STAT 3006: Statistical Computing Lecture 8*

5 March

For a transition matrix P, if there exists a distribution $\{\pi_k^*\}$ such that $\pi_i^* p_{ij} = \pi_j^* \pi_{ji}$ for $\forall i, j$, then we say π^* satisfies the detailed balance condition. Furthermore, if π^* satisfies the detailed balance condition, it is also stationary. Noteworthy, a stationary distribution does not necessarily satisfy the detailed balance condition.

7 Markov Chain Monte Carlo (MCMC) Algorithm

For a distribution $\pi(x)$, the idea of MCMC algorithm to sample from $\pi(x)$ is constructing a Markov chain (or equivalently finding a transition distribution) with $\pi(x)$ being the chain's limiting distribution. When the chain $\{X_n : n \geq 1\}$ proceeds long enough, samples $\{X_n : n \geq B\}$ (B is a large number) can approximately be treated as samples from $\pi(x)$. In the following, we give the definition of the MCMC method.

Definition 7.1. A Markov chain Monte Carlo (MCMC) method for simulating f is any method producing an irreducible, aperiodic and positive recurrent Markov chain $\{X_n\}_{n=1}^{\infty}$ whose stationary distribution is f.

Remarks and Problems:

- A stationary distribution f in the irreducible and aperiodic Markov chain is also the chain's limiting distribution.
- The samples are not independent.
- If you want to draw n samples from f, no need to generate n chains and obtain one sample from each chain.
- How long should we stop the Markov chain?

*If you have any question about the note, please send an email to xyluo@link.cuhk.edu.hk

7.1 Metropolis-Hasting Algorithm

The Metropolis-Hasting (MH) algorithm as a MCMC method provides us with a general approach to constructing a Markov chain $\{X^{(t)}\}$ with $f(x)$ being the limiting distribution.

Algorithm: MH algorithm to sample from f.

Input: the pdf (or pmf) f, a starting point $x^{(0)}$ and a proposal distribution $q(y|x)$.
Initialize: $t \leftarrow 0$.
Repeat
generate $y_t \sim q(\cdot|x^{(t)})$;
calculate $\rho(x^{(t)}, y_t) = \min\left\{\frac{f(y_t)}{f(x^{(t)})}\frac{q(x^{(t)}|y_t)}{q(y_t|x^{(t)})}, 1\right\}$;
accept y_t as $x^{(t+1)}$ with probability $\rho(x^{(t)}, y_t)$;
otherwise, reject y_t and let $x^{(t+1)}$ be $x^{(t)}$.
$t \leftarrow t + 1$;
Until some criteria are met.
Output:Given a large number B, $\{x^{(B)}, x^{(B+1)}, \ldots\}$ are samples from the distribution f.

We can prove that the transition distribution $k(x, y)$ in MH algorithm satisfies the detailed balance condition $f(x)k(x, y) = f(y)k(y, x)$. Therefore, $f(x)$ is the limiting distribution of the Markov chain in MH algorithm.

Proof. When $x = y$, $f(x)k(x, y) = f(x)k(x, x) = f(y)k(y, x)$. When $x \neq y$,

$$\begin{aligned}
f(x)k(x, y) &= f(x)q(y|x)\rho(x, y) \\
&= f(x)q(y|x)\min\left\{\frac{f(y)q(x|y)}{f(x)q(y|x)}, 1\right\} \\
&= \min\left\{f(y)q(x|y), f(x)q(y|x)\right\} \\
&= \min\left\{f(x)q(y|x), f(y)q(x|y)\right\} \\
&= f(y)k(y, x).
\end{aligned}$$

Sometimes, the proposal distribution $q(y|x) = g(y)$ does not depend on x. It leads to the independent MH algorithm.

Algorithm: Independent MH algorithm to sample from f.

Input: the pdf (or pmf) f, a starting point $x^{(0)}$ and a proposal distribution $g(y)$.
Initialize: $t \leftarrow 0$.
Repeat
generate $y_t \sim g(\cdot)$;
calculate $\rho(x^{(t)}, y_t) = \min\left\{\frac{f(y_t)}{f(x^{(t)})}\frac{g(x^{(t)})}{g(y_t)}, 1\right\}$;
accept y_t as $x^{(t+1)}$ with probability $\rho(x^{(t)}, y_t)$;
otherwise, reject y_t and let $x^{(t+1)}$ be $x^{(t)}$.
$t \leftarrow t + 1$;
Until some criteria are met.
Output:Given a large number B, $\{x^{(B)}, x^{(B+1)}, \ldots\}$ are samples from the distribution f.

Remark 1. In this algorithm, y_ts are independent, but x_ts are still dependent.

Moreover, when $q(y|x) = q(x|y)$, we have Metropolis algorithm.

Algorithm: MH algorithm to sample from f.

Input: the pdf (or pmf) f, a starting point $x^{(0)}$ and a proposal distribution $q(y|x)$.
Initialize: $t \leftarrow 0$.
Repeat
  generate $y_t \sim q(\cdot|x^{(t)})$;
  calculate $\rho(x^{(t)}, y_t) = \min\left\{\frac{f(y_t)}{f(x^{(t)})}, 1\right\}$;
  accept y_t as $x^{(t+1)}$ with probability $\rho(x^{(t)}, y_t)$;
  otherwise, reject y_t and let $x^{(t+1)}$ be $x^{(t)}$.
  $t \leftarrow t + 1$;
Until some criteria are met.
Output:Given a large number B, $\{x^{(B)}, x^{(B+1)}, \ldots\}$ are samples from the distribution f.

As you can see, the computation of the acceptance probability becomes easier due to the symmetry of $q(y|x)$.

7.2 Gibbs Sampler

When $f(\mathbf{x})$ has a large number of variates, the transition distribution in the MH algorithm is high dimensional. A state space with a high dimensionality and a poor transition distribution often lead to very slow convergence of Markov chain to its limiting distribution. Gibbs sampler can get around the problem by iteratively sampling from *full conditional functions*.

Algorithm: Gibb sampler to sample from $f(\mathbf{x})$.

Input: the pdf (or pmf) $f(\mathbf{x})$ and a starting point $\mathbf{x}^{(0)} = (x_1^{(0)}, \ldots, x_p^{(0)})$.
Initialize: $t \leftarrow 0$.
Repeat
  generate $x_1^{(t+1)} \sim f_1(x_1|x_2^{(t)}, x_3^{(t)}, \ldots, x_p^{(t)})$;
  generate $x_2^{(t+1)} \sim f_2(x_2|x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)})$;
  $\vdots$
  generate $x_p^{(t+1)} \sim f_p(x_p|x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{p-1}^{(t+1)})$;
  $t \leftarrow t + 1$;
Until some criteria are met.
Output:Given a large number B, $\{\mathbf{x}^{(B)}, \mathbf{x}^{(B+1)}, \ldots\}$ are samples from the distribution f.

In the algorithm, univariate conditional functions $f_1, \ldots, f_p$ are called full conditional functions.

Theorem 7.1. *The Gibbs sampler is equivalent to the composition of p-MH algorithms with acceptance rate equal to one.*

Example (bivariate normal distribution): use Gibbs sampler to sample from $N(\mu, \Sigma)$, where

$\mu = (\mu_1, \mu_2)$ and $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$. If we derive the distribution's full conditional functions, it becomes easy for us to implement Gibbs sampler:

$$f(x_1|x_2) = N(x_1; \mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(x_2 - \mu_2), \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2}),$$
$$f(x_2|x_1) = N(x_2; \mu_2 + \frac{\sigma_{12}}{\sigma_1^2}(x_1 - \mu_1), \sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^2}),$$

where $N(x; \mu, \sigma^2)$ represents the density function of the normal distribution with mean μ and variance σ^2.

Example: use MH algorithm to sample from $N(\mu, \Sigma)$. We just need to specify the proposal distribution. Given (x_1, x_2), sample y_1 from $N(x_1, \sigma_1^2)$ and sample y_2 from $N(x_2, \sigma_2^2)$. The (y_1, y_2) is the proposal. We can see that the proposal distribution $q(\mathbf{y}|\mathbf{x})$ is symmetric, so the MH algorithm becomes Metropolis algorithm listed in the last subsection.