Markov Chain Monte Carlo.

Gibbs Sampler.

Anatoli Iambartsev
IME-USP

[CG] Introduction.

Explaining the Gibbs Sampler

GEORGE CASELLA and EDWARD I. GEORGE*

Computer-intensive algorithms, such as the Gibbs sampler, have become increasingly popular statistical tools, both in applied and theoretical work. The properties of such algorithms, however, may sometimes not be obvious. Here we give a simple explanation of how and why the Gibbs sampler works. We analytically establish its properties in a simple case and provide insight for more complicated cases. There are also a number of examples.

KEY WORDS: Data augmentation; Markov chains; Monte Carlo methods; Resampling techniques.

applications of the Gibbs sampler have been in Bayesian models, it is also extremely useful in classical (likelihood) calculations [see Tanner (1991) for many examples]. Furthermore, these calculational methodologies have also had an impact on theory. By freeing statisticians from dealing with complicated calculations, the statistical aspects of a problem can become the main focus. This point is wonderfully illustrated by Smith and Gelfand (1992).

In the next section we describe and illustrate the application of the Gibbs sampler in bivariate situations. Section 3 is a detailed development of the underlying theory, given in the simple case of a 2×2 table with

[CG] Introduction.

"The Gibbs sampler is a technique for generating random variables from a (marginal) distribution indirectly, without having to calculate the density. Although straightforward to describe, the mechanism that drives this scheme may seem mysterious. The purpose of this article is to demystify the workings of these algorithms by exploring simple cases. In such cases, it is easy to see that Gibbs sampling is based only on elementary properties of Markov chains.

Through the use of techniques like the Gibbs sampler, we are able to avoid difficult calculations, replacing them instead with a sequence of easier calculations."

[RC] Introduction.

The name Gibbs sampling comes from the landmark paper by Geman and Geman (1984), which first applied a Gibbs sampler on a Gibbs random field. For good or bad, it then stuck despite this weak link. Indeed, it is in fact a special case of the Metropolis–Hastings algorithm as detailed in Robert and Casella (2004, Section 10.6.1). The work of Geman and Geman (1984), built on that of Metropolis et al. (1953), Hastings (1970) and Peskun (1973), influenced Gelfand and Smith (1990) to write a paper that sparked new interest in Bayesian methods, statistical computing, algorithms, and stochastic processes through the use of computing algorithms such as the Gibbs sampler and the Metropolis–Hastings algorithm. It is interesting to see, in retrospect, that earlier papers such as Tanner and Wong (1987) and Besag and Clifford (1989) had proposed similar solutions (but did not receive the same response from the statistical community).

[CG] Illustrating the Gibbs Sampler.

"Suppose we are given a joint density $f(x, y_1, \ldots, y_p)$, and are interested in obtaining characteristics of the marginal density

$$f(x) = \int \cdots \int f(x, y_1, \ldots, y_p) dy_1 ... dy_p, \qquad (1)$$

such as the mean or variance. Perhaps the most natural and straightforward approach would be to calculate $f(x)$ and use it to obtain the desired characteristic. However, there are many cases where the integrations in (1) are extremely difficult to perform, either analytically or numerically. In such cases the Gibbs sampler provides an alternative method for obtaining $f(x)$."

[CG] Illustrating the Gibbs Sampler.

"Rather than compute or approximate $f(x)$ directly, the Gibbs sampler allows us effectively to generate a sample $X_1, \ldots, X_m \sim f(x)$ *without requiring* $f(x)$. By simulating a large enough sample, the mean, variance, or any other characteristic of $f(x)$ can be calculated to the desired degree of accuracy."

[CG] Illustrating the Gibbs Sampler.

"It is important to realize that, in effect, the end result of any calculations, although based on simulations, are the population quantities. For example, to calculate the mean of $f(x)$, we could use $(1/m)\sum X_i$, and the fact that

$$\lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} X_i = \int_{-\infty}^{\infty} x f(x) dx = \mathbb{E}(X).$$

Thus, by taking m large enough, any population characteristic, even the density itself, can be obtained to any degree of accuracy."

[CG] Two-stage Gibbs Sampler.

"To understand the workings of the Gibbs sampler, we first explore it in the two-variable case. Starting with a pair of random variables (X, Y), the Gibbs sampler generates a sample from $f(x)$ by sampling instead from the conditional distributions $f(x \mid y)$ and $f(y \mid x)$, distributions that are often known in statistical models."

[CG] Two-stage Gibbs Sampler.

"This is done by generating a *Gibbs sequence* of random variables

$$Y_0, X_0, Y_1, X_1, Y_2, X_2, \ldots, Y_k, X_k. \tag{2}$$

The initial value $Y_0 = y$ is specified, and the rest of (2) is obtained iteratively by alternately generating values from

$$X_j \sim f(x \mid Y_j = y_j), \quad Y_{j+1} \sim f(y \mid X_j = x_j)$$

We refer to this generation of (2) as Gibbs sampling. It turns out that under reasonably general conditions, the distribution of X_k converges to $f(x)$ (the true marginal of X) as $k \to \infty$. Thus, for k large enough, the final observation in (2), namely $X_k = x_k$, is effectively a sample point from $f(x)$.

[RC] Two-stage Gibbs Sampler.

7.2 The two-stage Gibbs sampler

The *two-stage Gibbs sampler* creates a Markov chain from a joint distribution in the following way. If two random variables X and Y have joint density $f(x, y)$, with corresponding conditional densities $f_{Y|X}$ and $f_{X|Y}$, the two-stage Gibbs sampler generates a Markov chain (X_t, Y_t) according to the following steps:

Algorithm 7 Two-stage Gibbs sampler

Take $X_0 = x_0$

For $t = 1, 2, \ldots,$ generate

1. $Y_t \sim f_{Y|X}(\cdot|x_{t-1})$;
2. $X_t \sim f_{X|Y}(\cdot|y_t)$.

[CG] Two-stage Gibbs Sampler. Example 1. ([RC] Example 7.2)

For the following joint distribution of X and Y,

$$f(x,y) \propto \binom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}, \tag{3}$$

with $x = 0, 1, \ldots, n$, $0 \leq y \leq 1$, suppose we are interested in calculating some characteristics of the marginal distribution $f(x)$ of X. The Gibbs sampler allows us to generate a sample from this marginal as follows. From (3) it follows (suppressing the overall dependence on n, α, and β) that

$$\begin{aligned} f(x \mid y) &\sim B(n,y), \\ f(y \mid x) &\sim Beta(x+\alpha, n-x+\beta). \end{aligned}$$

[CG] Two-stage Gibbs Sampler. Example 1. ([RC] Example 7.2)

Gibbs sampling is actually not needed in this example, since $f(x)$ can be obtained analytically from (3) as

$$f(x) = \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha+\beta+n)}, \tag{4}$$

with $x = 0, 1, \ldots, n$, the beta-binomial distribution. Here, characteristics of $f(x)$ can be directly obtained from (4), either analytically or by generating a sample from the marginal and not fussing with the conditional distributions. However, this simple situation is useful for illustrative purposes.

[CG] Two-stage Gibbs Sampler. Example 1. ([RC] Example 7.2)

One feature brought out by Example 1 is that the Gibbs sampler is really not needed in any bivariate situation where the joint distribution $f(x, y)$ can be calculated, since

$$f(x) = f(x, y)/f(y \mid x).$$

However, as the next example shows, Gibbs sampling may be indispensable in situations where $f(x, y), f(x)$, or $f(y)$ cannot be calculated.

[CG] Two-stage Gibbs Sampler. Example 2.

"Suppose X and Y have conditional distributions that are exponential distributions restricted to the interval $(0, B)$, that is,

$$\begin{array}{lcl} f(x \mid y) & \propto & ye^{-yx},\ 0 < x < B < \infty \\ f(y \mid x) & \propto & xe^{-xy},\ 0 < y < B < \infty \end{array} \tag{5}$$

where B is a known positive constant. The restriction to the interval $(0, B)$ ensures that the marginal $f(x)$ exists. Although the form of this marginal is not easily calculable, by applying the Gibbs sampler to the conditionals in (5) any characteristic of $f(x)$ can be obtained."

[CG] A simple convergence proof. It is not immediately obvious that a random variable with distribution $f(x)$ can be produced by the Gibbs sequence of (2)

$$Y_0, X_0, Y_1, X_1, Y_2, X_2, \ldots, Y_k, X_k. \tag{6}$$

or that the sequence even converges. That this is so relies on the Markovian nature of the iterations, which we now develop in detail for the simple case of a 2×2 table with multinomial sampling. Suppose X and Y are each (marginally) Bernoulli random variables with joint distribution

		X	
		0	1
Y	0	p_1	p_2
	1	p_3	p_4

$p_i \geq 0, p_1 + p_2 + p_3 + p_4 = 1,$

[CG] A simple convergence proof.

In terms of the joint probability function,

$$\begin{pmatrix} f_{x,y}(0,0) & f_{x,y}(1,0) \\ f_{x,y}(0,1) & f_{x,y}(1,1) \end{pmatrix} = \begin{pmatrix} p_1 & p_2 \\ p_3 & p_4 \end{pmatrix}$$

The marginal distribution of X is

$$f_x = [f_x(0), f_x(1)] = [p_1 + p_3, p_2 + p_4], \ X \sim B(p_2 + p_4).$$

The conditional distributions of $X \mid Y = y$ and $Y \mid X = x$ are straightforward to calculate All of the conditional probabilities can be expressed in two matrices

$$A_{y|x} = \begin{pmatrix} \frac{p_1}{p_1+p_3} & \frac{p_3}{p_1+p_3} \\ \frac{p_2}{p_2+p_4} & \frac{p_4}{p_2+p_4} \end{pmatrix} \text{ and } A_{x|y} = \begin{pmatrix} \frac{p_1}{p_1+p_2} & \frac{p_2}{p_1+p_2} \\ \frac{p_3}{p_3+p_4} & \frac{p_4}{p_3+p_4} \end{pmatrix}$$

where $A_{y|x}$ has the conditional probabilities of Y given $X = x$, and $A_{y|x}$ has the conditional probabilities of X given $Y = y$.

[CG] A simple convergence proof.

We are interested in simulations of the sequence of X's. Note that to go from X_k to X_{k+1} we pass through Y_{k+1}. Thus the transition probability, for any $k \geq 0$,

$$\begin{aligned}&\mathbb{P}(X_{k+1} = x_{k+1} \mid X_k = x_k)\\ &= \sum_y \mathbb{P}(Y_{k+1} = y \mid X_k = x_k)\mathbb{P}(X_{k+1} = x_{k+1} \mid Y_{k+1} = y)\end{aligned}$$

Thus the transition probability matrix for (X_k) is given by

$$A_{x|x} = A_{y|x}A_{x|y} \text{ and } \mathbb{P}(X_k = x_k \mid X_0 = x_0) = (A_{x|x})^k_{x_0,x_k}$$

[CG] A simple convergence proof.

It is straightforward to check that $f_x = [p_1 + p_3, p_2 + p_4]$ is stationary distribution for the matrix $A_{x|x}$:

$$[p_1{+}p_3, p_2{+}p_4] \begin{pmatrix} \frac{p_1}{p_1{+}p_3} & \frac{p_3}{p_1{+}p_3} \\ \frac{p_2}{p_2{+}p_4} & \frac{p_4}{p_2{+}p_4} \end{pmatrix} \begin{pmatrix} \frac{p_1}{p_1{+}p_2} & \frac{p_2}{p_1{+}p_2} \\ \frac{p_3}{p_3{+}p_4} & \frac{p_4}{p_3{+}p_4} \end{pmatrix} = [p_1{+}p_3, p_2{+}p_4]$$

"The algebra for the 2×2 case immediately works for any $n \times m$ joint distribution of X's and Y's. We can analogously define the $n \times n$ transition matrix $A_{x|x}$ whose stationary distribution will be the marginal distribution of X."

[CG] A simple convergence proof.

"If either (or both) of X and Y are continuous, then the finite dimensional arguments will not work. However, with suitable assumptions, all of the theory still goes through, so the Gibbs sampler still produces a sample from the marginal distribution of X. The conditional density of X_{k+1} given X_k could be written

$$f_{X_{k+1}|X_k}(x_{k+1} \mid x_k) = \int f_{X_{k+1}|Y_{k+1}}(x_{k+1} \mid y) f_{Y_{k+1}|X_k}(y \mid x_k) dy."$$

The density $f_{X_{k+1}|X_k}(x_{k+1} \mid x_k)$ represents a one-step transition. Observe, that the following relationship holds true

$$f_{X_{k+1}|X_0}(x_{k+1} \mid x_0) = \int f_{X_{k+1}|X_k}(x_{k+1} \mid t) f_{X_k|X_0}(t \mid x_0) dt, \qquad (7)$$

where $f_{X_{k+1}|X_0}(x_{k+1} \mid x_0)$ plays the role of f_{k+1}, and $f_{X_k|X_0}(x_k \mid x_0)$ plays the role of f_k. As k goes to infinity, it again follows that the stationary point of (7) is the marginal density of X.

[CG] Conditionals determine marginals.

"Gibbs sampling can be thought of as a practical implementation of the fact that knowledge of the conditional distributions is sufficient to determine a joint distribution (if it exists!). In the bivariate case, the derivation of the marginal from the conditionals is fairly straightforward. Complexities in the multivariate case, however, make these connections more obscure. We begin with some illustrations in the bivariate case and then investigate higher dimensional cases."

[CG] Conditionals determine marginals. Bivariate case.

Suppose that, for two random variables X and Y,we know the conditional densities $f_{X|Y}(x \mid y)$ and $f_{Y|X}(y \mid x)$. We can determine the marginal density of X, $f_X(x)$, and hence the joint density of X and Y, through the following argument.

$$\begin{aligned} f_X(x) &= \int f_{XY}(x,y)dy = \int f_{X|Y}(x \mid y) f_Y(y) dy \\ &= \int f_{X|Y}(x \mid y) \int f_{Y|X}(y \mid t) f_X(t) dt dy \\ &= \int \left(\int f_{X|Y}(x \mid y) f_{Y|X}(y \mid t) dy \right) f_X(t) dt =: \int h(x,t) f_X(t) dt, \end{aligned}$$

defines a *fixed point integral equation* for which $f_X(x)$ is a solution. The fact that it is a unique solution is explained by Gelfand and Smith (1990). (PS: Gelfand and Smith (1990): ‘‘Exploiting standard theory of such integral operators, Tanner and Wong (1987) showed that under mild regularity conditions this iterative process has the following properties: uniqueness, monotone convergence in L_1, geometrical rate.)

[CG] Conditionals determine marginals. Bivariate case.

$$f_X(x) = \int h(x,t) f_X(t) dt, \;\; h(x,t) := \int f_{X|Y}(x \mid y) f_{Y|X}(y \mid t) dy.$$

This equation is limiting form of Gibbs iteration scheme. As $k \to \infty$

$$f_{X_k|X_0}(x \mid x_0) \to f_X(x) \text{ and } f_{X_{k+1}|X_k}(x \mid t) \to h(x,t).$$

"Although the joint distribution of X and Y determines all of the conditionals and marginals, it is not always the case that a set of proper conditional distributions will determine a proper marginal distribution (and hence a proper joint distribution). The next example shows this."

[CG] Conditionals determine marginals. Bivariate case.

Consider the previous example of exponential distribution supposing now that $B = \infty$

$$\begin{array}{lcl} f(x \mid y) & \propto & ye^{-yx},\ 0 < x < \infty \\ f(y \mid x) & \propto & xe^{-xy},\ 0 < y < \infty \end{array} \tag{8}$$

Applying fixed point integral equation defined above the marginal distribution of X is the solution to

$$f_X(x) = \int \left[\int ye^{-yx} te^{-ty} dy \right] f_X(t) dt = \int \frac{t}{(x+t)^2} f_X(t) dt$$

Observe that $f_X(t) = 1/t$ provides the solution

$$\frac{1}{x} = \int \frac{t}{(x+t)^2} \frac{1}{t} dt$$

but not density function.

[CG] Conditionals determine marginals. Bivariate case.

"When the Gibbs sampler is applied to the conditional densities, convergence breaks down. It does not give an approximation to $1/x$, in fact, we do not get a sample of random variables from a marginal distribution. ...

The Gibbs sampler fails when $B = \infty$ above because $\int f_X(x)dx = \infty$, and there is no convergence as described in $f_{X_{k+1}|X_k}(x \mid t) \to h(x,t)$. In a sense, we can say that a sufficient condition for the convergence to occur is that $f_X(x)$ is a proper density, that is $\int f_X(x)dx < \infty$. One way to guarantee this is to restrict the conditional densities to lie in a compact interval, as was done in Example 2. General convergence conditions needed for the Gibbs sampler (and other algorithms) are explored in detail by Schervish and Carlin (1990), and rates of convergence are also discussed by Roberts and Polson (1990)."

[CG] Conditionals determine marginals. More than two variables.

As the number of variables in a problem increase, the relationship between conditionals, marginals, and joint distributions becomes more complex. For example, the relationship

$$\text{conditional} \times \text{marginal} = \text{joint}$$

does not hold for all of the conditionals and marginals. This means that there are many ways to set up a fixed-point equation, and it is possible to use different sets of conditional distributions to calculate the marginal of interest. Such methodologies are part of the general techniques of substitution sampling (see Gelfand and Smith 1990, for an explanation).

[CG] Conditionals determine marginals. More than two variables.

"Suppose we would like to calculate the marginal distribution $f_X(x)$ in a problem with random variables X, Y, and Z. A fixed-point integral equation can be derived if we consider the pair (Y, Z) as a single random variable. We have

$$f_X(x) = \int \left(\int \int f_{X|YZ}(x \mid y, z) f_{YZ|X}(y, z \mid t) dy dz \right) f_X(t) dt.$$

Cycling between $f_{X|YZ}(x \mid y, z)$ and $f_{YZ|X}(y, z \mid t)$ would again result in a sequence of random variables converging in distribution to $f_X(x)$. This is the idea behind the Data Augmentation Algorithm of Tanner and Wong (1987). By sampling iteratively from $f_{X|YZ}(x \mid y, z)$ and $f_{YZ|X}(y, z \mid t)$, they show how to obtain successively better approximations to $f_X(x)$."

[CG] Conditionals determine marginals. More than two variables.

"In contrast, the Gibbs sampler would sample iteratively from $f_{X|YZ}, f_{Y|XZ}$, and $f_{Z|XY}$. That is, the j-th iteration would be

$$\begin{array}{rcl} X_j & \sim & f(x \mid Y_j = y_j, Z_j = z_j) \\ Y_{j+1} & \sim & f(y \mid X_j = x_j, Z_j = z_j) \\ Z_{j+1} & \sim & f(z \mid X_j = x_j, Y_{j+1} = y_{j+1}). \end{array} \tag{9}$$

The iteration scheme of (9) produces a Gibbs sequence

$$Y_0, Z_0, X_0, Y_1, Z_1, X_1, Y_2, Z_2, X_2, \ldots,$$

with the property that, for large k, $X_k = x_k$, is effectively a sample point from $f(x)$. Although it is not immediately evident, the iteration in (9) will also solve the fixed-point equation."

[CG] Conditionals determine marginals. More than two variables. Generalization of Example 1.

In the distribution of Example 1 (3), we now let n be the realization of a Poisson random variable with mean λ, yielding the joint distribution

$$f(x,y,n) \propto \binom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1} e^{-\lambda}\frac{\lambda^n}{n!},$$
$$x = 0,1,\ldots,n,\ 0 \le y \le 1,\ n = 1,2,\ldots$$

Again, suppose we are interested in the marginal distribution of X. Unlike Example 1, here we cannot calculate the marginal distribution of X in closed form.

[CG] Conditionals determine marginals. More than two variables. Generalization of Example 1.

$$f(x,y,n) \propto \binom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1} e^{-\lambda}\frac{\lambda^n}{n!},$$
$$x = 0,1,\ldots,n,\ \ 0 \le y \le 1,\ \ n = 1,2,\ldots$$

However, it is reasonably straightforward to calculate the three conditional densities. Suppressing dependence on λ, α, and β,

$$\begin{aligned} f(x \mid y,n) &\sim B(n,y), \\ f(y \mid x,n) &\sim Beta(x+\alpha, n-x+\beta), \\ f(n \mid x,y) &\propto e^{-(1-y)\lambda}\frac{((1-y)\lambda)^{n-x}}{(n-x)!}, \quad n = x, x+1,\ldots. \end{aligned}$$

[CG] Conditionals determine marginals. More than two variables. Generalization of Example 1.

This model can have practical applications. For example, conditional on n and y, let x represent the number of successful hatchings from n insect eggs, where each egg has success probability y. Both n and y fluctuate across insects, which is modeled in their respective distributions, and the resulting marginal distribution of X is a typical number of successful hatchings among all insects.

[CG] Detecting Convergence.

"The Gibbs sampler generates a Markov chain of random variables which converge to the distribution of interest $f(x)$. Many of the popular approaches to extracting information from the Gibbs sequence exploit this property by selecting some large value for k, and then treating any X_j, for $j \geq k$ as a sample from $f(x)$. The problem then becomes that of choosing the appropriate value of k."

[CG] Detecting Convergence.

"A general strategy for choosing such k is to monitor the convergence of some aspect of the Gibbs sequence. ... For example, monitoring density estimates from m independent Gibbs sequences, and choosing k to be the first point at which these densities appear to be the same under a "felt-tip pen test." Tanner (1991) suggests monitoring a sequence of weights that measure the discrepancy between the sampled and the desired distribution. Geweke (in press) suggests monitoring based on time series considerations. Unfortunately, such monitoring approaches are not foolproof, illustrated by Gelman and Rubin(1991). An alternative may be to choose k based on theoretical considerations, as in Raftery and Banfield(1990). M.T.Wells (personal communication) has suggested a connection between selecting k and the cooling parameter in simulated annealing."

References.

[CG] Casella G. and George E.I. *Explaining the Gibbs Sampler*.

[RC] Cristian P. Robert and George Casella. *Introducing Monte Carlo Methods with R*. Series "Use R!" . Springer

Gelfand, A. E., and Smith, A. F. M. (1990), *Sampling-Based Approaches to Calculating Marginal Densities,* Journal of the American StatisticalAssociation, 85, 398-409.

Tanner, M. A., and Wong, W. (1987), *The Calculation of Posterior Distributions by Data Augmentation (with discussion)*, Journal of the American Statistical Association, 82, 528-550.