PART ONE

Subjects in Modeling, Simulation, and Estimation

CHAPTER SIXTEEN

Markov Chain Monte Carlo

James C. Spall
Johns Hopkins University, Baltimore, MD

16.1 Introduction/Purpose of the Chapter

The previous two chapters considered the interface of simulation and optimization. This chapter on Markov chain Monte Carlo (MCMC) continues the study of simulation-related methods, but with a different focus. MCMC is a powerful means for generating random samples that can be used in computing statistical estimates, numerical integrals, and marginal and joint probabilities. The approach is especially useful in statistical applications where one is forming an estimate based on a multivariate probability distribution or density function that would be hopeless to obtain analytically. In particu-

lar, MCMC provides a means for generating samples from joint distributions based on easier sampling from conditional distributions. The approach has had a large impact on the theory and practice of statistical modeling. In fact, MCMC sometimes applies in problems where it is hard to imagine any other approach working.

16.2 Vignette/Historical Notes

Markov chain Monte Carlo (MCMC) is a powerful means for generating random samples that can be used in computing statistical estimates and in computing marginal and conditional probabilities. MCMC methods rely on a dependent (Markov) sequence with a limiting distribution corresponding to a distribution interest.[1]

Markov chain Monte Carlo (MCMC) is a powerful means for generating random samples that can be used in computing statistical estimates and in computing marginal and conditional probabilities.

Although MCMC has general applicability, one area where MCMC has had a revolutionary impact is Bayesian analysis. MCMC has greatly expanded the range of problems for which Bayesian methods can be applied.

$$E[f(X)] = f(x)p(x)\,dx, \tag{16.1}$$

where the integral is over the domain for X. The density $p(x)$ is sometimes called the *target density*. More generally, the target density (distribution) represents the distribution for the random variables of interest for the analysis. In some cases, for example, the target will pertain to a subset of the elements in X (e.g., it may represent the marginal distribution for only the first component of X).

16.3 Theory and Applications

16.3.1 GIBBS SAMPLING

Gibbs sampling represents an implementation of the M-H algorithm on an element-by-element basis for the components in X. The term Gibbs sampling was introduced by Geman and Geman (1984) in a specific implementation of a Gibbs distribution for sampling on lattices . The term is now used more generally (and casually) to refer to the special case where the proposal distribution is built directly from the density of interest $p(0)$. Gibbs sampling

[1]The term *Marcov Chain* is sometimes reserved for use with processes having discrete outcomes. In general applications of MCMC, the revelant processes may have discrete, continuous, or hybrid outcomes.

is especially important in Bayesian implementations of the M-H algorithm. Gibbs sampling is uniquely designed for multivariate problems.

16.3.2 THEORETICAL FOUNDATION FOR GIBBS SAMPLING

While our focus is Gibbs sampling, given the close connection between Gibbs sampling and the M-H algorithm introduced earlier, the arguments here also provide some flavor of the basis for M-H, although the detailed arguments are somewhat different. This relatively informal discussion is a simplified version of the discussion in Gelfand and Smith (1990) and Robert and Casella (1999, Sect. 7.1.3).

EXAMPLE 16.1 Gibbs sampling for a normal distribution

Suppose that $XN(M-5)$ for some mean vector and covariance matrix S. Note that, the Gibbs sampler may not be the most efficient method of generating samples from a multivariate normal distribution. This example serves to illustrate more general principles, where Gibbs sampling is used to generate samples from non-standard distributions.

(16.2) $$E[f(X)] = f(x)p(x)\,dx$$

A standard result from multivariate normality is that the distribution of any selection of components within X conditioned on the remaining components is also normal (e.g., Mardia, 1979, pp. 62-63). Specifically, the distribution of the ith component conditioned on the remaining components provides the sampling distribution.

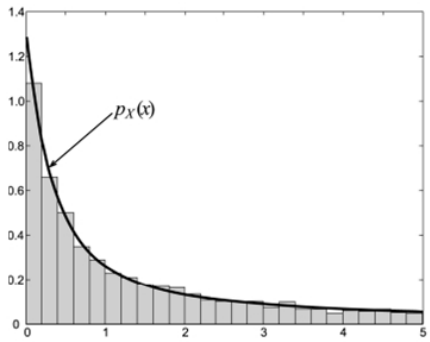

FIGURE 16.1 shows a histogram of output for a Gibbs sampler based on $n = 40$. The histogram is constructed from the terminal output of the chain using 5000 independent replications. The histogram closely matches the marginal density, indicating that the chain output has a distribution close to the desired distribution.

TABLE 16.1 Examples of two popular general forms for proposal distributions.

General Form of Proposal Distribution	$q(w\|x)$	$q(\|w)$
Normal with covariance matrix Σ	$N(x, \sigma)$	$N(w, \Sigma)$
Uniform of width 2δ for each component	$U_m(x - \delta l_m, x + \delta l_m)$	$U_m(w - \delta l_m, w + \delta l_m)$

Although MCMC has general applicability, one area where MCMC has had a revolutionary impact is Bayesian analysis. MCMC has greatly expanded the range of problems for which Bayesian methods can be applied.

16.4 Algorithms and Formulae

16.4.1 M-H ALGORITHM FOR ESTIMATING $E[F(X)]$

Step 0. (Initialization) Choose the length of the .burn-in. period M and an arbitrary initial state $X0$. Set $k = 0$.

Step 1. Generate a candidate point W according to the proposal distribution $q(|Xk)$.

Step 2. Generate a point U from a $U(0, 1)$ distribution. Set $Xk + 1 = WifU = (Xk, W)$ from (16.3). Otherwise set $Xk + 1 = Xk$.

Step 3. Repeat Steps 1 and 2 until XM is available. Terminate .burn-in. process and proceed to step 4 with $Xk = XM$.

Step 4. Carry out step 1.

Step 5. Carry out step 2.

Step 6. Repeat steps 4 and 5 until it is possible to compute the ergodic average of $n - M$ evaluations in (16.2). (Of course, if desired, this average can be computed recursively without storing all of $f(XM+1)$, $f(XM+2)$, $f(Xn)$. This ergodic average is the estimate of $E[f(X)]$ under the target density $p()$.

There are, of course, many ways in which the M-H algorithm can be implemented. The most obvious variation in implementation is in the choice of the proposal distribution $q(a|b)$. Although almost any choice of $q(a|b)$ will

work in the sense that the ergodic average in (16.2) will converge to $E[f(X)]$, there are clear differences in the rate of convergence depending on the nature of the problem.

16.4.1.1 Review of Bayesian Framework

One variation is to run many independent chains, each chain terminating at $XM + 1$. In this way, $E[f(X)]$ is estimated by forming a sample mean of independent values f.

To represent a 4 head. This creates independent blocks of iterations, allowing for the proposal distribution to be adapted at each block to improve the sampling. Let us present a simple example where the target density is bivariate.[2]

While we saw that Gibbs may be considered a special case of M-H, the techniques have developed largely independently of each other.

$$pX(x) = pX|Y,Z(x|y,z)pY,Z(y,z)\,dydz, \tag{16.3}$$

$$pY(y) = pY|X,Z(y|x,z)pX,Z(x,z)\,dxdz, \tag{16.4}$$

$$pZ(z) = pZ|X,Y(z|x,y)pX,Y(x,y)\,dxdy. \tag{16.5}$$

Note the presence of a full conditional in each of the integrands above. The full conditionals form the basis for the Markov aspect of the sampling because the next random variate is generated based on only the most recent conditioning.

1. *To represent a list internal head.* The techniques have developed largely independently of each other.
2. Recognizing this, we discuss M-H and Gibbs as separate approaches.
3. As with other stochastic search methods, no one approach is to be universally preferred.

While we saw that Gibbs may be considered a special case of M-H, the techniques have developed largely independently of each other. Recognizing this, we discuss M-H and Gibbs as separate approaches.

While we saw that Gibbs may be considered a special case of M-H, the techniques have developed largely independently of each other.

Recognizing this, we discuss M-H and Gibbs as separate approaches.

[2]The term Markov chain is sometimes reserved for use with processes having discrete outcomes. In general applications of MCMC, the relevant processes may have discrete, continuous, or hybrid outcomes. For consistency with standard terminology in the MCMC area, we follow suit in this chapter in using the term Markov chain under the more general application to discrete or continuous outcomes.

Gibbs sampling derives its name from the physicist Josiah W. Gibbs, 1839-1903, based on the connection to Gibbs random fields identified in Geman and Geman (1984).

As with other stochastic search methods, no one approach.

While we saw that Gibbs may be considered a special case of M-H, the techniques have developed largely independently of each other.

I. While we saw that Gibbs may be considered a special case of M-H, the techniques have developed largely independently of each other.

A. Recognizing this, we discuss M-H and Gibbs as separate approaches.

1. As with other stochastic search methods, no one approach is to be universally preferred.

a. To indicate the fourth level of outline list.

II. While we saw that Gibbs may be considered a special case of M-H, the techniques have developed largely independently of each other.

16.5 Summary

The discussion above summarizes the motivation, theory, implementation, and connection to Bayesian analysis for MCMC. The focus is on the M-H and Gibbs sampling versions of MCMC. Many large-scale practical implementations of MCMC borrow aspects from both M-H and Gibbs sampling.

While we saw that Gibbs may be considered a special case of M-H, the techniques have developed largely independently of each other. Recognizing this, we discuss M-H and Gibbs as separate approaches. As with other stochastic search methods, no one approach is to be universally preferred. One strong aspect of both M-H and Gibbs is the theory supporting the methods and guaranteeing convergence under modest conditions.

KEY TERMS

Burn-in Period: The first M iterations in a Markov chain.

Proposal Distribution: sometimes called an instrumental distribution or a candidate- generating, the proposal distribution may be chosen arbitrarily, although there may be efficiency advantages to one form over another in some applications. The proposal distribution satisfies the key condition for density functions.

EXERCISES

16.1 Discuss why the ergodic average of a given number of samples in typical applications of the M-H (and other) algorithms will have greater

variability than a corresponding average of the same number of independent samples of f (X). In the demonstration of this point in Example 16.1, verify that the standard deviation for the independent samples case is 0.0159.

16.2 **(a)** Based on 50 independent replications of the M-H algorithm in Example 16.1 with the uniform proposal distribution depicted in Figure 16.1, test whether the mean of the terminal estimate is statistically indistinguishable from the true value of 0.

(b) Repeat the test with a normal proposal distribution but other aspects of Example 16.1 unchanged. In particular, assume W | X = x N(x, I2/12) (note that this proposal distribution has the same mean and variance as the original uniform distribution).

(c) Finally, do the same test above, but with a U2 (x -212, x + 212) proposal distribution (see Table 16.1). Comment on the observed differences in the performance for the three proposal distributions.

16.3 Suppose X is bivariate normally distributed where the marginal distribution for the two components is $N(0,1)$. Present the two sampling distributions, $p1(x|Xk2)$ and $p2(x|Xk1)$, for use in step 1 (and 3) of the Gibbs sampling algorithm.

ANSWERS

16.3 Let X be the current state value and W be the candidate point. The candidate point W is accepted with probability (X, W). This probability is random as it depends on X and W. For convenience, let $p(x,w) = p(w)q(x)/[p(x)q(w)]$ (so $(x,w) = min\{p(x,w),1\}$ according to (16.3)).

(a) After several steps involving the re-expression of the integrals above (Reader should show these steps), it is found that $E[(X,W)] = 2$, $p(x,w) = 1$, $p(x)q(w)\ dxdw$. The result to be proved then follows in several more steps (reader to show) by invoking the given inequality $p(x) = Cq(x)$ (Incidentally, the form of M-H where $q(w|x)q(w)$ is sometimes called the independent M-H sampler.)

16.6 For the bivariate setting here, the general expression in (16.9) simplifies considerably. In particular, $i = 0$, $I, i = \cdot$, and $i, |I = 1$. Hence, the bivariate sampling for the Gibbs sampling procedure.

[You can label and ref the exercises and answers. For example, \ref{exer}, \ref{subexer}, \ref{answer}, \ref{subanswer} produces:
Ex. 16.1, Ex. 16.2.b, Ans. 16.3, Ans. 16.3.a.]

REFERENCES

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995), *Bayesian Computation and Stochastic Systems*, Statistical Science, vol. 10, pp. 3-66.

Cappé, O. and Robert, C. P. (2000), *Markov Chain Monte Carlo: 10 Years and Still Running!*, Journal of the American Statistical Association, vol. 95, pp. 1282-1286.

Casella, G. and George, E. I. (1992), *Explaining the Gibbs Sampler*, The American Statistician, vol. 46, pp. 167-174.

Chen, M.-H, Shao, Q.-M., and Ibrahim, J. G. (2000), *Monte Carlo Methods in Bayesian Computation*, Springer-Verlag, New York.

Chib, S. (1995), Marginal Likelihood from the Gibbs Sampler, Journal of the American Statistical Association, vol. 90, pp. 1313-1321.

Chib, S. and Greenberg, E. (1995), *Understanding the Metropolis-Hastings Algorithm*, The American Statistician, vol. 49, pp. 327-335.

Evans, M. and Swartz, T. (1995), *Methods for Approximating Integrals in Statistics with Special Emphasis on Bayesian Integration Problems*, Statistical Science, vol. 10, pp. 254-272.

Frigessi, A., Gasemyr, J., and Rue, H. (2000), *Antithetic Coupling of Two Gibbs Sampling Chains*, Annals of Statistics, vol. 28, pp. 1128-1149.

Gelfand, A. E. (2000), *Gibbs Sampling*, Journal of the American Statistical Association, vol. 95, pp. 1300-1304.

Gelfand, A. E. and Smith A. F. M (1990), *Sampling-Based Approaches to Calculating Marginal Densities*, Journal of the American Statistical Association, vol. 85, pp. 399-409.

Geman, S. and Geman, D. (1984), *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-6, pp. 721-741.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (eds.) (1996), Markov Chain Monte Carlo in Practice, Chapman and Hall, London.

Gilks, W. R., Roberts, G. O., and Sahu, S. K. (1998), *Adaptive Markov Chain Monte Carlo Through Regeneration*, Journal of the American Statistical Association, vol. 93, pp. 1045-1054.

Hastings, W. K. (1970), *Monte Carlo Sampling Methods Using Markov Chains and their Applications*, Biometrika, vol. 57, pp. 97-109.

Liu, J. S. (2001), Monte Carlo Strategies in Scientific Computing, Springer, New York.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), Multivariate Analysis, Academic,New York.

Metropolis, N., Rosenbluth, A., Rosenbluth, M. Teller, A., and Teller, E. (1953), *Equation of State Calculations by Fast Computing Machines*, Journal of Chemical Physics, vol. 21, pp. 1087-1092.

Parzen, E. (1962), Stochastic Processes, Holden-Day, New York.

Robert, C. P. and Casella, G. (1999), Monte Carlo Statistical Methods, Springer-Verlag, New York.