Monte Carlo and Empirical Methods for Stochastic Inference (MASM11/FMSN50)

Magnus Wiktorsson
Centre for Mathematical Sciences
Lund University, Sweden

Lecture 10
Markov chain Monte Carlo III
February 20, 2020

Plan of today's lecture

1. Last time: The Metropolis-Hastings algorithm (Ch. 7.1)
2. The Gibbs sampler (Ch. 7.2)
3. Variance of MCMC samplers

We are here $\longrightarrow \bullet$

The Metropolis-Hastings (MH) algorithm

Assuming that we can simulate from a transition density $r(z|x)$ (referred to as the proposal kernel) on X, the MH algorithm goes as follows.

Simulate a sequence of values (X_k), forming a Markov chain on X, with the following mechanism: Given X_k,

- generate $X^* \sim r(z|X_k)$ and
- set

$$X_{k+1} = \begin{cases} X^* & \text{w. pr. } \alpha(X_k, X^*) = 1 \wedge \dfrac{f(X^*)r(X_k|X^*)}{f(X_k)r(X^*|X_k)}, \\ X_k & \text{otherwise.} \end{cases}$$

The scheme is initialized by setting X_1 to an arbitrary value or drawing the same from some initial distribution χ.

The MH algorithm: Pseudo-code

draw $X_1 \sim \chi$
for $k = 1 \rightarrow (N-1)$ **do**
  draw $X^* \sim r(z|X_k)$
  set $\alpha(X_k, X^*) \leftarrow 1 \wedge \frac{f(X^*)r(X_k|X^*)}{f(X_k)r(X^*|X_k)}$
  draw $U \sim \mathcal{U}(0,1)$
  if $U \leq \alpha$ **then**
    $X_{k+1} \leftarrow X^*$
  else
    $X_{k+1} = X_k$
  end if
end for
set $\tau_N \leftarrow \frac{1}{N}\sum_{k=1}^{N} \phi(X_k)$
return τ_N

different types of proposal kernels

We also considered different classes of proposal kernels r, namely the

- independent proposals,
- symmetric proposals, and
- multiplicative proposals.

The transition kernel of the Metropolis-Hastings algorithm

Lemma (MH transition kernel)

The MH algorithm is a Markov chain with the following transition kernel:

$$q(z|x) = \alpha(x,z)r(z|x) + p_R(x)\delta_x(z),$$

where

$$p_R(x) = 1 - \int \alpha(x,z)r(z|x)dz$$

with

$$\alpha(x,z) = 1 \wedge \frac{f(z)r(x|z)}{f(x)r(z|x)}.$$

Note that $\delta_x(z) = \delta(z-x) = \delta(x-z) = \delta_z(x)$ where δ is the so called Dirac delta function.

Proof of lemma (sketch)

First note that the conditional density if we accepted the step given $X_k = x$ is

$$\mathbb{P}(U \leq \alpha(x,z)|X^* = z, X_k = x) f_{X^*|X_k=x}(z) = \alpha(x,z) r(z|x).$$

If we do not accept the step we stay at x giving a point mass at x. The probability for this to happen (i.e. the size of the point mass) is

$$\begin{aligned} & \int \mathbb{P}(U > \alpha(x,y)|X^* = y, X_k = x) f_{X^*|X_k=x}(y) \mathrm{d}y \\ = & \int (1 - \alpha(x,y)) r(y|x) \mathrm{d}y \\ = & \ 1 - \int \alpha(x,y) r(y|x) \mathrm{d}y = p_R(x) \end{aligned}$$

Putting the two parts together we get

$$q(z|x) = \alpha(x,z) r(z|x) + p_R(x) \delta_x(z). \quad \square$$

Convergence of the MH algorithm

The following result is fundamental.

Theorem

The chain (X_k) generated by the MH sampler has f as stationary distribution.

In addition, one may prove, under weak assumptions, that the MH algorithm is also geometrically ergodic, implying that

$$\tau_N = \frac{1}{N}\sum_{k=1}^{N}\phi(X_k) \to \tau \quad \text{as} \quad N \to \infty.$$

Proof of theorem (sketch)

In order to show that f is the stationary distribution we should check the global balance equation

$$\int f(x)q(z|x)\mathrm{d}x = f(z).$$

Proof of theorem (sketch)

or the local balance equation

$$f(x)q(z|x) = f(z)q(x|z), \quad \forall x, z \in \mathsf{X}.$$

Proof of theorem (sketch)

or the local balance equation

$$f(x)q(z|x) = f(z)q(x|z), \quad \forall x, z \in \mathsf{X}.$$

Using the lemma above we insert the transition kernel for the MH-algorithm yielding

$$f(x)q(z|x) = f(x)\alpha(x,z)r(z|x) + f(x)p_R(x)\delta_x(z).$$

Examining the first term we get

$$\begin{aligned} f(x)\alpha(x,z)r(z|x) &= f(x)r(z|x)\left(1 \wedge \frac{f(z)r(x|z)}{f(x)r(z|x)}\right) \\ &= f(x)r(z|x) \wedge f(z)r(x|z) = f(z)r(x|z)\left(1 \wedge \frac{f(x)r(z|x)}{f(z)r(x|z)}\right) \\ &= f(z)\alpha(z,x)r(x|z). \end{aligned}$$

proof (cont)

We now plug this into to the global balance equation

$$\begin{aligned}\int f(x)q(z|x)\mathrm{d}x &= \int \left(f(z)\alpha(z,x)r(x|z) + f(x)p_R(x)\delta_x(z)\right)\mathrm{d}x \\ &= f(z)\int \alpha(z,x)r(x|z)\mathrm{d}x + f(z)p_R(z)\end{aligned}$$

Using the expression for $p_R(z)$ we get

$$\begin{aligned}&= f(z)\int \alpha(z,x)r(x|z)\mathrm{d}x + f(z)\left(1 - \int \alpha(z,x)r(x|z)\mathrm{d}x\right) \\ &= f(z)\left(1 + \int \alpha(z,x)r(x|z)\mathrm{d}x - \int \alpha(z,x)r(x|z)\mathrm{d}x\right) = f(z),\end{aligned}$$

which concludes the proof. □

Example of particle MCMC kernel for smoothing distribution

INPUT: Conditional trajectory $X_f^{0:n}$, measurements $y^{0:n}$, parameter θ.
OUTPUT: Trajectory $X_*^{0:n}$.
draw $X_i^0 \sim \chi,\ i = 1, \cdots, N-1$ (intial distribution).
set $X_N^0 \leftarrow X_f^0$ (initilize fixed trajectory).
set $w_i^0 \leftarrow 1,\ i = 1, \cdots, N-1$ (initilize weights).
for $k = 1 \rightarrow (n-1)$ **do**
  draw a_k^i with $P(a_k^i = j) \propto w_{k-1}^j$ for $i = 1, \cdots, N-1$ (index resampling).
  draw $X_i^k \sim p_\theta(x_k | X_{a_k^i}^{k-1})$ for $i = 1, \cdots, N-1$ (move particles).
  draw a_k^N with $P(a_k^N = j) \propto w_{k-1}^j p_\theta(X_f^k | X_j^{k-1})$ (index resampling fixed trajectory).
  set $X_N^k \leftarrow X_f^k$ (update fixed trajectory).
  set $w_i^k \leftarrow p_\theta(y^k | X_i^k)$ for $i = 1, \cdots, N$ (update weights).
end for
draw I with $P(I = j) \propto w_n^j$ (initilize index for backward step).
set $X_*^n = X_I^n$ (initilize trajectory for backward step).
for $k = (n-1) \rightarrow 0$ **do**
  set $I \leftarrow a_{k+1}^I$ (track index backwards).
  set $X_*^k \leftarrow X_I^k$ (update trajectory for backward step).
end for
return $X_*^{0:n}$

We are here $\longrightarrow \bullet$

The Gibbs sampler

In the following,

- assume that the space X can be divided into m blocks, i.e. $x = (x^1, \ldots, x^m) \in \mathsf{X}$, where each block may itself be vector-valued.
- assume that we want to sample a multivariate distribution f on X.
- denote by x^i the ith component of x and by $x^{-i} = (x^\ell)_{\ell \neq i}$ the set of remaining components.
- denote by $f_i(x^i|x^{-i}) = f(x)/\int f(x)\,dx^i$ the conditional distribution of X^i given the other components $X^{-i} = x^{-i}$ and
- assume that it is easy to simulate from $f_i(x^i|x^{-i})$ for all $i = 1, \ldots, m$.

The Gibbs sampler (cont.)

The Gibbs sampler goes as follows.

Simulate a sequence of values (X_k), forming a Markov chain on X, with the following mechanism: Given X_k,

- draw $X_{k+1}^1 \sim f_1(x^1|X_k^2, \ldots, X_k^m)$,
- draw $X_{k+1}^2 \sim f_2(x^2|X_{k+1}^1, X_k^3, \ldots, X_k^m)$,
- draw $X_{k+1}^3 \sim f_3(x^3|X_{k+1}^1, X_{k+1}^2, X_k^4, \ldots, X_k^m)$,
- ...
- draw $X_{k+1}^m \sim f_m(x^m|X_{k+1}^1, X_{k+1}^2, \ldots, X_{k+1}^{m-1})$.

In other words, at the ℓth round of the cycle generating X_{k+1}, the ℓth component of X_{k+1} is updated by simulation from its conditional distribution given all other components.

Convergence of the Gibbs sampler

As for the MH algorithm, the following holds true.

Theorem

The chain (X_k) generated by the Gibbs sampler has f as stationary distribution.

In addition, one may prove, under weak assumptions, that the Gibbs sampler is also geometrically ergodic, implying that

$$\tau_N = \frac{1}{N}\sum_{k=1}^{N}\phi(X_k) \to \tau \quad \text{as} \quad N \to \infty.$$

Example: A tricky bivariate distribution

Suppose that we want to sample the distribution on $\{0, 1, 2, \ldots, n\} \times (0, 1)$ given by

$$f(x, y) \propto \frac{n!}{(n-x)!x!} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}.$$

This density is very complex and hard to sample from. The conditional distributions are however simple; indeed

- $X|Y = y \sim \text{Bin}(n, y)$,
- $Y|X = x \sim \text{Beta}(x + \alpha, n - x + \beta)$.

Thus, the problem of sampling $f(x, y)$ can be perfectly cast into the framework of the Gibbs sampler.

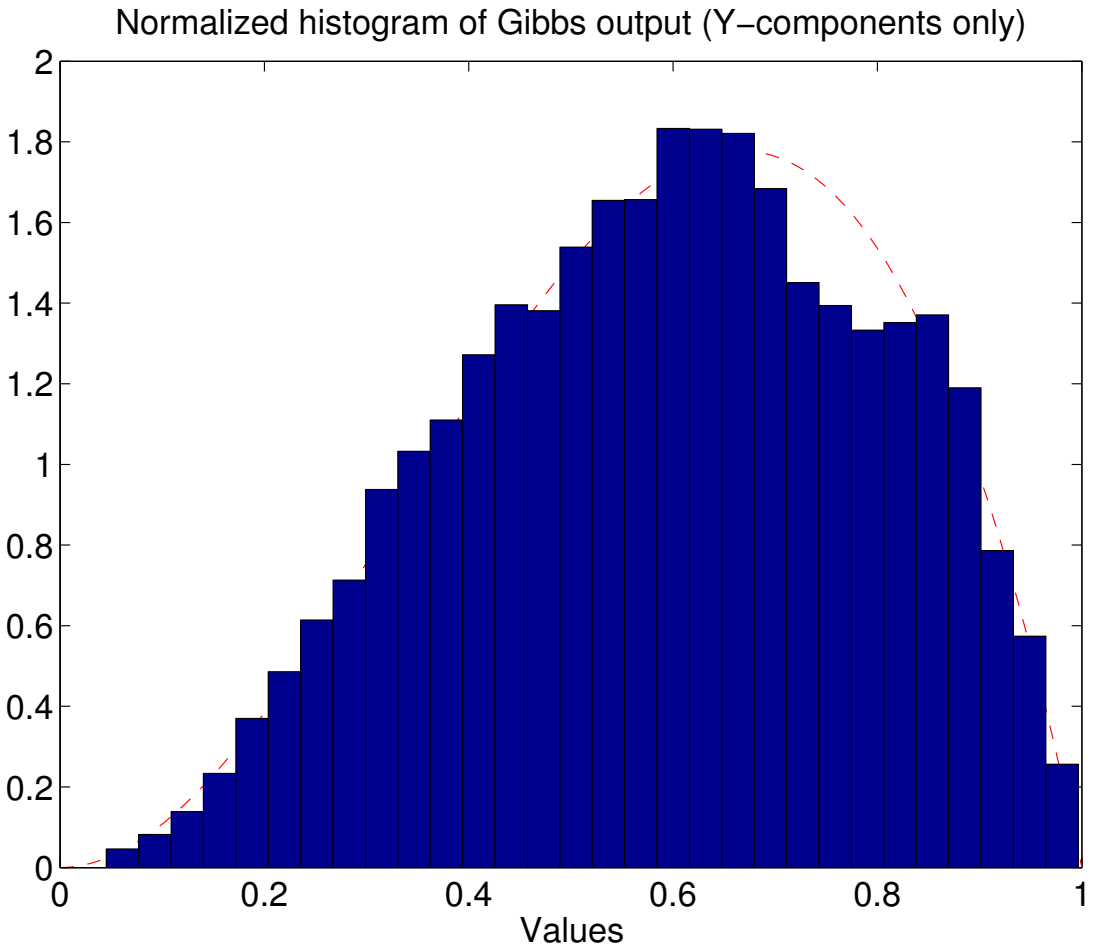
Example: A tricky bivariate distribution (cont.)

In Matlab:

```
burn_in = 1000;
M = N + burn_in;
X = zeros(1,M);
Y = X;
X(1) = 5;
Y(1) = 0.5;
for k = 1:(M - 1),
    x = binornd(n,Y(k));
    X(k + 1) = x;
    Y(k + 1) = betarnd(x + alpha,n - x + beta);
end
```

Example: A tricky bivariate distribution (cont.)

Comparison between the true density and the histogram of Y_k, $k = 1001, \ldots, 11000$.



We are here $\longrightarrow \bullet$

Variance of MCMC estimators

As mentioned, the MH and Gibbs samplers are geometrically ergodic, implying a LLN for the resulting estimators. In addition, one may establish the following CLT. Let

$$r(\ell) = \lim_{n\to\infty} \mathbb{C}(\phi(X_{n+\ell}), \phi(X_n))$$

be the covariance function of the MCMC chain at stationarity.

Theorem

For the MCMC samplers discussed above it holds that

$$\sqrt{N}(\tau_N - \tau) \xrightarrow{d.} \mathcal{N}(0, \sigma^2) \quad \text{as} \quad N \to \infty,$$

where

$$\sigma^2 = r(0) + 2\sum_{\ell=1}^{\infty} r(\ell).$$

Estimating the variance of MCMC samplers

For the ordinary Monte Carlo integration we used the sample variance (Matlab: `var`) to estimate $\mathbb{V}(\phi(X))$. However, now we need the entire covariance function $r(\ell)$. A number of different approximative solutions are possible, e.g.

- assume a parametric form of the covariance function, usually that of an AR process of low order, and estimate it,
- use only use samples that are far apart, ensuring approximate independence,
- block the samples into blocks that are large enough to be approximately independent. Then calculate averages of each block and use these to estimate the standard deviation.

Next week we will discuss the last solution in some detail.