Lecture Notes VIII – Markov Chain Monte Carlo

Marina Meila
mmp@stat.washington.edu

Department of Statistics
University of Washington

April 2025

Notation

Gibbs sampling

The detailed balance

Metropolis-Hastings sampling

Notation

- $V = \{X_1, \dots X_n\}$ set of random variables (nodes of a graphical model) also known as Markov network
- $X_{1:n} \in \{\pm 1\}$ (for simplicity)
- $E = \{(i,j),\ 1 \leq i < j \leq n\}$ edges of graph
- graph is not complete, some edges are missing
- We write $i \sim j$ for $(i,j) \in E$ or $(j,i) \in E$
- neigh_i = neighbors of X_i
- **Markov property** $\boxed{X_i \perp \text{all other variables} \,|\, \mathrm{neigh}_i}$
- $x = (x_1, \dots x_n) \in \{\pm 1\}^n = S$ an assignment to all variables in V

Notation

- $V = \{X_1, \dots X_n\}$ set of random variables (nodes of a graphical model) also known as Markov network
- $X_{1:n} \in \{\pm 1\}$ (for simplicity)
- $E = \{(i,j),\, 1 \leq i < j \leq n\}$ edges of graph
- graph is not complete, some edges are missing
- We write $i \sim j$ for $(i,j) \in E$ or $(j,i) \in E$
- neigh_i = neighbors of X_i

- **Markov property** $\boxed{X_i \perp \text{all other variables} \,|\, \mathrm{neigh}_i}$

- $\mathrm{x} = (x_1, \dots x_n) \in \{\pm 1\}^n = S$ an assignment to all variables in V
- Distribution over S

$$P(\mathrm{x}) = \frac{1}{Z} e^{-\phi(\mathrm{x})} \quad \text{with } \phi(\mathrm{x}) = \sum_{i=1}^{n} h_i x_i + \sum_{(i,j)\in E} h_{ij} x_i x_j \tag{1}$$

  (and $h_{ij} > 0$ for all $(i,j) \in E$)
- $Z = \sum_{\mathrm{x} \in S} e^{-\phi(\mathrm{x})}$
- Usually, intractable to compute Z

Wanted samples $\mathrm{x}^1, \mathrm{x}^2, \dots$ from P

An example

Gibbs sampling idea

$$P(\mathrm{x}) = \frac{1}{Z} e^{-\phi(\mathrm{x})} \quad \text{with } \phi(\mathrm{x}) = \sum_{i=1}^{n} h_i x_i + \sum_{(i,j)\in E} h_{ij} x_i x_j$$

- We cannot sample directly from P
- But we can sample each $X_i \sim P_{X_i|X_{-i}} = P_{X_i|\mathrm{neigh}_i}$ for any $i = 1 : n$

Gibbs sampling idea

$$P(\mathrm{x}) = \frac{1}{Z}e^{-\phi(\mathrm{x})} \quad \text{with } \phi(\mathrm{x}) = \sum_{i=1}^{n} h_i x_i + \sum_{(i,j)\in E} h_{ij} x_i x_j$$

- We cannot sample directly from P
- But we can sample each $X_i \sim P_{X_i|X_{-i}} = P_{X_i|\mathrm{neigh}_i}$ for any $i = 1:n$
- Why? For any x_{-i} let

$$\pi_i^+ = Pr[X_i = +|\mathrm{x}_{-i}], \quad \pi_i^- = Pr[X_i = -|\mathrm{x}_{-i}]. \tag{2}$$

$$\pi_i^+ = \frac{P(X_i = +1, \mathrm{x}_{-i})}{P(\mathrm{x}_{-i})} = \frac{P(X_i = +1, \mathrm{x}_{-i})}{P(X_i = +1, \mathrm{x}_{-i}) + P(X_i = -1, \mathrm{x}_{-i})} \tag{3}$$

$$\pi_i^- = \frac{P(X_i = -1, \mathrm{x}_{-i})}{P(X_i = +1, \mathrm{x}_{-i}) + P(X_i = -1, \mathrm{x}_{-i})} \quad \text{hence} \tag{4}$$

$$\frac{\pi_i^+}{\pi_i^-} = \frac{P(X_i = +1, \mathrm{x}_{-i})}{P(X_i = -1, \mathrm{x}_{-i})} = \frac{e^{-\phi(X_i=+1,\mathrm{x}_{-i})}}{e^{-\phi(X_i=-1,\mathrm{x}_{-i})}} = e^{-\phi(X_i=+1,\mathrm{x}_{-i})+\phi(X_i=-1,\mathrm{x}_{-i})} \tag{5}$$

$$= e^{-2h_i - 2\sum_{j\sim i} h_{ij}} \tag{6}$$

$$1 = \pi_i^+ + \pi_i^- \quad \text{hence } \pi_i^+ = \frac{e^{-2h_i-2\sum_{j\sim i} h_{ij}}}{1 + e^{-2h_i-2\sum_{j\sim i} h_{ij}}} = \frac{e^{-h_i-\sum_{j\sim i} h_{ij}}}{e^{h_i+\sum_{j\sim i} h_{ij}} + e^{-h_i-\sum_{j\sim i} h_{ij}}} \tag{7}$$

- $X_i \,|\, \mathrm{neigh}_i \sim \mathrm{Bernoulli}(\pi_i^+)$

Gibbs sampling algorithm

1. Initialize x^0 with some arbitrary values
2. For $t = 1, 2, \ldots$ we will sample sequentially $x^t | x^{t-1}$ as follows
   2.1 Pick $i \in 1:n$ uniformly at random
   2.2 Sample $X_i^t \mid \text{neigh}_i \sim \text{Bernoulli}(\pi_i^+)$
   2.3 Every T steps (where T is a LARGE number), output x^t

Q Why does this work?
A1 We are sampling from a Markov chain on S (transition probability matrix P on next page)
A2 If we take enough steps T, the distribution converges to the stationary distribution of this chain, let's call it P. We take a sample from it x^T
A3 If we continue for another T steps, the chain has "forgotten" about x^T; the new sample x^{2T} is independent of x^T. Etc, samples $x^{T,2T,\ldots NT}$ are i.i.d. from P^∞
TODO To show that $P^\infty = P$ the distribution we wanted to sample from.

The transition probability P of Gibbs sampling

- The transitions are between states x, x' that only differ in one variable i. All the other transition probabilities are 0.
- If x^t and x' differ only in variable i, then

$$P(x^{t+1} = x'|x) = \begin{cases} \pi_i^+(x_{-i}) & x'_i = +1 \\ \pi_i^-(x_{-i}) & x'_i = -1 \end{cases} \tag{8}$$

The detailed balance

Theorem

Let π be a distribution over S, and P a transition matrix of a Markov chain. Then if the following **detailed balance** *holds, π is the stationary distribution of P.*

$$\pi(x)\mathrm{P}(x, x') = \pi(x')\mathrm{P}(x', x) \tag{9}$$

Metropolis-Hastings (MH) idea

- MH is a **rejection sampling** algorithm
- We sample $x' \mid x^{t-1} \sim S$ a **proposal distribution**
- Then we **accept** $x^t = x'$ with some **acceptance probability** $a(x, x')$ that ensures the detail balance
- (if we don't accept, $x^t = x^{t-1}$)

- With MH, we have more flexibility in exploring the sample space around x^{t-1} than with Gibbs

Metropolis-Hastings algorithm

In Proposal distribution $S(x, x') \propto$ transition probability $x \to x'$
no need to be normalized, no need to be symmetric

1. Initialize x^0 with some arbitrary values
2. For $t = 1, 2, \ldots$ we will sample sequentially $x^t | x^{t-1}$ as follows

   2.1 Sample $x' \sim S(x^{t-1}, x')$

   2.2 Compute acceptance probability

$$a(x^{t-1}, x') = \min\left(1, \frac{P(x')S(x', x^{t-1})}{P(x^{t-1})S(x^{t-1}, x')}\right). \quad (10)$$

   2.3 $x^t = \begin{cases} x' & \text{w.p. } a \\ x^{t-1} & \text{w.p. } 1 - a \end{cases}$

   2.4 Every T steps (where T is a LARGE number), output x^t

Does it satisfy the detailed balance?

- If x' rejected ✓
- If x' accepted

$$
\begin{aligned}
\mathrm{P}(x', x) &= S(x', x)a(x', x) && (11)\\
P(x)\mathrm{P}(x, x') &= P(x)S(x, x')\min\left(1, \frac{P(x')S(x', x)}{P(x)S(x, x')}\right) && (12)\\
&= \min\left(P(x')S(x', x), P(x)S(x, x')\right) && (13)\\
&= P(x')\mathrm{P}(x', x) \qquad \text{by symmetry} && (14)
\end{aligned}
$$

Does it satisfy the detailed balance?

- If x' rejected ✓
- If x' accepted

$$\begin{aligned}
\mathrm{P}(x', x) &= S(x', x)a(x', x) && (11)\\
P(x)\mathrm{P}(x, x') &= P(x)\mathrm{S}(x, x')\min\left(1, \frac{P(x')\mathrm{S}(x', x)}{P(x)\mathrm{S}(x, x')}\right) && (12)\\
&= \min\left(P(x')\mathrm{S}(x', x), P(x)\mathrm{S}(x, x')\right) && (13)\\
&= P(x')\mathrm{P}(x', x) \qquad \text{by symmetry} && (14)
\end{aligned}$$

Recap: What we need to be able to do MH sampling

- To calculate $\mathrm{P}(x)/\mathrm{P}(x')$ but not P itself (okay not to have Z)
- To calculate $\mathrm{S}(x, x')/\mathrm{S}(x', x)$
- To sample from $\mathrm{S}(x, x')$