As Casella points out the conditionals f_1 and f_2 are only functionally compatible, corresponding to an improper joint density $f(x,y) \propto e^{-xy}$. Here too, the Gibbs chain cannot be positive recurrent. However, here the Gibbs chain $(x_1,y_1),(x_2,y_2),\ldots$ is obtained by successive substitution into

$$x_i = \epsilon_i^x / y_{i-1} \quad \text{and} \quad y_i = \epsilon_i^y / x_i \tag{4}$$

where ϵ_i^x and ϵ_i^y are independent exponential variables with mean 1. Thus, the derived Markov chain $z_1, z_2, \ldots$ where $z_i \equiv x_i y_i = \epsilon_i^y$ is simply an *iid* exponential sequence, again positive recurrent.

In both of the above examples, a positive recurrent chain $z_1, z_2, \ldots$ was constructed from the non positive recurrent chain (x_1,y_1), $(x_2,y_2),\ldots$. It is interesting to consider how the distribution of z arises through formal transformation of the improper density $f(x,y)$ corresponding to the Gibbs conditionals. In the first example, where $f(x,y) \propto e^{-(x+y)^2/2}$, the joint distribution of $z = x+y$ and $w = y$ is obtained as $f(z,w) \propto e^{-z^2/2}$. In the second example, where $f(x,y) \propto e^{-xy}$, the joint distribution of $z = xy$ and $w = y$ is obtained as $f(z,w) \propto \frac{1}{w} e^{-z}$. In both of these examples, an improper joint distribution has been transformed into the product of a proper distribution on z and an improper distribution on w. Thus, in both of these examples $f(x,y)$ contains a proper one-dimensional component which can be extracted from the output of a Gibbs sampler.

In light of these examples, I would like to ask Casella about the Gibbs subsequence of overall means $\beta^{(j)}, j \geq 1$ from Example 4 where $a = b = 0$. When (if ever) is this subsequence a positive recurrent component of the Gibbs chain? I have a hunch that it will be positive recurrent when $\pi(\beta|y)$, the posterior of β, is proper, in which case the subsequence will converge to $\pi(\beta|y)$. Can this be checked for the Gibbs output from Example 4?

JUN S. LIU *(Stanford University, USA)*

Professor Casella has provided us with a timely exposition of an important aspect of modern Monte Carlo methods. Stimulated by this reading, I would like to take the liberty of bringing up a few ideas on two interesting issues.

Rao-Blackwellizing an Importance Sampler. Consider an importance sampling scheme for a two-component random vector. Following no-

tations of Professor Casella, we let the target distribution of (X, Y) be $f(x, y)$ and let the trial sampling distribution be $g(x, y)$. Of interest is the estimation of, say, $\tau = E_f\{h(X, Y)\}$, for a given integrable function h. This can be achieved by using either rejection sampling, as demonstrated by Professor Casella, or importance sampling (IS). Suppose that we have drawn samples $(x_1, y_1), \ldots, (x_n, y_n)$ from $g(x, y)$. A standard IS estimate of τ is

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} w(x_i, y_i)h(x_i, y_i), \quad \text{where} \quad w(x, y) = \frac{f(x, y)}{g(x, y)}.$$

A rescaled estimate, as illustrated in Section 4.2 and used in Casella and Robert (1996b), Kong et al. (1994), Liu (1996) etc., is

$$\tilde{\tau} = \frac{1}{W}\sum_{i=1}^{n} w(x_i, y_i)h(x_i, y_i), \quad \text{where} \quad W = \sum_{i=1}^{n} w(x_i, y_i).$$

Besides the advantage mentioned by Professor Casella, using the rescaled estimate $\tilde{\tau}$ allows us the flexibility of knowing f and g only up to a normalizing constant. This advantage is much more pronounced in complicated problems (Kong et al. 1994). Because asymptotically the two estimates are equivalent and also because $\hat{\tau}$ is much more approachable mathematically, we will use $\hat{\tau}$ for theoretical discussions, although practically we advocate using $\tilde{\tau}$ all the time.

There are two ways of Rao-Blackwellizing: conditioning on either X or Y. If conditioned on Y, for example, we have

$$\begin{aligned} E_g\{w(X, Y)h(X, Y) \mid Y = y\} &= \int h(x, y)\frac{f(x, y)}{g(x, y)}g(x \mid y)dx \\ &= w_y(y)E_f\{h(X, Y) \mid Y = y\}, \end{aligned}$$

where $w_y(y) = f_y(y)/g_y(y)$. A more efficient estimate than $\hat{\tau}$ results:

$$\hat{\tau}_{rby} = \frac{1}{n}\sum_{i=1}^{n} w_y(y_i)E_f\{h(X, Y)) \mid Y = y_i\}.$$

When h is a function of one component alone, say $h(x, y) = h(y)$, the estimate $\hat{\tau}_{rby}$ is reduced to

$$\hat{\tau}_{rby} = \frac{1}{n}\sum_{i=1}^{n} w_y(y_i)h(y_i).$$

A quite different intuitive interpretation of this R-B effect is that *marginalization* reduces importance sampling variation. MacEachern, Clyde, and Liu (1996) derived one special case of this fact, and Rubinstein (1981, Section 4.3.7) recorded another.

Under this formulation, the importance sampling can be treated approximately as a Rao-Blackwellized rejection sampling; hence, it is statistically more efficient. This fact has been established by Casella and Robert (1996b) in a sophisticated setting and will be re-derived here more directly and heuristically. Let (I_i, y_i), $i = 1, \ldots, n$, be jointly drawn according to the acceptance-rejection rule; that is, the y_i are iid from a trial distribution $g(y)$, and the conditional distribution $[I_i \mid y_i]$ is Bernoulli$(r(y_i))$ with $r(y) = f(y)/Mg(y)$. Suppose the stopping effect of this rejection sampling can be safely ignored. Then I_i plays the role of x_i in the foregoing argument; and the R-B counterpart of δ_{AR} in (10) of Casella is

$$\delta_{IS} = \frac{1}{n}\sum_{i=1}^{n} w(y_i)h(y_i).$$

Without loss of generality we assume that $\tau = 0$. Then, since $M \geq \max_y\{w(y)\}$,

$$\begin{aligned} n\text{var}(\delta_{AR}) &\approx M\text{var}_f\{h(Y)\} \geq \int w_{\max} h^2(y) f(y)dy \\ &\geq \int \frac{f(y)}{g(y)} h^2(y) f(y)dy = E_g\{w^2(y)h^2(y)\} \\ &= \text{var}_g\{w(y)h(y)\} = n\text{var}(\delta_{IS}). \end{aligned}$$

An effort of comparing the two samplers with the Metropolized independence sampling was made in Liu (1996). Since the advantage of the rejection method is that exact draws from f can be obtained, it is sometimes useful to combine the two samplers when one wants to reduce importance sampling variations (Liu, Chen, and Wong 1996).

In many practical problems, the marginal weight $w_y(y)$ is difficult to compute, whereas the conditional expectation $E_f\{h(X) \mid Y = y\}$ is relatively easy to obtain. In such cases, as shown in Kong et al. (1994), one can use a partial RB-estimate

$$\hat{\tau}_{prb} = \frac{1}{n}\sum_{i=1}^{n} w(x_i, y_i)E_f\{h(X, Y) \mid Y = y_i\},$$

which is easily seen to be unbiased and consistent. Although many numerical results show that significant improvements can be obtained, optimality properties of $\hat{\tau}_{prb}$ are difficult to come by.

Imagine that a partial R-B is applied twice; then each summand of $\hat{\tau}_{prb}$, $E_f\{h(X,Y) \mid Y = y_i\}$, is substituted by $E_f[E_f\{h(X,Y)|Y\}| X = x_i]$. By applying partial R-B repeatedly, each summand has the form of iterative conditional expectations:

$$E_f[\cdots E_f\{E_f\{h(X,Y) \mid Y\} \mid X\} \cdots \mid \cdot],$$

whose limit converges to the true value τ. This form alludes to the Gibbs sampling structure (Liu, Wong and Kong 1994, 1995). When analytical evaluation of these iterative conditional expectations is not feasible, one is naturally reminded of the Gibbs sampler. A suggestion thus derived is that incorporating a Gibbs sampler or any MCMC step into an importance sampling scheme can be useful (MacEachern et al. 1996).

The Gibbs Sampler for Incompatible Conditionals

An impressive result of Hobert and Casella (1996) is concerned with the stochastic instability of Gibbs sampling with incompatible — but functionally compatible — conditionals. I would like to venture on the functionally incompatible case. Consider the following example: suppose that the two conditionals $f_1(y|x)$ and $f_2(x|y)$ are given as follows:

$f_1(y\|x)$:		$y=1$	$y=2$
	$x=1$	0.9	0.1
	$x=2$	0.3	0.7

$f_2(x\|y)$:		$x=1$	$x=2$
	$y=1$	0.4	0.6
	$y=2$	0.2	0.8

It is easy to show that f_1 and f_2 are not functionally compatible using Besag's (1974) criterion. When running a systematic-scan Gibbs sampler, the concept of "limiting distribution" becomes a little complicated. In fact, the sampler has two limiting distributions depending on whether stopping at x or at y, i.e., whether (x,y) or (y,x) is defined as a joint state. The two limiting distributions are

$\pi_1(x,y)$:		$y=1$	$y=2$
	$x=1$	0.26591	0.02955
	$x=2$	0.21136	0.49318

$$\pi_2(x,y): \begin{array}{c|cc} & y=1 & y=2 \\ \hline x=1 & 0.19091 & 0.10455 \\ x=2 & 0.28636 & 0.41818 \end{array}$$

The sampler is, therefore, a combination of two positive recurrent Markov chains; and depending on how to define the joint state, the sampler converges into two different, though very close, distributions. When running a random-scan Gibbs sampler, however, a proper limiting distribution — that is the mixture of the two distributions given above — exists.

Under some regularity conditions that are satisfied in most practical situations, $T_x(x_0, x_1) = \int f_1(y|x_0) f_2(x_1|y) dy$ defines a positive recurrent transition function for the X space, and $T_y(y_0, y_1) = \int f_2(x|y_0) f_1(y_1|x) dx$ defines that for the Y space. Hence two limiting distributions $\pi_1(x)$ and $\pi_2(y)$, for T_x and T_y, respectively, are uniquely determined. In the incompatible case, we observe that

$$\pi_1(x,y) \equiv \pi_1(x) f_1(y \mid x) \neq \pi_2(y) f_2(x \mid y) \equiv \pi_2(x,y).$$

But

$$\int \pi_1(x) f_1(y \mid x) dx = \pi_2(y) \quad \text{and} \quad \int \pi_2(y) f_2(x \mid y) dy = \pi_1(x).$$

Let $\mathcal{P}_1$ be the set of all probability distributions compatible with $f_1(y|x)$, and let $\mathcal{P}_2$ be that for $f_2(x|y)$. Then $\pi_1(x,y) \in \mathcal{P}_1$, $\pi_2(x,y) \in \mathcal{P}_2$, and π_1 and π_2 have identical marginal distributions. On the other hand, if two distributions $p_1(x,y) \in \mathcal{P}_1$ and $p_2(x,y) \in \mathcal{P}_2$ have identical marginal distributions, they have to be the same as π_1 and π_2.

Due to numerical approximation in practice, we may end up having slightly incompatible conditionals. If the numerical error is small, the resulting T_x will be very close to the one, say, T_x^*, resulting from the compatible conditionals. This implies that the eigenvalues and eigenvectors of T_x and T_x^* are close to each other (true in the finite state space case); hence, the resulting limiting distributions are similar. It further suggests that no disasters are to be expected as long as the numerical approximation is reasonably accurate. The argument may be extended to a Gibbs sampler with more than two components. For a k component sampler, a systematic scan with a particular sweeping order will have

k limiting distributions, depending on which component the sampler stops. The total number of such limiting distributions is $k!$. The limiting distribution for a random-scan sampler is then a mixture of these $k!$ distributions.

XIAO-LI MENG *(The University of Chicago, USA)*

Posterior Checking. My discussion will focus on only one issue: checking the propriety of a posterior resulting from the Gibbs-sampler specifications. Professor Casella's article is much broader, touching on many issues that are of current interest to me (e.g., the emphasis on being receptive to both frequentist and Bayesian perspectives; the interplay of algorithms and inferences; the connection between EM-type algorithms and the Gibbs sampler). However, due to stringent time constraints (being a father of a newborn and a 16-month-old, I had to prepare this discussion in between frequent posterior checking; no impropriety was found, though I did learn why it is a good idea to avoid a sensitive posterior), I have to skip this great opportunity for advertising several related papers that I authored or co-authored. Nevertheless, I want to thank the Editor, and of course the author, for providing me with such an opportunity.

Recursive De-conditioning and Conditional Compatibility. The need for checking the compatibility of conditional distributions reminds me of an identity I learned more than a year ago. Let $p(x_1, x_2)$ be a probability density function with respect to a product measure $\mu = \mu_1 \times \mu_2$ and with a support in the form $\Omega_1 \times \Omega_2$; we thus are assuming the *positivity* assumption of Hammersley and Clifford (c.f., Besag, 1974). Then

$$p(x_1) = \left[\int_{\Omega_2} \frac{p(x_2 \mid x_1)}{p(x_1 \mid x_2)} \mu_2(dx_2) \right]^{-1}, \tag{1}$$

which is a trivial consequence of the well-known identity

$$\frac{p(x_2 \mid x_1)}{p(x_1 \mid x_2)} = \frac{p(x_2)}{p(x_1)}. \tag{2}$$

While identity (1) also provides an explicit formula showing how $p(x_1 \mid x_2)$ and $p(x_2 \mid x_1)$ uniquely determine $p(x_1, x_2)$, it seems to be much