_i[T_j = \infty] = \alpha > 0$. Then we have $\mathbf{P}_i[T_j = \infty|N_i = \infty] = \alpha$ since $\mathbf{P}_i[N_i = \infty] = 1$. Let T_i^n be the n^{th} time that the chain visits state i. Then

  $\mathbf{P}_i[T_j = \infty] = \mathbf{P}_i[T_j = \infty \mid N_i = \infty] \cdot \mathbf{P}_i[N_i = \infty] + \mathbf{P}_i[T_j = \infty \mid N_i < \infty] \cdot \mathbf{P}_i[N_i < \infty] = \mathbf{P}_i[T_j = \infty \mid N_i = \infty]$.

  $$\forall n > 0,\ \mathbf{P}_i[T_j > T_i^n|N_i = \infty] \geq \mathbf{P}_i[T_j = \infty|N_i = \infty] = u$$

  On the otherhand, we have $\lim_{n\to\infty} \mathbf{P}_i[T_j > T_i^n|N_i = \infty] = \lim_{n\to\infty} \mathbf{P}_i[T_j > T_i^n] = \lim_{n\to\infty}(1-q)^n = 0$. This is a contradiction. Thus, $\mathbf{P}_i[T_j = \infty] = 0$.

- If $\mathbf{P}_j[T_i = \infty] = p > 0$, then we have that $\mathbf{P}_i[T_i = \infty] \geq q \cdot p > 0$. This is in contradiction with the fact that i is recurrent.

- If $\mathbf{P}_j[T_j = \infty] = r > 0$, then $\mathbf{P}_i[T_j = \infty] \geq q \cdot r > 0$. This is in contradiction with the first item of this proposition.

□

References

[JVV86] Mark R Jerrum, Leslie G Valiant, and Vijay V Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical computer science*, 43:169–188, 1986. 3