Chapter 7
Markov chain Monte-Carlo inference

I) Introduction

Remember that in the preceding chapters, we sometimes had to estimate a parameter θ by MLE. This is OK if the posterior density has an analytical representation but this cannot be always the case. In the general case, we can sample from the posterior (using MCMC) and somehow set $\hat{\theta}$ as the mode of our sampled density.
This method is popular because it works well in high dimension. In a survey by SIAM News (http://www.siam.org/pdf/news/637.pdf), MCMC was placed in the top 10 most important algorithms of the 20-th century.

II) Gibbs sampling

1) Basic idea

We sample each variable in turn, conditioned on the values of all the other variables. Given a joint sample $x^{(s)}$ of all the variables, we generate a new sample $x^{(s+1)}$ by sampling each component in turn, based of the most recent values of the other variables. For example, if $D=3$, we have:

$$\begin{cases} x_1^{(s+1)} \sim p(x_1 | x_2^{(s)}, x_3^{(s)}) \\ x_2^{(s+1)} \sim p(x_2 | x_1^{(s+1)}, x_3^{(s)}) \\ x_3^{(s+1)} \sim p(x_3 | x_1^{(s+1)}, x_2^{(s+1)}) \end{cases}$$

The chain $x^{(s)}$ will have (asymptotically) the desired law ($p(\cdot)$). For some reasons, it is necessary to discard some of the initial samples until the Markov chain has <u>burned in</u>, or entered its stationnary distribution.

2) Example: Gibbs sampling for the Ising model

We have $\Lambda = \{-N, -N+1, \dots, 0, \dots, N\}^2$ (a square in $\mathbb{Z}^2$). We set the state space to be: $E = \{0,1\}^\Lambda$ (functions from Λ to $\{0,1\}$).

For $x \in E$, m, m' in Λ (m and m' are neighbors), the interaction potential is $|x(m) - x(m')|^2$. We set:

$$H(x) = \frac{1}{2} \sum_{\substack{m,m' \in \Lambda \\ m \sim m'}} |x(m) - x(m')|^2$$

$$\begin{cases} \Pi(x) = \dfrac{1}{Z(\beta)} e^{-\beta H(x)} \quad (\beta > 0) \\ Z(\beta) = \displaystyle\sum_{x \in E} e^{-\beta H(x)} \end{cases}$$

We would like to sample according to Π.

We set $\begin{cases} \Lambda^+ = \{(m_1, m_2) \in \Lambda : m_1 + m_2 \text{ even}\} \\ \Lambda^- = \{(m_1, m_2) \in \Lambda : m_1 + m_2 \text{ odd}\}. \end{cases}$

For $x \in E$, we set $\begin{cases} x^+ = \{x(m), m \in \Lambda^+\} \\ x^- = \{x(m), m \in \Lambda^-\} \end{cases}$

So we can write: $x = (x^+, x^-)$.

We then have:

$$\Pi(x^+ | x^-) = \frac{\Pi(x^+, x^-)}{\Pi(x^-)} \qquad \text{(no more } \mathcal{Z}(\beta)\text{)}$$

$$= \frac{\Pi(x^+, x^-)}{\sum_{\substack{y \in E \\ y^- = x^-}} \Pi(y)}$$

$$= \frac{\exp\left(-\beta \sum_{m \in \Lambda^+} \sum_{\substack{m' \in \Lambda^- \\ m \sim m'}} (x^+(m) - x^-(m'))^2\right)}{\sum_{\substack{y \in E \\ y^- = x^-}} \exp\left(-\beta \sum_{m \in \Lambda^+} \sum_{\substack{m' \in \Lambda^- \\ m \sim m'}} (y^+(m) - y^-(m'))^2\right)}$$

$$= \frac{\exp\left(-\beta \sum_{m \in \Lambda^+} \sum_{\substack{m' \in \Lambda^- \\ m \sim m'}} (1 + 1 - 2x^+(m)x^-(m'))\right)}{\sum_{\substack{y \in E \\ y^- = x^-}} \exp\left(-\beta \sum_{m \in \Lambda^+} \sum_{\substack{m' \in \Lambda^- \\ m \sim m'}} (1 + 1 - 2y^+(m)y^-(m'))\right)}$$

$$= \frac{\exp\left(2\beta \sum_{m \in \Lambda^+} \sum_{\substack{m' \in \Lambda^- \\ m \sim m'}} x^+(m)x^-(m')\right)}{\sum_{\substack{y \in E \\ y^- = x^-}} \exp\left(2\beta \sum_{m \in \Lambda^+} \sum_{\substack{m' \in \Lambda^- \\ m \sim m'}} y^+(m)x^-(m')\right)}$$

$$= \frac{\prod_{m \in \Lambda^+} \exp\left(2\beta x^+(m) \sum_{\substack{m' \in \Lambda^- \\ m \sim m'}} x(m')\right)}{\sum_{\substack{y \in E \\ y^- = x^-}} \exp\left(2\beta \sum_{m \in \Lambda^+} \sum_{\substack{m' \in \Lambda^- \\ m \sim m'}} y^+(m)x^-(m')\right)}$$

So under $\Pi(\cdot | x^-)$, the components of $X^+(m)$ are independant and of law:

$$\mathbb{P}(X^+(n)=x^+(n) \mid X^-=x^-) \propto \exp\left(2\beta x^+(n) \sum_{\substack{n' \in \Lambda^- \\ n' \sim n}} x^-(n')\right)$$ (SG)

Wet set $M = \sum_{\substack{n' \in \Lambda^- \\ n \sim n'}} x^-(n')$ and we get :

$$\begin{cases} \mathbb{P}(X^+(n)=1 \mid X^-=x^-) = \dfrac{e^{2\beta M}}{e^{2\beta M}+e^{-2\beta M}} \\ \mathbb{P}(X^+(n)=-1 \mid X^-=x^-) = \dfrac{e^{-2\beta M}}{e^{2\beta M}+e^{-2\beta M}} \end{cases}$$

So it is easy to sample from $\pi(\cdot \mid x^-)$ (the case $\pi(x^- \mid x^+)$ is very similar).

III) Metropolis-Hastings algorithm (MH)

1) Basic idea

Suppose we want to sample according to some law p^*.

When in current state x, we propose a new state with probability $q(x' \mid x)$ (q is the proposal kernel) (if $q(x' \mid x) = q(x')$, the method is known as independent samples).

Having proposed a move to x', we decide whether to accept it or not, according to some formula which ensures that the fraction of time spent in each state

is proportional to $p^*(x)$.

We accept the new state x' with probability :

$$\begin{cases} \pi = \min(1, \alpha) \\ \alpha = \dfrac{p^*(x')\, q(x \mid x')}{p^*(x)\, q(x' \mid x)} \end{cases} \xrightarrow{\text{if } q \text{ is symmetric}} \text{this simplifies} \quad \frac{p^*(x')}{p^*(x)}$$

An important reason why MH is useful is that, when evaluating α, we only need $p^*(\cdot)$ up to a normalizing constant.

Result: (Under some conditions) If we call $x_1, x_2, \dots$ the chain generated by this algorithm:
- it has invariant law p^*
- $x_n \xrightarrow{\text{law}} p^* \quad (n \to \infty)$
- $\frac{1}{n} \sum_{i=1}^{n} \varphi(x_i) \longrightarrow p^*(\varphi)$

(see the handout « Introduction aux méthodes de Monte-Carlo par MH », on my teaching webpage for the details)

2) Gibbs sampling is a special case of MH

If we use MH with a sequence of proposals:

$q_i(x'|x) = p(x'_i | x_{-i}) \mathbb{1}_{x'_{-i} = x_{-i}}$

(x_{-i} means all the components of x, except x_i)

(instead of using always the same proposal, we cycle through $q_1, q_2, \dots, q_D$ (D is the dimension of x)).

the acceptance ratio is now:

$$\alpha = \frac{p(x') \, q(x|x')}{p(x) \, q(x'|x)} = \frac{p(x'_i | x'_{-i}) \, p(x'_{-i}) \, p(x_i | x'_{-i}) \, \mathbb{1}_{x'_{-i} = x_{-i}}}{p(x_i | x_{-i}) \, p(x_{-i}) \, p(x'_i | x_{-i}) \, \mathbb{1}_{x'_{-i} = x_{-i}}}$$

$$= 1.$$

So we move with the proposal $q_i(x'|x) = p(x'_i | x_{-i} = x'_{-i})$ and that is it.

3) Proposal distribution

For a given target p^*, a proposal q is valid if it gives a non-zero probability of moving to the states that have non-zero

probability on the target. Formally:

$$\text{supp}(p^*) \supseteq \bigcup_x \text{supp}(q(\cdot|x))$$

4) Why MH1 works

We first compute the matrix transition of the proposed chain.

If $x \neq x'$: probability of moving from x to x' is

$$p(x'|x) = \underbrace{q(x'|x)}_{\text{we propose } x'} \underbrace{a(x'|x)}_{\text{and } x' \text{ is accepted}}$$

From which we deduce: for $x = x'$

$$p(x|x) = 1 - \sum_{x' \neq x} p(x'|x).$$

Theorem: p^* is a stationnary distribution for the MH1 chain (it is unique if the chain is irreducible)

Proof: i) Let us show that $p^*(x)\,p(x'|x) = p^*(x')\,p(x|x')$.

For $x = x'$, it is easy. And for $x \neq x'$:

$$\begin{aligned} p^*(x)\,p(x'|x) &= p^*(x)\,q(x'|x)\,a(x'|x) \\ &= p^*(x)\,q(x'|x)\min\left(1, \frac{p^*(x')\,q(x|x')}{p^*(x)\,q(x'|x)}\right) \\ &= \min\left(p^*(x)\,q(x'|x),\ p^*(x')\,q(x|x')\right) \\ &\qquad \hookrightarrow \text{symmetric in } x, x' \\ &\vdots \\ &= p^*(x')\,p(x|x') \end{aligned}$$

ii) Let us show that $p(x'|x)\,p^*(x)$ has marginal $p^*(x')$.

For any φ:

$$\sum_{x',x} p(x'|x)\, p^*(x)\, \varphi(x') = \sum_{x',x} p^*(x')\, p(x|x')\, \varphi(x') \quad \text{(by (i))}$$

$$= \sum_{x'} p^*(x')\, \varphi(x')$$

□

5) Example : deciphering (already done in M1 IM)

Each text is written with a certain number of characters :

␣ (space), a, b, c, ... (letters), ., ; , : , ... (punctuation signs).

A simple cipher method is to attribute a number to each character : a text is then coded by a list of number. Then we choose a permutation $\sigma \in \mathfrak{S}_n$ (n = number of characters).

Original text → Ciphered text

$(a_1, a_2, \dots) \longrightarrow (\sigma(a_1), \sigma(a_2), \dots)$

↳ sequence in $\{1, 2, \dots, n\}$

(To de-cipher : apply σ^{-1})

A text is seen as a Markov chain (in the space of characters) with transition matrix M. We are given a ciphered text : $(b_1, b_2, \dots, b_N)$. To each candidate $s \in \mathfrak{S}_n$ to de-ciphering, we attribute a score :

$$\prod_{i=1}^{N-1} M(s(b_i), s(b_{i+1}))$$

(The higher the score, the more likely it is that s is the correct deciphering).

The idea is to sample according to the law

$$\pi(s) = \frac{1}{Z} \prod_{i=1}^{N-1} M(s(b_i), s(b_{i+1}))$$

↳ unknown constant

this can be done by MCMC.

We use the proposal kernel Q described below.

* When in $f \in \mathcal{F}_n$, we sample $\{x, y\}$ unif. in $\mathcal{P}_2(\{1, \dots, n\})$.
* Jump in $f^{(x,y)}$ defined by: $f^{(x,y)}(\eta) = \begin{cases} = f(\eta) & \text{if } \eta \notin \{x, y\} \\ = f(x) & \text{if } \eta = y \\ = f(y) & \text{if } \eta = x \end{cases}$