Markov Chains and Algorithmic Applications: WEEK 9

1 Metropolis-Hastings algorithm: applications

Reminder. We would like to sample from a distribution $\pi = (\pi_i,\, i \in S)$ on state space S. One option for this is the Metropolis-Hastings algorithm:

1. Consider a base chain on S with transition probabilities ψ_{ij}
(irreducible, aperiodic and such that $\psi_{ij} > 0$ if and only if $\psi_{ji} > 0$)
2. Define acceptance probabilities $a_{ij} = \min\left(1, \dfrac{\pi_j\,\psi_{ji}}{\pi_i\,\psi_{ij}}\right)$
3. Define then
$$p_{ij} = \begin{cases} a_{ij}\,\psi_{ij} & \text{if } j \neq i \\ \psi_{ii} + \sum_{k \in S \setminus i} \psi_{ik}(1 - a_{ik}) & \text{if } j = i \end{cases}$$
4. Assume there exists i with $\psi_{ii} > 0$, or there exists a pair $i \neq k$ s.t $\psi_{ik} > 0$ and $a_{ik} < 1$. Then, the Markov chain on S with transition probabilities p_{ij} is such that $p_{ij}(n) \underset{n\to\infty}{\longrightarrow} \pi_j$ and detailed balance holds. Running then the Markov chain from an arbitrary initial state $i \in S$ for a sufficiently large amount of time (so that $p_{ij}(n)$ is indeed close to π_j for all $j \in S$) is a way to (approximately) sample from π.

We will see in the following two applications of this algorithm.

1.1 Optimization of a function

Let $f : \mathbb{Z} \to \mathbb{R}$ be a function to be minimized, which is assumed to be bounded from below and such that $\lim_{i\to\pm\infty} f(i) = +\infty$ (so that at least one global minimum exists).

Problem: If the function f is complicated and has many local minima, then (greedy) algorithms usually fail to converge to a global minimum[1].

Our aim: to sample from the distribution

$$\pi_\infty(i) = \frac{\mathbb{I}_{\{i \text{ is a global minimum of } f\}}}{Z_\infty}, \quad i \in \mathbb{Z}$$

where $Z_\infty = \sharp(\text{global minima of } f)$. Sampling from π_∞ is a difficult task because

1. we have to compute Z_∞, and
2. the global minima may be very isolated on the state space, hence checking the neighborhood of i is not sufficient to compute $\mathbb{I}_{\{i \text{ is a global minimum of } f\}}$.

Instead, we will sample from the distribution π_β:

$$\pi_\beta(i) = \frac{e^{-\beta f(i)}}{Z_\beta}, \quad i \in \mathbb{Z}$$

where $\beta > 0$ is a fixed parameter and $Z_\beta = \sum_{i\in\mathbb{Z}} e^{-\beta f(i)}$ is the normalization constant (that might still be difficult to compute). The idea is that as β increases, distribution π_β concentrates around the global minima of f, hence $\pi_\beta \overset{\beta\to\infty}{\longrightarrow} \pi_\infty$.

[1] This typically also happens when $\mathbb{Z}$, the domain of the function, is replaced by a finite but high-dimensional domain.

To avoid computing Z_β, we will use the Metropolis-Hastings algorithm to construct a Markov chain having π_β as its stationary distribution:

1. We choose a simple irreducible base chain (such that $\psi_{ij} > 0$ iff $\psi_{ji} > 0$), the symmetric random walk on $\mathbb{Z}$: $\psi_{i,i\pm1} = \frac{1}{2}$ (remember that this chain has no stationary distribution, yet this does not influence the algorithm in any way).
2. The acceptance probabilities are
$$a_{ij} = \min\left(1, \frac{\pi_j}{\pi_i}\right) = \begin{cases} \min\left(1, e^{-\beta(f(j)-f(i))}\right) & j = i \pm 1, \\ 0 & \text{otherwise.} \end{cases}$$
In words, we always accept a transition to a state with a lower value of f, but we still accept some non-favorable transitions to avoid getting stuck in a local minimum.
3. The constructed chain having transition probabilities
$$p_{ij} = \begin{cases} \psi_{ij} a_{ij} & j \neq i, \\ 1 - \sum_{k \neq i} \psi_{ik} a_{ik} & j = i, \end{cases}$$
is such that $p_{ij}(n) \overset{n\to\infty}{\longrightarrow} \pi_\beta(j) \quad \forall j \in S$.

1.1.1 How to choose β ?

Let us give a ballpark estimate to choose β correctly. Note that this is just a qualitative idea which can only serve as a first guide when these ideas are applied to specific problems. To choose β, we decide that we want to spend a $1-\epsilon$ fraction of time in global minima. Recall that π_i is the average fraction of time that the chain spends in state i when it has reached the stationary distribution. Thus we set

$$1 - \epsilon \approx \sum_{i \text{ global minimum}} \pi_\beta(i)$$

Let $f_0 = \min_{i\in\mathbb{Z}} f(i)$ be the global minimum and $f_1 = \min_{i\in\mathbb{Z}, f(i)\neq f_0} f(i)$, $f_2 = \min_{i\in\mathbb{Z}, f(i)\neq f_0, f_1} f(i)$, $\ldots$ be the local minima. Let N_0, N_1, N_2, $\ldots$ be the number of points were the minima f_0, f_1, f_2, $\ldots$ are reached. We have

$$\sum_{i \text{ global minimum}} \pi_\beta(i) = \frac{N_0 e^{-\beta f_0}}{Z} \quad \text{and} \quad Z = \sum_{i\in\mathbb{Z}} e^{-\beta f(i)} = \sum_{k\geq 0} N_k e^{-\beta f_k} \approx N_0 e^{-\beta f_0} + N_1 e^{-\beta f_1}$$

(as we think of β being reasonably large and $f_0 < f_1 < f_2 < \ldots$). Therefore:

$$\sum_{i \text{ global minimum}} \pi_\beta(i) \approx \frac{N_0 e^{-\beta f_0}}{N_0 e^{-\beta f_0} + N_1 e^{-\beta f_1}} = \frac{1}{1 + \frac{N_1}{N_0} e^{-\beta(f_1 - f_0)}} \approx 1 - \frac{N_1}{N_0} e^{-\beta(f_1 - f_0)}$$

Remembering that we want this term to be approximately equal to $1-\epsilon$, we obtain the following rough estimate for β:

$$\beta \approx \frac{1}{f_1 - f_0} \log\left(\frac{N_1}{\epsilon N_0}\right)$$

1.1.2 In practice: simulated annealing

The choice of β can influence the output of the Metropolis algorithm significantly:

- If we choose β large, then π_β is close to π_∞, but the chain produced by the algorithm converges very slowly due to the high probability given to self-loops. In essence, the chain can almost become reducible.

- If we choose β small, then the chain produced by the algorithm converges quickly to the stationary distribution π_β at the cost of potentially being very far from π_∞.

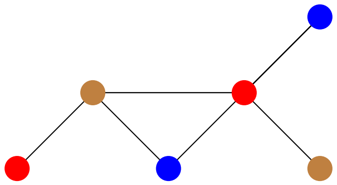
The ideal solution would be to combine the best of both worlds, similarly to creating certain alloys: simply mixing the metals at high temperature then immediately bringing the system down to room temperature does not give the alloy the desired properties. Instead, the temperature should be decreased at a slow speed for the metals to bond appropriately.

Consider β as representing an inverse temperature. Then the annealing approach detailed above gives us a good algorithm to find a global minimum:

1. Start with β small (i.e. high temperature regime): the algorithm will then visit all the states of S quite uniformly at the beginning. After a sufficiently high number of iterations, the Metropolized chain is roughly distributed as π_β.
2. Increase then β (i.e. lower the temperature) and rerun the algorithm from the state found in the previous step.
3. Repeat step 2 until β is sufficiently large, so that one can hope to have reached a global minimum.

1.2 Graph coloring

Let $G = (V, E)$ be a graph with vertex set V ($|V| = N$) 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 S be the set of all possible color configurations on G and $x = (x_v,\ v \in V) \in S$ a particular color configuration. A *proper q-coloring* of G is any configuration x such that $\forall v, w \in V$, if $(v, w) \in E$ then $x_v \neq x_w$.

Our aim: to sample uniformly amongst the proper q-colorings of G. In other words, we want to sample from the distribution

$$\pi(x) = \frac{\mathbb{I}_{\{x \text{ is a proper } q\text{-coloring}\}}}{Z}, \quad x \in S$$

where $Z = \sharp(\text{ proper } q\text{-colorings})$

Remark 1.1. Let $\Delta = \max_{v \in V} \deg(v)$. If $q \geq \Delta + 1$, then there exists at least one proper q-coloring.

In what follows, we are going to restrict our analysis to graphs satisfying $q > 3\Delta$.

One way to sample from π is by using the following algorithm:

1. Start from a proper q-coloring $x \in S$.
2. Select a vertex $v \in V$ uniformly at random.
3. Select a color $c \in \{1, \ldots, q\}$ uniformly at random.
4. If c is an allowed color at v, then recolor v (i.e. set $x_v = c$); do nothing otherwise.

5. Repeat steps 2, 3 and 4.

Remark 1.2. Since the algorithm started from a proper q-coloring $x \in S$, every visited state is also a proper q-coloring.

Remark 1.3. The algorithm could also be used to *find* a proper q-coloring on G. Indeed, if we start the algorithm in a state $x' \in S$ that is not a proper coloring, the algorithm ensures that eventually a proper coloring will be reached.

Definition 1.4. Let $x, y \in S$ be two color configurations. We write $x \sim y$ if x and y differ in at most one vertex.

Remark 1.5. The algorithm is actually an instance of the Metropolis-Hastings algorithm:

1. Let the base chain be $\psi_{xy} = \begin{cases} \frac{1}{Nq} & y \sim x, y \neq x, \\ \frac{1}{q} & y = x, \\ 0 & \text{otherwise.} \end{cases}$

   ψ is aperiodic (due to self-loops) and satisfies $\psi_{xy} > 0$ iff $\psi_{yx} > 0$ (due to symmetry). Moreover, the condition $q > 3\Delta$ ensures that ψ is irreducible.[2]

2. $a_{xy} = \min\left(1, \frac{\pi_y}{\pi_x}\right) = \min\left(1, \frac{\mathbb{I}_{\{y \text{ is a proper } q\text{-coloring}\}}/Z}{1/Z}\right) = \mathbb{I}_{\{y \text{ is a proper } q\text{-coloring}\}}$.
   (NB: we already know that x is a proper q-coloring).

3.
$$p_{xy} = \begin{cases} \psi_{xy} a_{xy} & y \neq x, \\ 1 - \sum_{z \in S \setminus x} \psi_{xz} a_{xz} & y = x \end{cases}$$
$$= \begin{cases} \frac{1}{Nq} \mathbb{I}_{\{y \text{ is a proper } q\text{-coloring}\}} & y \sim x, \ y \neq x, \\ 1 - \frac{1}{Nq} \sharp\{z \sim x, \ z \neq x, \ z \text{ proper } q\text{-coloring}\} & y = x, \\ 0 & \text{otherwise.} \end{cases}$$

1.2.1 Convergence rate analysis

The *mixing time* of this chain is $T_\epsilon = \inf\{n \geq 1 : \max_{x \text{ proper } q\text{-coloring}} \|P_x^n - \pi\|_{\text{TV}} \leq \epsilon\}$.

Theorem 1.6. If $q > 3\Delta$, then for all proper q-colorings x, $\|P_x^n - \pi\|_{\text{TV}} \leq N e^{-\frac{n}{N}\left(1 - \frac{3\Delta}{q}\right)}$, implying that

$$T_\epsilon \leq \frac{1}{1 - \frac{3\Delta}{q}} N \left(\log(N) + \log\left(\frac{1}{\epsilon}\right)\right)$$

Remark 1.7. The proof given below is completely constructive in the sense that it avoids using the general structure theorems seen so far. This is often the case when one wants to derive concrete estimates about mixing times.

Proof. Let $(X_n, n \geq 0)$ be a Markov chain on S starting at $X_0 = x$ (a proper q-coloring) and evolving according to P. Let $(Y_n, n \geq 0)$ be a Markov chain on S starting at $Y_0 \sim \pi$ and also evolving according to P.

We will couple X and Y as follows:

1. Select a vertex $v \in V$ uniformly at random.
2. Select a color $c \in \{1, \ldots, q\}$ uniformly at random.

[2] We do not prove this fact here. The analysis of the mixing time in the next section is self-contained and independent of this proof.

3. Update X at vertex v if c is an allowed color.
   Update Y at vertex v if c is an allowed color.

Definition 1.8. The *Hamming distance* between two colorings x and y is the number of positions in which x and y disagree:

$$d(x,y) = \sum_{v \in V} \mathbb{I}_{\{x_v \neq y_v\}}$$

By a coupling argument seen in previous lectures, we have

$$\|P_x^n - \pi\|_{\mathrm{TV}} \leq \mathbb{P}\left(X_n \neq Y_n\right) = \mathbb{P}\left(d\left(X_n, Y_n\right) \geq 1\right) \leq \mathbb{E}(d\left(X_n, Y_n\right)),$$

where the last inequality is obtained by using the Markov inequality.

All that is left to do now is to upper bound $\mathbb{E}(d\left(X_n, Y_n\right))$. We will do so using two inductions:

1. Assume first that $d\left(X_0, Y_0\right) = 1$, i.e. X_0 and Y_0 differ at one vertex only, and let v be that vertex. Due to the coupling, at most one vertex can change color per transition, hence $d\left(X_1, Y_1\right) \in \{0, 1, 2\}$ and

$$\begin{aligned}\mathbb{E}(d\left(X_1, Y_1\right)) &= 0 \cdot \mathbb{P}\left(d\left(X_1, Y_1\right) = 0\right) + 1 \cdot \mathbb{P}\left(d\left(X_1, Y_1\right) = 1\right) + 2 \cdot \mathbb{P}\left(d\left(X_1, Y_1\right) = 2\right) \\ &= \left(1 - \mathbb{P}\left(d\left(X_1, Y_1\right) = 0\right)\right) + \mathbb{P}\left(d\left(X_1, Y_1\right) = 2\right)\end{aligned}$$

   $d\left(X_1, Y_1\right) = 0$ if and only if vertex v is chosen (with probability $\frac{1}{N}$) and that the color c chosen is allowed in both chains X and Y, hence

$$\mathbb{P}\left(d\left(X_1, Y_1\right) = 0\right) = \frac{1}{N} \cdot \frac{\sharp \text{ allowed colors at } v}{q} \geq \frac{1}{N} \cdot \frac{q - \Delta}{q}$$

   $d\left(X_1, Y_1\right) = 2$ if and only if the vertex w chosen is a neighbor of v and that either X or Y is recolored (but not both). The latter only happens when the chosen color c satisfies $c = x_v$ or $c = y_v$, so we have

$$\mathbb{P}\left(d\left(X_1, Y_1\right) = 2\right) \leq \frac{\Delta}{N} \cdot \frac{2}{q}$$

   Gathering both estimates together, we obtain

$$\mathbb{E}(d\left(X_1, Y_1\right)) \leq \left(1 - \frac{1}{N}\frac{q - \Delta}{q}\right) + \frac{\Delta}{N}\frac{2}{q} = 1 - \frac{1}{N}\left(1 - \frac{3\Delta}{q}\right)$$

2. Suppose now that $d\left(X_0, Y_0\right) = r$. Since P describes an irreducible Markov chain, there exists a sequence of $r - 1$ states $Z_0^{(1)}, \ldots, Z_0^{(r-1)}$ such that

$$p_{X_0 Z_0^{(1)}} p_{Z_0^{(1)} Z_0^{(2)}} \ldots p_{Z_0^{(r-1)} Y_0} > 0,$$

$$d\left(X_0, Z_0^{(1)}\right) = d\left(Z_0^{(1)}, Z_0^{(2)}\right) = \cdots = d\left(Z_0^{(r-1)}, Y_0\right) = 1$$

   This implies that

$$\begin{aligned}\mathbb{E}(d\left(X_1, Y_1\right)) &\leq \mathbb{E}(d\left(X_1, Z_1^{(1)}\right)) + \mathbb{E}(d\left(Z_1^{(1)}, Z_1^{(2)}\right)) + \cdots + \mathbb{E}(d\left(Z_1^{(r-1)}, Y_1\right)) \\ &= r\left(1 - \frac{1}{N}\left(1 - \frac{3\Delta}{q}\right)\right)\end{aligned}$$

3. This inequality is valid between times 0 and 1, but by time-homogeneity of the chain, it also holds between times $n-1$ and n for $n \geq 1$:

   $$\mathbb{E}(d\left(X_n, Y_n\right) | d\left(X_{n-1}, Y_{n-1}\right) = r) \leq r\left(1 - \frac{1}{N}\left(1 - \frac{3\Delta}{q}\right)\right)$$

   From the above (averaging over r) we deduce that

   $$\begin{aligned} \mathbb{E}(d\left(X_n, Y_n\right)) &\leq \left(1 - \frac{1}{N}\left(1 - \frac{3\Delta}{q}\right)\right)\mathbb{E}(d\left(X_{n-1}, Y_{n-1}\right)) \\ \Longrightarrow \mathbb{E}(d\left(X_n, Y_n\right)) &\leq \mathbb{E}(d\left(X_0, Y_0\right))\left(1 - \frac{1}{N}\left(1 - \frac{3\Delta}{q}\right)\right)^n \\ &\leq Ne^{-\frac{n}{N}\left(1 - \frac{3\Delta}{q}\right)} \end{aligned}$$

which completes the proof. □