PERFORMANCE OF THE GIBBS, HIT-AND-RUN, AND METROPOLIS SAMPLERS

by

Ming-Hui Chen and Bruce Schmeiser
Purdue University

Technical Report #92-18

Department of Statistics
Purdue University

April 1992

Performance of the Gibbs, Hit-and-Run, and Metropolis Samplers *

Ming-Hui Chen
Department of Statistics
Purdue University

Bruce Schmeiser
Department of Industrial Engineering
Purdue University

Abstract

We consider the performance of three Monte Carlo Markov chain samplers: the Gibbs sampler, which cycles through coordinate directions; the Hit-and-Run sampler, which randomly moves in any direction; and the Metropolis sampler, which moves with a probability that is a ratio of likelihoods.

We obtain several analytical results. We provide a sufficient condition of the geometric convergence on a bounded region S for the Hit-and-Run sampler. For a general region S, we overview the sufficient geometric convergence condition for the Gibbs sampler, which is provided by Schervish and Carlin [1990]. We show that for a bivariate normal distribution this Gibbs sufficient condition holds, and the Gibbs marginal sample paths are each an AR(1) process, and we obtain the standard errors of sample means and sample variances, which we later use to verify empirical Monte Carlo results.

We empirically compare the Gibbs and Hit-and-Run samplers on bivariate normal examples. For zero correlation, the Gibbs sampler provides independent data, resulting in better performance than H&R. As the absolute value of the correlation increases, H&R performance improves, with H&R substantially better for correlations above 0.9.

We also suggest and study methods for choosing the number of replications, and for starting replications to reduce bias, for estimating the standard error of point estimators, and for reducing point-estimator variance. We suggest using one single long run instead of using multiple *i.i.d.* separate runs. To reduce the initial-condition bias of the sample mean, we suggest the antithetic starting idea of Deligonul [1987]. We suggest using overlapping batch statistics (obs) to get the standard errors of estimates; Additional empirical results show that obs is accurate.

Finally, we review the geometric convergence of the Metropolis algorithm, and develop a Metropolised Hit-and-Run sampler. This sampler works well for high-dimensional and complicated integrands or Bayesian posterior densities.

Keywords: antithetic variates, AR(1) process, Bayesian posteriors, geometric convergence, Markov chain, Monte Carlo, multidimensional integration, overlapping batch statistics, simulation.

*This research was supported in part by NSF grant DMS-8717799 at Purdue University.

1 Introduction

Evaluating the k-dimensional integral $\int_S h(\underline{x})f(\underline{x})d\underline{x}$, where f is a density function with support $S \subset R^k$, arises in many contexts. For example, if f is a Bayesian posterior distribution, then choosing $h(\underline{x}) = x_i$ yields the i^{th} marginal posterior mean. Monte Carlo methods estimate the integral's value by averaging many observations of $h(\underline{x})$, and therefore require a method for generating pseudo-random points $\underline{x} \in S$ with density f. Classical random-variate methods such as composition and acceptance/rejection are difficult to implement and/or inefficient in multiple dimensions. An alternative is Markov chain samplers, which produce an infinite sequence of autocorrelated points $\underline{X}$ whose stationary distribution is f.

One such Markov chain sampler is the Hit-and Run (H&R) sampler. In its original form, as proposed independently by Boneh and Golan [1979], Smith [1980, 1984], the H&R sampler generates uniformly distributed points on S. Belisle, Romeijn and Smith [1990] proposed a more-general version that generates a sample of points from an arbitrary continuous probability density function f over a bounded support S. Kaufman and Smith [1990] show that the H&R sampler converges geometrically to the stationary distribution f; their proofs restrict the support S to be an open bounded set and the density function f to be bounded from above and away from zero. Schmeiser and Chen [1991] propose a variation of the H&R sampler, and have shown that the distribution of iterates converge to the target distribution in total variation for any density function f.

Other Markov chain samplers include the Gibbs sampler (Geman and Geman [1984], Gelfand and Smith [1990]), sometimes called the Successive-Substitution sampler, and the general Metropolis sampler (Hastings [1970], Tierney [1991], Müller [1991]).

Tierney [1991] and Schervish and Carlin [1990] discuss geometrical convergence for various algorithms. Schervish and Carlin [1990] show that

$$\int\int |k(\underline{x},\underline{y})|^2 d\mu(\underline{x})d\mu(\underline{y}) < \infty. \tag{1.1}$$

is sufficient to imply geometric convergence for a sampler with positive kernal k applied to any density f; the measure μ is discussed in Section 4.

In Section 2 we review the Hit-and-Run sampler and several existing results from by Schmeiser and Chen [1991].

In Section 3 we note that the Schervish and Carlin sufficient condition does not hold for H&R. We then develop another sufficient condition for H&R that does hold.

In Section 4, we review the Gibbs sampler and the Schervish and Carlin [1990] sufficient condition for the geometric convergence. We then verify that this sufficient condition is satisfied for independent sampling and for any bivariate normal f. We conjecture that the condition holds in general for the Gibbs sampler.

In Section 5 we compare performance of the Gibbs and H&R samplers. In Section 5.1 we discuss four types of error: systematic bias, point-estimator bias due to the initial transient, point-estimator bias due to autocorrelation, and sampling error. We note possible sensitivity of the Gibbs sampler to bad random-number generators, advocate a single long run to minimize both types of point-estimator bias, and overlapping batch statistics (obs) for estimating point-estimator standard errors. In Section 5.2 we consider the general bivariate normal f. We show that the Gibbs sampler creates two intertwined AR(1) processes. We empirically investigate the Gibbs and H&R samplers, finding that large bivariate-normal correlation results in large sampler autocorrelations, which in turn result in point-estimator bias and high sampling error. H&R is less sensitive thans the Gibbs sampler to high bivariate-normal correlation.

We investigate estimating standard errors with obs in Section 6. We demonstrate that obs standard-error estimators are accurate for suitable long runs.

In Section 7 we discuss reducing sample-mean bias caused by initial conditions. We advocate antithetic initial conditions.

In section 8, we discuss the general Metropolis algorithm and the Metropolisized H&R sampler, which are especially good for sampling from a high-dimensional and complicated density f. For the general Metropolis algorithm, we discuss its convergence properties and how to accelerate Metropolis Markov chain.

2 The Hit-and-Run Sampler

Let f be an arbitrary density with any support $S \subset R^k$. We consider the following Monte Carlo sampler which is given in Romeijn and Smith [1990] and Schmeiser and Chen [1991].

Algorithm Hit-and-Run Sampler

step 0. Choose a starting point $\underline{x}_0 \in S$, and set i=0.

step 1. Generate a uniformly distributed unit-length direction $\underline{d}_i \overset{def}{=} (d_i^1, d_i^2, \cdots, d_i^k)$.

step 2. Find the set $S_i = S_i(\underline{d}_i, \underline{x}_i) \overset{def}{=} \{\lambda \in R \mid \underline{x}_i + \lambda \underline{d}_i \in S\}$.

step 3. Generate a signed distance λ_i from density

$$f_i(\lambda) = \frac{f(\underline{x}_i + \lambda \underline{d}_i)}{\int_{S_i} f(\underline{x}_i + u\underline{d}_i)du}, \quad \lambda_i \in S_i. \tag{2.1}$$

step 4. Set $\underline{x}_{i+1} = \underline{x}_i + \lambda_i \underline{d}_i$ and set i=i+1. Go to step 1.

A random unit-length direction $\underline{d}_i$ can be generated in Step 1 by independently generating $z_l \sim N(0,1), l = 1, 2, ..., k$, and setting $d_i^l = z_l \left(\sum_{j=1}^k z_j^2\right)^{-\frac{1}{2}}, l = 1, 2, ..., k$ (e.g. see Devroye [1986, Section 4.2]).

Let $\{\underline{X}_i, i \geq 0\}$ be the homogeneous Markov chain generated by the H&R sampler. Then, this Markov chain has one-step transition probability density at $\underline{X}_{i+1} = \underline{y}$ given $\underline{X}_i = \underline{x}$

$$p(\underline{y} \mid \underline{x}) = \frac{2}{C_k \parallel \underline{x} - \underline{y} \parallel^{k-1}} \cdot \frac{f(\underline{y})}{\int_{S_i(\frac{\underline{y}-\underline{x}}{\|\underline{y}-\underline{x}\|}, \underline{x})} f(\underline{x} + u\frac{\underline{y}-\underline{x}}{\|\underline{y}-\underline{x}\|})du}, \quad for\ all \ \ \underline{y} \neq \underline{x} \in S, \tag{2.2}$$

where $C_k = \frac{2\pi^{\frac{k}{2}}}{\Gamma(\frac{k}{2})}$ is the surface area of the k-dimensional unit hypersphere. Since the Lebesgue measure of one single point $\underline{x}$ is zero, we are free to give $p(\underline{x}|\underline{x})$ a positive value. We use the following results, which are proven in Schmeiser and Chen [1991].

Lemma 2.1 *The Markov chain $\{\underline{X}_i, i \geq 0\}$ is time reversible, i.e.,*

$$p(\underline{y}|\underline{x})f(\underline{x}) = p(\underline{x}|\underline{y})f(\underline{y}), \quad \textit{for every} \quad \underline{x}, \underline{y} \in S, \tag{2.3}$$

and $p(\underline{y}|\underline{x}) > 0$ for every $\underline{x}, \underline{y} \in S$.

Let $\mathcal{B}_s^k$ denote the Borel sets of S. For every $A \in \mathcal{B}_s^k$, let probability measure ϕ, defined by f, be

$$\phi(A) \overset{def}{=} \int_A f(\underline{x})d\underline{x}. \tag{2.4}$$

Then, we have

Lemma 2.2 *The probability measure ϕ is invariant for the Markov chain $\{\underline{X}_i, i \geq 0\}$, i.e.,*

$$\phi(A) = \int_S \int_A p(\underline{y}|\underline{x})d\underline{y}f(\underline{x})d\underline{x}, \tag{2.5}$$

for every $A \in \mathcal{B}_s^k$.

A Markov chain is called *ergodic* if it is positive *Harris* recurrent and aperiodic.

Proposition 2.1 *The Markov chain $\{\underline{X}_i, i \geq 0\}$ is ergodic.*

From Schmeiser and Chen [1991], we also have

Proposition 2.2 *The probability measure after the n-th iteration of the Markov chain $\{\underline{X}_i, i \geq 0\}$ converges to the invariant probability measure ϕ in total variation regardless of the starting point $\underline{x}_0$.*

Furthermore, we have

Proposition 2.3 *If h is integrable with respect to f, i.e., $\int_S |h(\underline{x})|f(\underline{x})d\underline{x} < \infty$, then for every fixed* $0 \leq j_0 < \infty$

$$\lim_{n\to\infty} \frac{1}{n - j_0 + 1} \sum_{j=j_0}^{n} h(\underline{X}_j) = E_f(h) \quad a.s., \tag{2.6}$$

where $E_f(h) = \int_S h(\underline{x})f(\underline{x})d\underline{x}$.

A fundamental problem of Markov chain sampling is to determine the speed at which the n^{th}-iteration distribution converges to the stationary distribution f. In the next section, we show that H&R converges geometrically when S is bounded.

3 Geometric Convergence of the Hit-and-Run Sampler

Applying the Schervish and Carlin sufficient condition (1.1) for geometric convergence to H&R yields

$$\int_S \int_S p(\underline{x}|\underline{y})p(\underline{y}|\underline{x})d\underline{x}d\underline{y} < \infty,$$

where, $p(\underline{y}|\underline{x})$ is the one-step transition probability density of the H&R sampler of Equation (2.2). Primarily due to the distance $||\underline{x} - \underline{y}||$ expression in the denominator of $p(\underline{y}|\underline{x})$, this sufficient condition does not hold for H&R, even for bounded support S.

Kaufman and Smith [1990] bound the H&R rate of convergence by assuming the density function f to be bounded from above and away from zero, and by assuming bounded support S. In this section, we relax the Kaufman and Smith sufficient conditions by substituting a weaker assumption on f.

Let $K(\underline{x}, A)$ be the H&R one-step transition probability kernel. Then,

$$K(\underline{x}, A) = \int_A p(\underline{y}|\underline{x})d\underline{y}, \quad for\ every\ \ \underline{x} \in S\ and\ A \in \mathcal{B}_s^k. \tag{3.1}$$

Let $K^n(\underline{x}, A)$ be the n^{th}-iteration transition probability kernel. We say that the kernel $K(\underline{x}, A)$ satisfies the *Minorization Condition* $M(n_0, \beta, s, \nu)$ if

$$K^{n_0}(\underline{x}, A) \geq \beta s(\underline{x})\nu(A), \quad for\ every\ \ \underline{x} \in S, A \in \mathcal{B}_s^k, \tag{3.2}$$

where $n_0 \geq 1$ is an integer, $\beta > 0$ is a constant, s is a positive real valued measurable function on $(S, \mathcal{B}_s^k)$ and ν is a positive measure on $\mathcal{B}_s^k$. Such an s is called a *Small Function.* A set $C \in \mathcal{B}_s^k$ is called a *Small Set* if the kernel $K(\underline{x}, A)$ satisfies a *Minorization Condition* $M(n_0, \beta, C, \nu)$ for an integer $n_0 \geq 1$, a constant $\beta > 0$, the *Small Function* $s(\underline{x}) = I_C(\underline{x})$, which is an indicator function, and a probability measure ν on $\mathcal{B}_s^k$, i.e.,

$$K^{n_0}(\underline{x}, A) \geq \beta\nu(A), \quad for\ every\ \ \underline{x} \in C, A \in \mathcal{B}_s^k, \tag{3.3}$$

Let B denote the k-dimensional unit open sphere centered at the origin and let ∂B denote its surface. Then the set of directions is

$$\partial B = \{\underline{d} \in R^k : \| \underline{d} \| = 1\}. \tag{3.4}$$

Let ρ be defined by

$$\rho = \sup_{\underline{d} \in \partial B,\ \underline{x} \in S} \left\{ \int_{S_i(\underline{d}, \underline{x})} f(\underline{x} + \lambda \underline{d}) d\lambda \right\}, \tag{3.5}$$

where $S_i(\underline{d}, \underline{x})$ is defined by Step 2 in the H&R sampler.

We now present an alternative sufficient condition for H&R geometric convergence.

Assumption A:

$$S\ \ is\ \ bounded, \ \ and\ \ \rho < \infty. \tag{3.6}$$

For many cases, the condition $\rho < \infty$ automatically holds. For example, if f is bounded above, then $\rho < \infty$ is satisfied since S is bounded.

The following lemma will result in the geometric convergence of the H&R sampler.

Lemma 3.1 *Under Assumption A, S is a Small Set. More specifically,*

$$K(\underline{x}, A) \geq \beta\phi(A), \quad for\ every\ \ \underline{x} \in S, A \in \mathcal{B}_s^k, \tag{3.7}$$

where $\beta = \frac{2}{C_k d(S) \rho}$, C_k *is the surface area of the k-dimensional unit hypersphere,* $d(S)$ *is the diameter of* S, *and* ϕ *is the stationary probability measure defined by (2.4).*

Proof: According to (2.2) and *Assumption A*, we have $p(\underline{y}|\underline{x}) \geq \frac{2}{C_k d(S)\rho} f(\underline{y}) = \beta f(\underline{y})$, for every $\underline{x}$, $\underline{y} \in S$. Thus,

$$K(\underline{x}, A) = \int_A p(\underline{y}|\underline{x}) d\underline{y} \geq \beta \int_A f(\underline{y}) d\underline{y} = \beta\phi(A),$$

for every $\underline{x} \in S$ and $A \in \mathcal{B}_s^k$. Therefore, S is a *Small Set.* ■

Before proving geometric convergence, we review two types of ergodicity. An ergodic Markov chain with invariant distribution ϕ is geometrically ergodic if there exists a nonnegative real-valued function M with $\int_S M(\underline{x})\phi(d\underline{x}) < \infty$ and a positive constant $r < 1$ such that

$$\| K^n(\underline{x}, \cdot) - \phi \| \leq M(\underline{x}) r^n, \tag{3.8}$$

for every $\underline{x} \in S$. The chain is uniformly ergodic if there is a positive constant M and a positive constant $r < 1$ such that

$$\sup_{\underline{x} \in S} \| K^n(\underline{x}, \cdot) - \phi \| \leq M r^n. \tag{3.9}$$

Uniform ergodicity implies geometric ergodicity. Under *Assumption A*, we prove that the Markov chain generated by the H&R sampler is uniformly ergodic.

Theorem 3.1 *Under Assumption A, the transition probability kernel $K(\underline{x}, A)$ of the H&R sampler satisfies a minorization condition $M(1, \beta, S, \phi)$, and the corresponding Markov chain is uniformly ergodic with the convergence rate $r \leq (1 - \beta)$, where β is defined in (3.7).*

Proof: Lemma 3.1 gives that under *Assumption A*, the kernel $K(\underline{x}, A)$ satisfies the minorization condition $M(1, \beta, S, \phi)$. Thus, the uniform ergodicity follows from Theorem 6.15 of Nummelin [1984], together with Proposition 2.1. According to Proposition 2 of Tierney [1991b], the rate r satisfies $r \leq 1 - \beta$. ■

The above theorem says that if S is bounded, then the H&R sampler has geometric convergence. S being bounded is not a strong condition, since a change of variables can (at least theoretically) transform an unbounded region into a bounded region; such transformations are common in numerical analysis. However, such transformations can be difficult to implement in high-dimensional problems, such as in Bayesian posterior analysis where unboundedness is common.

In the next section we empirically investigate the performance of the H&R sampler on unbounded S. We compare the performance of the Gibbs sampler, Metropolis algorithm and the H&R sampler, three Markov chain sampling schemes to generate a dependent random series from a multidimensional distribution. Furthermore, we discuss sources of error, using overlapping batch statistics for estimating standard errors, and using antithetic initial points for bias reduction.

4 The Gibbs sampler

In this section, we review the Gibbs sampler and the Schervish and Carlin [1990] sufficient condition for geometric convergence. This sufficient condition is investigated for the independent components and the bivariate normal distribution.

We denote densities generically by square brackets, so that joint, conditional and marginal forms for random variables $\underline{X}, \underline{Y}$, appear as $[\underline{X}, \underline{Y}]$, $[\underline{X}|\underline{Y}]$ and $[\underline{Y}]$, respectively. Let $\underline{X}$ be distributed as f. Then, the basic scheme of the Gibbs sampler is as follows:

Algorithm Gibbs Sampler

```
step 0. Choose an arbitrary starting point $\underline{X}^{(0)} = (X_1^{(0)}, X_2^{(0)}, \cdots, X_k^{(0)})$, and set i=0.
step 1. Generate $\underline{X}^{(i+1)} = (X_1^{(i+1)}, X_2^{(i+1)}, \cdots, X_k^{(i+1)})$.
        • Generate $X_1^{(i+1)} \sim [X_1 | X_2^{(i)}, \cdots, X_k^{(i)}]$;
        • Generate $X_2^{(i+1)} \sim [X_2 | X_1^{(i+1)}, X_3^{(i)}, \cdots, X_k^{(i)}]$;
        • Generate $X_3^{(i+1)} \sim [X_3 | X_1^{(i+1)}, X_2^{(i+1)}, X_4^{(i)}, \cdots, X_k^{(i)}]$;
          ...   ...   ...
        • Generate $X_k^{(i+1)} \sim [X_k | X_1^{(i+1)}, X_2^{(i+1)}, \cdots, X_{k-1}^{(i+1)}]$.
step 2. Set $i = i + 1$, and go to step 1.
```

Thus each component of $\underline{X}$ is visited in the natural order and a cycle in this scheme requires generation of k random variates. Under the regularity conditions, Gelfand and Smith [1990] give the following limiting result of the sample average.

Proposition 4.1 *For any measurable function T of $X_1, \cdots, X_k$ whose expectation exists,*

$$\lim_{i \to \infty} \frac{1}{i} \sum_{l=1}^{i} T(X_1^{(l)}, \cdots, X_k^{(l)}) \overset{a.s.}{=} E(T(X_1, \cdots, X_k)). \tag{4.1}$$

Now, we look at the sufficient condition of geometric convergence for Gibbs sampler. We use the notation of Schervish and Carlin [1990]. Let $\underline{X}$ be a random vector with coordinates $X_1, X_2, \cdots, X_k$. Let $f^{(i)}(\underline{X}) = f(X_i | X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_k)$. For two vectors $\underline{X}$ and $\underline{X}'$, define for each $i = 1, \cdots, k-1$, $\underline{X}^{(i')} = (X_1, \cdots, X_i, X'_{i+1}, \cdots, X'_k)$. We adopt the convention that $\underline{X}^{(0')} = \underline{X}'$, $\underline{X}^{(k')} = \underline{X}$, and $f^{(0)}(\underline{X}) = f(\underline{X})$. Define the measure μ by

$$\mu(A) = \int_A \frac{1}{f(\underline{x})} d\underline{x}, \quad \text{for } A \in \mathcal{B}_s^k.$$

Denote the transition kernel for the Gibbs sampler $k(\underline{x}', \underline{x})$ by

$$k(\underline{x}', \underline{x}) = \prod_{i=0}^{k} f^{(i)}(\underline{x}^{(i')}). \tag{4.2}$$

Define $\mathcal{H}$ be the Hilbert space of functions that are square integrable with respect to the measure μ. The inner product in $\mathcal{H}$ is $\int_S g_1(\underline{y}) g_2(\underline{y}) d\mu(\underline{y})$, for every $g_1, g_2 \in \mathcal{H}$. Thus, a sufficient condition of the geometric convergence of the Gibbs sampler, which is given by Schervish and Carlin [1990], is

Assumption B:

$$\int_S \int_S |k(\underline{x}', \underline{x})|^2 d\mu(\underline{x}') d\mu(\underline{x}) < \infty. \tag{4.3}$$

Let f_n be the density of the observation $\underline{X}^{(n)}$ after the n^{th} iteration of the Gibbs sampler with the starting density $f_0 \in \mathcal{H}$. Schervish and Carlin give the following geometric convergence result for the Gibbs sampler.

Proposition 4.2 *Under Assumption B, there exists a number $c \in [0,1)$ such that for every density $f_0 \in \mathcal{H}$,*

$$||f_n - f|| \leq ||f_0|| c^n, \quad for \quad all \quad n, \tag{4.4}$$

where $|| \cdot ||$ is the inner product in $\mathcal{H}$.

Let F_n and F be the probability measures corresponding to the densities f_n and f, respectively. Then, a natural result follows Proposition 4.2.

Corollary 4.1 *Under Assumption B, there exists a number $c \in [0,1)$ such that for every density $f_0 \in \mathcal{H}$, and for every $A \in \mathcal{B}_s^k$,*

$$|F_n(A) - F(A)| \leq ||f_0|| c^n, \quad for \quad all \quad n. \tag{4.5}$$

Usually, it is more convenient to start the Gibbs sampler with a point $\underline{x}_0$. In this case, we still have above two results for almost every $\underline{x}_0 \in S$. The only difference is that now we begin the Markov chain at $\underline{X}_1$ rather than at $\underline{x}_0$ by substituting c^{n-1} for c^n on the left of Equations 4.4 and 4.5 and similarly replacing f_0 by f_0^*, where $f_0^*(\underline{x}) = k(\underline{x}_0, \underline{x})$.

Assumption B is quite general in guaranteeing geometric convergence for the Gibbs sampler, but often it is hard to verify. However, this sufficient condition holds at least for the following two cases.

Result 4.1 *Let $\underline{X} = (X_1, X_2, \cdots, X_k) \sim f$. If X_1, X_2, $\cdots$, X_k are independent, then Assumption B holds.*

The above result is easy to verify, since for this special case, the transition kernel $k(\underline{x}', \underline{x})$ is $k(\underline{x}', \underline{x}) = f(\underline{x}')f(\underline{x})$.

Result 4.2 *Let $\underline{X} = (X_1, X_2) \sim N(\underline{\mu}, \Sigma)$, where $\underline{\mu} = (\mu_1, \mu_2)$ and*

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

with $|\rho| < 1$. Then, Assumption B holds.

Proof: This proof uses only basic calculus. To simplify notation, let $(X, Y) \sim N(\underline{\mu}, \Sigma)$. Then, the transition kernel $k((x', y'), (x, y)) = f_{X,Y}(x', y') f_{X|Y}(x|y') f_{Y|X}(y|x)$, where $f_{X,Y}$ is the joint *p.d.f.* of $N(\underline{\mu}, \Sigma)$, $f_{X|Y}(x|y')$ is the *p.d.f.* of $N(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y' - \mu_2), \sigma_1^2(1 - \rho^2))$, and $f_{Y|X}(y|x)$ is the *p.d.f.* of $N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1), \sigma_2^2(1 - \rho^2))$. So, for this case, ignoring constants, the left side of Equation (4.3) is proportional to

$$\begin{aligned} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \quad & \exp\{-\frac{1}{2(1-\rho^2)}[\frac{(x'-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x'-\mu_1)}{\sigma_1}\frac{(y'-\mu_2)}{\sigma_2} + \frac{(y'-\mu_2)^2}{\sigma_2^2}]\} \\ & \exp\{-\frac{(x-\mu_1-\rho\frac{\sigma_1}{\sigma_2}(y'-\mu_2))^2}{\sigma_1^2(1-\rho^2)}\}\exp\{-\frac{(y-\mu_2-\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1))^2}{\sigma_2^2(1-\rho^2)}\} \\ & \exp\{\frac{1}{2(1-\rho^2)}[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)}{\sigma_1}\frac{(y-\mu_2)}{\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}]\}dx'dy'dxdy. \end{aligned}$$

By changing variables $\frac{x-\mu_1}{\sigma_1\sqrt{1-\rho^2}} \to x$, $\frac{y-\mu_2}{\sigma_2\sqrt{1-\rho^2}} \to y$, $\frac{x'-\mu_1}{\sigma_1\sqrt{1-\rho^2}} \to x'$, and $\frac{y'-\mu_2}{\sigma_2\sqrt{1-\rho^2}} \to y'$, then ignoring constants, the above integral is proportional to

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\{-\frac{1}{2}[x'^2 - 2\rho x'y' + y'^2]\}\exp\{-(x-\rho y')^2\}$$
$$\exp\{-(y-\rho x)^2\}\exp\{\frac{1}{2}[x^2 - 2\rho xy + y^2]\}dx'dy'dxdy.$$

Then integrating x', y' and y out, the above integral is proportional to

$$\int_{-\infty}^{\infty} \exp\{-\frac{(1-\rho^2)^2}{2(1+\rho^2)}x^2\}dx. \tag{4.6}$$

Since Integral (4.6) is finite, *Assumption B* holds. ■

Results 4.1 and 4.2 imply that if the components of $\underline{X}$ are independent or if $\underline{X}$ is distributed as a bivariate normal distribution, then the distribution of the n^{th} Gibbs sampler iteration converges to the stationary distribution at a geometric rate. The examples demonstrate that none of independence, normality, or dimensionality are required for geometric convergence. We conjecture that *Assumption B* holds for any multinormal distribution. More boldly, we conjecture that if f is a density in k dimensions, then the Gibbs sampler converges geometrically.

5 Comparing the Gibbs and Hit-and-Run Samplers

In this section, we discuss Markov chain sampling versus *i.i.d.* sampling and discuss sources of error for Markov chain samplers. For the bivariate normal stationary distribution we show that the marginal sample paths of the Gibbs sampler are $AR(1)$ processes; we also derive the standard errors of sample mean and sample variance. For this special case, we empirically compare the Gibbs and Hit-and-Run samplers.

5.1 Controlling Sampling Error

If samples are generated by an ergodic Markov chain sampling scheme, the factors that influence the quality of point estimators include *standard error*, *estimator bias*, and *systematic bias*. Standard error arises from randomly sampling, estimator bias arises from using either a biased estimator or from nonrepresentative samples (such as caused by the initial transient), and systematic bias arises from programming error, numerical error, or the use of pseudo-random numbers. So,

true value = estimate − standard error − estimate bias − systematic bias,

where "-" means "eliminating".

In our experience the Gibbs sampler seems more sensitive to systematic error. Using the same (uniform) random-number and (non-uniform) random-variate generators, we have had examples where the empirical Gibbs sampler behavior didn't match known asymptotic behavior while the H&R sampler behaved as expected; changing generators caused both algorithms to behave as expected. The straight-forward logic of the Gibbs sampler might explain this sensitivity. The $(i(k-1)+j)^{\text{th}}$ random number $U_{i(k-1)+j}$ is transformed into $X_j^{(i)}$, so lack of k-dimensional uniformity in the pseudorandom number generator is passed directly to the Gibbs sampler results. Other samplers, such as H&R, use more-complicated transformations, which might ameliorate this effect. Systematic bias can be tested by comparing to known solutions for simple problems or by comparing to the known asymptotic performance. In the following discussion, we assume that the experiment has no systematic error.

We consider the following mean and variance estimation for a simple Markov process to look at how the correlation structure reflects standard error and estimator bias. Suppose $\{Y_i, i \geq 0\}$ is a stationary Markov process. Let $\mu = E(Y_i)$, $\sigma^2 = Var(Y_i)$ and $\rho_h = Corr(Y_i, Y_{i+h})$. Then, the usual estimates of mean and variance of Y_i are

$$\hat{\mu} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2. \tag{5.1}$$

If $\{Y_i, 1 \leq i \leq n\}$ are *i.i.d.*, then the standard error of $\bar{Y}$ is $\sigma/\sqrt{n}$. But, now, because of dependence,

$$nVar(\bar{Y}) = \sigma^2 \left[1 + 2\sum_{h=1}^{n-1}(1 - \frac{h}{n})\rho_h\right]. \tag{5.2}$$

Since often ρ_h is positive (Tierney [1991], Schmeiser and Chen [1991]), then the standard error for dependent observations is bigger than that for *i.i.d.* observations. Furthermore, for estimating variances

$$\begin{aligned} E(S^2) &= \frac{1}{n-1}E\left(\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right) = \frac{1}{n-1}\left\{n\sigma^2 - nVar(\bar{Y})\right\} \\ &= \frac{1}{n-1}\left\{n\sigma^2 - \sigma^2\left[1 + 2\sum_{h=1}^{n-1}(1 - \frac{h}{n})\rho_h\right]\right\} = \sigma^2\left[1 - \frac{2}{n-1}\sum_{h=1}^{n-1}(1 - \frac{h}{n})\rho_h\right]. \qquad (5.\ 3) \end{aligned}$$

Thus, S^2 is a biased estimate of σ^2, but for *i.i.d.* observations, S^2 is an unbiased estimate of σ^2. So, the correlation of dependent observations obtained by a Markov chain sampling causes the bigger standard error of the mean estimation, and also causes S^2 to be a biased estimate of σ^2.

We now discuss the choice of experimental design (which $\underline{X}$ observations to generate) and the choice of point estimator (primarily, which $\underline{X}$ observations to include when computing the point estimator).

We now consider three experimental designs, including their point-estimator bias and standard error.

First, at one extreme, as described by Tierney [1991], we can generate n *i.i.d.* Markov chain realizations, using only the last observation $\underline{X}_{i,m}$ of realization i in the point estimator. The disadvantage is the difficulty of determining a value of m that balances initial-transient point-estimator bias with the wasted data.

Second, at the other extreme, we can generate a single long run of length nm, using all observations $\underline{X}_1, \ldots, \underline{X}_{nm}$ in the point estimator, possibly after discarding the j_0 initial observations $\underline{X}_{-j_0+1}, \cdots, \underline{X}_0$. The disadvantage is that the autocorrelation between observations complicates standard-error estimation.

Third, an approach between these two extremes, we can use each of n *i.i.d.* runs of length m to obtain n point estimates, each based on $\underline{X}_{i,1}, \underline{X}_{i,2} \cdots, \underline{X}_{i,m}$. These n independent point estimates are averaged to obtain a single point estimator with an easy to compute standard-error estimate. The advantage and the disadvantage of the third approach is that it combines the best and worst of the first two approaches. Consider, for example, estimating the marginal variance σ^2. Let S_i^2 be the estimate of σ^2 from ith *i.i.d.* run of length m. Then the estimate of σ^2 is $\frac{1}{m}\sum_{i=1}^{m} S_i^2$, which has the same bias as a single S_i^2, since the bias depends only on run length m.

Although exceptions can occur, a single long run usually provides the best point estimator for a fixed number of observations nm, since a single long run simultaneously reduces the standard error and the bias. (Whitt [1991] discusses this issue, focusing on queueing simulations.)

The single long run does complicate standard-error estimation. Estimating standard errors of general point estimators from stationary autocorrelated data requires (either implicitly or explicitly) including the autocorrelations in the standard-error estimate. A variety of methods are discussed in system simulation textbooks, for example, Bratley, Fox and Schrage [1987] or Law and Kelton [1991], with emphasis on sample averages. General point estimators are considered in Schmeiser, Avramidis and Hashem [1990], who discuss overlapping batch statistics.

Since the standard error depends on the autocorrelation structure of the data, in the next section we investigate the correlation structure of the Gibbs sampler applied to bivariate normal models and compare the H&R sampler with the Gibbs sampler empirically.

5.2 Bivariate Normal Model

In this example, we apply both the H&R and Gibbs samplers to estimate the means, variances, and correlation of the bivariate normal distribution $N(\underline{\mu}, \Sigma)$, where $\underline{\mu} = (\mu_1, \mu_2)$, and

$$\Sigma = \left(\begin{array}{cc} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{array} \right),$$

where $|\rho| < 1$. Let $\{\underline{X}_i, i \geq 0\}$ and $\{\underline{Y}_i, i \geq 0\}$ denote Markov chains generated from the above bivariate normal distribution by the H&R and the Gibbs samplers, respectively. If we start from the stationary distribution, i.e., $\underline{X}_0$ and $\underline{Y}_0 \sim N(\underline{\mu}, \Sigma)$, then both $\{\underline{X}_i, i \geq 0\}$ and $\{\underline{Y}_i, i \geq 0\}$ are stationary stochastic processes. For this particular example, the Gibbs sampler has the following nice property.

Proposition 5.1 *Let $\underline{Y}_i = (Y_i^{(1)}, Y_i^{(2)})$. If $\underline{Y}_i$ is generated from a bivariate normal distribution by using the Gibbs sampler, then each of $Y_i^{(1)}$ and $Y_i^{(2)}$ is corresponding to an AR(1) process.*

Proof: Let $\{Z_i^{(1)}, Z_i^{(2)}, i \geq 0\}$ be an *i.i.d.* $N(0,1)$ random variable sequence. Then the structure of the Gibbs sampler implies

$$\left\{ \begin{array}{rcl} Y_i^{(1)} &=& \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(Y_{i-1}^{(2)} - \mu_2) + \sigma_1\sqrt{1-\rho^2}Z_i^{(1)} \\ Y_i^{(2)} &=& \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(Y_i^{(1)} - \mu_1) + \sigma_2\sqrt{1-\rho^2}Z_i^{(2)}, \quad for \ \ i \geq 1 \end{array} \right. \tag{5. 4}$$

$$\left\{ \begin{array}{rcl} Y_0^{(1)} &=& \mu_1 + \sigma_1 Z_0^{(1)} \\ Y_0^{(2)} &=& \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(Y_0^{(1)} - \mu_1) + \sigma_2\sqrt{1-\rho^2}Z_0^{(2)}. \end{array} \right. \tag{5. 5}$$

Now, we consider the first component $Y_i^{(1)}$. ¿From Equation 5.4, for $i \geq 1$

$$\begin{array}{rcl} Y_i^{(1)} &=& \mu_1 + \rho\frac{\sigma_1}{\sigma_2}\left[\rho\frac{\sigma_2}{\sigma_1}(Y_{i-1}^{(1)} - \mu_1) + \sigma_2\sqrt{1-\rho^2}Z_{i-1}^{(2)}\right] + \sigma_1\sqrt{1-\rho^2}Z_i^{(1)} \\ &=& \mu_1 + \rho^2(Y_{i-1}^{(1)} - \mu_1) + \rho\sigma_1\sqrt{1-\rho^2}Z_{i-1}^{(2)} + \sigma_1\sqrt{1-\rho^2}Z_i^{(1)}.. \end{array} \tag{5. 6}$$

Let $\psi = \rho^2$, $\sigma_1^{*2} = \sigma_1^2(1-\rho^4)$. Let $\{Z_i^*, i \geq 0\}$ denote an *i.i.d.* $N(0,1)$ random variable sequence. Since $Z_{i-1}^{(2)}$ and $Z_i^{(1)}$ are independently and identically distributed as $N(0,1)$, then we can rewrite Equation 5.6 as

$$\begin{aligned} Y_i^{(1)} &= \mu_1 + \psi(Y_{i-1}^{(1)} - \mu_1) + \sigma_1^* Z_i^*, \quad for \;\; i \geq 1, & (5.\ 7) \\ Y_0^{(1)} &= \mu_1 + \sigma_1 Z_0^*. & (5.\ 8) \end{aligned}$$

Thus, $\{Y_i^{(1)}, i \geq 0\}$ is an AR(1) process with lag-one autocorrelation $\psi = \rho^2$. Similarly, we can prove that $\{Y_i^{(2)}, i \geq 0\}$ is also an AR(1) process with lag-one autocorrelation $\psi = \rho^2$. The only difference is that we use $\sigma_2^* = \sigma_2\sqrt{1-\rho^4}$ instead of σ_1^* in Equation 5.7, and we use μ_2 and σ_2 instead of μ_1 and σ_1 in Equation 5.8. ■

Since $\{Y_i^{(1)}\}$ is an AR(1) process, then Schmeiser and Song [1990], for example, give the following result.

Result 5.1 *Let* $\bar{Y}^{(1)} = \frac{1}{n}\sum_{i=1}^n Y_i^{(1)}$. *Then* $Var(Y_i^{(1)}) = \sigma_1^2$, $\rho_h = Corr(Y_i^{(1)}, Y_{i+h}^{(1)}) = \psi^h$ *and*

$$nVar(\bar{Y}^{(1)}) = \sigma_1^2\left[\frac{1+\psi}{1-\psi} - \frac{2\psi(1-\psi^n)}{n(1-\psi)^2}\right]. \tag{5.9}$$

Therefore, if we use a single long run, then after the n^{th} iteration, for the Gibbs sampler we have

Result 5.2 *By using Equation 5.1 to estimate* μ_1 *and* σ_1^2, *then the standard error (ste) of* $\hat{\mu}_1$ *is*

$$ste(\hat{\mu}_1) = \frac{\sigma_1}{\sqrt{n}}\sqrt{\frac{1+\psi}{1-\psi} - \frac{2\psi(1-\psi^n)}{n(1-\psi)^2}} \tag{5.10}$$

and the bias of $\hat{\sigma}_1^2$ *is*

$$|bias(\hat{\sigma}_1^2)| = |E(\hat{\sigma}_1^2) - \sigma_1^2| = \frac{\sigma_1^2}{n-1}\left[\frac{2\psi}{1-\psi} - \frac{2\psi(1-\psi^n)}{n(1-\psi)^2}\right]. \tag{5.11}$$

The proof of Result 5.2 follows from Equations 5.2 and 5.3 and Result 5.1. Furthermore, from Equations 5.2 and 5.3 again, for general stationary processes we have the asymptotical result, that for large n

$$(ste(\hat{\mu}_1))^2 \doteq |bias(\hat{\sigma}_1^2)|. \tag{5.12}$$

We used these results to verify our simulation experiments with respect to programming , numerical, and initial-transient errors.

Deriving exact expressions for standard errors is more difficult when the point estimators are not means or when the sampler has a structure more complicated than the Gibbs sampler. An important case is the standard error of the Gibbs sampler variance estimator, as given in Proposition 5.2. We first give a well-known result of covariance of quadratic form without proof.

Lemma 5.1 *Let* $\underline{Y}$ *and* $\underline{\theta}$ *be two* $1 \times n$ *vectors. Let* Σ *denote an* $n \times n$ *positive definite matrix. Then if* $\underline{Y} \sim N(\underline{\theta}, \Sigma)$, *and if* A *and* B *are any two* $n \times n$ *symmetric matrices,*

$$Cov\left(\underline{Y}A\underline{Y}^t, \underline{Y}B\underline{Y}^t\right) = 2tr\left(A\Sigma B\Sigma\right) + 4\underline{\theta}A\Sigma B\underline{\theta}^t, \tag{5.13}$$

where $tr(A)$ *is the trace of* A.

Since $\hat{\sigma}_1^2 = S_{(1)}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i^{(1)} - \bar{Y}^{(1)})^2$, $\bar{Y}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} Y_i^{(1)}$, then let I be the identity matrix of order n, $\underline{1} = (1, 1, \cdots, 1)$, $\underline{\mu}_1 = \mu_1\ \underline{1}$, and $\underline{Y}^{(1)} = (Y_1^{(1)}, Y_2^{(1)}, \cdots, Y_n^{(1)})$,

$$S_{(1)}^2 = \frac{1}{n-1}\underline{Y}^{(1)}\left(I - \frac{1}{n}\underline{1}^t\ \underline{1}\right)\underline{Y}^{(1)^t}. \tag{5.14}$$

Proposition 5.2 gives the result of $Var(\hat{\sigma}_1^2)$.

Proposition 5.2

$$(n-1)^2 Var(\hat{\sigma}_1^2) = 2n\sigma_1^4\left[\frac{1+\psi^2}{1-\psi^2} - \frac{2\psi^2(1-\psi^{2n})}{n(1-\psi^2)^2} - \frac{(1+\psi)^2 + 4\psi^{n+1}}{n(1-\psi)^2}\right.$$
$$\left. - \frac{4\psi^2(1-\psi^{2n})}{n^2(1-\psi)^2(1-\psi^2)} + \frac{4\psi(1+\psi)(1-\psi^n)}{n^2(1-\psi)^3} + \frac{4\psi^2(1-\psi^n)^2}{n^3(1-\psi)^4}\right] \tag{5. 15}$$

Proof: See Appendix A. ■

Corollary 5.1 *The standard error (ste) of $\hat{\sigma}_1^2$ can be obtained by Equation 5.15, and*

$$ste(\hat{\sigma}_1^2) = \frac{\sqrt{2}\sigma_1^2}{\sqrt{n}}\sqrt{\frac{1+\psi^2}{1-\psi^2}} + O(\frac{1}{n}). \tag{5.16}$$

Similarly, we can get the expressions of $Var(\hat{\sigma}_2^2)$ and $ste(\hat{\sigma}_2^2)$ by using σ_2 instead of σ_1 in Equations 5.15 and 5.16.

Unlike the Gibbs sampler, analysis of the H&R sampler is less tractable even in this special bivariate normal case because of its sampling structure, i.e., generating a random direction plus a signed distance. Lacking theoretical results, we compare the empirical performance for both samplers, using both multiple runs and a single long run. As before, we let $\{\underline{X}_i, i \geq 0\}$ and $\{\underline{Y}_i, i \geq 0\}$ denote Markov chains generated from the above bivariate normal distribution by the H&R and the Gibbs samplers, respectively. Let $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$, and $\hat{\rho}$ denote the usual estimators of the means μ_1, μ_2, variances σ_1^2, σ_2^2, and correlation ρ of the bivariate normal distribution using either $\{\underline{X}_i, i \geq 0\}$ or $\{\underline{Y}_i, i \geq 0\}$. By Proposition 2.3 and Proposition 4.1, $\hat{\mu}_i \xrightarrow{\text{a.s.}} \mu_i$, $\hat{\sigma}_i^2 \xrightarrow{\text{a.s.}} \sigma_i^2$, $i = 1, 2$, and $\hat{\rho} \xrightarrow{\text{a.s.}} \rho$, as $n \to \infty$. We avoid initial bias by starting both samplers from the known stationary distribution. We use Markov chain $n = 1000$ iterations and $m = 500$ *i.i.d.* macro-replications.

Table 5.1: Empirical performance of the bivariate normal model

parameter	true value	Gibbs Sampler		H&R	
		estimate (ste)	bias	estimation (ste)	bias
μ_1	0.0	-0.013(0.014)	-0.013	-0.012(0.007)	-0.012
μ_2	0.0	-0.018(0.020)	-0.018	-0.017(0.010)	-0.017
σ_1^2	1.0	0.908(0.012)	-0.092	0.972(0.009)	-0.028
σ_2^2	2.0	1.816(0.025)	-0.184	1.943(0.018)	-0.057
ρ	0.99	0.98792(0.00017)	-0.00208	0.98927(0.000083)	-0.00073

Comment 5.1: In Table 5.1, the estimated bias is the parameter estimate minus the true parameter value.

Comment 5.2: One validation of the empirical results is to compare the estimates to the known true values for the Gibbs sampler obtained from Equations 5.10, 5.11, and 5.15: $ste(\hat{\mu}_1) = 0.0138$, $ste(\hat{\mu}_2) = 0.0194$, $ste(\hat{\sigma}_1^2) = 0.0126$, $ste(\hat{\sigma}_2^2) = 0.0251$, $bias(\hat{\sigma}_1^2) = -0.094$, and $bias(\hat{\sigma}_2^2) = -0.187$. The estimated values are close to the corresponding true values. For example, setting $\psi = 0$ (since macro-replications are independent) and simplifying in Equation 5.15 gives the value of the standard error of the sample variance of the Gibbs sampler estimator of the mean $\hat{\mu}_1$ as $[Var(\hat{\mu}_{1,j})/m]\sqrt{2/(m-1)} = 1.2 \times 10^{-5}$, where $Var(\hat{\mu}_{1,j})$, the j^{th} macro-replication estimator of μ_1, is from Equation 5.9. Therefore the difference is well within sampling error.

Table 5.1 shows that for this example the Gibbs sampler estimators of σ_1^2 and σ_2^2 have larger biases than the H&R estimators, and the H&R standard errors (ste) are smaller than those for the Gibbs sampler. Since both bias and large standard errors are caused by the the sum of autocorrelations, the empirical results imply that the Gibbs sampler has larger autocorrelations than the H&R sampler.

Figure 1 shows component-1 autocorrelations, both the known values for the Gibbs sampler and estimates (with 3-σ limits) for H&R. The empirical results are based on 30 independent replications of 50000 iterations each. At all lags H&R has substantially smaller autocorrelations. The sum of autocorrelations $\gamma_0 = \sum_{h=-\infty}^{\infty} \rho_h$ is about 40 for H&R and $(1+\psi)/(1-\psi) = 99.5$ for the Gibbs sampler. The constant γ_0 can be thought of as the number of dependent observations with information equivalent to a single independent observation, since the variance of the sample mean is $Var(X_i^{(1)})/(n/\gamma_0)$. So for this example the Gibbs sampler needs about 2.5 times more observations than H&R for the same precision.

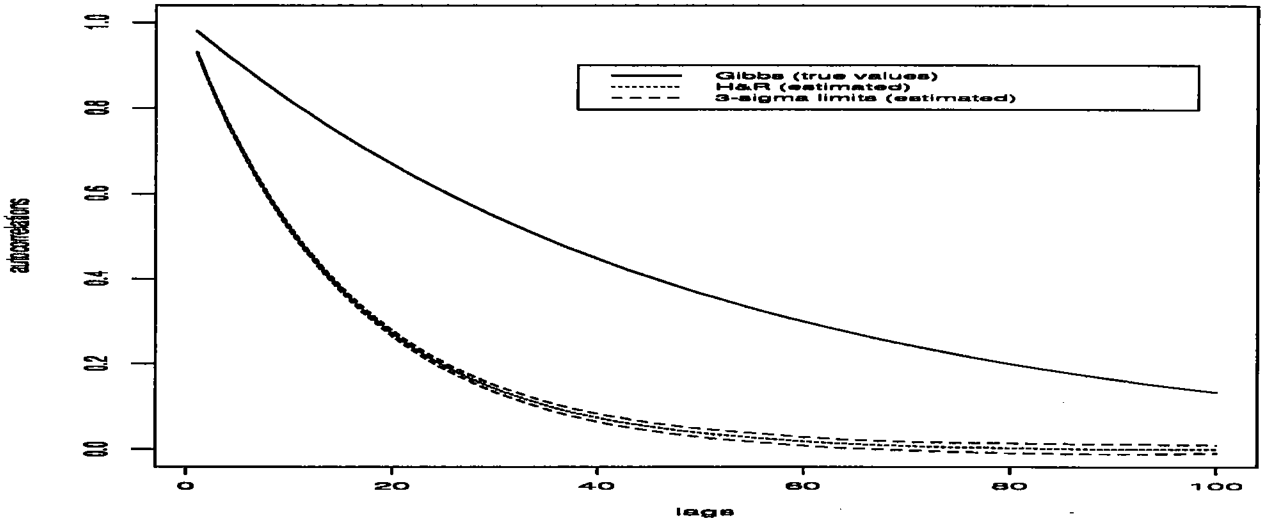

Figure 1: **Gibbs and H&R autocorrelations for component 1 of the bivariate model with $\rho = 0.99$, $\sigma_1^2 = 1$, and $\sigma_2^2 = 2$.**

An alternative to many short runs is a single long run. Figure 2 shows point estimates and associated (overlapping-batch-means) standard errors for μ_1 from a single run of 200,000 iterations for correlations $\rho = .99$ and $\rho = .01$. The results are similar to those for several short runs. For the extreme case $\rho = 0.99$, graphs (a) and (b) show that the H&R estimates are closer to the true value than the Gibbs estimates, and the H&R standard errors are smaller than those of the Gibbs sampler. Both samplers are more efficient in the nearly independent case of $\rho = 0.01$, as shown in graphs (c) and (d), with the Gibbs sampler having the smaller standard error.

In the next section, we discuss the overlapping batch statistics, and we provide empirical evidence that the estimated standard errors obtained by the overlapping batch statistics are good. We also discuss antithetic-variates sampling.

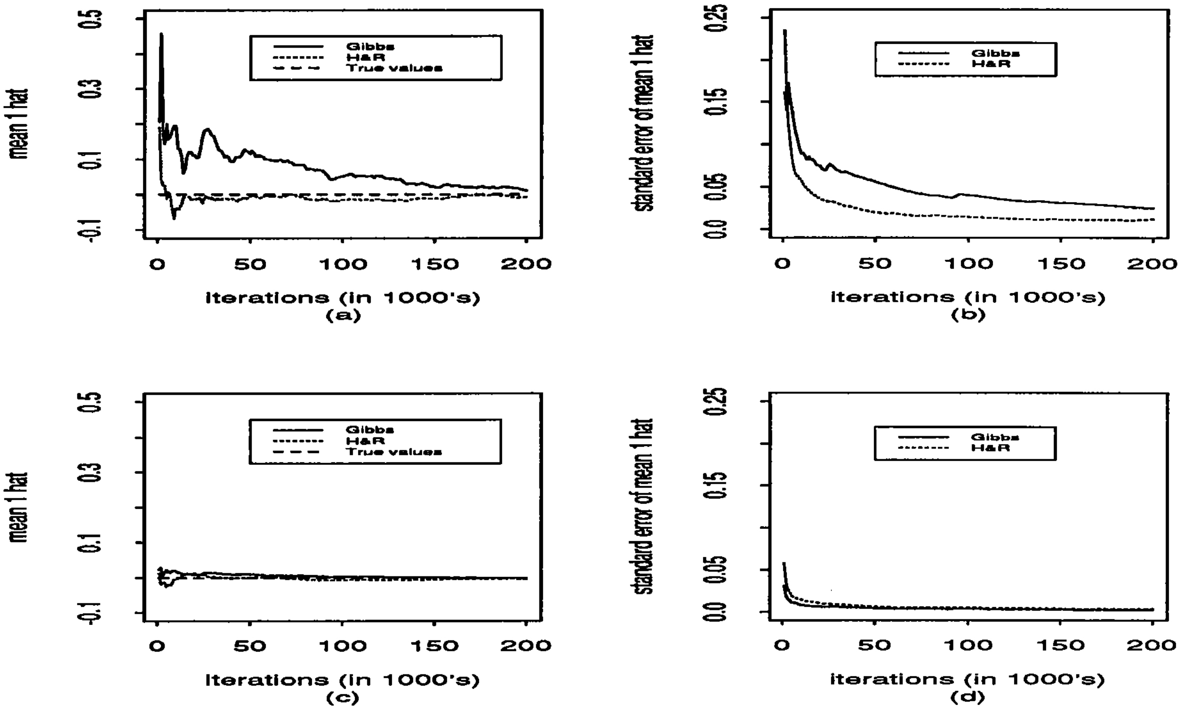

Figure 2: **A single trajectory for estimates and standard errors of the first bivariate normal mean with** $\mu_1 = \mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$, **(a) and (b)** $\rho = 0.99$, **(c) and (d)** $\rho = 0.01$.

6 Estimating Standard Errors Using Overlapping Batch Statistics

As noted in Section 5.1, when we use a single run to using Markov chain to sample from a multidimension distribution, a complication that arises from the dependence in using a single series is that variances of estimates are harder to obtain. Geweke [1991] suggested using the asymptotic variance of the sample mean obtained from spectral analysis with a Daniell window. The spectral analysis method for estimating variance of the sample mean of a dependent sample can be found in [Hannan, 1970, pp. 207-210] and [Bratley, Fox, and Schrage, 1987, pp. 100-103]. In this section, we discuss using overlapping batch statistics (Schmeiser, Avramidis, and Hashem, 1990) to estimate the variances of point estimators, including means, variances, and quantiles.

Suppose that the simulation experiment produces an output sequence $\{Y_i, 1 \leq i \leq n\}$, from which the point estimator, $\hat{\theta}$, is computed. As described in Schmeiser, Avramidis, and Hashem [1990], the overlapping batch statistics (obs) estimate of the variance of $\hat{\theta}$ is

$$\hat{V}(m) = \left[\frac{m}{n-m}\right] \frac{\sum_{j=1}^{n-m+1}(\hat{\theta}_j - \hat{\theta})^2}{(n-m+1)}, \tag{6.1}$$

where $\hat{\theta}_j$ is defined analogously to $\hat{\theta}$ but is a function of only $Y_j, Y_{j+1}, \cdots, Y_{j+m-1}$, the data in j^{th} batch of size m. Sufficient conditions for obs estimators to be unbiased and to have variance inversely proportional to n are given in Schmeiser, Avramidis, and Hashem [1990].

Specializing $\hat{\theta}$ to the sample mean yields the overlapping batch means (obm) estimator, which is based on $\hat{\theta}_j = m^{-1} \sum_{i=j}^{j+m-1} Y_i$. Song [1987] shows that the obm estimator is a spectral estimator with the spectral window

$$\omega_m^{(0)}(h) = \begin{cases} n^2(n-m+1)^{-1}(n-m)^{-1}(1-\frac{|h|}{m}) & \text{if } h = 1, \cdots, m, \\ 0 & \text{otherwise,} \end{cases}$$

which is essentially the Bartlett window when $n^2(n-m+1)^{-1}(n-m)^{-1} \simeq 1$. Therefore, ignoring end effects, obm can be viewed as an efficient computational method for the Barlett window at zero frequency. Computation is $O(n)$, as discussed in Meketon and Schmeiser [1984]; Schmeiser and Song [1987] give a Fortran subroutine for computing the obm estimate $\hat{V}_{obm}(m)$.

The statistical properties of $\hat{V}_{obm}(m)$ depend upon the batch size m. Bias decreases and variance increases with m. Asymptotically the bias of obm is that of the nonoverlapping-batch-means (nbm) estimator and has two-thirds the variance of nbm. Schmeiser and Song (1990) show that the mse-optimal asymptotic batch size is $m* = 1+((9/8)c_g n)^{1/3}$, where c_g is the center of gravity of absolute value of the autocorrelation lags.

Specializing $\hat{\theta}$ to the sample variance yields the overlapping batch variance (obv) estimator, which is based on $\hat{\theta}_j = S_{j,m}^2$, the sample variance of $\{Y_j, \cdots, Y_{j+m-1}\}$. The obv estimator is

$$\hat{V}_{obv}(m) = \frac{m}{n-m}\left[\frac{1}{n-m+1}\sum_{j=1}^{n-m+1}(S_{j,m}^2 - S^2)^2\right], \tag{6.2}$$

where m (satisfying $2 \leq m \leq n-1$) is the batch size. Schmeiser, Avramidis, and Hashem [1990] provide a Fortran subroutine for the obv estimator $\hat{V}_{obv}(m)$ requires in $O(n)$ time for any given value of m.

The primary difficulty in using the obs estimator is the choice of the batch size m to balance bias and variance, since no optimal batch size formula is known for obs estimators other than for obm. Schmeiser [1982] discusses the trade-off in the context of confidence intervals for nonoverlapping batch means. Song and Schmeiser [1988a,b] consider estimator variances and covariances of variance estimators for the standard error of the sample means. Limiting behavior for sample means is discussed by Goldsman and Meketon [1986] and Schmeiser and Song [1990]. For many situations, choosing m so that $10 \leq \frac{n}{m} \leq 20$ is reasonable.

We now study the empirical performance of obs estimators for the bivariate normal model discussed in Section 5.2, where we showed that the Gibbs sampler sample paths $Y_i^{(1)}$ and $Y_i^{(2)}$ are each an $AR(1)$ process with lag-one autocorrelation ρ^2. We directly check whether or not the obm and obv standard error estimates are consistent with the analytical values $Var(\hat{\mu}_j)$ and $Var(\hat{\sigma}_j^2)$, $j = 1, 2$ given in Equations 5.9 and 5.15.

Figure 3 shows the obm and obv estimates and the true standard errors for $\hat{\mu}_1$ and $\hat{\sigma}_1^2$ from a single Gibbs run of 200,000 iterations for the bivariate normal distribution with $\mu_1 = \mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$ and $\rho = .01, .5, .85, .99$. Graphs (a), (b), (c), and (d) show obm estimates and the true values; graphs (e) and (f) show obv estimates and the true values. Here we choose batch size $m = \max(\frac{n}{20}, \frac{1+\rho^2}{1-\rho^2})$.

This crude batch-size heuristic works reasonably well in this case. For the smaller correlations $|\rho| = .01, .5, .85$, the obm and obv estimators are consistent with the analytical values (see graphs (a), (b), (c), and (e)), even for small numbers of iterations n. For the large correlation $|\rho| = 0.99$, the obm and obv estimators also are consistent with the true standard errors (see graphs (d) and (f)) except for the small values of n. The difficulty with large correlation ρ is the large autocorrelations *psubh*, which decrease the effective sample size, as discussed earlier.

The alternative of using several shorter independent replications yields a consistent standard error for any number of iterations, but then the bias in $\hat{\theta}$ is a concern. Thus, with either a single long run or severl short runs, one needs to be careful not to stop too soon based on a bad standard-error estimate caused by small sample size.

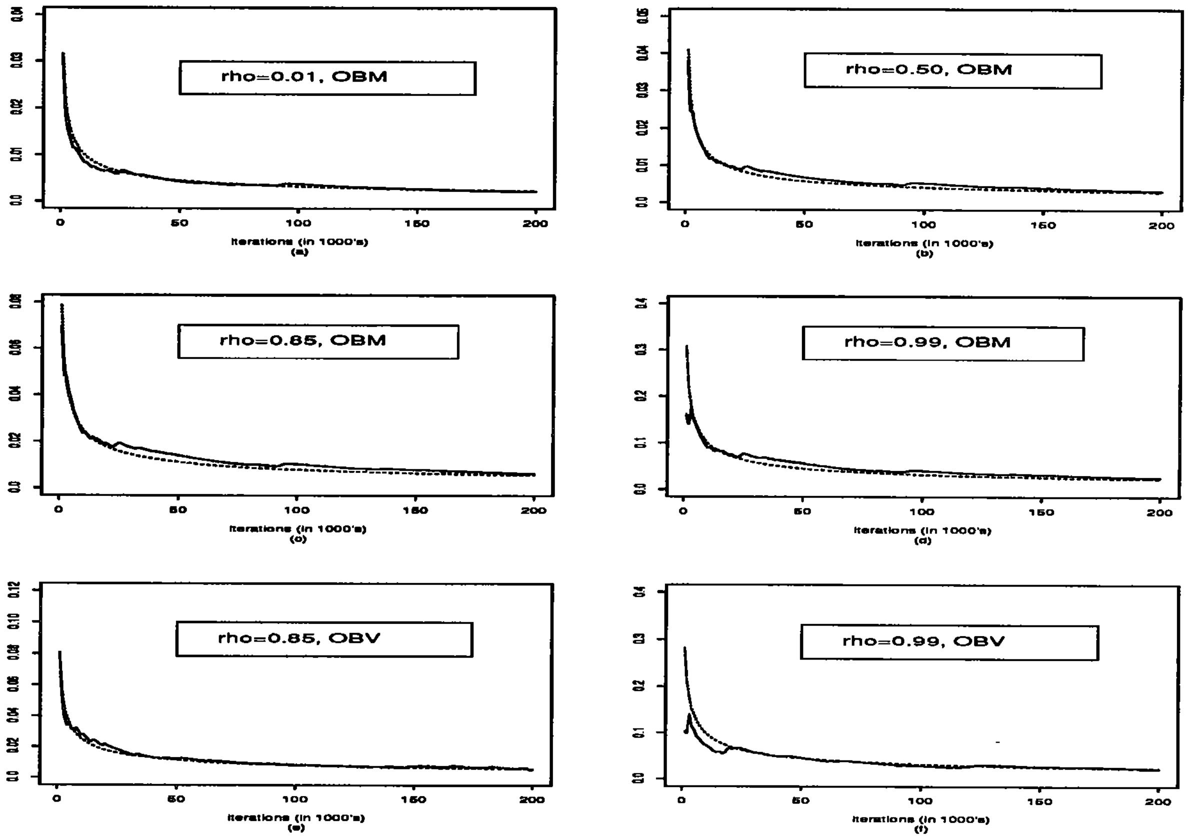

Figure 3: A single Gibbs trajectory for standard errors of $\hat{\mu}_1$ and $\hat{\sigma}_1^2$. The dashed curves are true standard errors; the solid curves are overlapping-batch-statistics standard-error estimates.

7 Antithetic Variates Sampling

Ideally, a Markov chain sampler is started by choosing a random point from the stationary distribution. Since such a choice is difficult, in practice the initial value is chosen from some tractable (often degenerate) distribution. The error caused by such sampling is called the initial bias.

Deligonul [1987] suggests reducing the initial bias by using antithetic initial points. The initial point of the second run is $2\underline{\bar{x}} - \underline{x}_0$, where $\underline{x}_0$ is the initial point of the first run and $\underline{\bar{x}}$ is the sample mean from the first run. The idea has intuitive appeal, at least for estimating means. Koksalan and Basoz [1991] discuss a modification, but we consider only the original idea.

If two runs are made, then inducing negative correlation between the two estimators reduces the variance of their average. The usual method, antithetic variates, uses a uniform random-number stream $\underline{U} = (U_1, U_2, \cdots)$ for the first run and the uniform random-number stream $\underline{W} = (W_1, W_2, \cdots)$ for the second run, where often $W_i = 1 - U_i$. Multiple streams are used if necessary to synchronize use of the random numbers for the two runs; in our case one stream is sufficient, since each iteration requires a constant number of random numbers. General discussion can be found in simulation textbooks, such as Bratley, Fox, and Schrage [1987].

Applying this approach to Markov chain sampling for estimating mean response, we have to use different strategy for each Markov chain sampling scheme. In particular, we must choose the transformation W_i to obtain the most negative correlation possible. This choice depends on the use for U_i. We discuss antithetic-variates sampling for the H&R and Gibbs samplers, individually.

7.1 Antithetic-variate sampling for the Gibbs sampler

We want to get an antithetic-variates estimate of a marginal mean of $\underline{Y}$ for the Gibbs sampler. Let $\underline{Y} \sim f$ denote a k-dimensional probability density function. We assume that Y_i is generated using the inverse transformations in each of the k coordinate directions with the k random numbers $U_{(i-1)k+1}, U_{(i-1)k+2}, \cdots, U_{ik}$, respectively.

Denote $\{\underline{Y}_i^{(j)}, 0 \leq i \leq n\}$, $j = 1, 2$, be two random variable strings, and define

$$\underline{\bar{Y}}^{(j)} = \frac{1}{n}\sum_{i=1}^{n} \underline{Y}_i^{(j)}, j = 1, 2.$$

Then antithetic-variate sampling for the Gibbs sampler is obtained using the same random-number stream U, but the second time using the uniform $(0, 1)$ random numbers $1 - U$. That is

Run 1 Start at $\underline{Y}_0^{(1)} = \underline{Y}_0$. Use the uniform random-number stream $\{U_1, U_2, \cdots, U_n, \cdots\}$ to get the first Gibbs sample $\{\underline{Y}_i^{(1)}, 1 \leq i \leq n\}$, and calculate $\underline{\bar{Y}}^{(1)}$.

Run 2 Start at $\underline{Y}_0^{(2)} = 2\underline{\bar{Y}}^{(1)} - \underline{Y}_0$. Use $\{1 - U_1, 1 - U_2, \cdots, 1 - U_n, \cdots\}$ to get the second Gibbs sample $\{\underline{Y}_i^{(2)}, 1 \leq i \leq n\}$, and calculate $\underline{\bar{Y}}^{(2)}$.

Thus, the antithetic-variate estimate of mean response is

$$\underline{\bar{Y}} = (\underline{\bar{Y}}^{(1)} + \underline{\bar{Y}}^{(2)})/2, \tag{7.1}$$

and the antithetic-variate path is

$$\underline{Y}_i = (\underline{Y}_i^{(1)} + \underline{Y}_i^{(2)})/2, \quad i = 1, 2, \cdots, n. \tag{7.2}$$

The above antithetic-variate sampling approach is common, made possible by the constant number of random numbers per iteration. More complicated sampling methods, such as acceptance/rejection or Metropolis ideas (Müller [1991]), require more than one stream to synchronize random-number use.

7.2 Antithetic-variate sampling for the Hit-and-Run sampler

Now, consider antithetic-variate sampling for the H&R sampler. Let $\underline{X} \sim f$. We assume that $\underline{X}_i$ is generated using $U_{(i-1)(k+1)+1}, U_{(i-1)(k+1)+2}, \cdots, U_{i(k+1)}$, with the first k random numbers generating the random k-dimensional direction and the $(k+1)^{\text{st}}$ random number used in the inverse transformation to generate the random signed distance λ.

Let $\{\underline{X}_i^{(j)}, 0 \leq i \leq n\}$, $j = 1, 2$, denote two random variable strings, and define $\bar{\underline{X}}^{(j)} = n^{-1} \sum_{i=1}^{n} \underline{X}_i^{(j)}, j = 1, 2$. The straight-forward transformation $W_i = 1 - U_i$ is not effective, since the direction is reversed and the signed difference is antithetic, causing positive correlation between the runs. Two alternatives seems reasonable: (a) *same* direction choices and and *antithetic* signed-distance choices and (b) *antithetic* direction choices and and *same* signed-distance choices. For our normal example, these methods produce the same sample path. Our implementation is of alternative (a).

Run 1 Start at $\underline{X}_0^{(1)} = \underline{X}_0$. Use uniform random-number stream $\{U_1, U_2, \cdots, U_n, \cdots\}$ to get the first H&R sample $\{\underline{X}_i^{(1)}, 1 \leq i \leq n\}$, and calculate $\bar{\underline{X}}^{(1)}$.

Run 2 Start at $\underline{X}_0^{(2)} = 2\bar{\underline{X}}^{(1)} - \underline{X}_0$. For sampling a random direction $\underline{d}$, use the same uniform random variables U_i as for **Run 1**, and for sampling a signed distance λ_i, use the antithetic variables $1 - U_i$ to get the second Hit-and-Run sample $\{\underline{X}_i^{(2)}, 1 \leq i \leq n\}$, and calculate $\bar{\underline{X}}^{(2)}$.

As with the Gibbs sampler, the antithetic-variates sample path and point estimator are the averages from the two runs. For each marginal mean, the point-estimator variance is reduced by half of the induced negative covariance. As we see in the next subsection, this strategy for this problem class induces perfect negative correlation in symmetric problems.

7.3 Empirical Results

Figure 4 shows antithetic sample paths of these antithetic-variate sampling schemes for the Gibbs and H&R sampler for the bivariate normal model with $\mu_1 = \mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$, and $\rho = 0.99$. The initial point of the first run is chosen from the arbitrary point (0.577, 0.537); the second run begins with Deligonul's antithetic reflection through the sample mean of the first run.

Graphs (a) and (b) give single run paths of Run 1, Run2, and the antithetic-variate estimates for $\hat{\mu}_1$. Graph (c) shows the enlarged Gibbs and H&R antithetic-variate estimate paths. For this problem, the antithetic-variate sampling estimates are extremely good, with the H&R sample paths, both original and antithetic, being a bit better than the Gibbs sampler, as before. Because of the almost perfect symmetry, the standard errors of the antithetic-variate point estimates for means are very small.

The two-run antithetic sampling scheme is quite inefficient for estimating marginal variances, since the two antithetic sample paths produce approximately the same distances from the mean. Therefore, an independent second run (or doubling the first run) would be better for estimating variances.

In addition, antithetic sampling is sometimes not appropriate even when estimating means. Often the inverse transformation is not tractable. And even when it is tractable, non-sysmmetric problems can reduce the induced correlation, since the sample paths will not be symmetric around the mean.

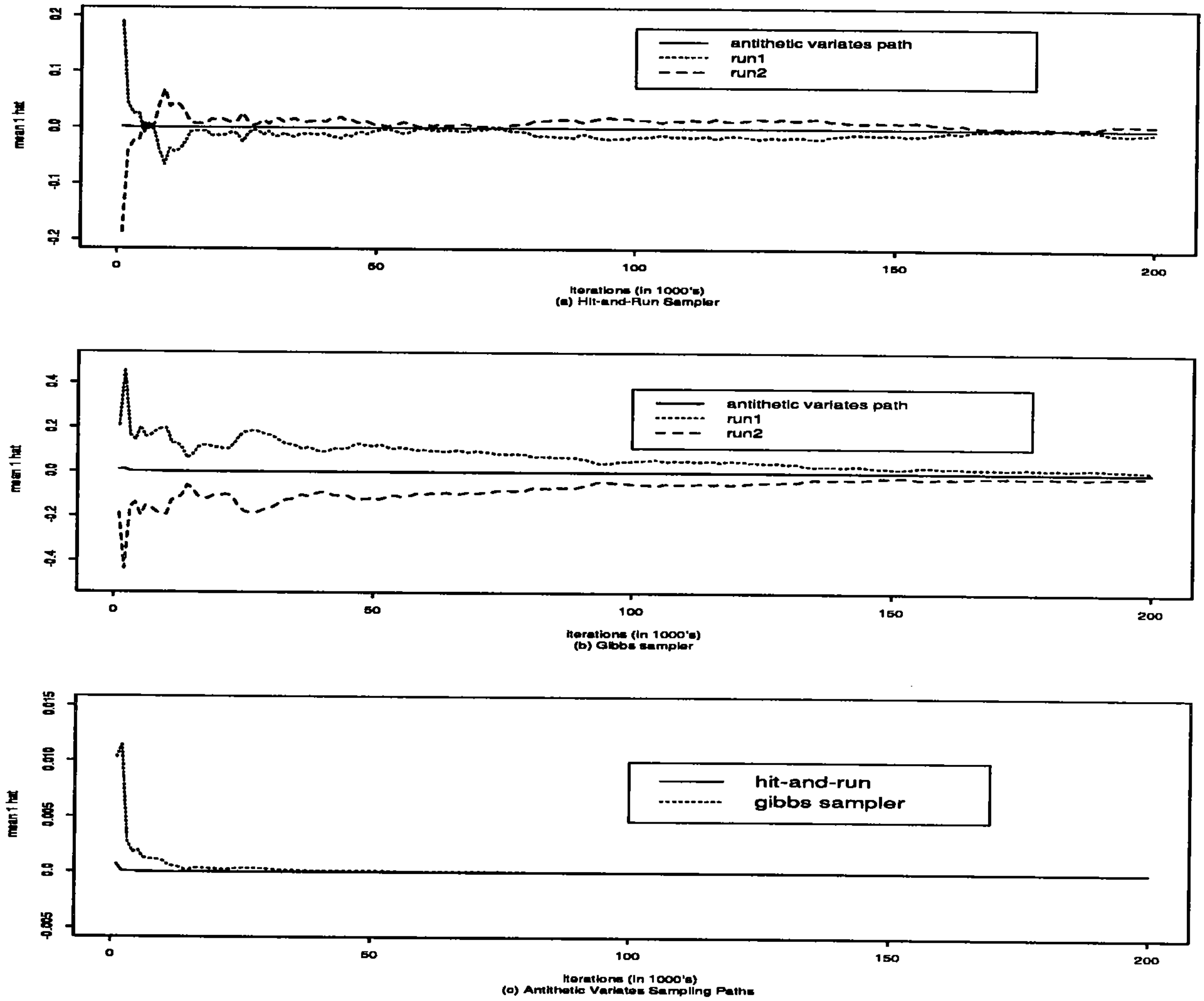

Figure 4: Antithetic sample paths for $\hat{\mu}_1$ for the bivariate normal example with $\mu_1 = \mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$, and $\rho = 0.99$.

8 The General Metropolis and Hit-and-Run Samplers

Hastings [1970] describes and discusses a general version of the Metropolis algorithm. Tierney [1991] shows that the general algorithm produces (under mild conditions) an ergodic Markov chain with stationary distribution f.

Let Q be a transition probability kernel of the form $Q(\underline{x}, d\underline{x}) = q(\underline{y}|\underline{x})d\underline{y}$, and assume $Q(\underline{x}, S)$ $= 1$ if $\underline{x} \in S$ and 0 otherwise.

Algorithm General Metropolis Sampler

step 0. Choose a starting point $\underline{x}_0 \in S$, and set i=0.

step 1. Generate a candidate $\underline{y}$ from distribution $Q(\underline{x}_i, \cdot)$.

step 2. Set

$$\underline{x}_{i+1} = \begin{cases} \underline{y}, & \text{with the probability } \; a(\underline{y}|\underline{x}_i) \\ \underline{x}_i, & \text{otherwise,} \end{cases}$$

where

$$a(\underline{y}|\underline{x}_i) = \begin{cases} \min\left\{\frac{f(\underline{y})q(\underline{x}_i|\underline{y})}{f(\underline{x}_i)q(\underline{y}|\underline{x}_i)}, 1\right\} & if \;\; f(\underline{x}_i)q(\underline{y}|\underline{x}_i) > 0 \\ \quad 1 & if \;\; f(\underline{x}_i)q(\underline{y}|\underline{x}_i) = 0. \end{cases} \tag{8.1}$$

step 3. Set $i = i + 1$, and go to step 1.

A nice feature of the general algorithm is that it is easy to implement even for complicated multidimension probability density function or posterior density function, because we sample from an arbitrary q rather than from the conditional distribution given $\underline{x}_i$. A disadvantage of this algorithm is that the Markov chain may stay the same point for more than one iteration, which creats positive autocorrelation and therefore slower point-estimator convergence. We may accelerate the Metropolis algorithm by choosing a good $q(\underline{y}|\underline{x})$.

For many multidimension distributions, generating random variates from the marginal conditional distribution $[X_j|X_1, X_2, \cdots, X_{j-1}, X_{j+1}, \cdots, X_k]$ for the Gibbs sampler, and a signed distance λ for the H&R sampler, is difficult. We may solve this problem by using acceptance/rejection or griddy method, but that may be inefficient. In this section, we focus on a special Metropolis algorithm, i.e., the independence chain sampler given by Tierney [1991], who also discusses other Metropolis samplers. The independence chain sampler has geometric convergence rate for a general density f on a general support S; we discuss how to accelerate the sampler. Finally, we consider a generalized H&R sampler mixed with Metropolis algorithm.

8.1 Independence Chain Sampler and Its Convergence

The independence chain sampler, proposed by Tierney [1991], is a special case of the general Metropolis sampler in which the candidate is generated independently of the current point $\underline{x}_i$.

Algorithm Independence chain Sampler

step 0. Choose a starting point $\underline{x}_0 \in S$, and set i=0.

step 1. Generate a candidate $\underline{y}$ from the distribution with a *p.d.f.* $g(\underline{y})$.

step 2. Set

$$\underline{x}_{i+1} = \begin{cases} \underline{y}, & \text{with the probability } \ a(\underline{y}|\underline{x}_i) \\ \underline{x}_i, & \text{otherwise,} \end{cases}$$

where

$$a(\underline{y}|\underline{x}_i) = \min\left\{\frac{\omega(\underline{y})}{\omega(\underline{x}_i)}, 1\right\}, \tag{8.2}$$

and the weight function $\omega = f/g$.

step 3. Set $i = i + 1$, and go to step 1.

The Markov chain generated by independence chain sampling has transition probability kernel

$$K(\underline{x}, A) = \left(1 - \int_S g(\underline{y})a(\underline{y}|\underline{x})d\underline{y}\right) I_A(\underline{x}) + \int_A g(\underline{y})a(\underline{y}|\underline{x})d\underline{y}, \tag{8.3}$$

where $A \subset S$ is an arbitrary Borel set in R^k, and $I_A(\underline{x}) = 1$ if $\underline{x} \in A$ and 0 otherwise. Under mild conditions, Tierney [1991] gives the following geometric convergence result.

Proposition 8.1 *An independence chain sampler kernel $K(\underline{x}, A)$ with density $g > 0$ on S and bounded weight function $\omega = f/g$ satisfies a minorization condition $M(1, \beta, S, \phi)$ with $\beta = (\sup\omega)^{-1}$, and is thus uniformly ergodic. The convergence rate is $r \leq (1-\beta) = (1 - (\sup\omega)^{-1})$.*

In practice, for a given density f, the problem is to find a easy-to-generate density g having a bounded weight function ω. Common choices of g include normal, split-normal, split-t, and mixtures of these. Here we introduce two other possibilities: the multivariate Cauchy (Johnson, [1987]) and, a spherically symmetric exponential distribution with density proportional to $e^{-\gamma||\underline{x}-\underline{\mu}||}$ with $\gamma > 0$. The following result says that these choices satisfy Tierney's conditions for ergodicity if the tails of g are sufficiently heavy for the f of interest.

Corollary 8.1 *Suppose f is continuous on R^k, and a density g is chosen one of the following densities: (a) multinormal, (b) split-normal, (c) split-t, (d) multivariate Cauchy, (e) spherically symmetric exponential distribution, and (f) mixtures of the above densities. If*

$$\lim_{||\underline{x}|| \to \infty} \sup_{\underline{d} \in \partial B} \frac{f(||\underline{x}||\underline{d})}{g(||\underline{x}||\underline{d})} < \infty, \tag{8.4}$$

where ∂B is defined in (3.4), then the independence chain sampler kernel $K(\underline{x}, A)$ is uniformly ergodic.

The proof of the above corollary follows from the boundedness of the weight function ω.

Generating random candidates from these distributions is easy. Algorithms to sample from split-normal and split-t can be founded in Geweke [1989]. The multivariate Cauchy distribution has density function

$$g(\underline{x}) = \pi^{-(k+1)/2}\Gamma(\frac{k+1}{2})(1 + \underline{x}\underline{x}^t)^{-(k+1)/2}, \qquad \underline{x} = (x_1, \cdots, x_k) \in R^k. \tag{8.5}$$

The k-variate Cauchy random variable $\underline{X}$ can be sampled as follows: generate the first component X_1 of $\underline{X}$ from the standard univariate Cauchy distribution, i.e., $X_1 = tan[\pi(U - \frac{1}{2})]$, where U is uniform $(0,1)$, then generate X_m from the conditional distribution of X_m given $X_1 = x_1, \cdots, X_{m-1} = x_{m-1}$, which is straight forward since $\left[m^{1/2}\left(1 + \sum_{i=1}^{m-1} X_i^2\right)^{-1/2}\right] X_m$ has a univariate Student's t distribution with m degrees of freedom, for $m = 2, \cdots, k$. The t variates can be sampled as $Y/\sqrt{(Z/m)}$, where Y is standard normal and Z is an independent $\Gamma(m/2, 2)$ variate. More details are given in Johnson [1987]. The spherically symmetric exponential variate $\underline{X}$ can be sampled as $\underline{X} = \underline{\mu} + \lambda \underline{d}$, where $\underline{d}$ is a k-dimensional uniformly distributed unit-length direction, and λ is a univariate $\Gamma(k, 1/\gamma)$.

In practice, Corollary 8.1 implies that if the support of f is unbounded, and the tail of f is no heavier than multivariate Cauchy, we can choose g as a multivariate Cauchy, and the independence chain sampler still has geometric convergence rate. To accelerate the Metropolis Markov chain, g should be chosen to be similar to f. Therefore, one could perform some pilot sampling, roughly estimate f, and use a mixture distribution of (a)-(e) in Corollary 6.1.

8.2 Metropolisized Hit-and-Run Sampler

The H&R sampler, described in Section 2, is a special case of the general algorithm with $[f(y)q(\underline{y}|\underline{x})]/[f(x)q(\underline{x}|\underline{y})] = 1$ and therefore $a(\underline{y}|\underline{x}) = 1$. The Romeijn and Smith [1990] generalization of H&R is also a special case, as is that of Schmeiser and Chen [1991].

The following algorithm modifies the Schmeiser and Chen [1991] generalization. We no longer require g to be symmetric about $\underline{x}_i$, but we now require that g depends on $\underline{x}_i$ only through the line segment S_i. This modification is a special case of the general algorithm. Since g depends on $\underline{x}_i$ through S_i, this algorithm is not a special case of independence-chain sampling.

Algorithm Metropolisized Hit-and-Run Sampler

Steps 0 through Step 2 are those of the H&R sampler, as given in Section 2. Steps 3 and 4 become

step 3. Generate a signed distance λ_i from density $g_i(\underline{x}_i + \lambda \underline{d}_i)$, where $\lambda_i \in S_i$.

step 4. Set $\underline{y} = \underline{x}_i + \lambda_i \underline{d}_i$. Then set

$$\underline{x}_{i+1} = \begin{cases} \underline{y}, & \text{with the probability } \; a(\underline{y}|\underline{x}_i) \\ \underline{x}_i, & \text{otherwise.} \end{cases}$$

step 5. Set i=i+1, and go to step 1.

Here

$$a(\underline{y}|\underline{x}_i) = \min\{\omega(\underline{y})/\omega(\underline{x}_i), 1\},$$

where $\omega(\underline{y}) = f(\underline{x}_i + \lambda_i \underline{d}_i)/g_i(\underline{x}_i + \lambda_i \underline{d}_i)$, $\omega(\underline{x}_i) = f(\underline{x}_i)/g_i(\underline{x}_i)$, and $g_i(\underline{x}_i + \lambda_i \underline{d}_i) > 0$ is an arbitrary density function on S_i, and *depends only on* the line segment S_i and not on the point $\underline{x}_i$.

The Metropolisized H&R Markov chain $\{\underline{X}_n, n \geq 0\}$ is time reversible and ergodic, and therefore has all the properties of H&R discussed in Section 2. Choosing g, the one-dimensional distribution of

λ, for the Metropolised H&R is often easier than choosing the k-dimensional g of the independence-chain sampler. For example, g could correspond to a piecewise-linear distribution function based on evaluating f at a few points on S_i, similar to the Griddy sampler of Ritter and Tanner [1991].

Appendix A: Proof of Proposition 5.2

From Equation 5.7 and Result 5.1, $\underline{Y}^{(1)} \sim N(\underline{\mu}_1, \sigma_1^2\Sigma)$, where

$$\Sigma = \begin{pmatrix} 1 & \psi & \psi^2 & \cdots & \psi^{n-1} \\ \psi & 1 & \psi & \cdots & \psi^{n-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \psi^{n-1} & \psi^{n-2} & \psi^{n-3} & \cdots & 1 \end{pmatrix}.$$

Using Equation 5.14, Lemma 5.1 with $\underline{\theta} = \underline{\mu}_1$ and $A = B = I - \frac{1}{n}\underline{1}^t\ \underline{1}$, and $\underline{\mu}_1 A = \underline{0}$ and $A\underline{\mu}_1^t = \underline{0}^t$, we have

$$\begin{aligned} (n-1)^2 Var(\hat{\sigma}_1^2) &= Var\left(\underline{Y}^{(1)}(I - \frac{1}{n}\underline{1}^t\ \underline{1})\underline{Y}^{(1)^t}\right) \\ &= 2\sigma_1^4 tr\left((I - \frac{1}{n}\underline{1}^t\ \underline{1})\Sigma(I - \frac{1}{n}\underline{1}^t\ \underline{1})\Sigma\right) \\ &= 2\sigma_1^4\left[tr(\Sigma^2) - \frac{2}{n}\underline{1}\Sigma^2\underline{1}^t + \frac{1}{n^2}(\underline{1}\Sigma\underline{1}^t)^2\right]. \end{aligned} \tag{A.1}$$

For easy writing, denote $\Sigma = (\sigma_{ij})_{n\times n}$, and $\rho_h = \psi^h$, for $h = 0, 1, \ldots$. Then $\sigma_{ij} = \rho_{|i-j|}$, and therefore

$$\underline{1}\Sigma\underline{1}^t = \sum_{i=1}^n\sum_{j=1}^n \sigma_{ij} = \sum_{i=1}^n\sum_{j=1}^n \rho_{|i-j|} = n\left[1 + 2\sum_{h=1}^{n-1}(1 - \frac{h}{n})\rho_h\right], \tag{A.2}$$

$$tr(\Sigma^2) = \sum_{i=1}^n\sum_{j=1}^n \sigma_{ij}^2 = \sum_{i=1}^n\sum_{j=1}^n \rho_{|i-j|}^2 = n\left[1 + 2\sum_{h=1}^{n-1}(1 - \frac{h}{n})\rho_h^2\right]. \tag{A.3}$$

Thus, by Equations 5.2 and 5.9,

$$\underline{1}\Sigma\underline{1}^t = n\left[\frac{1+\psi}{1-\psi} - \frac{2\psi(1-\psi^n)}{n(1-\psi)^2}\right], \tag{A.4}$$

$$tr(\Sigma^2) = n\left[\frac{1+\psi^2}{1-\psi^2} - \frac{2\psi^2(1-\psi^{2n})}{n(1-\psi^2)^2}\right]. \tag{A.5}$$

Since $\Sigma^2 = (\sum_{k=1}^n \sigma_{ik}\sigma_{kj})_{n\times n}$, then

$$\underline{1}\Sigma^2\underline{1}^t = \sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n \sigma_{ik}\sigma_{kj} = \sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n \rho_{|i-k|}\rho_{|k-j|}$$

$$= \sum_{k=1}^n\left(\sum_{i=1}^n \rho_{|i-k|}\right)\left(\sum_{j=1}^n \rho_{|j-k|}\right) = \sum_{k=1}^n\left(\sum_{i=1}^n \rho_{|i-k|}\right)^2 = \sum_{k=1}^n\left(\sum_{i=1}^k \rho_{|i-k|} + \sum_{i=k+1}^n \rho_{|i-k|}\right)^2$$

$$
\begin{aligned}
&= \sum_{k=1}^{n}\left(\sum_{i=1}^{k}\psi^{k-i}+\sum_{i=k+1}^{n}\psi^{i-k}\right)^{2}=\sum_{k=1}^{n}\left[\psi^{k-1}\frac{1-(\frac{1}{\psi})^{k}}{1-\frac{1}{\psi}}+\psi\frac{1-\psi^{n-k}}{1-\psi}\right]^{2}\\
&= \sum_{k=1}^{n}\left[\frac{1-\psi^{k}}{1-\psi}+\frac{\psi-\psi^{n-k+1}}{1-\psi}\right]^{2}=\frac{1}{(1-\psi)^{2}}\sum_{k=1}^{n}\left(1+\psi-\psi^{k}-\psi^{n-k+1}\right)^{2}\\
&= \frac{1}{(1-\psi)^{2}}\sum_{k=1}^{n}\left[(1+\psi)^{2}+\psi^{2k}+\psi^{2(n-k+1)}-2(1+\psi)\psi^{k}-2(1+\psi)\psi^{n-k+1}+2\psi^{n+1}\right]\\
&= \frac{1}{(1-\psi)^{2}}\left[n\left((1+\psi)^{2}+2\psi^{n+1}\right)+2\sum_{k=1}^{n}\psi^{2k}-4(1+\psi)\sum_{k=1}^{n}\psi^{k}\right]\\
&= \frac{1}{(1-\psi)^{2}}\left[n\left((1+\psi)^{2}+2\psi^{n+1}\right)+2\psi^{2}\frac{1-\psi^{2n}}{1-\psi^{2}}-4(1+\psi)\psi\frac{1-\psi^{n}}{1-\psi}\right]\\
&= n\left[\frac{(1+\psi)^{2}}{(1-\psi)^{2}}+\frac{2\psi^{n+1}}{(1-\psi)^{2}}+\frac{2\psi^{2}(1-\psi^{2n})}{n(1-\psi)^{2}(1-\psi^{2})}-\frac{4\psi(1+\psi)(1-\psi^{n})}{n(1-\psi)^{3}}\right]. \qquad \text{(A.6)}
\end{aligned}
$$

Thus, by Equations A.1, A.4, A.5, and A.6,

$$
\begin{aligned}
&(n-1)^{2}Var(\hat{\sigma}_{1}^{2}) = 2\sigma_{1}^{4}\left[tr(\Sigma^{2})-\frac{2}{n}\underline{1}\Sigma^{2}\underline{1}^{t}+\frac{1}{n^{2}}(\underline{1}\Sigma\underline{1}^{t})^{2}\right]\\
&= 2n\sigma_{1}^{4}\left[\left(\frac{1+\psi^{2}}{1-\psi^{2}}-\frac{2\psi^{2}(1-\psi^{2n})}{n(1-\psi^{2})^{2}}\right)+\frac{1}{n}\left(\frac{1+\psi}{1-\psi}-\frac{2\psi(1-\psi^{n})}{n(1-\psi)^{2}}\right)^{2}\right.\\
&\left.-\frac{2}{n}\left(\frac{(1+\psi)^{2}}{(1-\psi)^{2}}+\frac{2\psi^{n+1}}{(1-\psi)^{2}}+\frac{2\psi^{2}(1-\psi^{2n})}{n(1-\psi)^{2}(1-\psi^{2})}-\frac{4\psi(1+\psi)(1-\psi^{n})}{n(1-\psi)^{3}}\right)\right]. \qquad \text{(A.7)}
\end{aligned}
$$

By simplifying the right hand side of Equation A.7, Proposition 5.2 follows. ■

References

[1] Belisle, Claude J.P., Romeijn, H. Edwin and Smith, Robert L. (1990), "Hit-and-Run Algorithms for Generating Multivariate Distributions," Technical Report 90-18, The University of Michigan, Department of Industrial and Operations Engineering.

[2] Boneh, A. and Golan A. (1979), "Constraints' Redundancy and Feasible Region Boundedness by Random Feasible Point Generator (RFPG)," Third European Congress on Operations Research, EURO III, Amsterdam (April 9-11).

[3] Bratley, P., Fox, B. L., and Schrage, L. E. (1987), *A Guide to Simulation*, Second Edition, Springer-Verlag, New York.

[4] Deligonul, Z.S. (1987), "Antithetic Bias Reduction for Discrete-Event Simulations," *J. Opl. Res. Soc. 38, 5,* 431-437.

[5] Devroye, Luc (1986), *Non-Uniform Random Variate Generation*, Springer-Verlag, New York.

[6] Gelfand, A. E. and Smith, A.F.M. (1990), "Sampling Based Approaches to Calculating Marginal Densities," *Journal of American Statistical Association, 85,* 398-409.

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

[8] Geweke, J. (1989), "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica, 57,* 1317-1339.

[9] Geweke, J. (1991), "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments," *Abstracts of the Fourth Valencia International Meeting on Bayesian Statistics,* Peñiscola, Spain, 21.

[10] Goldsman, D. and Meketon, M.S. (1986), "A Comparison of Several Variance Estimators," Technical Report J-85-12, Georgia Institute of Technology, School of Industrial and Systems Engineering.

[11] Hannan, E.J. (1970), *Multiple Time Series*, Wiley, New York.

[12] Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika, 57*, 97 -109.

[13] Johnson, M.E. (1987), *Multivariate Statistical Simulation*, John Wiley & Sons, New York.

[14] Kaufman, David E. and Smith, Robert L. (1990), "Optimal Direction Choice for Hit-and-Run Sampling," Technical Report 90-08, University of Michigan, Department of Industrial and Operations Engineering.

[15] Koksalan, M.M. and Basoz, N. (1991), "A Replication Approach to Interval Estimation in Simulation," In *Proceedings of the 1991 Winter Simulation Conference*, 1023-1029.

[16] Law, A.M. and Kelton, W.D. (1991), *Simulation Modeling & Analysis*, Second Edition, McGraw-Hill, Inc., New York.

[17] Meketon, M.S. and Schmeiser, B.W. (1984), "Overlapping Batch Means: Something for Nothing?" In *Proceedings of the Winter Simulation Conference*, 227-230.

[18] Müller, P. (1991), "A Generic Approach to Posterior Integration and Gibbs Sampling," Technical Report # 91-09, Purdue University, Department of Statistics.

[19] Nummelin, E. (1984), *General Irreducible Markov Chains and Non-negative Operators*, Cambridge University Press, Cambridge.

[20] Ritter, C. and Tanner, M. A. (1991), "Facilitating the Gibbs Sampler: the Gibbs Stopper and the Griddy Gibbs Sampler," Technical Report, Division of Biostatistics, University of Rochester.

[21] Romeijn, H. Edwin and Smith, Robert, L. (1990), "Sampling Through Random Walks," Technical Report 90-2, The University of Michigan, Department of Industrial and Operations Engineering.

[22] Schervish, Mark J. and Carlin, Bradley P. (1990) "On the Convergence of Successive Substitution Sampling," Technical Report No. 492, Carnegie-Mellon University, Department of Statistics.

[23] Schmeiser, B.W. (1982), "Batch Size Effects in the Analysis of Simulation Output," *Operations Research 30, 3*, 556-568.

[24] Schmeiser, Bruce W., Avramidis, Thanos N., and Hashem, Sherif (1990), "Overlapping Batch Statistics," In *Proceedings of the 1990 Winter Simulation Conference*, 395-398.

[25] Schmeiser, Bruce W. and Chen, Ming-Hui (1991),"On Hit-and-Run Monte Carlo Sampling for Evaluating Multidimensional Integrals," revision of Technical Report 91-39, Purdue University, Department of Statistics.

[26] Schmeiser, B.W. and Song, W.-M. T. (1987), "Correlation among Estimators of the Variance of the Sample Mean," In *Proceedings of the Winter Simulation Conference*, 309-317.

[27] Schmeiser, Bruce W. and Song, W.-M T. (1990), "Optimal Mean-Squared-Error Batch Sizes," revision of Technical Report SMS-89-3, Purdue University, School of Industrial Engineering.

[28] Smith, R. L. (1980), "A Monte Carlo Procedures for Generating Random Feasible Solutions to Mathematical Programs," A Bulletin of the ORSA/TIMS Joint National Meeting, Washington, D.C., 101.

[29] Smith, R. L. (1984), "Efficient Monte Carlo Procedures for Generating Points Uniformly Distributed over Bounded Regions," *Operations Research, 32, 6*, 1297-1308.

[30] Song, W.-M. T. (1988), "Estimators of the Variance of the Variance of the Sample Mean: Quadratic Forms, Optimal Batch Sizes, and Linear Combinations," unpublished Ph.D. Dissertation, Purdue University, School of Industrial Engineering.

[31] Song, W.-M. T. and Schmeiser, B.W. (1988a), "On the Dispersion Matrix of Estimators of the Variance of the Sample Mean in the Analysis of Simulation Output," *Operations Research Letters, 7, 5*, 259-266.

[32] Song, W.-M. T. and Schmeiser, B.W. (1988b), "Minimal-Mse Linear Combinations of Variance Estimators of the Sample Mean," In *Proceedings of the Winter Simulation Conference*, 414-421.

[33] Tierney, Luke (1991), "Markov Chains for Exploring Posterior Distributions," Technical Report No. 560, University of Minnesota, School of Statistics.

[34] Whitt, W. (1991), "The Efficiency of One Long Run Versus Independent Replications in Steady-State Simulation," *Management Science, 37, 6*, 645-666.