Markov Chains and Algorithmic Applications: WEEK 9

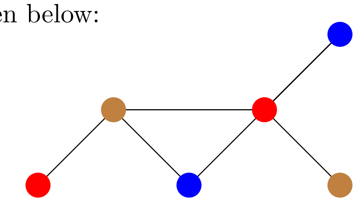
1 Markov Chain Monte Carlo (MCMC) Sampling

The idea behind the MCMC method to obtain samples of a distribution π on S is to construct a Markov chain on S with transition matrix P having π as its stationary distribution. The samples of π are then obtained by iterating P long enough to reach the stationary distribution π, then sampling among the states of the Markov chain. The advantage here is that a) we do not have to sample directly from π, and b) we do not even need to know everything about π, as we will see below.

For practical reasons, we want P to have certain properties:

1. π should be the unique limiting distribution of P.
2. Convergence to the stationary distribution π should be fast, so as to obtain samples within a reasonable amount of time.

Example 1.1 (Graph Coloring)**.** Let $G = (V, E)$ be a graph with vertex set V and edge set E. We want to color each vertex of the graph with one of the q colors at our disposal such that a vertex's color differs from that of all its neighbors, as seen below:



More formally, let $x = (x_v,\ v \in V)$ be a particular color configuration of the vertex set V. A *proper q-coloring* of G is any configuration x such that $\forall v, w \in V$, if $(v, w) \in E$ then $x_v \neq x_w$.

If S represents the set of all possible color configurations, then the uniform distribution π over all proper q-colorings is given by

$$\pi(x) = \frac{1}{Z}\, \mathbb{1}\{x \text{ is a proper } q\text{-coloring}\}$$

where Z is the total number of proper q-colorings in G.

Computing Z would require enumerating all possible proper q-colorings which is non-trivial depending on G. Still, we would like to sample from π without computing Z explicitly.

1.1 Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm is a procedure to construct a Markov chain on S having as limiting distribution π (for convenience, we assume that $\pi_i > 0$ for all $i \in S$). Here is the algorithm:

1. Select an easy-to-simulate irreducible and aperiodic Markov chain ψ on S with the constraint that $\psi_{ij} > 0$ if and only if $\psi_{ji} > 0$.[1] We call ψ the *base chain*.
2. Design acceptance probabilities $a_{ij} = \mathbb{P}(\text{transition from } i \text{ to } j \text{ is accepted})$ such that the matrix P given below has limiting distribution π.
3. Construct the matrix P as such:

$$\begin{cases} p_{ij} &= \psi_{ij}\, a_{ij}, \quad j \neq i \\ p_{ii} &= \psi_{ii} + \sum_{k \neq i} \psi_{ik}\, (1 - a_{ik}) = 1 - \sum_{k \neq i} \psi_{ik} a_{ik} \end{cases}$$

   In other words, we are adding self-loops of different weights to each state.

[1]If S is finite, then these conditions imply positive-recurrence, hence ψ is ergodic and has a unique limiting distribution, but this limiting distribution is of no interest to the algorithm.

We must now choose the weights a_{ij} so that $p_{ij}(n) \underset{n\to\infty}{\longrightarrow} \pi_j$. Moreover, we were able to upper-bound the mixing time of chains satisfying detailed balance in the previous lectures, so we would like P to satisfy this condition too: $\pi_i p_{ij} = \pi_j p_{ji}$

Theorem 1.2 (Metropolis-Hastings)**.** If $a_{ij} = \min\left(1, \frac{\pi_j \psi_{ji}}{\pi_i \psi_{ij}}\right)$, then the matrix P constructed above is ergodic with stationary distribution π. Moreover, P satisfies detailed balance.

Proof. By assumption, ψ is irreducible and aperiodic, and $\forall i, j \in S, \psi_{ij} > 0$ iff $\psi_{ji} > 0$. So if $\psi_{ij} > 0$, then $a_{ij} > 0$ and $p_{ij} > 0$ also. Therefore, P is also irreducible and aperiodic. We then have

$$\pi_i p_{ij} = \pi_i \psi_{ij} a_{ij} = \pi_i \psi_{ij} \min\left(1, \frac{\pi_j \psi_{ji}}{\pi_i \psi_{ij}}\right) = \min\left(\pi_i \psi_{ij}, \pi_j \psi_{ji}\right)$$

whose expression is symmetric in i, j. It is therefore also equal to $\pi_j\, p_{ji}$: detailed balance holds and P has π as stationary distribution.

Finally, since P is irreducible and has a stationary distribution π, then by a previously seen theorem, P must be positive-recurrent and π must be unique. therefore P is ergodic and π is also a limiting distribution. ☐

Remark 1.3. If $\psi_{ij} = \psi_{ji}$, then the expression for a_{ij} simplifies to $a_{ij} = \min\left(1, \frac{\pi_j}{\pi_i}\right)$.
The intuition behind choosing a_{ij} as such is the following: if $\pi_j > \pi_i$ the transition $i \to j$ should be taken with probability 1 since the chain is heading towards the more probable state j. However if $\pi_j < \pi_i$, then the move $i \to j$ should be taken with probability $\frac{\pi_j}{\pi_i} < 1$. In other words, the chain should tend towards the states having high probability, but it should be able to return to less probable states in order not to get stuck in a state that locally maximizes π.

Remark 1.4. The advantage of the Metropolis-Hastings algorithm is that the acceptance probabilities a_{ij} depend on π only through the ratios $\frac{\pi_j}{\pi_i}$, which can be significantly easier to compute than π_i and π_j separately! In the graph coloring example given previously, $\frac{\pi_j}{\pi_i} = \frac{\mathbb{1}\{j \text{ is a proper } q\text{-coloring}\}}{\mathbb{1}\{i \text{ is a proper } q\text{-coloring}\}}$, so we can avoid computing the expensive normalization constant Z entirely.

Example 1.5 (Metropolized Independent Sampling)**.** To obtain samples of distribution π on S, we choose the base chain ψ such that $\psi_{ij} = \psi_j > 0\ \forall i, j \in S$ (i.e. the process realizations are just sequences of i.i.d. random variables).

The acceptance probabilities are $a_{ij} = \min\left(1, \frac{w_j}{w_i}\right)$ with $w_i = \frac{\pi_i}{\psi_i}$, so the transition probabilities of P are given by

$$\begin{cases} p_{ij} &= \psi_{ij} a_{ij} = \psi_j \min\left(1, \frac{w_j}{w_i}\right), \quad j \neq i \\ p_{ii} &= 1 - \sum_{k \neq i} \psi_{ik} a_{ik} = 1 - \sum_{k \neq i} \psi_k \min\left(1, \frac{w_k}{w_i}\right) \end{cases}$$

In this particular example, one can show the following (no proof given here):

Theorem 1.6 (Liu)**.** Let $\lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_{N-1}$ be the eigenvalues of P, and $\lambda_* = \max(\lambda_1, -\lambda_{N-1})$. Then

$$\lambda_* = 1 - \frac{1}{w_*}, \quad \text{where } w_* = \max_{i \in S} \frac{\pi_i}{\psi_i} > 1$$

Correspondingly, the spectral gap $\gamma = \frac{1}{w_*}$.

From the above and the previous lectures, we find that

$$\|P_i^n - \pi\|_{\text{TV}} \leq \frac{\lambda_*^n}{2\sqrt{\pi_i}} \leq \frac{1}{2\sqrt{\pi_i}} e^{-\gamma n} = \frac{1}{2\sqrt{\pi_i}} e^{-\frac{n}{w_*}}$$

Therefore, if w_* is large (i.e. if the distance between π and ψ is large), then convergence to the stationary distribution π is slow (this resembles the situation we already encountered with rejection sampling).