Possible biases induced by MCMC convergence diagnostics

by

Mary Kathryn Cowles*, Gareth O. Roberts**, and Jeffrey S. Rosenthal***

(December, 1997.)

1. Introduction.

Markov chain Monte Carlo (MCMC) methods (Metropolis et al., 1953; Hastings, 1970; Geman and Geman, 1984; Tanner and Wong, 1987; Gelfand and Smith, 1990; Smith and Roberts, 1993; Neal, 1993) are very commonly used to estimate the expected value $\mu = \mathbf{E}(g)$ of a functional g with respect to a complicated probability measure $\pi(\cdot)$. The procedure involves setting up a Markov chain $P(\cdot, \cdot)$ having stationary distribution $\pi(\cdot)$, running this chain on a computer to obtain output $X_0, X_1, \ldots, X_n$, and then using this sample to estimate μ. The estimator used is typically of the form $\widehat{\mu} = \frac{1}{n-r} \sum_{i=r+1}^{n} g(X_i)$, corresponding to throwing away the first r runs and then averaging the functional over the remaining $n - r$ runs.

However, it is not at all clear how to choose the burn-in time "r". Theoretical analysis has sometimes succeeded in providing useful values of r (Meyn and Tweedie, 1994; Rosenthal, 1995b, 1996; Cowles and Rosenthal, 1996), but only in certain restricted cases. More often, convergence diagnostics (Roberts, 1992, 1994; Raftery and Lewis, 1992; Geweke, 1992; Gelman and Rubin, 1992; Geyer, 1992; Cowles and Carlin, 1996; Brooks and Roberts, 1995) are used. That is, the value of r is itself random, and is obtained by a statistical analysis of the initial sample values of the chain. Very roughly, convergence diagnostics instruct the user to wait until the values of one or several runs of the Markov chain have "settled down" in some sense.

* Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242-1419, U.S.A. Internet: kcowles@stat.uiowa.edu.

** Statistical Laboratory, University of Cambridge, Cambridge CB2 1SB, U.K. Internet: G.O.Roberts@statslab.cam.ac.uk.

*** Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3. Internet: jeff@utstat.toronto.edu. Partially supported by NSERC of Canada.

It is well known (see e.g. Cowles and Carlin, 1996 for a review) that for certain target distributions (e.g. the "witch's hat" example, cf. Matthews, 1993), convergence diagnostics will prematurely claim convergence, i.e. will provide a value of r which is too small and therefore leads to severely biased estimators $\widehat{\mu}$. However, another possible problem with convergence diagnostics is less well studied. Specifically, the dependence of r on the actual sample values can lead to further biasing of $\widehat{\mu}$. Most dramatically, even if the chain was in fact *started* in the stationary distribution $\pi(\cdot)$, so that for any *non-random* choice of r the estimator $\widehat{\mu}$ would be perfectly unbiased, it is *still* possible that $\widehat{\mu}$ may have bias for natural choices of r which depend on the actual sample values. Such "diagnostic biases" are the subject of the present paper.

This paper is organised as follows. In Section 2, we present an over-simplified version of a convergence diagnostic, and study analytically its performance on certain simple Markov chains. We restrict ourselves primarily to chains which in fact produce i.i.d. samples from $\pi(\cdot)$, i.e. which mix extremely rapidly. For such chains, any non-random choice of r would result in a perfectly unbiased estimator; however, the convergence diagnostic introduces biases of its own, which we are able to compute explicitly.

In Section 3, we present numerical simulations of variants of the Geweke (1992) diagnostic to two simple Bayesian inference problems. We demonstrate that in certain cases, significant biases can be caused by the naive use of these diagnostics. Diagnostic bias is thus an important practical issue.

In Section 4, we consider the alternative option of choosing r to be non-random, i.e. completely independent of the sample values. This avoids the biases of dependent r; however it raises the question of what r to choose. We concentrate on a simple two-state Markov chain in which we can explicitly compute the standard error of the estimator $\widehat{\mu}$. We show that the optimal choice of r depends very strongly on the relative sizes of the transition probabilities: it is sometimes very small, but is sometimes a substantial fraction of n.

Finally, in Section 5, we discuss the implications of our results for applied users of convergence diagnostics. In particular, we argue that the ideal approach *combines* diagnostics with non-random choices of r, by first analysing some initial runs to determine a

reasonable choice of r, and then creating an estimator based on a fresh new run of the Markov chain, which has r fixed in advance.

2. Analytic study of biases from simple diagnostics.

In this section, we consider an over-simplified convergence diagnostic, applied to i.i.d. chains. For this simple case, we are able to analytically compute the biases involved.

Specifically, suppose that $X_0, X_1, X_2, \ldots$ are the output of a Markov chain, and that in fact the X_n are independent and each distributed according to $\pi(\cdot)$ (so that the true burn-in time for the chain is 0).

Suppose the convergence is diagnosed by way of batch means. Specifically, suppose we choose a fixed positive integer m and a fixed $\epsilon > 0$. We then compute the quantities $A_{m,i} = \frac{1}{m} \sum_{t=mi+1}^{m(i+1)} X_t$, for $i = 0, 1, 2, \ldots$, and let i^* be the smallest value of i such that $|A_{m,i^*} - A_{m,i^*-1}| < \epsilon$. We then take our estimator to be $\widehat{\mu} = \frac{1}{m} \sum_{t=mi^*+1}^{m(i^*+1)} g(X_t)$. In words, we assess "convergence" as the first time that two consecutive batches (each of length m) give the same mean to within a tolerance of ϵ; we then use the most recent m sample values to estimate μ.

We show that even in this simple case, biases are introduced through the use of this sampler. Indeed, suppose the law of $\frac{1}{m} \sum_{t=1}^{m} X_t$ has density f with respect to Lebesgue measure. Assume that f is piecewise continuous. Then, as $\epsilon \searrow 0$, it is easily seen that the law of $\widehat{\mu}$ will converge weakly to the distribution having density proportional to f^2. That is, for small $\epsilon > 0$, the probability of having two independent batch means *both* very close to a particular value is approximately proportional to the *square* of having just one batch mean close to that value. In other words, the very act of demanding that two (independent) consecutive batch means be very close to each other, in fact replaces their probability density by (approximately) its square.

Is this change significant? Well, as $m \to \infty$ the distribution of the batch mean will (by the weak law of large numbers) be close to a point mass at the true mean μ, so squaring the density will have little impact, at least on the mean of the distribution sampled. However, for finite values of m this squaring will have a noticeable effect. Moreover, if the bias

is calculated in other ways (for instance total variation distance between the true target density and the density actually simulated) it is possible as we shall see that the bias does not decrease at all with sample size

We consider some examples here.

Example 1. Suppose that each X_i has the standard normal distribution $N(0,1)$. Then the law of $\frac{1}{m}\sum_{t=1}^{m} X_t$ is the normal distribution $N(0,1/m)$, so f is the density function for $N(0,1/m)$. But then it is easily seen that f^2 is proportional to the density function for $N(0,1/2m)$. That is, in this case, demanding that two consecutive batch means be very close to each other in fact means the sampled batch mean will have distribution $N(0,1/2m)$ instead of $N(0,1/m)$. Of course, by symmetry the mean of each of these distributions is 0, so there is no bias in the mean. However, the variance is off by $1/m$. Moreover the error in total variation distance is constant as a function of m; it does *not* go to 0 as $m \to \infty$.

Example 2. Suppose that each X_i has the gamma distribution Gamma(a,b), with density proportional to $x^{a-1}e^{-bx}$. Then the law of $\frac{1}{m}\sum_{t=1}^{m} X_t$ is the gamma distribution Gamma(ma,mb). Now, the square of the density of this distribution is proportional to the density of Gamma$(2ma-1,2mb)$. Hence, in this case the diagnostic replaces the Gamma(ma,mb) distribution by the Gamma$(2ma-1,2mb)$ distribution. Among other things, this leads to a bias in the mean: It replaces the mean $\frac{ma}{mb}$ by the mean $\frac{2ma-1}{2mb}$, giving a bias of $-\frac{1}{2mb}$.

Example 3. Suppose that each X_i has the standard Cauchy distribution, with density $\frac{1}{\pi(1+x^2)}$. Then $\frac{1}{m}\sum_{t=1}^{m} X_t$ again has the standard Cauchy distribution. Hence, in this case the diagnostic replaces the standard Cauchy distribution by a distribution having density $\frac{2}{\pi(1+x^2)^2}$, regardless of the choice of m. (Note that here the mean μ is not defined, so the weak law of large numbers does not apply.) Furthermore, we compute the total variation distance (between the target and the sampled densities) to be $\frac{1}{2}\int_{-\infty}^{\infty}\left|\frac{1}{\pi(1+x^2)}-\frac{2}{\pi(1+x^2)^2}\right|dx = 1/\pi \doteq 0.318$, again regardless of the choice of m.

We thus see that, in all three of these examples, demanding that two consecutive batch means be very similar significantly affects the distribution that is actually sampled. In cases where μ is finite (i.e. the first two examples above), the errors for the mean and

variance decrease to 0 as $m \to \infty$, i.e. as the batch size gets large.

We close this section with a number of remarks.

Remarks.

1. We have stated our diagnostic in terms of requiring two consecutive batch means (of the same Markov chain run) to be similar. However, our results are clearly identical if we instead imagine running two independent copies of the chain, and taking the smallest i such that the $i^{\mathbf{th}}$ batch mean of the first chain is similar to the $i^{\mathbf{th}}$ batch mean of the second chain. This version is closer in spirit to the multiple independent chain approach to convergence diagnostics (e.g. Gelman and Rubin, 1992).
2. More generally, we might require that k (instead of just 2) consecutive batch means are similar. In this case, the density f would be approximately replaced by its $k^{\mathbf{th}}$ power, $f^{\boldsymbol{k}}$. This makes the resulting errors even *worse*. For example, for the Gamma(a, b) example above, this "k equal" diagnostic would replace the Gamma(ma, mb) distribution by the Gamma$(kma-(k-1), kmb)$ distribution. Hence, the bias in the mean would be $\frac{\boldsymbol{kma-(k-1)}}{\boldsymbol{kmb}} - \frac{\boldsymbol{ma}}{\boldsymbol{mb}} = -\frac{\boldsymbol{k-1}}{\boldsymbol{kmb}}$, which is increasing as a function of k (but is never more than twice the bias of the $k = 2$ case).
3. Of course, if the Markov chain is really providing i.i.d. samples, then we can avoid all diagnostic-related problems by simply using the *following* batch mean for our estimator, rather than using one of the batch means we actually used to diagnose convergence. However, in a more realistic situation in which the Markov chain only slowly changes values, this "solution" would not be a solution at all, and the diagnostic biasing problems would be virtually unchanged.
4. There may be times when we actually *want* to sample from the distribution with density proportional to $f^{\boldsymbol{2}}$ (or to $f^{\boldsymbol{k}}$), but it is easier (for some reason) to construct and run a Markov chain having stationary distribution with density f. In that case, our analysis above suggests that one way to proceed is to use our simple diagnostic rule to convert a sampler for f into one for $f^{\boldsymbol{2}}$ (or $f^{\boldsymbol{k}}$ as in Remark 2 above). However, this is likely not a practical approach in general, since as $\epsilon \searrow 0$ the waiting time until "convergence" goes to infinity.
5. For a Markov chain which does *not* converge immediately to the stationary distri-

bution, there will be some interplay between diagnostic bias as discussed here, and more traditional convergence bias (i.e. because the chain hasn't converged yet). It is even possible in rare cases that these biases will cancel each other out. For example, consider the Markov chain on the two-point space $\{1,2\}$, with initial probabilities $(\frac{1}{2},\frac{1}{2})$, with transition probabilities $p_{12} = 2\delta$ and $p_{21} = \delta$, and hence with stationary distribution $(\frac{1}{3},\frac{2}{3})$. For very small values of δ, and with a small batch size m, our diagnostic would likely claim convergence immediately, and hence the sampled distribution would be $(\frac{1}{2},\frac{1}{2})$. (Note that for the multiple-chain diagnostic as in Remark 1 above, the situation is slightly more complicated.) On the other hand, if $\delta = \alpha/k$ for the k equal rule (as in Remark 2 above), then the probability that our sampled point will be 1 is

$$e^{-\alpha}\ \frac{2 - e^{-\alpha}}{2(1 + e^{-\alpha} - e^{-2\alpha})}\,;$$

hence, for an appropriate choice of α this can be made equal to $\frac{2}{3}$, as desired. Thus, in this example, it is indeed possible for the two types of bias (diagnostic and convergence) to cancel each other out.

6. On the other hand, it is not possible in general for the diagnostic bias to cancel out from *every* choice of initial state. To see this, consider the following simple example together with the "two equal" stopping rule. Let $K(\cdot,\cdot)$ be the transition probabilities for a discrete state space Markov chain. Suppose that K is irreducible and aperiodic with stationary distribution f. Also let $p(x,y)$ denote the probability that starting at x our terminal value (ie the first repeated value) is y. Suppose we assume that X_0 has the stationary distribution f (so as to eliminate the effect of the convergence bias). Now if we want the bias to be non-existent, we will also require that the probability of terminating in the state y should be $f(y)$. Therefore f solves

$$f(x) = \sum_y f(y)p(y,x)$$

also. That is, p becomes a transition matrix with stationary distribution f. However because of the particular construction of the stopping time, for $y \neq x$,

$$p(y,x) = \sum K(y,z)p(z,x) - K(y,y)p(y,x)$$

whereas

$$p(x,x) = \sum K(y,z)p(z,x) + (1 - p(x,x))K(x,x) \ .$$

Summing these two equations over y gives

$$\sum_{y} f(y)K(y,y)P(y,x) = f(x)K(x,x)$$

so that by uniqueness of the invariant measure f for p (p is irreducible and strongly aperiodic by inspection) $f(y)K(y,y)$ must be a constant multiple of f and so $K(y,y)$ is necessarily constant. To summarise, it is impossible for the bias to be completely removed unless $K(y,y)$ is constant, and it is easy to check that under this condition, the "k equal" stopping rule introduces no bias.

7. If the bias is measured in tems of the χ^2 distance, it has a natural interpretation in terms of the coefficient of variation of the random variable $f(X)$. Specifically, recall the χ^2 distance with respect to the density f of two densities g and h:

$$\|g - h\| = \mathbf{E}_{\boldsymbol{f}} \left[\left(\frac{g(X)}{f(X)} - \frac{h(X)}{f(X)} \right)^2 \right] \ .$$

Then a simple calculation yields that

$$\|f - Cf^2\| = (cv_{\boldsymbol{f}}(f(X)))^2 \ ,$$

where C is chosen so that Cf^2 is a density, and where $cv(Y)$ is the coefficient of variation of a non-negative random variable: $cv(Y) = \mathbf{Var}(Y)^{1/2}/\mathbf{E}(Y)$.

3. Simulation study of biases from realistic diagnostics.

To investigate whether bias may be introduced by convergence diagnostics that are used in common practice, we performed a simulation study of two variants of the convergence diagnostic proposed by Geweke (1992). One of these variants is built into the BUGS ("Bayesian inference Using Gibbs Sampling") software package (Spiegelhalter, Thomas, Best and Gilks, 1995), and another is included in CODA (Best, Cowles, and Vines, 1995), a suite of Splus routines distributed with BUGS that performs convergence diagnosis and output analysis on Gibbs sampler output.

The basic idea of Geweke's diagnostic is as follows. Suppose an MCMC sampler has been run for n iterations. For each individual parameter of interest, one should compute the sample mean $\bar{x}_{early}$ of a sequence of iterates from the beginning of the chain (he suggests the first 10%) and the sample mean $\bar{x}_{late}$ of the iterates from a longer sequence at the end of the chain (he suggests the last 50% of the chain). If a large number of iterations separates the "early" and "late" segments of the chain, then $\bar{x}_{early}$ and $\bar{x}_{late}$ should be approximately independent, and

$$Z = \frac{\bar{x}_{early} - \bar{x}_{late}}{\sqrt{\mathbf{Var}(\bar{x}_{early}) + \mathbf{Var}(\bar{x}_{late})}}$$

should have approximately a standard normal distribution. Geweke suggests that if $|Z| > 1.96$, iterates from the "early" segment were not yet drawn from the target distribution and should be discarded.

Care is needed in computing $\mathbf{Var}(\bar{x}_{early})$ and $\mathbf{Var}(\bar{x}_{late})$ because of autocorrelations between iterates in the MCMC output for any individual parameter. Geweke's solution is to regard the "early" and "late" segments as separate time series with respective spectral densities $S_g(\omega)_{early}$ and $S_g(\omega)_{late}$. Then if the number of iterates in the "early" segment is n_{early}, the asymptotic variance of $\bar{x}_{early}$ is $S_g(0)_{early}/n_{early}$ and similarly for the "late" segment. (See Geweke, 1992, for further details.) Geweke's diagnostic with this method of variance estimation is included in CODA, with user-modifiable default values of .1 for the proportion of iterates in the "early" segment and .5 for the proportion "late."

The "diag" command in BUGS invokes a variant of the Geweke diagnostic with a less computationally-intensive method of estimating $\mathbf{Var}(\bar{x}_{early})$ and $\mathbf{Var}(\bar{x}_{late})$ called the "batch means" method. In the early and late segments separately, values are divided into 25 bins and the mean calculated within each bin. $\mathbf{Var}(\bar{x}_{early})$ is estimated as the empirical variance of the 25 bin means from the early segment, divided by 25, and similarly for $\mathbf{Var}(\bar{x}_{late})$. The "diag" command assigns the first 25% of iterates to the "early" segment and the last 50% to the "late" segment. The $|Z|$ statistic is then computed as above.

For our simulation study, we used two simple two-parameter models for which computing the true posterior joint and marginal distributions is straightforward; thus estimates from MCMC runs could be compared to the "truth." The first is the normal mean model

with likelihood $f(x) = N(\mu, \sigma^2)$ and noninformative prior $\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$, which we fit to the data on observed breaking strength of 20 samples of yarn given in Box and Tiao (1973), Table 2.2.1.

We ran 300 parallel Gibbs samplers, all started in stationarity (i.e. initial values were random draws from the true joint posterior distribution) and run for 1000 iterations. We then applied the Geweke diagnostic with both variance-estimation methods (designated "Batch" and "Spectral" in the tables) and two choices of proportion of iterates in the "early" segment (.1 and .25) to the output from each sampler for each parameter in the following sequential manner. [This type of procedure is often used in applied practice, especially when samplers are run in one software package (e.g. BUGS) and convergence is assessed in another (e.g. CODA), so that it is easy to remove burn-in iterations but less convenient to extend the length of the chain.]

If the value of the $|Z|$ statistic was greater than 1.96 for either parameter, the first 100 (or 250 when the second definition of "early" was being used) iterates were discarded for all parameters and the diagnostic was reapplied to the remaining portion of the chain. This process was continued until either the $|Z|$ statistic was ≤ 1.96 for both parameters (in which case the number of iterations retained, the sample mean for each parameter, and the width of the interval between the .025 and .975 quantiles for each parameter, were stored) or the next discarding would have resulted in fewer than 250 iterates being retained (in which case the sampler run was deemed to have failed to converge and no statistics were stored).

Table 1 presents the results of the simulation study. For each parameter, the first section of the table presents the analytically computed posterior mean and width of the 95% HPD region. The next section shows the mean across all 300 chains of the sample mean, the mean squared error, and the width of the interval between the .025 and .975 quantiles for each parameter. The "t" statistics are for univariate tests of the difference between the true value and the corresponding value estimated from the samplers.

Because the samplers were started in stationarity and mixing is rapid for this model (i.e., autocorrelations within the sampler output for each parameter, as well as cross-correlations between the parameters, are negligible), no burn-in iterations should have

been discarded. However, application of the four versions of the diagnostic resulted in discarding initial iterations from 41 to 49 of the 300 chains. The section of Table 1 for each version of the diagnostic presents the means, MSEs, interval widths, and number of iterations retained averaged over only those chains for which some initial iterations were discarded. There is no evidence of systematic upward or downward bias in the estimates of the means for either parameter. However, when diagnostics result in discarding iterates, the estimation of μ is affected in two ways: the mean squared errors are inflated by 33% – 125%, and the widths of the 95% HPD region are slightly (by about 1.2%), but statistically significantly, underestimated. Even for the "Batch" diagnostic with "proportion early = .25," the estimated interval width for μ is smaller than for the full run with no diagnostics; the t-statistic is less extreme only because of the small number of chains on which it is based.

To investigate the effect of these diagnostics when Gibbs samplers mix slowly due to high correlations, we next tried a bivariate normal model with both marginal distributions N(0,1) and known correlation coefficient $\rho = .95$. We ran 200 parallel chains, each started in stationarity and run for 1500 iterations. The four versions of the diagnostic were applied in the same manner as for the first model. The results are presented in Table 2. Here the interval widths for both parameters are significantly though slightly underestimated (by 1%) even in the complete chains; for a slow-mixing model like this, many more than 1500 iterations are needed, or a different MCMC algorithm should be used. Discarding iterations in order to obtain a more homogeneous chain makes the problem worse. When diagnostics result in discarding iterations, interval widths are underestimated by up to 3.5% and mean squared errors are two to four times as large as when the full run is used. The "batch" means method appears to introduce less bias in interval widths than the "spectral" method, which suggests that the former may be providing more accurate estimates of $\mathbf{Var}(\bar{x}_{\boldsymbol{early}})$ and $\mathbf{Var}(\bar{x}_{\boldsymbol{late}})$. For each method of variance estimation, the magnitudes of the mean squared errors and of the bias in estimated interval widths are greater when the proportion early is .25 than .10.

4. Choosing $r = cn$, with c non-random.

One way to avoid the bias problems of the previous sections is to consider simply setting $r = cn$, where c is some fixed constant, independent of the sample run. In this case, if the Markov chain is started in stationarity, then the resulting estimator will clearly be unbiased. Indeed, if starting in stationarity, then it is optimal to take $c = 0$, i.e. to have no burn-in period at all. However, if the chain is *not* started in stationarity, then it may be desirable to choose some $c > 0$. We investigate this question here. For concreteness, we consider the very simple case of a two-state Markov chain. This will allow us to compute the standard error of $\widehat{\mu}$ explicitly.

Consider the two-state Markov chain with $\mathcal{X} = \{0, 1\}$, and with

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} .$$

Furthermore let g be the identity function, i.e. suppose that we are interested in estimating the mean of the stationary distribution $\pi(\cdot)$. To fix ideas, imagine that the chain is in fact started at the point mass $\delta_0(\cdot)$.

We are interested in the estimator $\widehat{\mu} = e/(n-r)$, where $e = \sum_{i=r+1}^{n} X_i$. Specifically, we shall consider the squared-error loss function

$$\mathbf{E}\left((\widehat{\mu} - \mu)^2\right) \;=\; \frac{1}{(n-r)^2}\mathbf{E}\left((e - \pi(e))^2\right) \;=\; \frac{1}{(n-r)^2}\left(\mathbf{E}(e^2) - 2\pi(e)\mathbf{E}(e) + \pi(e)^2\right) .$$

We set $a = \frac{p}{q+p}$, $a' = \frac{q}{q+p} = 1 - a$, and $\lambda = 1 - p - q$. We recall (cf. Hoel, Port, and Stone, 1972, Section 1.2; Rosenthal, 1995c) that

$$P^i = \begin{pmatrix} a' + a\lambda^i & a - a\lambda^i \\ a' - a'\lambda^i & a + a'\lambda^i \end{pmatrix} ,$$

and that furthermore $\pi(0) = a'$ and $\pi(1) = a$. We then have that

$$\pi(e) \;=\; (n-r)a \, ,$$

that

$$\mathbf{E}(e) \;=\; \sum_{i=r+1}^{n} P^i(0,1) \;=\; \sum_{i=r+1}^{n} (a - a\lambda^i) \;=\; a\left[(n-r) - \left(\frac{\lambda^{r+1} - \lambda^{n+1}}{1-\lambda}\right)\right] ,$$

and (since $(X_i)^2 = X_i$) that

$$\mathbf{E}(e^2) \;=\; \sum_{i,j=r+1}^{n} \mathbf{E}(X_i X_j) \;=\; \sum_{i=r+1}^{n} \mathbf{E}(X_i^2) \;+\; 2 \sum_{\substack{i,j=r+1 \\ i<j}}^{n} \mathbf{E}(X_i X_j) \;=\; \mathbf{E}(e) + 2s\,,$$

where

$$\begin{aligned} s \;&=\; \sum_{r+1 \le i < j \le n} \mathbf{E}(X_i X_j) \;=\; \sum_{r+1 \le i < j \le n} P^i(0,1) P^{j-i}(1,1) \\ &=\; \sum_{i=r+1}^{n-1} \sum_{k=1}^{n-i} P^i(0,1) P^k(1,1) \;=\; \sum_{i=r+1}^{n-1} \sum_{k=1}^{n-i} [a - a\lambda^i][a + a'\lambda^k]\,. \end{aligned}$$

We then have that

$$\mathbf{E}\left[(\widehat{\mu} - \mu)^2\right] \;=\; \left([1 - 2\pi(e)]\mathbf{E}(e) + 2s + [\pi(e)]^2\right)/(n-r)^2\,,$$

with $\pi(e)$, $\mathbf{E}(e)$, and s as above.

Now, the explicit analytic computation of these quantities is possible but messy. In particular, if p, q, and p/q are all $O(1)$, then most of the terms are negligible, and we get the approximation

$$\mathbf{E}\left[(e - \pi(e))^2\right] \;\approx\; an(1-c)\left[1 + 2a'\frac{\lambda}{1-\lambda}\right] \;+\; o(n)\,,$$

valid as $n \to \infty$, with $r = cn$ and with c, p, and q fixed. (Note that the $O(n^2)$ terms cancel, as expected.) It then follows that

$$\mathbf{E}\left[(\mu - \widehat{\mu})^2\right] \;\approx\; an^{-1}(1-c)^{-1}\left[1 + 2a'\frac{\lambda}{1-\lambda}\right]\,.$$

Hence, this squared error is minimised when $(1-c)$ is maximised, i.e. when c is minimised. Hence, as expected, we should simply take $c = 0$, i.e. $r = 0$, i.e. no burn-in period, in this case.

However, if p and q are both very small (say, $O(n^{-1})$), which corresponds more closely to a typical diagnostic situation on a more complicated state space, then the question gets more interesting. In this case, terms like λ^n cannot necessarily be neglected.

We have investigated this question for different values of n, p, and q, using the Mathematica computation system (Wolfram, 1988). If p and q are at all large (i.e., if the chain

is mixing quickly), or if $q \geq p$ (i.e., if the stationary distribution is not so far from the starting distribution), then the expected squared error is minimised when c is very small, i.e. it is optimal to let $\widehat{\mu}$ average over nearly all of the available sample values. However, if p and q are very small and $q \ll p$ (i.e., if the chain is mixing slowly and our starting distribution is far from the stationary distribution), then this is not necessarily the case.

For a specific example, suppose that $n = 1000$, with $p = 0.01$ and $q = 0.001$ (so that the stationary distribution puts mass $1/11$ at 0, and mass $10/11$ at 1). Then the squared error, as a function of the burn-in fraction c, is near 0.0257 for c close to 0; dips down to 0.0186 for c near 0.2; then climbs up steadily to 0.0693 when $c = 0.95$. Hence, in this example it is optimal to choose c close to 0.2, i.e. to discard the first 20% of the sample values.

Keeping $n = 1000$ fixed, we can vary the values of p and q to see how the optimal c value changes. If $p = 0.1$ and $q = 0.01$ (or, if $q = 0.001$), then the optimal c value is very close to 0; convergence is so quick (due to the large value of p) that hardly any sample values should be discarded. On the other hand, if $p = 0.01$ and $q = 0.0001$, then the optimal c value is near 0.45, i.e. in this case it is optimal to discard the first 45% of the sample values. For $p = 0.001$ and $q = 0.0001$, the optimal c value is very close to 1; convergence is so slow that later sample values are much more helpful than earlier ones.

To summarise: In the case of slow convergence when starting far from stationary (i.e. $p, q \leq O(1/n)$ and $q \ll p$), the optimal choice of burn-in fraction c is not close to 0; indeed, it increases towards 1 as p, q, and q/p all go to 0.

5. Discussion and conclusion.

In this paper, we have considered biases that may result from the stochastic dependence of convergence diagnostics on the actual sample values used for estimating. We have argued that such dependence can significantly affect the distribution of the sample values used.

The results of Section 2 indicate that, if the diagnostic makes use of batch means of size m, then diagnostic biases may indeed occur (and, in simple cases, they can be computed explicitly). As $m \to \infty$ the biasing errors for the mean and variance go to 0, at least when

the diagnostic random variable has finite mean. (However, the total variation distance errors may not decrease with m.) This suggests that, for our over-simplified diagnostic at least, it is better to choose a sufficiently *large* batch size.

The results of Section 3 show that, when standard diagnostics are applied to the output of MCMC samplers, the diagnostic biases can sometimes be significant. Geweke's diagnostic was selected for illustration only because of its similarity to the batch means approach explored analytically in Section 2; similar results could be expected from other standard convergence diagnostics that seek to detect an initial transient. The example problems showed that the mixing rate of the sampler affects the meaning of convergence diagnostics. If a chain mixes slowly, differences in batch means for different segments of the chain may simply indicate that different regions of the parameter space are being traversed slowly; in this case discarding initial iterations may severely bias estimation because the discarded values may be the only samples from a high-probability region of the parameter space. In this case, since the total length of the chain is finite and fixed, basing the diagnostic on larger batch sizes is likely to worsen the resulting bias because a larger proportion of the iterates used for estimation are required to come from the same limited region of the parameter space.

The results of Section 4 suggest that, if we wish to choose our burn-in time non-randomly to avoid biasing the results, then the optimal choice of burn-in time depends heavily on the nature of the Markov chain and of the starting distribution. In particular, the farther we start from the stationary distribution, and the more slowly the Markov chain is mixing, the larger a burn-in time we require.

There is a way to combine random and non-random burn-in times, which avoids the biases of diagnostics but also avoids the uncertainty of fixed burn-in times. Specifically, it is possible to *first* apply a diagnostic procedure to one or more *initial* runs of the chain to estimate a burn-in time r. This diagnostic procedure should include inspection of the sample paths of quantities of interest and assessment of autocorrelations and cross-correlations, as well as use of a convergence diagnostic. The choice of r from the initial run(s) is then *fixed* and used for a fresh *final* run on which all estimation and inference are based. Starting values for the final run must be generated from the same distribution

as was used to start the initial runs. In fact, the same starting values may be used as long as the random number seed is different for the final run so that the sample paths are different.

We believe that this combined approach achieves the best of both of these alternatives, and it is what we recommend in the absence of any theoretical knowledge about convergence rate of the Markov chain being used.

Acknowledgements. We thank Radford Neal for suggestions related to Section 4 herein.

REFERENCES

N.G. Best, M.K. Cowles, and K. Vines (1995), CODA: Convergence diagnosis and output analysis software for Gibbs sampling output, Version 0.30. Tech. Rep., MRC Biostatistics Unit, University of Cambridge.

C.E.P. Box and G.C. Tiao (1973), Bayesian inference in statistical analysis (Chapter 2). Addison-Wellesley, Reading, Massachusetts.

S.P. Brooks and G.O. Roberts (1996), Diagnosing convergence of Markov chain Monte Carlo algorithms. Preprint.

M.K. Cowles and B.P. Carlin (1996), Markov chain Monte Carlo convergence diagnostics: a comparative review. J. Amer. Stat. Assoc. **91**, 883–904.

M.K. Cowles and J.S. Rosenthal (1996), A simulation approach to convergence rates for Markov chain Monte Carlo algorithms. Stat. and Comput., to appear.

A.E. Gelfand and A.F.M. Smith (1990), Sampling based approaches to calculating marginal densities. J. Amer. Stat. Assoc. **85**, 398–409.

A.E. Gelfand, S.E. Hills, A. Racine-Poon, and A.F.M. Smith (1990), Illustration of Bayesian inference in normal data models using Gibbs sampling. J. Amer. Stat. Soc. **85**, 972–985.

A. Gelman and D.B. Rubin (1992), Inference from iterative simulation using multiple sequences. Stat. Sci., Vol. **7**, No. **4**, 457–472.

S. Geman and D. Geman (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. on pattern analysis and machine intelligence **6**, 721–741.

J. Geweke (1992), Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In Bayesian Statistics **4** (J.M. Bernardo et al., eds.), 169–193. Oxford University Press.

C. Geyer (1992), Practical Markov chain Monte Carlo. Stat. Sci., Vol. **7**, No. **4**, 473–483.

W.K. Hastings (1970), Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97–109.

P. Matthews (1993), A slowly mixing Markov chain with implications for Gibbs sampling. Stat. Prob. Lett. **17**, 231-236.

N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953), Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1091.

S.P. Meyn and R.L. Tweedie (1994), Computable bounds for convergence rates of Markov chains. Ann. Appl. Prob. **4**, 981–1011.

R.M. Neal (1993), Probabilistic inference using Markov chain Monte Carlo methods. Tech. Rep. CRG-TR-93-1, Dept. of Computer Science, University of Toronto.

A.E. Raftery and S. Lewis (1992), How many iterations in the Gibbs sampler? In Bayesian Statistics **4** (J.M. Bernardo et al., eds.), 763–773. Oxford University Press.

G.O. Roberts (1992), Convergence diagnostics of the Gibbs sampler. In Bayesian Statistics **4** (J.M. Bernardo et al., eds.), 777–784. Oxford University Press.

G.O. Roberts (1994), Methods for estimating L^2 convergence of Markov chain Monte Carlo. In Bayesian Statistics and Econometrics: Essays in Honor of Arnold Zellner (D. Berry et al., eds.). North Holland, Amsterdam.

J.S. Rosenthal (1995a), Rates of convergence for Gibbs sampling for variance components models. Ann. Stat. **23**, 740–761.

J.S. Rosenthal (1995b), Minorization conditions and convergence rates for Markov chain Monte Carlo. J. Amer. Stat. Assoc. **90**, 558–566.

J.S. Rosenthal (1995c), Convergence rates of Markov chains. SIAM Review **37**, 387–405.

J.S. Rosenthal (1996), Analysis of the Gibbs sampler for a model related to James-Stein estimators. Stat. and Comput. **6**, 269–275.

A.F.M. Smith and G.O. Roberts (1993), Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). J. Roy. Stat. Soc. Ser. B **55**, 3–24.

D.J. Spiegelhalter, A. Thomas, N.G. Best, and W.R. Gilks (1995), BUGS: Bayesian inference using Gibbs sampling, Version 0.30. Tech. Rep., MRC Biostatistics Unit, University of Cambridge.

M.A. Tanner and W.H. Wong (1987), The calculation of posterior distributions by data augmentation (with discussion). J. Amer. Stat. Assoc. **82**, 528-550.

S. Wolfram (1988), Mathematica: A system for doing mathematics by computer. Addison-Wesley, New York.

Table 1: Normal Means Model

Diagnostics applied to all parameters
300 parallel chains, 1000 iterations each

	mu	tau
True posterior values		
mean	50	0.0546
width	4.0059	0.0688
Full run, no diagnostics applied		
mean	50.0001	0.0547
t	0.0282	1.7071
MSE	0.0012	0
width	3.9944	0.0688
t	−1.444	−0.436
Batch, proportion early = .1		
mean	50.0034	0.0546
t	0.5991	0.2458
MSE	0.0016	0
width	3.9582	0.0682
t	−2.8612	−1.5381

Chains with initial iterations discarded: 49
Mean iterations retained: 873
Failures to converge: 0

Batch, proportion early = .25		
mean	49.9889	0.0546
t	−1.6156	0.3522
MSE	0.0021	0
width	3.9893	0.0685
t	−0.7665	−0.7971
Chains with initial iterations discarded: 42		
Mean iterations retained: 720		
Failures to converge: 0		
Spectral, proportion early = .1		
mean	50.0011	0.0546
t	0.1756	−0.1669
MSE	0.0018	0
width	3.9631	0.0683
t	−2.3912	−1.3365
Chains with initial iterations discarded: 48		
Mean iterations retained: 873		
Failures to converge: 0		
Spectral, proportion early = .25		
mean	49.9993	0.0546
t	−0.0884	0.0235
MSE	0.0027	0
width	3.9559	0.0689
t	−2.0845	0.2221
Chains with initial iterations discarded: 41		
Mean iterations retained: 701		
Failures to converge: 1		

Table 2: Bivariate Normal, rho = 0.95

Diagnostics applied to all parameters
200 parallel chains, 1500 iterations each

	mu1	mu2
True posterior values		
mean	0	0
width	3.9199	3.9199
Full run, no diagnostics applied		
mean	0.0036	0.0034
t	0.4379	0.4148

MSE	0.0134	0.0135
width	3.8801	3.8761
t	−2.528	−2.799

Batch, proportion early = .1

mean	0.0337	0.0321
t	1.2661	1.2014
MSE	0.0259	0.0261
width	3.8903	3.892
t	−0.767	−0.6757

Chains with initial iterations discarded: 36
Mean iterations retained: 1267
Failures to converge: 0

Batch, proportion early = .25

mean	0.0273	0.028
t	0.446	0.4576
MSE	0.0495	0.0494
width	3.828	3.8124
t	−1.6536	−1.6729

Chains with initial iterations discarded: 14
Mean iterations retained: 1018
Failures to converge: 0

Spectral, proportion early = .1

mean	0.0106	0.0101
t	0.6462	0.6128
MSE	0.0227	0.0228
width	3.8369	3.8416
t	−2.9355	−2.7723

Chains with initial iterations discarded: 85
Mean iterations retained: 1196
Failures to converge: 0

Spectral, proportion early = .25

mean	0.0568	0.0555
t	2.0729	2.0048
MSE	0.0483	0.0491
width	3.7822	3.7849
t	−2.8077	−2.6567

Chains with initial iterations discarded: 61
Mean iterations retained: 928
Failures to converge: 5