RAFTERY, A. E. and LEWIS, S. (1992). How many iterations in the Gibbs sampler? In *Bayesian Statistics 4* (J. O. Berger, J. M. Bernardo, A. P. Dawid and A. F. M. Smith, eds.) 763–773. Oxford Univ. Press.

REVUZ, D. (1975). *Markov Chains*. North-Holland, Amsterdam.

RIPLEY, B. D. (1987). *Stochastic Simulation*. Wiley, New York.

RITTER, C. and TANNER, M. A. (1991). The griddy Gibbs sampler. Technical Report, Div. Biostatistics, Univ. Rochester.

ROBERTS, G. and POLSON, N. (1990). A note on the geometric convergence of the Gibbs sampler. Unpublished manuscript.

SCHERVISH, M. J. and CARLIN, B. P. (1992). On the convergence of successive substitution sampling. *Journal of Computational and Graphical Statistics* **1** 111–127.

SCHMEISER, B. and CHEN, M. H. (1991). On random-direction Monte-Carlo sampling for evaluating integrals. Technical Report SMS 91-1, School of Industrial Engineering, Purdue Univ.

SMITH, R. L. (1984). Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Oper. Res.* **32** 1296–1308.

STEWART, L. T. (1979). Multiparameter univariate Bayesian inference. *J. Amer. Statist. Assoc.* **74** 684–693.

SWENDSEN, R. H. and WANG, J. S. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.* **58** 86–88.

TANNER, M. A. (1991). *Tools for Statistical Inference: Observed Data and Data Augmentation Methods. Lecture Notes in Statist.* **67**. Springer, New York.

TIERNEY, L. (1991). Exploring posterior distributions using Markov chains. In *Computer Science and Statistics: 23rd Symposium on the Interface* (E. M. Keramidas, ed.) 563–570.

TIERNEY, L., KASS, R. E. and KADANE, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Amer. Statist. Assoc.* **84** 710–716.

TÓTH, B. (1986). Persistent random walks in random environment. *Probab. Theory Related Fields* **71** 615–625.

WHITT, W. (1991). The efficiency of one long run versus independent replications in steady-state simulation. *Management Sci.* **37** 645–666.

ZEGER, S. L. and KARIM, M. R. (1991). Generalized linear models with random effects: a Gibb's sampling approach. *J. Amer. Statist. Assoc.* **86** 79–86.

ZELLNER, A. and ROSSI, P. E. (1984). Bayesian analysis of dichotomous quantal response models. *J. Econometrics* **25** 365–393.

SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
270 VINCENT HALL
206 CHURCH STREET
MINNEAPOLIS, MINNESOTA 55455

DISCUSSION

HANI DOSS[1]

Ohio State University

0. Introduction. In my comments I discuss two topics, the basic convergence theorem (Theorem 1) and the importance-weighted Gibbs sampler, in particular, the question of assessing the variability of estimates formed by this method.

[1] Research supported by Air Force Office of Scientific Research Grant 94-1-0028.

1. The basic convergence theorem. Theorem 1 of the paper may be stated as follows. If the chain has an invariant probability distribution π, is aperiodic, and has the property that

(1.1) for every set B with $\pi(B) > 0$, the probability that the chain eventually enters B is positive, no matter what is the starting point of the chain,

then

$$\|P^n(x, \cdot) - \pi(\cdot)\| \to 0 \quad \text{for } [\pi]\text{-almost all } x. \tag{1.2}$$

This theorem contrasts with the most standard results on the asymptotic behavior of the n-step transition probabilities for general state space Markov chains in that the standard results are stated under the assumption that there exists a recurrent set A (recall that a set A is recurrent if from any starting point x, the probability that the chain eventually enters A is 1). See, for example, the main result in the classical paper by Athreya and Ney (1978) or the main result in the treatment of asymptotics in the textbook by Durrett (1991) (cf. Theorem 6.8). Now verifying that a set is recurrent is extremely difficult in the problems that arise in Bayesian analysis (at least in my experience). On the other hand, in Theorem 1 in Tierney's paper, the basic assumption (1.1) requires only the verification that certain probabilities are merely *positive*, which is in practice much easier.

Actually, (1.1) seems like a rather benign condition, but in fact is not quite so, because to check it requires that we check a certain condition for all sets to which π gives positive mass. Of course, if we are doing a standard Bayesian analysis in which π is obtained via the formula "posterior is proportional to the likelihood times the prior," then π will be mutually absolutely continuous with respect to Lebesgue measure (on some subset of $\mathbb{R}^n$) and therefore identifying the sets in (1.1) is trivial. The problem is that in some complicated models, we do not have enough of a handle on the unknown π to even identify those sets to which it gives positive mass.

Here is an example which involves a Bayesian nonparametric analysis of censored data. [This example is developed in Doss (1994); here we give only enough detail to make the point raised in the paragraph above.] Suppose that there are random variables $X_1, \dots, X_n \sim_{\text{iid}} F$, but that we do not necessarily see the X_i's. For each i we know only a set A_i within which X_i is known to lie. (The case where for each i, A_i is either a singleton or an interval of the form $[c_i, \infty)$ corresponds to the model of right censorship.) Thus, our data, which will be denoted by the subscript data, consists of the sequence $A_1, \dots, A_n$. Suppose we put a Dirichlet prior on the unknown F and we want to obtain the posterior distribution of F given the data. For the purpose of showing how the Gibbs sampler can be used to do this, we use the following special construction of $\mathcal{D}_\alpha$, the Dirichlet prior with parameter measure α. Let $\alpha = \alpha_0 \cdot \alpha(\mathbb{R})$, so that α_0 is a probability measure. Generate $B_1, B_2, \dots \sim_{\text{iid}} \text{Beta}(1, \alpha(R))$, generate $V_1, V_2, \dots \sim_{\text{iid}} \alpha_0$, with the sequences $\{B_j\}$ and $\{V_j\}$ mutually independent, let $P_j = B_j \prod_{r=1}^{j-1}(1 - B_r)$ and form the random distribution function

$$F = \sum_{j=1}^{\infty} P_j \delta_{V_j}, \tag{1.3}$$

where δ_a denotes the probability measure giving unit mass to the point a. This random F has the Dirichlet distribution with parameter α.

Let $\mathcal{L}_{\text{data}}$ and $\mathcal{L}$ denote conditional distribution given the data and unconditional distribution, respectively. We wish to obtain $\mathcal{L}_{\text{data}}(F)$. To implement the Gibbs sampler, we need $\mathcal{L}_{\text{data}}(\mathbf{X} \mid F)$ and $\mathcal{L}_{\text{data}}(F \mid \mathbf{X})$, the latter being simply $\mathcal{L}(F \mid \mathbf{X})$, which is $\mathcal{D}_{\alpha + \Sigma_{i=1}^{n} \delta_{X_i}}$. From construction (1.3), we see that it is easy to generate a vector from $\mathcal{L}_{\text{data}}(\mathbf{X} \mid F)$. To generate X_1 we generate U_1, a $U(0,1)$ random variable, and determine the index J_1 such that $\Sigma_{j=1}^{J_1-1} P_j < U_1 \leq \Sigma_{j=1}^{J_1} P_j$. If $V_{J_1} \in A_1$, set $X_1 = V_{J_1}$; otherwise, repeat using an independent uniform and continue until the corresponding "V-value" is in the set A_1. This whole process is repeated independently for $i = 1, \dots, n$, and this generates the random variables $X_1, \dots, X_n$. (Note that we never need to generate all of F, but only the part of F that we need.)

Note that the sequences $B_1, B_2, \dots$ and $V_1, V_2, \dots$ are held fixed throughout this process. For this reason, the random variables $X_1, \dots, X_n$ are (unconditionally) dependent. Look now at the Markov chain $(\mathbf{X}^{(l)}, F^{(l)})$ formed by running the Gibbs sampler. For $\pi = \mathcal{L}_{\text{data}}(\mathbf{X}, F)$, it is clear that it is difficult to identify the sets to which π gives positive probability, the problem being both the complicated dependence structure of the X's and the fact that the Markov chain is running in a high-dimensional space. In Theorem 1 of Athreya, Doss and Sethuraman (1992), which includes Theorem 1 of Tierney's paper as a special case, the conditions that one needs to check are phrased only in terms of the transition function $P(\cdot, \cdot)$, and so this theorem is useful for establishing convergence when one does not want to deal directly with the unknown π.

2. Estimating variability when dealing with importance weighted output of Markov chains. A very important point discussed in Hastings (1970) is that by properly reweighing the output of a Markov chain corresponding to a distribution $\pi^{(0)}$, one can study features of another distribution $\pi^{(1)}$. The basic requirement is knowledge of $(d\pi^{(1)}/d\pi^{(0)})(\theta)$ up to a multiplicative constant. [If $\pi^{(i)}$ are posterior distributions corresponding to priors $\rho^{(i)}$, and $\pi^{(i)}$ are proportional to the likelihood times the prior, then $(d\pi^{(1)}/d\pi^{(0)})(\theta)$ is a constant times $(d\rho^{(1)}/d\rho^{(0)})(\theta)$.] An important problem is to assess the accuracy of estimates obtained by reweighing the Markov chain.

To gain a clear understanding of what is involved, let us review the reweighing method. Suppose that $(d\pi^{(1)}/d\pi^{(0)})(\theta) = cl(\theta)$, where l is known. Let f be a function of θ, suppose we wish to estimate $\int f(\theta)\, d\pi^{(1)}(\theta)$ and suppose that we have available the output $\theta_1, \dots, \theta_n$ of a Markov chain which corresponds to the distribution $\pi^{(0)}$. To estimate

$$\int f(\theta)\, d\pi^{(1)}(\theta) = \frac{\int f(\theta) l(\theta)\, d\pi^{(0)}(\theta)}{\int l(\theta)\, d\pi^{(0)}(\theta)},$$

denote $f_i = f(\theta_i), v_i = l(\theta_i)$ and the weights $w_i = v_i/\sum_{j=1}^n v_j$ and form

$$\frac{\sum_{i=1}^n f_i v_i}{\sum_{j=1}^n v_j} = \sum_{i=1}^n f_i w_i. \tag{2.1}$$

To study the asymptotic distribution of $\sum_{i=1}^n f_i v_i/\sum_{j=1}^n v_j$, we need to establish conditions under which $(\sum_{i=1}^n f_i v_i, \sum_{i=1}^n v_i)$ is asymptotically normal.

According to Theorem 1.5 of Ibragimov (1962), if a sequence of random variables $\{\xi_i\}$ is ϕ-mixing [a definition is given on page 349 of Ibragimov (1962)] and

$$\text{the sequence } \{\phi(k)\} \text{ of mixing coefficients satisfies } \sum_{k=1}^\infty \phi(k)^{1/2} < \infty, \tag{2.2}$$

$$E(\xi_1^2) < \infty, \tag{2.3}$$

then

$$\text{the series } \operatorname{var}(\xi_0) + 2\sum_{j=1}^\infty \operatorname{cov}(\xi_0, \xi_j) \text{ converges absolutely,} \tag{2.4}$$

$$\frac{(\sum_{i=1}^n \xi_i - E(\xi_0))}{n^{1/2}} \to_d \mathcal{N}(0, \sigma^2) \quad \text{where } \sigma^2 = \operatorname{var}(\xi_0) + 2\sum_{j=1}^\infty \operatorname{cov}(\xi_0, \xi_j). \tag{2.5}$$

Now, if the chain $\{\theta_i\}_{i=0}^\infty$ is stationary and uniformly ergodic, then for any function $h(\theta)$, if $\xi_i = h(\theta_i)$, then the sequence $\{\xi_i\}$ is ϕ-mixing, and (2.2) is satisfied. Thus, if we apply the Cramér–Wold device, we see that

$$\text{if the chain } \{\theta_i\} \text{ is stationary and uniformly ergodic,} \tag{2.6}$$

$$\int (f_1 v_1)^2 \, d\pi^{(0)} < \infty, \tag{2.7}$$

$$\int (v_1)^2 \, d\pi^{(0)} < \infty, \tag{2.8}$$

then

$$\begin{aligned} &\text{each of the three series, } \gamma_{11} = \operatorname{var}(f_0 v_0) + 2\sum_{j=1}^\infty \operatorname{cov}(f_0 v_0, f_j v_j), \\ &\gamma_{12} = \gamma_{21} = \operatorname{cov}(f_0 v_0, v_0) + \sum_{j=1}^\infty \left[\operatorname{cov}(f_0 v_0, v_j) + \operatorname{cov}(v_0, f_j v_j)\right] \text{ and} \\ &\gamma_{22} = \operatorname{var}(v_0) + 2\sum_{j=1}^\infty \operatorname{cov}(v_0, v_j), \text{ converges absolutely,} \end{aligned} \tag{2.9}$$

$$n^{-1/2} \begin{pmatrix} \sum_{i=1}^n f_i v_i - n \int f_1 v_1 \, d\pi^{(0)} \\ \sum_{i=1}^n v_i - n \int v_1 \, d\pi^{(0)} \end{pmatrix} \to_d \mathcal{N}(\mathbf{0}, \Gamma), \tag{2.10}$$

where

$$\Gamma = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix}.$$

The matrix Γ can be estimated using standard techniques for estimation of the cross-spectrum in multiple time series. Statement (2.10) implies, of course, the asymptotic normality of (2.1). Conditions (2.7) and (2.8) make precise the statement that unbounded weight functions should be used with caution. These conditions place limits on how much the prior $\rho^{(1)}$ can deviate from $\rho^{(0)}$. (Note that the assumption that $\{\theta_i\}_{i=0}^{\infty}$ is stationary is not necessary by Theorem 5 of Tierney's paper.)

To give an illustration, consider a variant of the example in Section 5. Tierney considers the model in which $\beta \sim G(\gamma_0, \delta_0)$ and, conditional on $\beta, \lambda_1, \dots, \lambda_n$ are iid $\sim G(\alpha_0, \beta)$. Here, $\gamma_0 = .01, \delta_0 = 1, \alpha_0 = 1.802$ and $n = 10$. If we consider the same model, but with hyperparameters γ_1, δ_1 and α_1, then the calculation in the Appendix below [adapted from Zhang (1993)] shows that (2.8) is equivalent to

$$2\delta_1 - \delta_0 > 0 \quad \text{and} \quad 2n\alpha_1 - n\alpha_0 + 2\gamma_1 - \gamma_0 > 0. \tag{2.11}$$

[When $f(\boldsymbol{\lambda}, \beta) = \lambda_1$ or when $f(\boldsymbol{\lambda}, \beta) = E_{\pi^{(1)}}(\lambda_1 \mid \beta)$, conditions (2.7) and (2.8) turn out to be equivalent.] Thus, even if we stay within the same parametric family, conditions (2.7) and (2.8) present nontrivial constraints.

APPENDIX

Proof that (2.8) is equivalent to (2.11) for the gamma model. Let $g_{a,b}(\cdot)$ denote the gamma density with parameters a and b. We have

$$I \equiv \int \left(\left[\frac{d\pi^{(1)}}{d\pi^{(0)}} \right] (\boldsymbol{\lambda}, \beta) \right)^2 d\pi^{(0)}(\boldsymbol{\lambda}, \beta) \propto \int \left(\left[\frac{d\rho^{(1)}}{d\rho^{(0)}} \right] (\boldsymbol{\lambda}, \beta) \right)^2 l_{s,t}(\boldsymbol{\lambda}, \beta)\, d\rho^{(0)}(\boldsymbol{\lambda}, \beta),$$

where $\rho^{(j)}$, the prior on $(\boldsymbol{\lambda}, \beta)$ corresponding to the hyperparameter values γ_j, δ_j and α_j, is given by

$$d\rho^{(j)}(\boldsymbol{\lambda}, \beta) = \left(\prod_{i=1}^{n} g_{\alpha_j, \beta}(\lambda_i) \right) g_{\gamma_j, \delta_j}(\beta)\, d\boldsymbol{\lambda}\, d\beta,$$

and the likelihood $l_{s,t}(\boldsymbol{\lambda}, \beta)$ is given by

$$l_{s,t}(\boldsymbol{\lambda}, \beta) = \prod_{i=1}^{n} \left(\frac{(\lambda_i t_i)^{s_i}}{s_i!} \exp(-\lambda_i t_i) \right).$$

We have

$$
\begin{aligned}
I &\propto \int \frac{\Big(\big(\prod_{i=1}^{n}\beta^{\alpha_1}\lambda_i^{\alpha_1-1}\exp(-\beta\lambda_i)\big)\cdot\beta^{\gamma_1-1}\exp(-\beta\delta_1)\Big)^2}{\big(\prod_{i=1}^{n}\beta^{\alpha_0}\lambda_i^{\alpha_0-1}\exp(-\beta\lambda_i)\big)\cdot\beta^{\gamma_0-1}\exp(-\beta\delta_0)} \\
&\qquad\times\left(\prod_{i=1}^{n}\lambda_i^{s_i}\exp(-\lambda_i t_i)\right)d\boldsymbol{\lambda}\,d\beta \\
&= \int \beta^{2n\alpha_1-n\alpha_0+2\gamma_1-\gamma_0-1}\exp\big(-\beta(2\delta_1-\delta_0)\big) \\
&\qquad\times\left\{\int\prod_{i=1}^{n}\Big(\lambda_i^{2\alpha_1-\alpha_0+s_i-1}\exp\big(-(\beta+t_i)\lambda_i\big)\Big)\,d\boldsymbol{\lambda}\right\}d\beta \\
&= \int \beta^{2n\alpha_1-n\alpha_0+2\gamma_1-\gamma_0-1}\exp\big(-\beta(2\delta_1-\delta_0)\big) \\
&\qquad\times\left\{\prod_{i=1}^{n}\frac{\Gamma(2\alpha_1-\alpha_0+s_i)}{(\beta+t_i)^{2\alpha_1-\alpha_0+s_i}}\right\}d\beta \\
&\propto \int_0^\infty \frac{\beta^{2n\alpha_1-n\alpha_0+2\gamma_1-\gamma_0-1}}{\prod_{i=1}^{n}(\beta+t_i)^{2\alpha_1-\alpha_0+s_\iota}}\exp\big(-\beta(2\delta_1-\delta_0)\big)\,d\beta.
\end{aligned}
\tag{A.1}
$$

It is easy to see that conditions (2.11) are necessary and sufficient for convergence of the last integral above.

In checking condition (2.7) for the function $f(\boldsymbol{\lambda},\beta)=\lambda_1$ the calculation is the same, except that in the third line of (A.1), the term

$$\frac{\Gamma(2\alpha_1-\alpha_0+s_1)}{(\beta+t_1)^{2\alpha_1-\alpha_0+s_1}}$$

is replaced by

$$\frac{\Gamma(2\alpha_1-\alpha_0+s_1+2)}{(\beta+t_1)^{2\alpha_1-\alpha_0+s_1+2}};$$

it is clear that this change does not affect the convergence of the integral. For the function $f(\boldsymbol{\lambda},\beta)=E_{\pi^{(1)}}(\lambda_1\mid\beta)$, we wind up with the last line in (A.1), except that we have the extra factor $((\alpha_1+s_1)/(\beta+t_1))^2$; again, this does not affect the convergence of the integral.

REFERENCES

ATHREYA, K. B. and NEY, P. (1978). A new approach to the limit theory of recurrent Markov chains. *Trans. Amer. Math. Soc.* **245** 493–501.

DOSS, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *Ann. Statist.* **22** 1763–1786.

DURRETT, R. (1991). *Probability: Theory and Examples.* Wadsworth and Brooks/Cole, Pacific Grove, CA.

IBRAGIMOV, I. A. (1962). Some limit theorems for stationary processes. *Theory Probab. Appl.* **7** 349–382.

ZHANG, R. H. (1993). Unpublished notes.

DEPARTMENT OF STATISTICS
OHIO STATE UNIVERSITY
COLUMBUS, OHIO 43210-1247

JULIAN BESAG[1]

University of Washington

It is a pleasure to add my congratulations to Luke Tierney on his important paper, which not only provides a sound theoretical basis for the use of Markov chain Monte Carlo (MCMC) methods in Bayesian inference but also gives valuable practical guidance. It is noteworthy that versions of the paper have been available for a couple of years now and have already proved to be highly influential. Subsequent developments, often involving the author himself, have been extremely rapid and I hope he will take the opportunity to tell us something about these in his rejoinder. For example, regeneration methods, which are only briefly discussed in the paper, have been the subject of considerable progress [e.g., Mykland, Tierney and Yu (1995)]. In the very recent work of Geyer and Thompson (1993), they are used cleverly on a succession of chains, ranging from "hot" (e.g., independence) to "cold" (the distribution of interest). The idea is that swaps into the hot chain, which can be sampled exactly and hence forgetfully, provide the regeneration points. These authors also show how to adapt their strategy to a single chain by subsampling from a randomly varying distribution between regenerations, so that no form of burn-in is required.

Markov random fields and Gibbs. I particularly welcome Tierney's survey of a wide variety of different MCMC algorithms, including hybrid implementations to which I shall return later. It is easy to be seduced into using the Gibbs sampler as one's only Bayesian inference machine, as I know only too well in spatial applications [Besag (1989), Besag and Mollié (1989), Besag and York (1989) and Besag, York and Mollié (1991)]. In fact, Gibbs has extra allure in spatial statistics. The reason is that a standard means of obtaining a distribution π for a random vector $X = (X_1, \dots, X_n)$, where each X_i is associated with a fixed spatial location (or *site*) i, is in terms of a Markov random field formulation [Besag (1974)]. This requires that one examines each site in turn and specifies the "full" conditional distribution $\pi(x_i \mid x_{-i})$ there; these conditionals are called *local characteristics* in spatial statistics. Such a conditional probability approach to spatial interaction was advocated by Bartlett (1967), as part of his presidential address to the Royal Statistical Society. There are two immediate questions. Do the local characteristics determine π and what

[1] Research supported in part by NSF Grant DMS-92-14497.