Markov Chain Monte Carlo

Galin L. Jones

School of Statistics

University of Minnesota

Draft: March 4, 2023

Contents

1 Introduction

Later.

2 Markov Chains

Let $(\mathsf{X}, \mathcal{B})$ be a measurable space. A sequence of X-valued random variables $\{X_1, X_2, X_3, \ldots\}$ is a Markov chain if for all g

$$E\left[g(X_{n+1}, X_{n+2}, \ldots) \mid X_n, X_{n-1}, \ldots, X_1\right] = E\left[g(X_{n+1}, X_{n+2}, \ldots) \mid X_n\right].$$

Then P is a *Markov kernel* if $P : \mathsf{X} \times \mathcal{B} \to \mathbb{R}$ satisfying (i) for each fixed $x \in \mathsf{X}$, $P(x, \cdot)$ is a probability measure and (ii) for each fixed $B \in \mathcal{B}$, $P(\cdot, B)$ is a measurable function.

When X is a discrete set a Markov kernel can be represented as a square matrix whose entries are nonnegative and whose rows sum to 1.

Example 2.1. Suppose

$$P = \begin{pmatrix} 1/2 & 1/2 \\ 1/3 & 2/3 \end{pmatrix}$$

Then P is a Markov matrix on two states $\{0, 1\}$, say. The first row for example, is interpreted as the probability of moving in one step from state 0 to state 0 is $1/2$ which is the same as the probability of moving in one step from state 0 to state 1.

Example 2.2. Let $\mathsf{X} = \mathbb{Z}$ and let $0 < \theta < 1$. If $x \geq 1$, then a Markov kernel is defined by the matrix P with elements

$$P(x, x+1) = P(-x, -x-1) = \theta, \qquad P(x, 0) = P(-x, 0) = 1 - \theta,$$

and $P(0, 1) = P(1, 0) = 1/2$.

Often X will be uncountable and $\mathcal{B}$ will be countably generated. If X is topological, then $\mathcal{B}$ will be the Borel σ-algebra generated by X.

Example 2.3. Let $\mathsf{X} = (0, 1)$ and consider the Markov chain that evolves as follows. Draw $U \sim$ Uniform$(0, 1)$. If $u \leq 0.5$, $X_{n+1} \sim$ Uniform$(0, X_n)$, but if $u > 0.5$, $X_{n+1} \sim$ Uniform$(X_n, 1)$. Then if $X_n = x$ and $B \in \mathcal{B}$

$$P(x, B) = \int_B \left[\frac{1}{2}\frac{1}{x} I_y((0, x)) + \frac{1}{2}\frac{1}{1-x} I_y(x, 1)\right] \mathrm{d}y.$$

In Example 2.3, the integrand in the Markov kernel is a conditional density on X. This is a setting that will be encountered repeatedly throughout. If there is a conditional density $k(y \mid x)$, with respect to a measure λ, such that the Markov kernel satisfies for $B \in \mathcal{B}$

$$P(x, B) = \int_B k(y \mid x) \lambda(\mathrm{d}y),$$

then say k is a *Markov transition density.*

Example 2.4. Suppose $f(x, y)$ is a joint density with support $\mathbb{R}^2$ and conditional densities $f_{X|Y}(x \mid y)$ and $f_{Y|X}(y \mid x)$. Then

$$k(x', y' \mid x, y) = f_{X|Y}(x' \mid y) f_{Y|X}(y' \mid x')$$

is a Markov transition density. The Markov chain evolves from $(X_k = x, Y_k = y)$ to (X_{k+1}, Y_{k+1}) by drawing $X_{k+1} \sim F_{X|Y}(\cdot \mid y)$ followed by $Y_{k+1} \sim F_{Y|X}(\cdot \mid X_{k+1})$. This is a special case of the so-called two-variable Gibbs sampler.

Suppose λ is a positive measure on $(\mathsf{X}, \mathcal{B})$, define

$$\lambda P(B) = \int_{\mathsf{X}} \lambda(\mathrm{d}x) P(x, B). \tag{1}$$

When λ is a probability measure, the encouraged interpretation is that $X_{n+1} \mid X_n \sim P(X_n, \cdot)$ and $X_n \sim \lambda$, the product $\lambda(\mathrm{d}x)P(x, \cdot)$ is the joint distribution of (X_n, X_{n+1}) and λP is the marginal distribution of X_{n+1}.

Since Markov kernels act to the left on measures (1),

$$P^2(x, B) = \int_{\mathsf{X}} P(x, \mathrm{d}x_k) P(x_k, B).$$

Continuing in this fashion obtain for every $n \geq 2$

$$P^n(x, B) \int_{\mathsf{X}} P(x, \mathrm{d}x_k) P(x_k, \mathrm{d}x_{k+1}) \cdots P(x_{k+n-2}, B).$$

More generally, the so-called Chapman-Kolmogorov equations hold for $n \geq m \geq 0$

$$P^n(x, B) \int_{\mathsf{X}} P^m(x, \mathrm{d}y) P^{n-m}(y, B).$$

If $\lambda = \lambda P$, then λ is *invariant* for P. Notice that if λ is invariant for P and $X_n \sim \lambda$, then $X_{n+1} \sim \lambda$. That is, the marginal distribution does not depend upon n in which case the Markov chain is *stationary*.

Come back to these examples.

Example 2.5.

Example 2.6. Recall the Markov chain defined in Example (2.4)

One common way of establishing invariance of MCMC Markov chains is to verify a *detailed balance condition*; see Exercise 3.1. Detailed balance holds if

$$\lambda(\mathrm{d}x)P(x, \mathrm{d}y) = \lambda(\mathrm{d}y)P(y, \mathrm{d}x). \tag{2}$$

When λ is a probability measure, one interpretation is that the joint distribution of (X_k, X_{k+1}) is the same as the distribution of (X_{k+1}, X_k) so that this is also often called the *reversibility condition*. Another name often encountered is that P is λ-*symmetric*.

2.1 Stability

MCMC applications typically are constructed so that a specific probability distribution F is invariant. However, in applications where MCMC is required it is typically difficult to simulate from the invariant distribution. The most that can be hoped for is that the simulation will eventually produce a representative sample from F. This long-run behavior is in no way guaranteed without additional assumptions. The following simple examples illustrate that the problems can arise due to the either the way the kernel is specified or the properties of the state space X.

Example 2.7. Suppose F lives on $\{1, 2\}$ with $F(1) = 1 - F(2) = 1/4$ and

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Since the Markov chain moves deterministically between the two states, it will over represent state 1 and underrepresent state 2 no matter how many iterations there are.

Example 2.8. Suppose F lives on $\{1, 2, 3\}$ with $F(1) = F(2) = F(3) = 1/3$ and

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Thus starting at state $\{3\}$ the chain remains there forever while starting from $\{1, 2\}$ the chain will never visit $\{3\}$. Thus the chain cannot represent F in the long run.

Example 2.9. Suppose for $i = 1, 2$, f_i is a pf on $\mathsf{X}_i \subseteq \mathbb{R}$ and g_i is a pf on $\mathsf{Y}_i \subseteq \mathbb{R}$. Set

$$f(x, y) = \frac{1}{2} f_1(x) g_1(y) + \frac{1}{2} f_2(x) g_2(y).$$

Then

$$f_{X|Y}(x \mid y) = \frac{f_1(x)g_1(y) + f_2(x)g_2(y)}{g_1(y) + g_2(y)}$$

and

$$f_{Y|X}(y \mid x) = \frac{f_1(x)g_1(y) + f_2(x)g_2(y)}{f_1(y) + f_2(y)}$$

and the Gibbs sampler MTD is

$$k(x', y' \mid x, y) = f_{X|Y}(x' \mid y) f_{Y|X}(y' \mid x').$$

When $\mathsf{X}_i = \mathsf{Y}_i = \mathbb{R}$ this Gibbs sampler will produce a representative sample eventually. However, complications may arise if the spaces are constrained.

Suppose $\mathsf{X}_1 = \mathsf{X}_2 = (0, 1)$ and that $f_1 = f_2$ is the Uniform density. Let $Y_1 = (0, 1)$ and $Y_2 = (2, 3)$ and g_1 and g_2 be Uniform densities. Easy calculation yields that

$$f_{X|Y}(x \mid y) = I(0 < x < 1) \left[I(0 < y < 1) + I(2 < y < 3) \right]$$

and

$$f_{Y|X}(y \mid x) = \frac{1}{2} I(0 < x < 1) \left[I(0 < y < 1) + I(2 < y < 3) \right]$$

and that, no matter which square, $\mathsf{X}_1 \times \mathsf{Y}_1$ or $\mathsf{X}_2 \times \mathsf{Y}_2$, the current state is in there is a positive probability of the next state being in either square. This Gibbs sapler will eventually produce a representative sample.

Now consider the setting with $\mathsf{X}_1 = \mathsf{Y}_1 = (0, 1)$ and $\mathsf{X}_2 = \mathsf{Y}_2 = (2, 3)$ so that

$$f_{X|Y}(x \mid y) = \frac{I(0 < x < 1)I(0 < y < 1) + I(2 < x < 3)I(2 < y < 3)}{I(0 < y < 1) + I(2 < y < 3)}$$

and

$$f_{Y|X}(y \mid x) = \frac{I(0 < x < 1)I(0 < y < 1) + I(2 < x < 3)I(2 < y < 3)}{I(0 < x < 1) + I(2 < x < 3)}.$$

If $y \in (0, 1)$, then $f_{X|Y}(x \mid y) = I(0 < x < 1)$ and, similarly, if $X \in (0, 1)$, then $f_{Y|X}(y \mid x) = I(0 < y < 1)$. Thus if the current state is in the square $\mathsf{X}_1 \times \mathsf{Y}_1$, then the next step will be in $\mathsf{X}_1 \times \mathsf{Y}_1$. That is, there is no chance for the chain to visit $\mathsf{X}_2 \times \mathsf{Y}_2$. This Gibbs sampler will not produce a representative sample from the target distribution.

The above examples demonstrate that one way problems arise is when the Markov chain cannot access all of the space eventually and hence properties which avoid this are required.

Let ϕ be a non-triviual positive measure on $\mathcal{B}$. Then $A \in \mathcal{B}$ is *ϕ-communicating* if for all $B \subseteq A$ such that $\phi(B) > 0$ and for all $x \in A$, there exists n such that $P^n(x, B) > 0$.

This is a weak property that does not ensure desirable long run properties. Consider the Gibbs sampler from Example 2.9 with $\mathsf{X}_1 = \mathsf{Y}_1 = (0, 1)$ and $\mathsf{X}_2 = \mathsf{Y}_2 = (2, 3)$. If ϕ denotes Lebesgue measure, then this Markov chain is ϕ-communicating but does not have desirable long run properties.

The Markov kernel, P, is *ϕ-irreducible* if for all $x \in \mathsf{X}$ and for all $A \in \mathcal{B}$ such that $\phi(A) > 0$ there exists n such that $P^n(x, A) > 0$. This is a key property in many MCMC settings, but is not enough to ensure desirable long run properties; consider the Markov chain in Example 2.7 which is irreducible.

There is some arbitrariness in the definition of ϕ-irreducibility, but if for some ϕ, P is ϕ-irreducible, then there exists a maximal irreducibility measure ψ (meaning that $\psi(A) = 0$ implies $\phi(A) = 0$ for all irreducibility mmeasure ϕ).

Proposition 2.1. *If λ is an invariant measure for the Markov kerenle P and, for some ϕ, P is ϕ-irreducible, then P is λ-irreducible.*

Corollary 2.1. *If P is λ-symmetric and ϕ-irreducible, then P is λ-irreducible.*

Often inspection of the kernel P and the underlying space X is enough to establish ϕ-irreducibility. For example, many kernels obviously satisfy a positivitiy condition so that $P(x, A) > 0$ for all $x \in \mathsf{X}$ and $A \in \mathcal{B}$ with $\phi(A) > 0$. The following is a special case of a more general result [1, Theorem 3], but is often useful.

Proposition 2.2. *Suppose X is a connected, separable metric space. If every nonempty, open set A satisfies $\phi(A) > 0$ and every point has a ϕ-communicating neighborhood, the Markov kernel is ϕ-irreducible.*

3 Constructing MCMC Algorithms

3.1 Metropolis-Hastings

Give credit to R?

The fundamental MCMC algorithm is Metropolis-Hastings [2, 3]. It serves as a building block for many, if not most, applications of MCMC. The following will be generalized later, but, for now, suppose $\mathsf{X} \subseteq \mathbb{R}^d$ and that $q(x, y)$ is a proposal conditional density that is easy to sample. This notation can take an adjustment for those accustomed to the alternative $q(y \mid x)$, which is more common in statistics. Define the *Hastings ratio*

$$r(x, y) = \frac{f(y)q(y, x)}{f(x)q(x, y)}$$

and note that the numerator defines a joint density for (X, Y) while the denominator defines a joint density for (Y, X).

Algorithm 1 Metropolis-Hastings

1: *Input:* Current value $X_n = x$.

2: Draw $Y \sim Qx, \cdot)$

3: Draw $U \sim \text{Uniform}(0, 1)$

4: If $u \leq r(x, y) \wedge 1$, accept y and set $X_{n+1} = y$, else set $X_{n+1} = x$.

Algorithm 1 is the formal definition, but is not at all how the algorithm should be implemented in practice.

Exercises

Exercise 3.1. Prove that if Equation 2 holds, then λ is invariant for P.

Exercise 3.2. What is the invariant distribution of the Markov chain defined in Example 2.7?

Exercise 3.3. Consider the Gibbs samplers in Example 2.9. establish that the Gibbs sampler on $\mathsf{X}_1 = \mathsf{Y}_1 = (0, 1)$ and $\mathsf{X}_2 = \mathsf{Y}_2 = (2, 3)$ is not irreducible while the one on $\mathsf{X}_1 = \mathsf{Y}_1 = \mathsf{X}_2 = (0, 1)$ and $\mathsf{Y}_2 = (2, 3)$ is irreducible.

Exercise 3.4. Establish Proposition 2.1.

Exercise 3.5. Establish Proposition 2.2.

Appendix

References

[1] Geyer, C. J. (2014). Stat 8501 lecture notes.

[2] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1).

[3] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machine. *Journal of Chemical Physics*, 21.