ON RANDOM-DIRECTION MONTE CARLO SAMPLING FOR EVALUATING MULTIDIMENSIONAL INTEGRALS*

by

Bruce Schmeiser	and	Ming-Hui Chen
Department of Industrial Engineering		Department of Statistics
Purdue University		Purdue University
West Lafayette, IN 47907		West Lafayette, IN 47907-1399

Technical Report # 91-39

Department of Statistics
Purdue University

July, 1991

*This research was supported in part by NSF Grants DMS–8717799 at Purdue University.

△

ON RANDOM-DIRECTION MONTE CARLO SAMPLING FOR EVALUATING MULTIDIMENSIONAL INTEGRALS

by

Bruce Schmeiser
Department of Industrial Engineering
Purdue University
West Lafayette, IN 47907

and

Ming-Hui Chen
Department of Statistics
Purdue University
West Lafayette, IN 47907-1399

Abstract

We elaborate on the hit-and-run algorithm, a Monte Carlo approach that estimates the value of a high-dimensional integral with integrand $h(\underline{x})f(\underline{x})$ by sampling from a time-reversible Markov chain over the support of the density f. The Markov chain transitions are defined by choosing a random direction and then moving to a new point $\underline{x}$ whose likelihood depends on f in that direction. The serially dependent observations of $h(\underline{x}_i)$ are averaged to estimate the integral. The algorithm applies directly to f being a nonnegative function with finite integral.

We generalize the convergence results of Belisle, Romeijn and Smith to unbounded regions and to unbounded integrands. Here convergence is of the point estimator to the value of the integral; this convergence is based on convergence in distribution of realizations to their limiting distribution f. In addition we discuss three variations that are intended to reduce point-estimator variance: conditional expectation in the random direction, sampling in multiple directions, and adaptive external control variates.

An important application is determining properties of Bayesian posterior distributions. Here f is proportional to the posterior density and h is chosen to indicate the property being estimated. Typical properties include means, variance, correlations, probabilities of regions, and predictive densities.

Keywords: Bayesian posterior distribution, Barker's method, conditional expectation, control variates, hit-and-run, Markov chain, Metropolis's method, simulation, stratified sampling.

1 Introduction

We consider Monte Carlo methods for evaluating the multi-dimensional integral $\int_{R^k} h(\underline{x})f(\underline{x})d\underline{x}$, where f is a density function. Methods that sample from f are applicable, since this integral is an expectation $E_f(h)$. Such methods are useful when no closed-form solution can be found, whether due to the complexity of the problem, due to the integrand being known only numerically, or due to a desire for automatic evaluation.

The Bayesian statistical community has shown a recent interest in multidimensional integration for determining properties of posterior distributions f. Methods include importance sampling [Geweke, 1989], approximation via Laplace's method [Tierney and Kadane, 1986], Gaussian quadrature [Naylor and Smith, 1982 and 1988], data augmentation [Tanner and Wang, 1987], and the Gibbs sampler [Gelfand and Smith, 1990].

Very recently, both Bayesian statisticians and those interested in the more-general numerical integration problem have developed the Markov chain sampling methods of Metropolis *et al.* [1953] and their generalization, as discussed in Hastings [1970]. Applegate, Kannan, and Polson [1990] showed polynomial-time convergence for Markov chain algorithms that move over a finite discrete state space. Müller [1991] investigates Metropolis' algorithm, with emphasis on fitting a multivariate normal distribution to the posterior density and on specializing the algorithm for use in Gibbs sampling.

Our work is closely related to Belisle, Romeijn and Smith [1990], who propose Markov chain "hit-and-run" algorithms. Quoting almost directly, their description follows. Let S be a bounded open subset of R^k and let ϕ be an absolutely continuous probability measure on S. Let f be a probability density function (*p.d.f.*) for ϕ and assume that it is bounded, almost everywhere continuous (with respect to Lebesgue measure on S), and strictly positive. Let B denote the k-dimensional unit open sphere centered at the origin and let ∂B denote its topological boundary. Thus

$$B = \{\underline{x} \in R^k : ||\underline{x}|| < 1\} \text{ and } \partial B = \{\underline{x} \in R^k : ||\underline{x}|| = 1\}.$$

Finally, let ν be an arbitrary probability measure on ∂B. The hit-and-run algorithm with direction distribution ν and with target distribution ϕ can be described as follows:

Algorithm Hit-and-Run

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

step 1. Choose a direction $\underline{d}_i$ on ∂B, with distribution ν.

step 2. Choose a signed distance $\lambda_i \in S_i(\underline{d}_i, \underline{x}_i) \overset{def}{=} \{\lambda \in R : \underline{x}_i + \lambda \underline{d}_i \in S\}$, from the distribution with density

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

step 3. Set $\underline{x}_{i+1} = \underline{x}_i + \lambda_i \underline{d}_i$ and set i=i+1. Go to step 1.

The proof of Belisle, Romeijn and Smith [1990] for the hit-and-run algorithm restricts the support of the probability density function f to be an open bounded set and the density function f to be bounded. However, in many real problems, for example Bayesian posterior estimation, the support of the probability density function is unbounded. In addition, the density function might also be unbounded.

We are interested in developing the theory to remove those restraints, as well as in discussing variance reduction ideas. We generalize the results of Belisle, Romeijn and Smith [1990] in the following directions: (a) f is any probability density function, a) the support of f can be any R^k subset, and c) three variance reduction sampling algorithms are proposed.

The outline of the paper is as follows. In Section 2, we introduce notation and provide a detailed description of the random-direction Monte Carlo sampling approaches that we consider. In Section 3, we introduce several candidates for transition probability kernels. Asymptotic convergence results and their proofs are given in Section 4. In Sections 5, 6, and 7 we consider variance reduction ideas: conditional sampling, multiple-direction stratified sampling, and adaptive external control variates, respectively. Discussion is given in Section 8.

2 Random-Direction Monte Carlo Sampler

We consider a variation of the hit-and-run algorithm: The random directions are specialized to be uniformly distributed; the random distances are generalized (as in Romeijn and Smith [1990]) to no longer require direct sampling from the conditional density. We also generalize to arbitrary density f with any support $S \subset R^k$.

Algorithm 1

```
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 $S_i(\underline{d}_i, \underline{x}_i) \overset{def}{=} \{\lambda \in R \mid \underline{x}_i + \lambda \underline{d}_i \in S\}$.
step 3.  Generate a signed distance $\lambda_i$ from density $g_i(\lambda | \underline{d}_i, \underline{x}_i)$, where $\lambda_i \in S_i$.
step 4.  Set $\underline{y} = \underline{x}_i + \lambda_i \underline{d}_i$. Then set
         $\underline{x}_{i+1} = \begin{cases} \underline{y}, & \text{with the probability } a(\underline{y} \mid \underline{x}_i) \\ \underline{x}_i, & \text{otherwise,} \end{cases} \qquad (2.1)$
step 5.  Set i=i+1, and 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]).

Candidates of $g_i(\lambda | \underline{d}_i, \underline{x}_i)$ and $a(\underline{y} \mid \underline{x}_i)$ that yield the asymptotic distribution f are discussed in the next section.

3 Transition Probability Kernels and Time Reversibility

Let $\{\underline{X}_n, n \geq 0\}$ be the homogeneous Markov chain generated by *Algorithm* 1. The statement of *Algorithm* 1 specifies neither the choice of the density $g_i(\lambda \mid \underline{d}_i, \underline{x}_i)$ nor the probability $a(\underline{y} \mid \underline{x}_i)$. In this section we discuss sufficient conditions on $g_i(\lambda \mid \underline{d}_i, \underline{x}_i)$ and $a(\underline{y} \mid \underline{x}_i)$ for the resulting Markov chain to have a probability transition kernel and to be time reversible.

Theorem 3.1 *For any density $g_i(\lambda \mid \underline{d}_i, \underline{x}_i)$, as used in Step 3 of Algorithm 1, such that*

$$g_i(-\lambda \mid -\underline{d}_i, \underline{x}_i) = g_i(\lambda \mid \underline{d}_i, \underline{x}_i), \tag{3.1}$$

and for any $0 < a(\underline{y} \mid \underline{x}_i) \leq 1$*, as used in Step 4, the Markov chain* $\{\underline{X}_n, n \geq 0\}$ *in Algorithm 1 has one-step transition probability density at* $\underline{X}_{i+1} = \underline{y}$ *given* $\underline{X}_i = \underline{x}$

$$p(\underline{y} \mid \underline{x}) = \frac{2}{C_k \| \underline{x} - \underline{y} \|^{k-1}} \cdot g_i(\| \underline{x} - \underline{y} \| \Big| \frac{\underline{y} - \underline{x}}{\| \underline{y} - \underline{x} \|}, \underline{x}) \cdot a(\underline{y} \mid \underline{x}), \text{ for all } \underline{y} \neq \underline{x} \in S, \tag{3.2}$$

where $C_k = 2\pi^{\frac{k}{2}} \Big/ \Gamma(\frac{k}{2})$ *is the surface area of the k-dimensional unit hypersphere. And at the point* $\underline{X}_{i+1} = \underline{x}$*, the one-step transition probability mass is*

$$p(\underline{x} \mid \underline{x}) = 1 - \int\limits_{S - \{\underline{x}\}} p(\underline{y} \mid \underline{x}) d\underline{y}. \tag{3.3}$$

Furthermore, the transition probability kernel is

$$K(\underline{x}, A) \overset{def}{=} P(\underline{X}_{i+1} \in A \mid \underline{X}_i = \underline{x}) = p(\underline{x} \mid \underline{x}) I_A(\underline{x}) + \int_A p(\underline{y} \mid \underline{x}) d\underline{y}, \tag{3.4}$$

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.*

Proof: See Appendix A. ■

Notice that we are using $p(\cdot \mid \cdot)$ as both a density and as a probability, with the interpretation being clear from the arguments.

Equation (3.2) is intuitively appealing. The numerator "2" arises from two directions passing from $\underline{x}$ through $\underline{y}$. The surface area C_k arises from choosing a unit-length random direction in Step 1. The denominator $\| \underline{x} - \underline{y} \|^{k-1}$ arises from distant points being "harder to hit" than close points. The density $g_i(\lambda \mid \underline{d}_i, \underline{x}_i)$ reflects the choice of candidate points generated in Step 3. The function $a(\underline{y} \mid \underline{x}_i)$ is the probability of moving to the candidate point in Step 4. The proof in Appendix A reflects a variety of intricacies that arise.

To obtain the convergence results of Section 4, we require choices of $g_i(\lambda \mid \underline{d}_i, \underline{x}_i)$ and $a(\underline{y} \mid \underline{x}_i)$ that yield time-reversibility

$$p(\underline{y} \mid \underline{x}) f(\underline{x}) = p(\underline{x} \mid \underline{y}) f(\underline{y}), \text{ for every } \underline{x}, \underline{y} \in S. \tag{3.5}$$

Corollary 3.1 *If* $g_i(\lambda \mid \underline{d}_i, \underline{x}_i)$ *satisfies Equation (3.1),* $g_i(\lambda \mid \underline{d}_i, \underline{x}_i) > 0$ *for* $\lambda \in S_i(\underline{d}_i, \underline{x}_i)$*,* $0 < a(\underline{y} \mid \underline{x}_i) \leq 1$*, and*

$$g_i(\| \underline{x} - \underline{y} \| \Big| \frac{\underline{y} - \underline{x}}{\| \underline{y} - \underline{x} \|}, \underline{x}) \cdot a(\underline{y} \mid \underline{x}) f(\underline{x}) = g_i(\| \underline{y} - \underline{x} \| \Big| \frac{\underline{x} - \underline{y}}{\| \underline{x} - \underline{y} \|}, \underline{y}) \cdot a(\underline{x} \mid \underline{y}) f(\underline{y}), \qquad (3.6)$$

for all $\underline{x} \neq \underline{y} \in S$*, then the Markov chain* $\{\underline{X}_n, n \geq 0\}$ *satisfies Theorem 3.1 and is time reversible.*

Now we consider two candidate sets for $g_i(\lambda \mid \underline{d}_i, \underline{x}_i)$ and $a(\underline{y} \mid \underline{x}_i)$ that satisfy the assumptions of Theorem 3.1 and time reversibility. The first samples from the density f directly, while the second samples indirectly.

Candidate Set I :

$$g_i^I(\lambda \mid \underline{d}_i, \underline{x}_i) = \frac{f(\underline{x}_i + \lambda \underline{d}_i)}{\int_{S_i(\underline{d}_i, \underline{x}_i)} f(\underline{x}_i + u\underline{d}_i) du}, \ for \ \lambda \in S_i(\underline{d}_i, \underline{x}_i), \qquad (3.7)$$

and

$$a^I(\underline{y} \mid \underline{x}_i) = a^I(\underline{x}_i \mid \underline{y}), \quad 0 < a^I(\underline{y} \mid \underline{x}_i) \leq 1, \quad \text{for all } \underline{x}_i, \underline{y} \in S. \qquad (3.8)$$

Typically $a^I(\underline{y} \mid \underline{x}_i) = 1$.

Candidate Set II:

Choose $g_i(\lambda \mid \underline{d}_i, \underline{x}_i)$ to be one of the following:

a) If S is bounded, then

$$g_i^{II}(\lambda \mid \underline{d}_i, \underline{x}_i) = \frac{1}{m(S_i(\underline{d}_i, \underline{x}_i))}, \ for \ \lambda \in S_i(\underline{d}_i, \underline{x}_i), \qquad (3.9)$$

where m is Lebesgue measure.

b) If S is unbounded, then choose $g_i^{II}(\lambda \mid \underline{d}_i, \underline{x}_i)$ to be any symmetric-about-zero, continuous distribution with unbounded support. For example, $g_i^{II}(\lambda \mid \underline{d}_i, \underline{x}_i)$ can be a normal distribution, Cauchy distribution, or double-exponential distribution with location parameter zero.

Independently of the choice a) or b), choose $a(\underline{y} \mid \underline{x}_i)$ to be either

c) Barker's method [Barker, 1965]

$$a^{II}(\underline{y} \mid \underline{x}_i) = \frac{f(\underline{y})}{f(\underline{x}_i) + f(\underline{y})}. \tag{3.10}$$

or

d) Metropolis's method [Metropolis *et al.*, 1953]

$$a^{II}(\underline{y} \mid \underline{x}_i) = \min(1, \frac{f(\underline{y})}{f(\underline{x}_i)}). \tag{3.11}$$

Those choices, which are motivated by Hastings [1970], are also used in Romeijn and Smith [1990].

4 Asymptotic Convergence Results

Let f denote a given *p.d.f.* in R^k with support S, an arbitrary subset of R^k. Let $\{\underline{X}_n, n \geq 0\}$ be the Markov chain in *Algorithm* 1. Then, under the conditions of Corollary 3.1, for any starting point $\underline{x}_0$ in Step 0 of *Algorithm* 1, we prove in Theorem 4.2 that the limiting density of $\underline{X}_n$ is f, and in Theorem 4.3 the average of sample $h(\underline{X}_i)$ converges to the integral $\int_S h(\underline{x})f(\underline{x})d\underline{x}$. We first state and prove Lemmas 4.1, 4.2 and 4.3 and Theorem 4.1.

Now, let $K(\underline{x}, A)$ be the transition probability kernel given in (3.4), and let $p(\underline{y} \mid \underline{x})$ denote the one-step transition probability density and mass, as given in (3.2) and (3.3), respectively. We assume here and in future sections that *Algorithm* 1 is subject to the conditions of Corollary 3.1; then we have

$$p(\underline{y} \mid \underline{x})f(\underline{x}) = p(\underline{x} \mid \underline{y})f(\underline{y}), \text{ and } p(\underline{y} \mid \underline{x}) > 0, \text{ for any } \underline{x}, \underline{y} \in S. \tag{4.1}$$

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

$$\phi(A) \overset{def}{=} \int_A f(\underline{x})d\underline{x}. \tag{4.2}$$

Then $\phi(S) = \int_S f(\underline{x})d\underline{x} = 1$. Thus, we have

Lemma 4.1 *The probability measure ϕ is invariant for the transition probability kernel $K(\underline{x}, A)$, i.e.,*

$$\phi(A) = \int_S K(\underline{x}, A) f(\underline{x}) d\underline{x}, \tag{4.3}$$

for every Borel set $A \subset S$.

Proof: By (3.4), $K(\underline{x}, A) = p(\underline{x} \mid \underline{x}) I_A(\underline{x}) + \int_A p(\underline{y} \mid \underline{x}) d\underline{y}$,
and by (3.3), $p(\underline{x} \mid \underline{x}) = 1 - \int\limits_{S-\{\underline{x}\}} p(\underline{y} \mid \underline{x}) d\underline{y} = 1 - \int_S p(\underline{y} \mid \underline{x}) d\underline{y}$, since $\int_{\{\underline{x}\}} p(\underline{x} \mid \underline{x}) d\underline{y} = 0$.
Thus

$$\int_S K(\underline{x}, A) f(\underline{x}) d\underline{x} = \int_S \left(1 - \int_S p(\underline{y} \mid \underline{x}) d\underline{y}\right) I_A(\underline{x}) f(\underline{x}) d\underline{x} + \int_S \left(\int_A p(\underline{y} \mid \underline{x}) d\underline{y}\right) f(\underline{x}) d\underline{x}$$

$$= \int_S I_A(\underline{x}) f(\underline{x}) d\underline{x} - \int_S \left(\int_S p(\underline{y} \mid \underline{x}) d\underline{y}\right) I_A(\underline{x}) f(\underline{x}) d\underline{x} + \int_S \left(\int_A p(\underline{y} \mid \underline{x}) f(\underline{x}) d\underline{y}\right) d\underline{x}$$

$$= \int_A f(\underline{x}) d\underline{x} - \int_A \left(\int_S p(\underline{y} \mid \underline{x}) f(\underline{x}) d\underline{y}\right) d\underline{x} + \int_S \left(\int_A p(\underline{y} \mid \underline{x}) f(\underline{x}) d\underline{y}\right) d\underline{x}.$$

By (4.1) and Fubini's Theorem, we have

$$\int_A \left(\int_S p(\underline{y} \mid \underline{x}) f(\underline{x}) d\underline{y}\right) d\underline{x} = \int_A \left(\int_S p(\underline{x} \mid \underline{y}) f(\underline{y}) d\underline{y}\right) d\underline{x}$$

$$= \int_S \left(\int_A p(\underline{x} \mid \underline{y}) f(\underline{y}) d\underline{x}\right) d\underline{y} = \int_S \left(\int_A p(\underline{y} \mid \underline{x}) f(\underline{x}) d\underline{y}\right) d\underline{x}.$$

Thus

$$\int_S K(\underline{x}, A) f(\underline{x}) d\underline{x} = \int_A f(\underline{x}) d\underline{x} = \phi(A).$$

■

Now, if $K_1(\underline{x}, A), K_2(\underline{x}, A)$ are two transition probability kernels, their product $K_1 K_2$ is defined by

$$K_1 K_2(\underline{x}, A) \overset{def}{=} \int_S K_1(\underline{x}, d\underline{y}) K_2(\underline{y}, A)$$

(e.g. see Nummelin [1984, p. 2] or Revuz [1975, p. 11]). Thus $K^j(\underline{x}, A)$ is well-defined, and we have the following result:

Lemma 4.2 *For every integer j and every Borel set $A \subset S$,*

$$\int_S K^j(\underline{x}, A) f(\underline{x}) d\underline{x} = \int_A f(\underline{x}) d\underline{x} = \phi(A). \tag{4.4}$$

Proof: By mathematical induction on j using Lemma 4.1. ■

Since the Markov chain $\{\underline{X}_n, n \geq 0\}$ has the transition probability kernel $K(\underline{x}, A)$, then the j-step transition probability is

$$P_{\underline{x}}(\underline{X}_j \in A) \overset{def}{=} P(\underline{X}_j \in A \mid \underline{X}_0 = \underline{x}) = K^j(\underline{x}, A). \tag{4.5}$$

Thus by Lemma 4.2, we have

$$\int_S P_{\underline{x}}(\underline{X}_j \in A) f(\underline{x}) d\underline{x} = \phi(A). \tag{4.6}$$

Lemma 4.3 *The transition probability kernel $K(\underline{x}, A)$ is ϕ-irreducible, i.e., for every Borel set $A \subset S$, if $\phi(A) > 0$, then*

$$\sum_{j=1}^{\infty} K^j(\underline{x}, A) > 0, \quad \textit{for every} \ \ \underline{x} \in S. \tag{4.7}$$

Proof: Since the transition probability kernel $K^j(\underline{x}, A), j = 1, 2, \cdots$, are nonnegative, we need only show that $K(\underline{x}, A) > 0$. For every $\underline{x} \in S$, $K(\underline{x}, A) = p(\underline{x} \mid \underline{x}) I_A(\underline{x}) + \int_A p(\underline{y} \mid \underline{x}) d\underline{y}$, from Equation (3.4). Since $p(\underline{x} \mid \underline{x}) \geq 0$ and $I_A(\underline{x}) \geq 0$, $K(\underline{x}, A) \geq \int_A p(\underline{y} \mid \underline{x}) d\underline{y}$. For every $\underline{y} \neq \underline{x}$, any choice of g and a satisfying the conditions of Corollary 3.1 yields $p(\underline{y} \mid \underline{x}) > 0$. Therefore, since $\phi(A) > 0$ implies $m(A) > 0$, we have $\int_A p(\underline{y} \mid \underline{x}) d\underline{y} > 0$. ■

The *potential kernel* associated with the transition probability kernel $K(\underline{x}, A)$ is

$$G(\underline{x}, A) \overset{def}{=} \sum_{j=0}^{\infty} K^j(\underline{x}, A), \tag{4.8}$$

where $K^0(\underline{x}, A) = I_A(\underline{x})$. The potential kernel $G(\underline{x}, A)$ is *proper* if S can be written as the union of an increasing sequence of S-subsets $\{S_n\}$ in $\mathcal{B}^k$ such that for every $\underline{x} \in S$, $G(\underline{x}, S_n)$ is finite for every n. Otherwise it is *improper*. See Revuz [1975, p. 42].

Lemma 4.4 *The potential kernel $G(\underline{x}, A)$ is improper.*

Proof: We prove by contradiction. Suppose $G(\underline{x}, A)$ is *proper*. Then by Proposition 1.15 of

Chapter 2 in Revuz [1975], there exists $\{S_n, S_n \in \mathcal{B}^k\}$ such that $\{S_n\}$ increasingly converges to S, and $G(\underline{x}, S_n)$ is finite for every $\underline{x} \in S$. But for every set $S_n \subset S$ and every $\underline{x} \in S$, the potential kernel satisfies

$$G(\underline{x}, S_n) = \sum_{j=0}^{\infty} K^j(\underline{x}, S_n) < \infty,$$

so

$$\lim_{j \to \infty} K^j(\underline{x}, S_n) = 0, \quad \text{for every} \quad \underline{x} \in S.$$

Since $\int_S f(\underline{x})d\underline{x} = \phi(S) = 1$ is finite and since $|K^j(\underline{x}, S_n)f(\underline{x})\, | \leq f(\underline{x})$ for every $\underline{x} \in S$, $j \geq 0$ and $n \geq 1$, the Lebesgue Dominated Convergence Theorem implies

$$\lim_{j \to \infty} \int_S K^j(\underline{x}, S_n)f(\underline{x})d\underline{x} = \int_S \lim_{j \to \infty} K^j(\underline{x}, S_n)f(\underline{x})d\underline{x} = 0, \tag{4.9}$$

for every $\underline{x} \in S$ and every $n = 1, 2, \cdots$.

Since $\{S_n\}$ increasingly converges to S and $\phi(S) = 1$, there exists an n^* such that for every $n \geq n^*$, $\phi(S_n) > 0$. Then from Lemma 4.2, for every $n \geq n^*$

$$\lim_{j \to \infty} \int_S K^j(\underline{x}, S_n)f(\underline{x})d\underline{x} = \lim_{j \to \infty} \int_{S_n} f(\underline{x})d\underline{x} = \phi(S_n) > 0, \tag{4.10}$$

which is a contradiction to Equation (4.9). ■

A Markov chain $\{\underline{X}_n, n \geq 0\}$ is said to be *Harris recurrent* if there exists a positive, σ-finite, invariant measure μ such that for every $A \in \mathcal{B}^k$, $\mu(A) > 0$ implies

$$P_{\underline{x}} \left[\sum_{j=1}^{\infty} I_A(\underline{X}_j) = \infty \right] = 1, \tag{4.11}$$

for every $\underline{x} \in S$ [Revuz, 1975, p. 75].

Theorem 4.1 *The Markov chain $\{\underline{X}_n, n \geq 0\}$ in Algorithm 1 with the transition probability kernel $K(\underline{x}, A)$ is Harris recurrent.*

Proof: According to Theorems 2.6 and 2.7 of Chapter 3 from Revuz [1975], we have that if the Markov chain $\{\underline{X}_n, n \geq 0\}$ is ϕ-irreducible and its potential kernel is improper, then it is Harris recurrent. But these two conditions are the consequences of Lemmas 4.3 and 4.4, respectively. ■

Now we prove that the n-step distribution of the Markov chain $\{\underline{X}_n, n \geq 0\}$ converges to the limiting distribution f regardless of the starting point $\underline{x}_0$. Let μ_0 be any starting probability measure on $\mathcal{B}^k$ in Step 0 of *Algorithm* 1. Then the n-step probability measure $\mu_0 P^n$ is

$$\mu_0 P^n(A) \overset{def}{=} \int_S K^n(\underline{x}, A)\mu_0(d\underline{x}), \quad \text{for every} A \in \mathcal{B}^k. \tag{4.12}$$

The total variation between $\mu_0 P^n$ and ϕ is

$$\| \mu_0 P^n - \phi \| \overset{def}{=} \sup_{A \in \mathcal{B}^k} (\mu_0 P^n(A) - \phi(A)) - \inf_{A \in \mathcal{B}^k} (\mu_0 P^n(A) - \phi(A)) \tag{4.13}$$

(e.g. see Nummelin [1984, pp. 108-109]).

Theorem 4.2 *The n-step probability measure of the Markov chain $\{\underline{X}_n, n \geq 0\}$ in Algorithm 1 converges to the invariant probability measure ϕ in total variation; that is,*

$$\lim_{n \to \infty} \| \mu_0 P^n - \phi \| = 0. \tag{4.14}$$

Proof: By Theorem 4.1, the Markov chain $\{\underline{X}_n, n \geq 0\}$ is Harris recurrent. Since it is aperiodic, then Theorem 2.8 of Chapter 6 in Revuz [1975] yields the result. ∎

Theorem 4.3 says that the sample mean of the observations $h(\underline{X}_i)$ converges to the value of the integral, regardless of the number of observations ignored to warm-up the Markov chain.

Theorem 4.3 *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 $0 \leq j_0 < \infty$*

$$\lim_{n \to \infty} \frac{1}{n - j_0 + 1} \sum_{j=j_0}^{n} h(\underline{X}_j) = E_f(h) \quad a.s., \tag{4.15}$$

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

Proof: Theorem 4.1 and Lemma 4.1 say that the Markov chain $\{\underline{X}_n, n \geq 0\}$ is Harris recurrent with invariant probability measure ϕ. Thus, for every h integrable with respect to f, Theorem 3.6 of Chapter 4 in Revuz [1975] gives

$$\lim_{n \to \infty} \frac{1}{n+1} \sum_{j=0}^{n} h(\underline{X}_j) = E_f(h) \quad a.s. \tag{4.16}$$

Therefore,

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n-j_0+1}\sum_{j=j_0}^{n} h(\underline{X}_j) \\ = &\lim_{n\to\infty}\left(\frac{n+1}{n-j_0+1}\cdot\frac{1}{n+1}\sum_{j=0}^{n}h(\underline{X}_j) - \frac{1}{n-j_0+1}\sum_{j=0}^{j_0-1}h(\underline{X}_j)\right) = E_f(h) \text{ a.s.} \end{aligned}$$

■

5 Conditional-Expectation Sampling

Let f denote a given *p.d.f.* in R^k with the support S, an arbitrary subset of R^k. We want to evaluate the k-dimension integral $E_f(h) = \int_S h(\underline{x})f(\underline{x})d\underline{x}$, where we continue to assume $E_f|h| < \infty$. The idea here is that if we use *Algorithm* 1, at the ith-iteration, we generate $\underline{x}_i$ and $\underline{d}_i$. At that iteration, the original quantity for estimating $E_f(h)$ is $h(\underline{x}_i)$. But now we use $y_{i+1} \overset{def}{=} E\left(h(\underline{X}_{i+1}) \mid \underline{d}_i, \underline{x}_i\right)$. Since $E\left(h(\underline{X}_{i+1}) \mid \underline{d}_i, \underline{x}_i\right)$ is a one-dimensional integral, numerical evaluation of y_{i+1} is reasonable. The expected value of each observation is unchanged since by the Markov property,

$$E(Y_{i+1} \mid \underline{X}_0 = \underline{x}_0) = E\left(E\left(h(\underline{X}_{i+1}) \mid \underline{D}_i, \underline{X}_i\right) \mid \underline{X}_0 = \underline{x}_0\right) = E\left(h(\underline{X}_{i+1}) \mid \underline{X}_0 = \underline{x}_0\right). \qquad (5.1)$$

From *Algorithm* 1, given $\underline{X}_i = \underline{x}_i$ and $\underline{D}_i = \underline{d}_i$, then $\underline{X}_{i+1} = \underline{x}_i + \Lambda\underline{d}_i$, where the random signed distance Λ has density and mass

$$q_i(\lambda \mid \underline{d}_i, \underline{x}_i) = \begin{cases} g_i(\lambda \mid \underline{d}_i, \underline{x}_i)a(\underline{x}_i + \lambda\underline{d}_i \mid \underline{x}_i), & \text{if } \lambda \neq 0, \lambda \in S_i(\underline{d}_i, \underline{x}_i), \\ 1 - \int_{S_i(\underline{d}_i, \underline{x}_i)} g_i(u \mid \underline{d}_i, \underline{x}_i)a(\underline{x}_i + u\underline{d}_i \mid \underline{x}_i)du & \text{if } \lambda = 0. \end{cases} \qquad (5.2)$$

Let $Q_i(\lambda \mid \underline{d}_i, \underline{x}_i)$ denote the *c.d.f.* of Λ.

Algorithm 2

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

step 1. Generate a random unit-length direction $\underline{d}_i$ uniformly in the k-dimension space.

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

step 3. Compute $y_{i+1} = \int_{S_i} h(\underline{x}_i + \lambda \underline{d}_i) dQ_i(\lambda \mid \underline{d}_i, \underline{x}_i)$.

step 4. Generate a candidate λ_i from $g_i(\lambda \mid \underline{d}_i, \underline{x}_i)$, where $\lambda_i \in S_i(\underline{d}_i, \underline{x}_i)$.

step 5. Set $\underline{y} = \underline{x}_i + \lambda_i \underline{d}_i$. Then set

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

step 6. Set i=i+1, and go to step 1.

The candidate sets for $g_i(\lambda \mid \underline{d}_i, \underline{x}_i)$, and $a(\underline{y} \mid \underline{x}_i)$ are the same as for *Algorithm* 1. We continue to assume that the conditions of Corollary 3.1 hold.

Let $Y_1, Y_2, \cdots, Y_n$ be the sample from *Algorithm* 2. Then, we can use

$$\hat{E}_f(h) \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{5.3}$$

as an estimator of $E_f(h) = \int_S h(\underline{x}) f(\underline{x}) d\underline{x}$. *Algorithm* 2 is valid in the following sense.

Theorem 5.1 *For the Markov chain $\{\underline{X}_n, n \geq 0\}$ in Algorithm 2, if $E_f|h\ |< \infty$, then for almost all $\underline{x}_0 \in S$ with respect to the invariant probability measure ϕ,*

$$\lim_{n\to\infty} E(\hat{E}_f(h) \mid \underline{X}_0 = \underline{x}_0) = E_f(h). \tag{5.4}$$

Proof: By Theorem 4.1 and Lemma 4.1, the Markov chain $\{\underline{X}_n, n \geq 0\}$ is Harris recurrent with invariant probability measure ϕ. Since

$$\begin{aligned} E(\hat{E}_f(h) \mid \underline{X}_0 = \underline{x}_0) &= \frac{1}{n}\sum_{i=1}^{n} E(Y_i \mid \underline{X}_0 = \underline{x}_0) \\ = \frac{1}{n}\sum_{i=1}^{n} E\left(E\left(h(\underline{X}_i) \mid \underline{D}_{i-1}, \underline{X}_{i-1}\right) \mid \underline{X}_0 = \underline{x}_0\right) &= \frac{1}{n}\sum_{i=1}^{n} E(h(\underline{X}_i) \mid \underline{X}_0 = \underline{x}_0), \end{aligned}$$

and $\int_S |h(\underline{x})\ | f(\underline{x}) d\underline{x} < \infty$, then the fact that for almost all $\underline{x}_0 \in S$ with respect to the invariant probability measure ϕ,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} E(h(\underline{X}_i) \mid \underline{X}_0 = \underline{x}_0) = E_f(h) \tag{5.5}$$

is a consequence of Theorem 3.5 of Chapter 4 in Revuz [1975]. ■

Therefore $n^{-1}\sum_{i=1}^{n} Y_i$ is an asymptotically unbiased estimator of $E_f(h)$. Theorem 5.2 says that under the nonnegative conditional correlation assumption, the variance of the estimator is reduced.

Theorem 5.2 *If for every $i, j \geq 1$, and for every given $\underline{X}_0 = \underline{x}_0 \in S$, $\underline{D}_0 = \underline{d}_0 \in \partial B \overset{def}{=} \{\underline{d} \in R^k \big| \; \| \underline{d} \| = 1 \; \}$, the conditional covariance of $h(\underline{X}_i)$ and $h(\underline{X}_j)$ is finite and nonnegative, then*

$$Var_{\underline{x}_0}\left(\sum_{i=1}^{n} E(h(\underline{X}_i) \mid \underline{D}_{i-1}, \underline{X}_{i-1})\right) \leq Var_{\underline{x}_0}\left(\sum_{i=1}^{n} h(\underline{X}_i)\right), \quad \textit{for every } \; \underline{x}_0 \in S, \tag{5.6}$$

where

$$Var_{\underline{x}_0}\left(\sum_{i=1}^{n} E(h(\underline{X}_i) \mid \underline{D}_{i-1}, \underline{X}_{i-1})\right) =$$
$$E\left(\left(\sum_{i=1}^{n} E(h(\underline{X}_i) \mid \underline{D}_{i-1}, \underline{X}_{i-1})\right)^2 \mid \underline{X}_0 = \underline{x}_0\right) - \left(\sum_{i=1}^{n} E(h(\underline{X}_i) \mid \underline{X}_0 = \underline{x}_0)\right)^2 \tag{5.7}$$

and

$$Var_{\underline{x}_0}\left(\sum_{i=1}^{n} h(\underline{X}_i)\right) = E\left(\left(\sum_{i=1}^{n} h(\underline{X}_i)\right)^2 \mid \underline{X}_0 = \underline{x}_0\right) - \left(\sum_{i=1}^{n} E(h(\underline{X}_i) \mid \underline{X}_0 = \underline{x}_0)\right)^2, \tag{5.8}$$

for every $\underline{x}_0 \in S$.

Proof: See Appendix B. ■

6 Multiple-Direction Stratified Sampling

In this section, we modify *Algorithm* 2 by considering multiple directions at each iteration. The modified algorithm is as follows:

Algorithm 3

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

step 1. Generate a random unit-length direction $\underline{d}_i$ uniformly in the k-dimension space. Denote $\underline{d}_i^1 = \underline{d}_i$.

step 2. Get any random unit-length direction set $\underline{d}_i^2, \underline{d}_i^3, \cdots, \underline{d}_i^k$ such that $\underline{d}_i^1, \underline{d}_i^2, \cdots, \underline{d}_i^k$ are orthogonal. Let $k^* = k$.

step 2′. Alternatively, use all of k! permutations of the direction components of $\underline{d}_i$. Let $k^* = k!$, and denote this random-direction set by $\underline{d}_i^1, \underline{d}_i^2, \cdots, \underline{d}_i^{k^*}$. (Notice that $k^* = k$ if step 2 is used, and $k^* = k!$ if step 2′ is used.)

step 3. Find the set $S_i^j(\underline{d}_i^j, \underline{x}_i) \overset{def}{=} \{\lambda \in R \mid \underline{x}_i + \lambda \underline{d}_i^j \in S\}$, $j = 1, 2, \cdots, k^*$.

step 4. Compute $y_{i+1}^j = \int_{S_i^j(\underline{d}_i^j, \underline{x}_i)} h(\underline{x}_i + \lambda \underline{d}_i^j) dQ_i^j(\lambda \mid \underline{d}_i^j, \underline{x}_i)$, where $Q_i^j(\lambda \mid \underline{d}_i^j, \underline{x}_i)$ is the *c.d.f.* of the random signed distance Λ_i^j, which for $j = 1, 2, \cdots, k^*$ has density and mass

$$q_i^j(\lambda \mid \underline{d}_i^j, \underline{x}_i) = \begin{cases} g_i^j(\lambda \mid \underline{d}_i^j, \underline{x}_i) a(\underline{x}_i + \lambda \underline{d}_i^j \mid \underline{x}_i) & \text{if } \lambda \neq 0, \lambda \in S_i^j(\underline{d}_i^j, \underline{x}_i) \\ 1 - \int_{S_i^j(\underline{d}_i^j, \underline{x}_i)} g_i^j(\lambda \mid \underline{d}_i^j, \underline{x}_i) a(\underline{x}_i + \lambda \underline{d}_i^j \mid \underline{x}_i) d\lambda & \text{if } \lambda = 0. \end{cases} \tag{6.1}$$

step 5. Generate a candidate λ_i from $g_i^1(\lambda \mid \underline{d}_i^1, \underline{x}_i)$, where $\lambda_i \in S_i^1(\underline{d}_i^1, \underline{x}_i)$.

step 6. Set $\underline{y} = \underline{x}_i + \lambda_i \underline{d}_i^1$. Then set

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

step 7. Set i=i+1, and go to step 1.

Let $\{\underline{X}_1, \underline{X}_2, \cdots, \underline{X}_n\}$ and $\{Y_1^j, Y_2^j, \cdots, Y_n^j, \ j = 1, 2, \cdots, k^*\}$ be the samples from *Algorithm* 3. The Markov chain $\{\underline{X}_n, n \geq 0\}$ from *Algorithm* 3 is the same as that from *Algorithm* 1. Thus, the Markov chain $\{\underline{X}_n, n \geq 0\}$ from *Algorithm* 3 is Harris recurrent with the invariant probability measure ϕ.

Now, we use

$$\hat{E}_f(h) \overset{def}{=} \frac{1}{nk^*} \sum_{i=1}^{n} \sum_{j=1}^{k^*} Y_i^j \tag{6.2}$$

as an estimator of $E_f(h)$. Similar to *Algorithm* 2, *Algorithm* 3 is valid in the following sense.

Theorem 6.1 *Under the conditions of Corollary 3.1, if $E_f|h| < \infty$, then for almost all $\underline{x}_0 \in S$*

with respect to the invariant probability measure ϕ,

$$\lim_{n\to\infty} E(\hat{E}_f(h) \mid \underline{X}_0 = \underline{x}_0) = E_f(h). \tag{6.3}$$

Proof: Since

$$\begin{aligned}
&E(\hat{E}_f(h) \mid \underline{X}_0 = \underline{x}_0) = E(\frac{1}{nk^*}\sum_{i=1}^{n}\sum_{j=1}^{k^*} Y_i^j \mid \underline{X}_0 = \underline{x}_0) \\
&= \frac{1}{nk^*}\sum_{i=1}^{n}\sum_{j=1}^{k^*} E\left(E(h(\underline{X}_i) \mid \underline{D}_i^j, \underline{X}_{i-1}) \mid \underline{X}_0 = \underline{x}_0\right) = \frac{1}{nk^*}\sum_{i=1}^{n}\sum_{j=1}^{k^*} E(h(\underline{X}_i) \mid \underline{X}_0 = \underline{x}_0) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(h(\underline{X}_i) \mid \underline{X}_0 = \underline{x}_0),
\end{aligned}$$

the result follows from Theorem 5.1. ■

In Step 2 (or 2′) we can substitute any other set of k^* directions that maintains the unconditional uniform distribution on ∂B.

7 Adaptive External Control-Variate Sampling

In this section, we modify the *Algorithm* to use external control-variate (e.g. see Bratley, Fox, and Schrage [1987] or Nelson [1990]). Let f_N denote the *p.d.f.* of the normal distribution $N(\underline{\mu}, \Sigma)$ with mean $\underline{\mu}$ and covariance Σ. Furthermore, assume $E_{f_N}(h)$ is available, i.e., $E_{f_N}(h)$ can be easily computed. Let the samples be $\{\underline{X}_i, 1 \le i \le n\}$ from f, and $\{\underline{Z}_i, 1 \le i \le n\}$ from f_N. Then the control-variate estimator of $E_f(h)$ is

$$\hat{\theta}^* \overset{def}{=} \hat{\theta} - \hat{\beta}(\hat{\theta}_N - \theta_N), \tag{7.1}$$

where

$$\hat{\theta} \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} h(\underline{X}_i), \qquad \hat{\theta}_N \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} h(\underline{Z}_i), \tag{7.2}$$

and $\hat{\beta}$ is some function of $\{\underline{X}_i\}$ and $\{\underline{Z}_i\}$.

Inducing positive correlation between $\hat{\theta}$ and $\hat{\theta}_N$ is central to obtain $var(\hat{\theta}^*) < var(\hat{\theta})$. We hope to obtain such positive correlation by obtaining a sample path $\{\underline{X}_i\}$ that is similar to the sample path $\{\underline{Z}_i\}$. Similar sample paths can arise by using the same Uniform $(0, 1)$ random numbers to obtain $\{\underline{X}_i\}$ and $\{\underline{Z}_i\}$, if we are careful to synchronize their use. Typically, in *Algorithm* 1 we might use one random-number stream to generate directions, another to generate g using the inverse transformation, and a third to accept or reject $\underline{y}$ based on $a(\underline{y} \mid \underline{x}_i)$. Synchronization is simplified and the sample paths more similar if $a(\underline{y} \mid \underline{x}_i) = 1$, since then the third stream is unnecessary.

In Sections 7.1, 7.2 and 7.3, we discuss an adaptive external control-variate sampling algorithm for which we first obtain the initialized estimates of β, the normal mean μ and covariance Σ, then present the asymptotical result. In Section 7.4, we discuss an alternative that requires stronger assumptions, but that is better when applicable.

7.1 Initialization

Choose a starting normal distribution $N(\underline{\mu}, \Sigma)$. Let n_0 denote the number of iterations of *Algorithm* 1. Let $\{\underline{X}_i^0, 1 \leq i \leq n_0\}$ be the sample from $f(\underline{x})$, and let $\{\underline{Z}_i^0, 1 \leq i \leq n_0\}$ be the sample from $f_N(\underline{x})$. Let $Y_i^0 = h(\underline{X}_i^0), C_i^0 = h(\underline{Z}_i^0), i = 1, 2, \cdots, n_0$, $\underline{Y}^0 = (Y_1^0, Y_2^0, \cdots, Y_{n_0}^0)'$, and $\underline{C}^0 = (C_1^0, C_2^0, \cdots, C_{n_0}^0)'$. Choose

$$\hat{\beta}^* \stackrel{def}{=} \frac{S_{\underline{C}^0\underline{Y}^0}}{S_{\underline{C}^0\underline{C}^0}}, \tag{7.3}$$

where

$$S_{\underline{C}^0\underline{C}^0} = (n_0 - 1)^{-1}\sum_{i=1}^{n_0}(C_i^0 - \bar{C}^0)^2, \quad S_{\underline{C}^0\underline{Y}^0} = (n_0 - 1)^{-1}\sum_{i=1}^{n_0}(C_i^0 - \bar{C}^0)(Y_i^0 - \bar{Y}^0), \tag{7.4}$$

$$\bar{C}^0 = \frac{1}{n_0}\sum_{i=1}^{n_0} C_i^0, \quad \bar{Y}^0 = \frac{1}{n_0}\sum_{i=1}^{n_0} Y_i^0. \tag{7.5}$$

7.2 Updating Sampling Scheme

Now, we choose

$$\hat{\underline{\mu}} \stackrel{def}{=} \frac{1}{n_0}\sum_{i=1}^{n_0} \underline{X}_i^0, \tag{7.6}$$

$$\hat{\Sigma} \stackrel{def}{=} \frac{1}{n_0^2}\sum_{i=1}^{n_0}\sum_{j=1}^{n_0}(\underline{X}_i^0 - \hat{\underline{\mu}})(\underline{X}_j^0 - \hat{\underline{\mu}})'. \tag{7.7}$$

Let $\hat{f}_N$ be the *p.d.f.* of $N(\underline{\hat{\mu}}, \hat{\Sigma})$, and let

$$\theta_N \overset{def}{=} E_{\hat{f}_N}(h) = \int_{R^k} h(\underline{x})\hat{f}_N(\underline{x})d\underline{x}. \tag{7.8}$$

We assume θ_N is known.

Therefore, for Step 1 in *Algorithm* 1, generate $\underline{Z}_0^1 \sim N(\underline{\hat{\mu}}, \hat{\Sigma})$, and set $\underline{X}_0^1 = \underline{Z}_0^1$, restart *Algorithm* 1. Let n be the number of iterations. Then, we have the samples $\{\underline{X}_i^1, 1 \le i \le n\}$ from f and $\{\underline{Z}_i^1, 1 \le i \le n\}$ from $\hat{f}_N$.

7.3 Estimation and Asymptotic Result

Let

$$\begin{aligned} \hat{\theta}_N &\overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} h(\underline{Z}_i^1), && (7.9)\\ \hat{\theta} &\overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} h(\underline{X}_i^1), && (7.10)\\ \hat{\theta}^* &\overset{def}{=} \hat{\theta} - \hat{\beta}^*(\hat{\theta}_N - \theta_N), && (7.11) \end{aligned}$$

where θ_N is given in (7.8). Hence, we have the following asymptotic result for the estimator $\hat{\theta}^*$.

Theorem 7.1 *If $\int_{R^k} |h(\underline{x})| f(\underline{x}) d\underline{x} < \infty$ and if $\int_{R^k} |h(\underline{x})| f_N(\underline{x}) d\underline{x} < \infty$, then given $\{\underline{X}_i^0, 1 \le i \le n_0\}$ and $\{\underline{Z}_i^0, 1 \le i \le n_0\}$,*

$$\lim_{n\to\infty} \hat{\theta}^* = E_f(h) \quad a.s., \tag{7.12}$$

where $\hat{\theta}^*$ is given in (7.11).

Proof: Similar to the proof of Theorem 4.3, given $\{\underline{X}_i^0, 1 \le i \le n_0\}$ and $\{\underline{Z}_i^0, 1 \le i \le n_0\}$, we have $\hat{\theta}_N \longrightarrow \theta_N$, as $n \to \infty$ *a.s.*, and $\hat{\theta} \longrightarrow E_f(h)$, as $n \to \infty$ *a.s.*, which yield the result. ■

For the above adaptive external control-variate sampling algorithm, we don't require that the second moment $E_f(h^2)$ exists. If $E_f(h^2)$ exists, the following alternative external control-variate sampling algorithm will be better.

7.4 Alternative Estimation and Corresponding Asymptotic Result

Pilot sampling again yields $\hat{f}_N$ and θ_N, where for Step 1 in *Algorithm* 1, generate $\underline{Z}_0^1 \sim N(\underline{\hat{\mu}}, \hat{\Sigma})$, and set $\underline{X}_0^1 = \underline{Z}_0^1$. $\hat{\theta}_N = n^{-1}\sum_{i=1}^n h(\underline{Z}_i^1)$ and $\hat{\theta} = n^{-1}\sum_{i=1}^n h(\underline{X}_i^1)$ are still determined by (7.9) and (7.10), respectively. But, we choose

$$\hat{\theta}^{**} \overset{def}{=} \hat{\theta} - \hat{\beta}^{**}(\hat{\theta}_N - \theta_N), \tag{7.13}$$

where

$$\hat{\beta}^{**} \overset{def}{=} \frac{S_{\underline{C}^1\underline{Y}^1}}{S_{\underline{C}^1\underline{C}^1}}, \tag{7.14}$$

and

$$S_{\underline{C}^1\underline{C}^1} = (n-1)^{-1}\sum_{i=1}^n (C_i^1 - \bar{C}^1)^2, \quad S_{\underline{C}^1\underline{Y}^1} = (n-1)^{-1}\sum_{i=1}^n (C_i^1 - \bar{C}^1)(Y_i^1 - \bar{Y}^1), \tag{7.15}$$

$$\bar{C}^1 = \frac{1}{n}\sum_{i=1}^n C_i^1, \quad \bar{Y}^1 = \frac{1}{n}\sum_{i=1}^n Y_i^1, \text{ and } Y_i^1 = h(\underline{X}_i^1), \quad C_i^1 = h(\underline{Z}_i^1), \quad i = 1, \cdots, n. \tag{7.16}$$

Let

$$\sigma_N^2 \overset{def}{=} E_{\hat{f}_N}(h^2) - \theta_N^2 \quad \text{and} \quad \mu_2^h \overset{def}{=} \int_{R^k} h^2(\underline{x}) f(\underline{x}) d\underline{x}. \tag{7.17}$$

Then, finite variances yield convergence, as stated in Theorem 7.2.

Theorem 7.2 *Given* $\{\underline{X}_i^0, \quad 1 \le i \le n_0\}$, *if* σ_N^2 *and* μ_2^h *are finite, then*

$$\lim_{n\to\infty} \hat{\theta}^{**} = E_f(h) \quad a.s. \tag{7.18}$$

Proof: Similar to the proof of Theorem 7.1, given $\{\underline{X}_i^0, 1 \le i \le n_0\}$, we have $\lim_{n\to\infty} \hat{\theta}_N = \theta_N \quad a.s.$, and $\lim_{n\to\infty} \hat{\theta} = E_f(h) \quad a.s.$ Now, it suffices to prove that $\lim_{n\to\infty} \hat{\beta}^{**}(\hat{\theta}_N - \theta_N) = 0 \quad a.s.$ Since

$$\begin{aligned} \mid \hat{\beta}^{**}(\hat{\theta}_N - \theta_N) \mid &= \mid \frac{\frac{1}{n-1}\sum_{i=1}^n (C_i^1 Y_i^1 - n\bar{C}^1\bar{Y}^1)}{S_{\underline{C}^1\underline{C}^1}} \mid \cdot \mid \hat{\theta}_N - \theta_N \mid \\ &\le \frac{\mid \hat{\theta}_N - \theta_N \mid}{S_{\underline{C}^1\underline{C}^1}} \cdot \frac{1}{2} \cdot \left(\frac{1}{n-1}\sum_{i=1}^n (C_i^1)^2 + \frac{1}{n-1}\sum_{i=1}^n (Y_i^1)^2 + \frac{2n}{n-1} \mid \bar{C}^1 \mid \cdot \mid \bar{Y}^1 \mid \right), \end{aligned}$$

and by Theorem 3.6 of Chapter 4 in Revuz [1975],

$$\frac{1}{n-1}\sum_{i=1}^{n}(C_i^1)^2 \;\xrightarrow{\text{a.s.}}\; \sigma_N^2+\theta_N^2 \;\text{ as }\; n\to\infty,$$
$$\frac{1}{n-1}\sum_{i=1}^{n}(Y_i^1)^2 \;\xrightarrow{\text{a.s.}}\; \mu_2^h \;\text{ as }\; n\to\infty,$$

and

$$S_{\underline{C}^1\underline{C}^1} \xrightarrow{\text{a.s.}} \sigma_N^2, \bar{C}^1 \xrightarrow{\text{a.s.}} \theta_N, \;\text{ and }\; \bar{Y}^1 \xrightarrow{\text{a.s.}} E_f(h) \;\text{ as }\; n\to\infty,$$

then

$$\begin{aligned} 0 \;&\leq\; \underline{\lim}_{n\to\infty} \mid \hat{\beta}^{**}(\hat{\theta}_N-\theta_N) \mid \leq \overline{\lim}_{n\to\infty} \mid \hat{\beta}^{**}(\hat{\theta}_N-\theta_N) \mid \\ &\leq\; \overline{\lim}_{n\to\infty} \frac{\mid \hat{\theta}_N-\theta_N \mid}{S_{\underline{C}^1\underline{C}^1}}\cdot\frac{1}{2}\cdot\left(\frac{1}{n-1}\sum_{i=1}^{n}(C_i^1)^2+\frac{1}{n-1}\sum_{i=1}^{n}(Y_i^1)^2+\frac{2n}{n-1}\mid \bar{C}^1 \mid\cdot\mid \bar{Y}^1 \mid\right) \\ &=\; 0 \;\text{ a.s.} \end{aligned}$$

Consequently, $\hat{\beta}^{**}(\hat{\theta}_N-\theta_N) \xrightarrow{\text{a.s.}} 0$ as $n\to\infty$. So, $\hat{\theta}^{**} \xrightarrow{\text{a.s.}} E_f(h)$ as $n\to\infty$. ■

Notice that if we use $\hat{\theta}^{**}$ for $E_f(h)$, we don't need to obtain the sample $\{\underline{Z}_i^0, 1\leq i\leq n_0\}$. Furthermore, if $\underline{\hat{\mu}}$ is chosen to be the mode of f and $\hat{\Sigma}$ is the minus inverse Hessian at $\underline{\hat{\mu}}$ when applicable, or if $\underline{\hat{\mu}}$ and $\hat{\Sigma}$ are any intuitive guesses, we don't need the sample $\{\underline{X}_i^0, 1\leq i\leq n_0\}$. In this case, the condition " given $\{\underline{X}_i^0,\ 1\leq i\leq n_0\}$ " in Theorem 7.2 is no longer needed.

8 Discussion

The uniform distribution of directions in Step 1 of *Algorithm* 1 can be modified by transformations. For example, suppose the integral is originally posed as $\int_S h(\underline{w})f(\underline{w})d\underline{w}$ and we linearly transform the variables using $\underline{x}=C\underline{w}$. Since in the algorithm $\underline{x}_{i+1}=\underline{x}_i+\lambda\underline{d}_i$, iterations are now

$$C^{-1}\underline{x}_{i+1}=C^{-1}(\underline{x}_i+\lambda\underline{d}_i)$$

or

$$\underline{w}_{i+1}=\underline{w}_i+\lambda C^{-1}\underline{d}.$$

The point $C^{-1}\underline{d}_i$ lies on the surface of the ellipsoid $\underline{w}^T C^T C \underline{w} \leq 1$. The unit direction is

$$\frac{C^{-1}\underline{d}_i}{||C^{-1}\underline{d}_i\,||}.$$

The marginal distribution is proportional to $||C^{-1}\underline{d}_i\,||$, the distance of the point to the origin, since $C^{-1}\underline{d}$ is the projection to the surface of points *in* the ellipsoid, which are uniformly distributed [Devroye, 1986, p. 567].

Thus, unless C scales all components equally and with no rotation, even a simple conversion such as meters to kilometers modifies the sample path and the performance of the algorithm. Intuitively, choosing directions uniformly on the sphere is good when the integrand is close to spherical, which might often arise by choosing units that are comfortable to the analyst. Kaufman and Smith [1991] discuss optimal distributions for directions.

Throughout this paper, the assumed form of the integral to be estimated has been $\int_S h(\underline{x}) f(\underline{x}) d\underline{x}$. Equivalently, one could consider the importance-sampling transformation to $\int_S h^*(\underline{x}) f^*(\underline{x}) d\underline{x}$ where f^* is a density on S and $h^*(\underline{x}) = h(\underline{x}) f(\underline{x}) / f^*(\underline{x})$ for all $\underline{x} \in S$. Such a transformation can be useful in two ways. First, the associated conditional density $f^*(\underline{x} + \lambda \underline{d})$ is more tractable than $f(\underline{x} + \lambda \underline{d})$; for example, when S is bounded choose f^* uniform over S, and when S is unbounded choose f^* to be some multivariate normal density. Second, the variance of the estimator based on $h^*(\underline{x})$ might be less than the variance based on $h(\underline{x})$; the ideal is for $h^*(\underline{x})$ to be constant over all $\underline{x} \in S$. The choice of the form of the integrand can dramatically affect performance for the algorithms in this paper, as well as for many other solution methods. However, the results of this paper do not depend on the particular choice.

In Section 5, we have proved that the algorithm is asymptotically unbiased and that if conditional correlations are nonnegative, the variance of the estimator is reduced. The desired result is that the algorithm converges (almost surely) and moreover that the variance of the estimator is reduced. In our empirical experience and from our intuition, we have no examples of negative conditional correlation.

Our interest in this problem is motivated by the need to determine properties of Bayesian posterior distributions, such as probabilities, means, variances, covariances, and higher-order moments. In this case h reflects the property of interest and f is the posterior density. However, often f is known only up to a multiplicative constant c, since the shape of f depends only on the product of prior density and the likelihood function. An advantage of *Algorithm* 1 is that the value of c is not required. The sampling algorithms discussed "see" f only via g and a, which are a density and a probability, respectively, for any value of c.

Empirical results on a variety of examples associated with posterior distributions are consistent with the results of this paper. The results indicate good computational performance. Based on empirical results and informal reasoning, we think that Markov chain sampling can also be used for some problems that are not easily posed as integration. For example, the ith component of observed values $\underline{x}$ might be sorted to estimate quantiles of the ith marginal distribution of f.

Appendix A: Proof of Theorem 3.1

For $k = 1$, the proof of Theorem 3.1 is straightforward, since the distribution of the random direction D_i is $P(D_i = 1) = P(D_i = -1) = \frac{1}{2}$. We consider now $k \geq 2$, where D_i has a density. We first prove Equation (3.2) and then Equation (3.3). For Equation (3.2), we have all $\underline{y} \neq \underline{x} \in S$.

Let $\underline{Y}$ denote the random candidate for $\underline{X}_{i+1}$ given $\underline{X}_i = \underline{x}$ determined in Step 4 of *Algorithm* 1. Let $\underline{D}_i = (D_i^1, D_i^2, \cdots, D_i^k)$ be the random direction of Step 1, which is uniformly distributed over the surface of the unit k-dimensional hypersphere. Thus, by Johnson [1987, pp. 125-127], the components of the direction $\underline{d}_i$ can be written

$$\begin{aligned}
d_i^1 &= \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{k-2} \sin\theta_{k-1} \\
d_i^2 &= \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{k-2} \cos\theta_{k-1} \\
d_i^3 &= \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{k-3} \cos\theta_{k-2} \\
&\vdots \\
d_i^{k-2} &= \sin\theta_1 \sin\theta_2 \cos\theta_3 \\
d_i^{k-1} &= \sin\theta_1 \cos\theta_2 \\
d_i^k &= \cos\theta_1,
\end{aligned} \tag{A.1}$$

where the random signed distance angles $(\Theta_1, \Theta_2, \cdots, \Theta_{k-1})$ have the distribution with density function

$$\begin{aligned}
f_\Theta(\theta_1, \theta_2, \cdots, \theta_{k-1}) = & \left(\frac{2\pi^{\frac{k}{2}}}{\Gamma(\frac{k}{2})}\right)^{-1} \sin^{k-2}\theta_1 \sin^{k-3}\theta_2 \cdots \sin\theta_{k-2}, \\
& 0 \leq \theta_j \leq \pi,\ j = 1, \cdots, k-2;\ 0 \leq \theta_{k-1} < 2\pi.
\end{aligned} \tag{A.2}$$

Thus, the joint probability density function of $(\Theta_1, \Theta_2, \cdots, \Theta_{k-1})$ and the random signed distance Λ_i is

$$f_{\Theta,\Lambda_i}(\theta_1, \theta_2, \cdots, \theta_{k-1}, \lambda) = f_\Theta(\theta_1, \theta_2, \cdots, \theta_{k-1})\; g_i(\lambda \mid \underline{d}_i(\theta_1, \theta_2, \cdots, \theta_{k-1}), \underline{x}), \tag{A.3}$$

where $\underline{d}_i(\theta_1, \theta_2, \cdots, \theta_{k-1}) \overset{def}{=} \underline{d}_i$ is used for clarity. Now, the random candidate $\underline{y} = (y_1, y_2, \cdots, y_k)$ is obtained by $\underline{y} = \underline{x} + \lambda \underline{d}_i$ in Step 4.

So

$$\begin{aligned}
y_1 &= x_1 + \lambda \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{k-2} \sin\theta_{k-1} \\
y_2 &= x_2 + \lambda \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{k-2} \cos\theta_{k-1} \\
y_3 &= x_3 + \lambda \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{k-3} \cos\theta_{k-2} \\
&\vdots \\
y_{k-2} &= x_{k-2} + \lambda \sin\theta_1 \sin\theta_2 \cos\theta_3 \\
y_{k-1} &= x_{k-1} + \lambda \sin\theta_1 \cos\theta_2 \\
y_k &= x_k + \lambda \cos\theta_1 .
\end{aligned} \tag{A.4}$$

Notice that if $\underline{y} = \underline{x} + \lambda \underline{d}_i(\theta_1, \theta_2, \cdots, \theta_{k-1})$, then $\underline{y} = \underline{x} + (-\lambda)\cdot(-\underline{d}_i(\theta_1, \theta_2, \cdots, \theta_{k-1}))$, and recall that the support of $f_\Theta(\theta_1, \theta_2, \cdots, \theta_{k-1})$ is $0 \le \theta_j \le \pi,\ j = 1, \cdots, k-2;\ 0 \le \theta_{k-1} < 2\pi$. Thus, for given $\underline{y}$ and $\underline{x}$, if $(\theta_1, \theta_2, \cdots, \theta_{k-1}, \lambda)$ is a solution of Equation (A.4), then $(\pi - \theta_1, \pi - \theta_2, \cdots, \pi - \theta_{k-2}, \theta_{k-1} \pm \pi, -\lambda)$ are also solutions of Equation (A.4). Therefore, the *p.d.f.* of the random candidate $\underline{Y}$ given $\underline{X}_i = \underline{x}$ is

$$\begin{aligned}
f_Y(\underline{y}) \;=\;& \left| \frac{\partial(y_1, y_2, \cdots, y_k)}{\partial(\theta_1, \theta_2, \cdots, \theta_{k-1}, \lambda)} \right|^{-1} f_\Theta(\theta_1, \theta_2, \cdots, \theta_{k-1}) g_i(\lambda \mid \underline{d}_i(\theta_1, \theta_2, \cdots, \theta_{k-1}), \underline{x}) \\
+& \left| \frac{\partial(y_1, y_2, \cdots, y_k)}{\partial(\pi - \theta_1, \cdots, \pi - \theta_{k-2}, \theta_{k-1} + \pi, -\lambda)} \right|^{-1} \cdot f_\Theta(\pi - \theta_1 \cdots, \pi - \theta_{k-2}, \theta_{k-1} + \pi) \\
& \cdot g_i(-\lambda \mid \underline{d}_i(\pi - \theta_1, \cdots, \pi - \theta_{k-2}, \theta_{k-1} + \pi), \underline{x}) \\
+& \left| \frac{\partial(y_1, y_2, \cdots, y_k)}{\partial(\pi - \theta_1, \cdots, \pi - \theta_{k-2}, \theta_{k-1} - \pi, -\lambda)} \right|^{-1} \cdot f_\Theta(\pi - \theta_1, \cdots, \pi - \theta_{k-2}, \theta_{k-1} - \pi) \\
& \cdot g_i(-\lambda \mid \underline{d}_i(\pi - \theta_1, \cdots, \pi - \theta_{k-2}, \theta_{k-1} - \pi), \underline{x}).
\end{aligned} \tag{A.5}$$

Similar to Kendall [1961, pp. 15-17], we have

$$\left| \frac{\partial(y_1, y_2, \cdots, y_k)}{\partial(\theta_1, \theta_2, \cdots, \theta_{k-1}, \lambda)} \right| = \left| \lambda \right|^{k-1} \cdot \sin^{k-2}\theta_1 \sin^{k-3}\theta_2 \cdots \sin\theta_{k-2}, \tag{A.6}$$

and

$$\begin{aligned}
& \mid \frac{\partial(y_1, y_2, \cdots, y_k)}{\partial(\pi - \theta_1, \cdots, \pi - \theta_{k-2}, \theta_{k-1} \pm \pi, -\lambda)} \mid \\
= & \mid \frac{\partial(y_1, y_2, \cdots, y_k)}{\partial(\theta_1, \theta_2, \cdots, \theta_{k-1}, \lambda)} \mid \cdot \mid \frac{\partial(\theta_1, \theta_2, \cdots, \theta_{k-1}, \lambda)}{\partial(\pi - \theta_1, \cdots, \pi - \theta_{k-2}, \theta_{k-1} \pm \pi, -\lambda)} \mid \\
= & \mid \frac{\partial(y_1, y_2, \cdots, y_k)}{\partial(\theta_1, \theta_2, \cdots, \theta_{k-1}, \lambda)} \mid . \qquad \text{(A.7)}
\end{aligned}$$

Since $d_i(\pi - \theta_1, \cdots, \pi - \theta_{k-2}, \theta_{k-1} \pm \pi) = -d_i(\theta_1, \theta_2, \cdots, \theta_{k-1})$,

$$\begin{aligned}
& g_i(-\lambda \mid \underline{d}_i(\pi - \theta_1, \cdots, \pi - \theta_{k-2}, \theta_{k-1} \pm \pi), \underline{x}) \\
= & g_i(-\lambda \mid -\underline{d}_i(\theta_1, \theta_2, \cdots, \theta_{k-1}), \underline{x}) = g_i(\lambda \mid \underline{d}_i(\theta_1, \theta_2, \cdots, \theta_{k-1}), \underline{x}), \qquad \text{(A.8)}
\end{aligned}$$

where the second equality is by assuming Equation (3.1). And from (A.2),

$$f_\Theta(\pi - \theta_1, \cdots, \pi - \theta_{k-2}, \theta_{k-1} + \pi) = f_\Theta(\theta_1, \theta_2, \cdots, \theta_{k-1}) I_{[0,\pi)}(\theta_{k-1}), \qquad \text{(A.9)}$$

and

$$f_\Theta(\pi - \theta_1, \cdots, \pi - \theta_{k-2}, \theta_{k-1} - \pi) = f_\Theta(\theta_1, \theta_2, \cdots, \theta_{k-1}) I_{[\pi,2\pi)}(\theta_{k-1}). \qquad \text{(A.10)}$$

Therefore, by (A.5) and (A.7) to (A.10), we have

$$\begin{aligned}
f_Y(\underline{y}) = & \mid \frac{\partial(y_1, y_2, \cdots, y_k)}{\partial(\theta_1, \theta_2, \cdots, \theta_{k-1}, \lambda)} \mid^{-1} f_\Theta(\theta_1, \theta_2, \cdots, \theta_{k-1}) \\
& \cdot g_i(\lambda \mid \underline{d}_i(\theta_1, \theta_2, \cdots, \theta_{k-1}), \underline{x})(1 + I_{[0,\pi)}(\theta_{k-1}) + I_{[\pi,2\pi)}(\theta_{k-1})) \\
= & 2 \mid \frac{\partial(y_1, y_2, \cdots, y_k)}{\partial(\theta_1, \theta_2, \cdots, \theta_{k-1}, \lambda)} \mid^{-1} f_\Theta(\theta_1, \cdots, \theta_{k-1})\ g_i(\lambda \mid \underline{d}_i(\theta_1, \cdots, \theta_{k-1}), \underline{x}). \qquad \text{(A.11)}
\end{aligned}$$

Thus, from (A.2) and (A.6),

$$f_Y(\underline{y}) = \frac{2}{\left(\frac{2\pi^{\frac{k}{2}}}{\Gamma(\frac{k}{2})}\right) \mid \lambda \mid^{k-1}} g_i(\lambda \mid \underline{d}_i(\theta_1, \theta_2, \cdots, \theta_{k-1}), \underline{x}). \qquad \text{(A.12)}$$

However, from (A.4) or from Step 4, we can find that $\mid \lambda \mid = \parallel \underline{x} - \underline{y} \parallel$; and if $\lambda = \pm \parallel \underline{x} - \underline{y} \parallel$, then

$$\underline{d}_i(\theta_1, \theta_2, \cdots, \theta_{k-1}) = \pm \frac{\underline{y} - \underline{x}}{\parallel \underline{y} - \underline{x} \parallel}.$$

Hence, the *p.d.f.* of the random candidate $\underline{Y}$ is

$$f_Y(\underline{y}) = \frac{2}{C_k \| \underline{x} - \underline{y} \|^{k-1}} g_i(\| \underline{x} - \underline{y} \| \Big| \frac{\underline{y} - \underline{x}}{\| \underline{y} - \underline{x} \|}, \underline{x}), \quad \text{for } \underline{y} \in S - \{\underline{x}\}, \tag{A.13}$$

where $C_k = \dfrac{2\pi^{\frac{k}{2}}}{\Gamma(\frac{k}{2})}$.

Now having the distribution of the random candidate $\underline{Y}$ given $\underline{X_i} = \underline{x}$, we find the distribution of $\underline{X_{i+1}}$, given $\underline{X_i} = \underline{x}$. Use the indicator variable W to express $\underline{X_{i+1}}$ as

$$\underline{X_{i+1}} = \underline{x} + (\underline{Y} - \underline{x})W, \tag{A.14}$$

where W has the conditional distribution

$$P\left(W = 1 \mid \underline{X_i} = \underline{x}, \underline{Y} = \underline{y}\right) = a(\underline{y} \mid \underline{x}), \tag{A.15}$$

$$P\left(W = 0 \mid \underline{X_i} = \underline{x}, \underline{Y} = \underline{y}\right) = 1 - a(\underline{y} \mid \underline{x}). \tag{A.16}$$

We use the vector inequality $(x_1^1, x_1^2, \cdots, x_1^k) \leq (x_2^1, x_2^2, \cdots, x_2^k)$ if and only if $x_1^j \leq x_2^j$, for all $j = 1, \cdots, k$. Then

$$P\left(\underline{X_{i+1}} \leq \underline{t} \mid \underline{X_i} = \underline{x}\right) = \int\limits_{S-\{\underline{x}\}} P\left(\underline{x} + (\underline{y} - \underline{x})W \leq \underline{t} \mid \underline{X_i} = \underline{x}, \underline{Y} = \underline{y}\right) f_Y(\underline{y}) d\underline{y}. \tag{A.17}$$

Thus, by considering the two cases $W = 1$ and $W = 0$ of Equations (A.15) and (A.16),

$$P\left(\underline{X_{i+1}} \leq \underline{t} \mid \underline{X_i} = \underline{x}\right) = \int\limits_{\{\underline{y} \leq \underline{t}, \underline{y} \in S - \{\underline{x}\}\}} f_Y(\underline{y}) a(\underline{y} \mid \underline{x}) d\underline{y} + \left(\int\limits_{S-\{\underline{x}\}} f_Y(\underline{y})(1 - a(\underline{y} \mid \underline{x})) d\underline{y} \right) I_{\{\underline{x} \leq \underline{t}\}}, \tag{A.18}$$

where $I_{\{\underline{x} \leq \underline{t}\}} = 1$ if $\underline{x} \leq \underline{t}$ and 0 otherwise.

Therefore the distribution of $\underline{X_{i+1}}$ has density

$$p(\underline{y} \mid \underline{x}) = f_Y(\underline{y}) a(\underline{y} \mid \underline{x}), \quad \text{for } \underline{y} \neq \underline{x}, \tag{A.19}$$

and the mass at the point $\underline{X}_{i+1} = \underline{x}$ is

$$\begin{aligned} p(\underline{x} \mid \underline{x}) &= P(\underline{X}_{i+1} = \underline{x} \mid \underline{X}_i = \underline{x}) \;=\; 1 - P(\underline{X}_{i+1} \in S - \{\underline{x}\} \mid \underline{X}_i = \underline{x}) \\ &= 1 - \int\limits_{S-\{\underline{x}\}} p(\underline{y} \mid \underline{x}) d\underline{y}. \end{aligned} \tag{A.20}$$

Thus, the transition probability kernel is

$$K(\underline{x}, A) = P(\underline{X}_{i+1} \in A \mid \underline{X}_i = \underline{x}) = p(\underline{x} \mid \underline{x}) I_A(\underline{x}) + \int_A p(\underline{y} \mid \underline{x}) d\underline{y}, \tag{A.21}$$

for any Borel set $A \subset S$. ■

Appendix B: Proof of Theorem 5.2

By Equations (5.7) and (5.8), in order to prove Inequality (5.6) it suffices to prove

$$E\left((\sum_{i=1}^{n} E(h(\underline{X}_i) \mid \underline{D}_{i-1}, \underline{X}_{i-1}))^2 \mid \underline{X}_0 = \underline{x}_0 \right) \leq E\left((\sum_{i=1}^{n} h(\underline{X}_i))^2 \mid \underline{X}_0 = \underline{x}_0 \right), \tag{B.1}$$

for every $\underline{x}_0 \in S$. And by the total-probability argument, it is sufficient to prove

$$E\left((\sum_{i=1}^{n} E(h(\underline{X}_i) \mid \underline{D}_{i-1}, \underline{X}_{i-1}))^2 \mid \underline{D}_0 = \underline{d}_0, \underline{X}_0 = \underline{x}_0 \right) \leq E\left((\sum_{i=1}^{n} h(\underline{X}_i))^2 \mid \underline{D}_0 = \underline{d}_0, \underline{X}_0 = \underline{x}_0 \right), \tag{B.2}$$

for every $\underline{x}_0 \in S, \underline{d}_0 \in \partial B$, where $\partial B \overset{def}{=} \{\underline{d} \in R^k \Big| \; \| \underline{d} \| = 1 \}$. To simplify notation, define the difference by

$$\Delta \overset{def}{=} \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_{ij}, \tag{B.3}$$

where

$$\begin{aligned} \Delta_{ij} \;&\overset{def}{=}\; E\left(h(\underline{X}_i) h(\underline{X}_j) \mid \underline{D}_0 = \underline{d}_0, \underline{X}_0 = \underline{x}_0\right) \\ &- \; E\left(E(h(\underline{X}_i) \mid \underline{D}_{i-1}, \underline{X}_{i-1}) E(h(\underline{X}_j) \mid \underline{D}_{j-1}, \underline{X}_{j-1}) \mid \underline{D}_0 = \underline{d}_0, \underline{X}_0 = \underline{x}_0\right). \end{aligned} \tag{B.4}$$

Notice that Inequality (B.2) is equivalent to $\Delta \geq 0$. Now either $i = j$ or $i \neq j$.

(a) For $i = j$, by Jensen's inequality for conditional expected values (e.g. see Billingsley [1986,

p. 470]),

$$\Delta_{ii} = E\left(h^2(\underline{X}_i) \mid \underline{D}_0 = \underline{d}_0, \underline{X}_0 = \underline{x}_0\right) - E\left((E(h(\underline{X}_i) \mid \underline{D}_{i-1}, \underline{X}_{i-1}))^2 \mid \underline{D}_0 = \underline{d}_0, \underline{X}_0 = \underline{x}_0\right) \geq 0. \tag{B.5}$$

(b) For $i > j$, by the Markov property,

$$E\left(h(\underline{X}_i)h(\underline{X}_j) \mid \underline{D}_0 = \underline{d}_0, \underline{X}_0 = \underline{x}_0\right) = E\left(E(h(\underline{X}_i)h(\underline{X}_j) \mid \underline{D}_{j-1}, \underline{X}_{j-1}) \mid \underline{D}_0 = \underline{d}_0, \underline{X}_0 = \underline{x}_0\right), \tag{B.6}$$

and

$$\begin{aligned} & E\left(E(h(\underline{X}_i) \mid \underline{D}_{i-1}, \underline{X}_{i-1})E(h(\underline{X}_j) \mid \underline{D}_{j-1}, \underline{X}_{j-1}) \mid \underline{D}_0 = \underline{d}_0, \underline{X}_0 = \underline{x}_0\right) \\ = \; & E\left(E\left(E(h(\underline{X}_i) \mid \underline{D}_{i-1}, \underline{X}_{i-1})E(h(\underline{X}_j) \mid \underline{D}_{j-1}, \underline{X}_{j-1}) \mid \underline{D}_{j-1}, \underline{X}_{j-1}\right) \mid \underline{D}_0 = \underline{d}_0, \underline{X}_0 = \underline{x}_0\right) \\ = \; & E\left(E(h(\underline{X}_j) \mid \underline{D}_{j-1}, \underline{X}_{j-1})E\left(E(h(\underline{X}_i) \mid \underline{D}_{i-1}, \underline{X}_{i-1}) \mid \underline{D}_{j-1}, \underline{X}_{j-1}\right) \mid \underline{D}_0 = \underline{d}_0, \underline{X}_0 = \underline{x}_0\right) \\ = \; & E\left(E(h(\underline{X}_i) \mid \underline{D}_{j-1}, \underline{X}_{j-1})E(h(\underline{X}_j) \mid \underline{D}_{j-1}, \underline{X}_{j-1}) \mid \underline{D}_0 = \underline{d}_0, \underline{X}_0 = \underline{x}_0\right). \end{aligned} \tag{B.7}$$

From (B.6) and (B.7),

$$\begin{aligned} \Delta_{ij} &= E\left(E(h(\underline{X}_i)h(\underline{X}_j) \mid \underline{D}_{j-1}, \underline{X}_{j-1}) \mid \underline{D}_0 = \underline{d}_0, \underline{X}_0 = \underline{x}_0\right) \\ &-E\left(E(h(\underline{X}_i) \mid \underline{D}_{j-1}, \underline{X}_{j-1})E(h(\underline{X}_j) \mid \underline{D}_{j-1}, \underline{X}_{j-1}) \mid \underline{D}_0 = \underline{d}_0, \underline{X}_0 = \underline{x}_0\right) \\ &= E\left(E(h(\underline{X}_i)h(\underline{X}_j) \mid \underline{D}_{j-1}, \underline{X}_{j-1})\right. \\ &\left.-E(h(\underline{X}_i) \mid \underline{D}_{j-1}, \underline{X}_{j-1})E(h(\underline{X}_j) \mid \underline{D}_{j-1}, \underline{X}_{j-1}) \mid \underline{D}_0 = \underline{d}_0, \underline{X}_0 = \underline{x}_0\right). \end{aligned} \tag{B.8}$$

Thus, by the homogeneity and the nonnegative conditional covariance assumption, $\Delta_{ij} \geq 0$. Since $\Delta_{ji} = \Delta_{ij}$, all terms of Δ are nonnegative. ■

References

[1] Applegate, D., Kannan, R. and Polson, N. (1990), "Random Polynomial Time Algorithms for Sampling from Joint Distributions," Technical Report # 500, Carnegie Mellon University, Department of Statistics .

[2] Barker, A. A. (1965), "Monte Carlo Calculations of the Radial Distribution Functions for a

Protonelectron Plasma," *Aust. J. Phys. 18*, pp. 119-133.

[3] Belisle, Claude J.P., Romeijn, H. Edwin, 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.

[4] Billingsley, Patrick (1986), *Probability and Measure*, Second Edition, John Wiley & Sons, New York.

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

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

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

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

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

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

[11] Kendall, M. G. (1961), *A Course in the Geometry of* n *Dimensions*, Hafner Publishing Company, New York.

[12] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953), "Equations of State Calculations by Fast Computing Machines," *J. Chem. Phys. 21*, pp. 1087-1091.

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

[14] Naylor, J. C. and Smith, A.F.M. (1988), "Econometric Illustrations of Novel Numerical Integration Strategies for Bayesian Inference," *Journal of Economics 38*, pp. 103 - 125.

[15] Naylor, J. C. and Smith, A.F.M. (1982), "Applications of a Method for the Efficient Computation of Posterior Distributions," *Appl. Statist. 31*, pp. 214 - 225.

[16] Nelson, B.L.(1990), "Control Variate Remedies," *Operations Research, 38*, pp. 974 - 992.

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

[18] Revuz, D. (1975), *Markov Chains*, North-Holland, Amsterdam.

[19] 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.

[20] Tanner, M.A. and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association, 82*, pp. 528 - 550.

[21] Tierney, L. and Kadane J. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association, 81*, pp. 82 - 86.