Stat260: Bayesian Modeling and Inference **Lecture Date: April 14, 2010**

Lecture 20

Lecturer: Michael I. Jordan *Scribe: Venkatesan Ekambaram*

1 MCMC Foundations

Definition 1 (Reversibility)**.** Let (X_n) be positive (i.e. ψ irreducible and has an invariant probability measure), then it is said to be reversible if

$$(X_{n+1}|X_{n+2} = X) \stackrel{d}{=} (X_{n+1}|X_n = X). \tag{1}$$

Definition 2. (X_n) is said to satisfy the detailed balance conditions if there existis a function f such that

$$k(y, x)f(y) = k(x, y)f(x), \tag{2}$$

corresponding to the measure $k(X, .)$.

Theorem 3. *Let (X_n) satisfy the detailed balance conditions, where f is the density corresponding to a probability measure π. We then have*

1. *π is the invariant measure of (X_n).*
2. *(X_n) is reversible.*

Theorem 4. *Let (X_n) be Harris positive. Suppose there exists an ergodic atom α, we then have*

$$\lim_{n\to\infty} ||K^n(X, .) - \pi(.)||_{TV} = 0, \tag{3}$$

where the norm is under the Total Variation norm.

Theorem 5 (Ergodicity)**.** *Let π be a $\sigma-$finite invariant measure for (X_n). Let $S_n(h) = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$, $h \in \mathcal{L}_\infty(\pi)$, we then have*

$$\lim_{n\to\infty} S_n(h) = \int h(x)\pi(dx), \tag{4}$$

iff (X_n) is Harris recurrent*.*

Ergodicity says that, even though there is memory in the system, the sample average goes to the expected value under the invariant measure.

2 Invariant measure for the Metropolis Hastings

Let's now analyze the Metropolis Hastings algorithm which is a MCMC sampling algorithm and see if the samples that we get are indeed from the desired density function. Let $f(.)$ be the density of interest from which we intend to sample and let $q(.|.)$ be the proposal density. The acceptance probability is given by

$$\rho(x, y) = \min\left(1, \frac{f(y)q(x|y)}{f(x)q(y|x)}\right). \tag{5}$$

The kernel for this is then given by

$$k(x,y) = q(y|x)\rho(x,y) + \delta_x(y)(1 - r(x)), \tag{6}$$

where $r(x) = \int \rho(x,y)q(y|x)dy$ and $\delta_x(y)$ is the standard *dirac-delta* function. Let's now see how this kernel comes about. We have the following

$$K(x^{(t)}, A) = P(X^{(t+1)} \in A | X^{(t)} = x^{(t)}) \tag{7}$$

Let

$$z_t = \begin{cases} 1 & \text{if } x^{(t)} \text{ is accepted} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

We can now express $K(x^{(t)}, A)$ as follows

$$\begin{aligned} K(x^{(t)}, A) &= P(X^{(t+1)} \in A, z_t = 1 | X^{(t)} = x^{(t)}) + P(X^{(t+1)} \in A, z_t = 0 | X^{(t)} = x^{(t)}) \\ &= P(z_t = 1 | X^{(t)} = x^{(t)}, X^{(t+1)} \in A) P(X^{(t+1)} \in A | X^{(t)} = x^{(t)}) + \\ &\quad P(z_t = 0 | X^{(t)} = x^{(t)}, X^{(t+1)} \in A) P(X^{(t+1)} \in A | X^{(t)} = x^{(t)}) \\ &= \int_A q(y|x^{(t)})\rho(x^{(t)}, y)dy + \delta_A(x^{(t)}) \int_A (1 - \rho(x^{(t)}, y)) q(y|x^{(t)})dy \end{aligned}$$

Thus we get $k(x,y)$ as defined before.

Let's now check if the detailed balance equations are satisfied. Consider the first term of $k(x,y)f(x)$. We have

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

The second term of $k(x,y)f(x)$ is given by

$$f(x)\delta_x(y)(1 - r(x)) = f(y)\delta_y(x)(1 - r(y)). \tag{9}$$

From the above two equations we can see that the detailed balance condition is satisfied. We are yet to show $\pi-$irreducibility where π is the measure obtained from f. However this is more problem specific.

2.1 Independence sampler

A special case of the Metropolis Hastings sampler is the *Independence sampler*. For this we have $q(y|x) = q(y)$ and $q(y)$ is typically taken to be heavy tailed. This sampler is closer to the rejection sampler. However one main difference is that, in rejection sampling if we reject the sample, we do not retain the sample. However, in this case if we reject, then we retain the sample. In rejection sampling the proposal distribution $q(x)$, is taken so that $f(x) \leq Mq(x)$. The probability of acceptance in rejection sampling is $\frac{1}{M}$. For independence sampling, the probability of acceptance depends on the current sample and is a random variable. It can be shown that $\mathbb{E}(P(\text{acceptance})) \geq \frac{1}{M}$. Hence in independence sampling we cannot lose more samples than that in rejection sampling.

2.2 More general kernels

In general, one could have cycles and mixtures of kernels. Let $K_1, K_2, ..., K_r$ be r kernels. We can have a composition of these kernels to get $\tilde{K}$.

$$\tilde{K} = K_1 \circ K_2 \circ .... \circ K_r. \tag{10}$$

Under some weak conditions, if any one of these kernels are irreducible, then a Metropolis algorithm using such a kernel can also be shown to converge. One can also show that if a kernel that is a mixture of other kernels is used i.e.

$$K* = d_1K_1 + d_2K_2 + ... + d_rK_r, \quad \sum d_i = 1, \quad d_i > 0, \tag{11}$$

then also the algorithm converges.

2.3 Random Gibbs

This algorithm is similar to the basic Gibbs algorithm. However as opposed to updating all the samples in the $(t+1)^{th}$ step, we would now select a component i randomly and update only the sample $X_i^{(t+1)} \sim p(X_i^{(t+1)}|X_{j\neq i}^{(t)})$. This is also a Metropolis Algorithm with the acceptance ratio

$$\rho = \min\left(1, \frac{p(X_i^{(t+1)}, X_{j\neq i}^{(t)})}{p(X_i^{(t)}, X_{j\neq i}^{(t)})} \frac{p(X_i^{(t)}|X_{j\neq i}^{(t)})}{p(X_i^{(t+1)}|X_{j\neq i}^{(t)})}\right) = 1. \tag{12}$$

Thus the Gibbs kernel gives an acceptance rate of one.

However one should be aware that the basic Gibbs algorithm in itself is not a Metropolis Hastings algorithm. By imposing an order in updating the samples, we lose on the irreducibility. But this can be shown to have an invariant distribution, if we use a deterministic sequence of distributions. Consider the following kernel

$$K(x, x') = p(X_1'|X_2, .., X_p)p(X_2'|X_1', X_3, ..., X_p)....p(X_p'|X_1', X_2', ..., X_{p-1}'). \tag{13}$$

We need to see if the given sequence of updates gives us samples X' that are from the true distribution $p(X' \in A)$. Consider the following

$$\begin{aligned}
\int \mathbf{1}_A(X')K(X, X')p(X)dXdX' &= \int \mathbf{1}_A(X')p(X_1'|X_2, .., X_p)p(X_2'|X_1', X_3, ..., X_p).... \\
&\quad p(X_p'|X_1', X_2', ..., X_{p-1}')p(X_1|X_2, ..., X_p)p(X_2, ..., X_p)dX_1...dX_pdX' \\
&= \int \mathbf{1}_A(X')p(X_1'|X_2, .., X_p)p(X_2'|X_1', X_3, ..., X_p).... \\
&\quad p(X_p'|X_1', X_2', ..., X_{p-1}')p(X_2, ..., X_p)dX_2...dX_pdX' \\
&= \int \mathbf{1}_A(X')p(X_1', X_2, .., X_p)p(X_2'|X_1', X_3, ..., X_p).... \\
&\quad p(X_p'|X_1', X_2', ..., X_{p-1}')dX_2...dX_pdX' \\
&= \int \mathbf{1}_A(X')p(X_2|X_1', X_3, ..., X_p).....dX_2...dX_pdX' \\
&= \int \mathbf{1}_A(X')p(X')dX' \\
&= p(X' \in A)
\end{aligned}$$

3 Rao-Blackwellization

Suppose the samples can be grouped as $(X^{(t)}, Y^{(t)})$ and $X^{(t)}$ are the only samples of interest and we intend to compute the mean value of some function $h(X)$. A naive estimator would get samples from both $(X^{(t)}, Y^{(t)})$ and compute the estimate $\frac{1}{T}\sum_{t=1}^{T} h(X^{(t)})$. A Rao-Blackwellized sampler would compute $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}(h(X)|Y^{(t)})$. Let's consider the variance of the samples. Then by the total variance rule we have $\mathrm{Var}(X) \geq \mathrm{Var}(\mathbb{E}(X|Y))$. However this is only true for two random variables (X, Y) and does not extend to vectors.

4 Improper priors

One should be careful that the theory holds true for any measure and not necessarily only the probability measure. Hence it is possible that we may converge to an improper posterior. For example consider the following system

$$\begin{aligned} Y_{ij} &= \mu + \alpha_i + \epsilon_{ij} \\ \alpha_i &\sim N(0, \sigma^2) \\ \epsilon_{ij} &\sim N(0, \tau^2) \\ \pi(\mu, \sigma, \tau) &\propto \frac{1}{\sigma\tau} \end{aligned}$$

We can prove that this system is ergodic and the Gibbs measure converges. However the posterior turns out to be flat and the samples obtained would not be useful.