Federal Reserve Bank of Minneapolis
Research Department Staff Report 192

May 1995

Monte Carlo Simulation and Numerical Integration*

John Geweke

Federal Reserve Bank of Minneapolis
and University of Minnesota

ABSTRACT

This is a survey of simulation methods in economics, with a specific focus on integration problems. It describes acceptance methods, importance sampling procedures, and Markov chain Monte Carlo methods for simulation from univariate and multivariate distributions and their application to the approximation of integrals. The exposition gives emphasis to combinations of different approaches and assessment of the accuracy of numerical approximations to integrals and expectations. The survey illustrates these procedures with applications to simulation and integration problems in economics.

*This is a chapter prepared for the *Handbook of Computational Economics*, edited by Hans Amman, David Kendrick, and John Rust, to be published by North-Holland. Comments from John Rust and two anonymous referees, who bear no responsibility for errors or omissions, are gratefully acknowledged. Kathleen Rolfe provided editorial assistance, and Maureen O'Connor created the charts. The work was supported in part by National Science Foundation Grant SES-9210070. The views expressed herein are those of the author and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

1. Introduction

Optimization problems in dynamic, stochastic environments are an increas ingly important part of economic theory and applied economics. Inspired by th e potential returns to richer and more realistic models of a variety of policy problems and the promise of ever-growing computational power, economists have turned more and more to mod els that can be simulated but not solved in closed form. Simulation methods can provide solutions for two related integration problems. One integration problem arises in model s olution, for agents whose expected utilities cannot be expressed as a closed function of sta te and decision variables. The other occurs when the investigator combines sources of u ncertainty about models to draw conclusions about policy.

This chapter concentrates on simulation methods that are both important and useful in the solution of these integration problems. In mathematics there is a l ong-standing use of simulation in the solution of integration problems, notably partial diff erential equations, where the form of the simulation is often suggested by the problem itsel f. The history of simulation methods to solve integration problems in economics is shorter , but these methods are appealing there for the same reason: integration generally i nvolves probability distributions in the integrand, which thereby suggests the simulation me thods to be employed.

This pervasive use of simulation methods in science persists despite the well-known asymptotic advantages of deterministic approaches to integration. This continued use of simulation methods occurs in part because astronomical computing time is often required to realize the promise of deterministic methods. A more important fact is that simulation methods are generally straightforward for the investigator to implement, relying on an understanding of a few principles of simulation and the structure of the problem at hand. By contrast, deterministic methods typically require much larger problem -specific investments in numerical methods. Simulation methods economize the use of that most valuable resource, the investigator's time.

The objective of this chapter is to convey an understanding of principle s for the practical application of simulation in economics, with a specific focus on integration problems. It begins with a discussion of circumstances in which determi nistic methods are preferred to simulation, in Section 2. The next section takes up genera l procedures for simulation from univariate and multivariate distributions, including acc eptance and adaptive methods. The construction and use of independent, identically distribut ed random vectors to solve the multidimensional integration problems that typically arise in economic models is taken up in Section 4, with special attention to combination of diffe rent approaches and

assessment of the accuracy of numerical approximations to the integral. Section 5 discusses some modifications of these methods to produce identically but not independently distributed random vectors, that often greatly reduce approximation erro r in applications in economics. Recently developed Markov chain Monte Carlo methods, which m ake use of samples that are neither independently nor identically distributed, have greatly expanded the scope of integration problems with convenient practical solutions. Thes e procedures are taken up in Section 6. The chapter concludes with some examples of rece nt applications of simulation to integration problems in economics.

2. Deterministic methods of integration

The evaluation of the integral $I = \int_a^b \mathrm{f}(x)dx$ is a problem as old as the calculus itself and is equivalent to solution of the differential equation $dy/dx = \mathrm{f}(x)$ subject to the boundary condition $y(a) = 0$. In well-catalogued instances, analytical solutions are available. (Gradshteyn and Ryzhik, 1965, is a useful standard reference.) The li terature on numerical approaches to each problem is huge, a review of any small part of which could occupy this entire volume. This section focuses on those procedures that provide th e most useful tools in economics and are readily available in commercial software. This mea ns neglecting the classical but dated approaches using equally spaced abscissas, like Newt on-Cotes; a useful overview of these methods is provided by Press *et al.* (1986, Chapter 4), and a more extended discussion may be found in Davis and Rabinowitz (1984, Chapter 2).

2.1 Unidimensional quadrature

The principle underlying most state-of-the-art deterministic evaluations of $I = \int_a^b \mathrm{f}(x)dx$ is Gaussian quadrature. If $\mathrm{f}(x) = \mathrm{p}(x)\mathrm{w}(x)$, where $\mathrm{p}(x)$ is any polynomial of degree $2n-1$ or lower and $w(x)$ is a chosen basis function, then there exist points $x_i \in [a,b]$ and a weight ω_i associated with each point such that

$$\int_a^b \mathrm{f}(x)dx = \int_a^b \mathrm{p}(x)\mathrm{w}(x)dx = \sum_{i=1}^n \omega_i \mathrm{p}(x_i).$$

The points and weights depend only on a, b, and the function $\mathrm{w}(x)$, and if they are known for a=0 and b=1, then it is straightforward to determine their values for any other c hoices of a and b. If $\mathrm{r}(x) = \mathrm{f}(x)/\mathrm{w}(x)$ is not a polynomial of degree $2n-1$ or lower, then

$$\sum_{i=1}^n \omega_i \mathrm{r}(x_i)$$

may be taken as an approximation to $I = \int_a^b \mathrm{f}(x)dx$. If $\mathrm{r}(x)$ is smooth relative to a polynomial of degree $2n-1$, then the approximation should be good. More precisely, one may show that if $\mathrm{r}(x)$ is $2n$-times differentiable, then

$$\int_a^b \mathrm{f}(x)dx - \sum_{i=1}^n \omega_i \mathrm{r}(x_i) = c_n \mathrm{r}^{(2n)}(\xi)$$

for some $\xi \in [a,b]$, where $\{c_n\}$ is a sequence of constants with $\lim_{n\to\infty} c_n = 0$. For example, if $\mathrm{w}(x) = 1$, $a = -1$, $b = +1$, then $c_n = 2^{2n+1}(n!)^4 \big/ \{(2n+1)![2n!]^3\}$ (Judd, 1991, pp. 6-7, 6-8).

This approach can be applied to any subinterval of $[a,b]$ as well. As long as $\mathrm{r}(x)$ is $2n$-times differentiable, one may satisfy prespecified convergence or error criteria through successive bisection. Error criteria are usually specified as the absol ute or relative

difference in the computed approximation to $I = \int_a^b \mathrm{f}(x)dx$ using n-point and m-point quadrature (Golub and Welsch, 1969).

Infinite and semi-infinite intervals can be treated through appropriate transformation of variable to a finite interval (Piessens *et al.*, 1983). Existence and boundedness of $\mathrm{r}^{(2n)}$ depend in part on the choice of basis function $\mathrm{w}(x)$. Some of the most useful are indicated in the following table.

$\mathrm{w}(x)$	Interval	Name
1	(-1,1)	Legendre
$1/\sqrt{1-x^2}$	(-1,1)	Chebyshev first kind
$\sqrt{1-x^2}$	(-1,1)	Chebyshev second kind
$\exp(-x^2)$	$(-\infty,+\infty)$	Hermite
$(1+x)^{\alpha}(1-x)^{\beta}$	(-1,1)	Jacobi
$\exp(-x)x^{\alpha}$	$(0,\infty)$	Generalized Laguerre
$1/\cosh(x)$	$(-\infty,+\infty)$	Hyperbolic cosine

For many purposes Gauss-Legendre rules are adequate, and there is a subs tantial stock of commercially supplied software to evaluate one-dimensional integrals up to specified tolerances. These methods have been adapted to include functions having singularities at identified points in the interval of integration (Piessens, *et al.*, 1983).

2.2 Multidimensional quadrature

Some multidimensional integration problems in fact reduce to an integrat ion in a single variable that must be carried out numerically. For example, all but one dimension may be integrable analytically, or the multidimensional integral may in fact be a product of integrals each in a single variable, perhaps after a suitable change of variable. In such cases quadrature for one-dimensional integrals usually provides a neat solutio n. Such cases are rare in economics and econometrics. If the dimension of the domain of i ntegration is not too high and the integrand is sufficiently smooth, then one-dimensional methods may be extended with practical results. These cases cover a small subset of in tegration problems in economics, but when they arise they deserve attention because quadrature -based methods are then often efficient and easy to use.

The straightforward extension of quadrature methods to higher dimensions shows both its strengths and weaknesses. Following Davis and Rabinowitz (1984, pp . 354-359), suppose that R is an m-point rule of integration over $B \subseteq \Re^r$, leading to the approximation

$$\mathrm{R}(\mathrm{f}) = \sum_{j=1}^{m} \omega_j \, \mathrm{f}(\mathbf{x}_j) \approx \int_B \mathrm{f}(\mathbf{x})d\mathbf{x}, \, \mathbf{x}_j \in B,$$

and that S is an n-point rule over $G \subseteq \Re^s$, leading to the approximation

$$S(f) = \sum_{k=1}^{n} v_k f(\mathbf{y}_k) \approx \int_G f(\mathbf{y}) d\mathbf{y}, \mathbf{y}_k \in G.$$

The product rule of R and S is the mn-point rule applicable to $B \times G$,

$$R \times S(f) = \sum_{j=1}^{m} \sum_{k=1}^{n} \omega_j v_k f(\mathbf{x}_j, \mathbf{y}_k) \approx \int_{B \times G} f(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}, \mathbf{x}_j \in B, \mathbf{y}_k \in G.$$

If $h(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{k} f_i(\mathbf{x}) g_i(\mathbf{y})$, and if R integrates $f_i(\mathbf{x})$ exactly over B and S integrates $g_i(\mathbf{y})$ exactly over G $(i = 1, \ldots, k)$, then $R \times S$ will integrate $h(\mathbf{x}, \mathbf{y})$ exactly over $B \times G$. The obvious extensions to the product of three or more rules can be made. These extensions can be expected to work well when (a) quadrature is adequate in the lower dimensional marginals of the function at hand, (b) $h(\mathbf{x}, \mathbf{y}) \approx f(\mathbf{x}) g(\mathbf{y})$, and (c) the product mn is small enough that computation time is reasonable. Condition (c) and perhaps (a) are violated when the support of h is concentrated on a set small relative to the Cartesian boundaries for that support, as illustrated in Figure 1(a). A more common occurrence in economics involves violations of (b) and (c): $B \times G = \Re^r \times \Re^s$, but the function is concentrated on a small subset of its support that cannot be expressed as a Cartesian product, as illustrated in Figure 1(b). Whether these difficulties are present or not, the number of function evaluations and products required in any product rule increases geometrically with the number of arguments of the function, a phenomenon sometimes dubbed "the curse of dimensionality."

These constitute the dominant problems for quadrature methods in economics. To a point, one may extend quadrature to higher dimensions using extensions more sophisticated than product rules. These extensions are usually specific to functions of a certain type, and for this reason the literature is large, but reliable software for a problem at hand may be hard to come by. For example, there has been considerable attention to monomials (polynomials for which the highest degree in any one product is bounded), e.g., McNamee and Stenger (1967), Genz and Malik (1983), Davis and Rabinowitz (1984, Section 5.7). Compound, or subregion, methods provide the most widely applied extensions of quadrature to higher dimensions. In these procedures, a finer and finer subdivision of the original integration region is dynamically constructed, with smaller subregions concentrated where the integrand is most irregular. Within each subregion, a local rule with a moderate number of points is used to approximate the integral. If, at a given step, a prespecified global convergence criterion is not satisfied, those regions for which the convergence criterion is farthest from being satisfied are subdivided, and the local rule is applied to the new subdivisions (van Dooren and de Ridder, 1976; Genz and Malik, 1980; Genz, 1991). For these procedures to work successfully, it is important to have a scheme for construction of subregions well suited to the problem at hand, as reconsideration of Figure 1(b) will make clear. For

example, Genz (1993) provides an algorithm that copes well with the is olated peaks in high-dimensional spaces often found in Bayesian multiparameter problems.

These extensions of quadrature are routinely successful for integrals th rough dimension four or five. Beyond four or five, success depends on whether the problem at hand is of a type for which existing subregion methods are well suited. Whereas the application of quadrature to a function of a single variable can be succ essful as a "black box" procedure, problems of dimensions three and four are more likely to require transformations or other analytical work before quadrature can be applie d. There are very few applications of quadrature-based methods to integrals of more than f ive dimensions in the literature.

2.3 Low discrepancy methods

A low discrepancy method defines a deterministic sequence of points $\{\mathbf{x}_j\}_{j=1}^{\infty}$ and a corresponding m-point integration rule $m^{-1}\sum_{j=1}^{m} \mathrm{f}(\mathbf{x}_j) \approx \int_B \mathrm{f}(\mathbf{x})d\mathbf{x}$. Gaussian quadrature organizes the choice of points to evaluate interactions of polynomials w ith basis functions exactly. Low discrepancy methods choose the sequence to minimize the di fference between the number of points in a set and its measure. (The discussion here cl osely follows parts of Niederreiter, 1992, Chapters 2 and 3.)

The canonical problem sets $B = \bar{I}^d$, the d-dimensional hypercube. (This stipulation is less restrictive than it might seem, and we shall return to this point i n an example in Section 4.4.) For arbitrary $S \subseteq B$ define

$$\mathrm{A}\left(S;\{\mathbf{x}_j\}_{j=1}^{m}\right) = \sum_{j=1}^{m} \chi_S(\mathbf{x}_j),$$

where $\chi_S(\mathbf{x})$ is the characteristic function of S, $\chi_S(\mathbf{x}) = 1$ if $\mathbf{x} \in S$ and $\chi_S(\mathbf{x}) = 0$ if $\mathbf{x} \notin S$. Thus $\mathrm{A}\left(S;\{\mathbf{x}_j\}_{j=1}^{m}\right)$ is the counting function that indicates the number of j with $1 \le j \le m$ for which $\mathbf{x}_j \in S$. If $\acute{S}$ is a nonempty family of Lebesgue measurable subsets of $\bar{I}^d$, then the *discrepancy* of the point set $\{\mathbf{x}_j\}_{j=1}^{m}$ is

$$\mathrm{D}_m\left(\acute{S};\{\mathbf{x}_j\}_{j=1}^{m}\right) = \sup_{S \in \acute{S}} \left| \mathrm{A}\left(S;\{\mathbf{x}_j\}_{j=1}^{m}\right)/m \; - \; \lambda_d(S) \right|,$$

where $\lambda_d(\cdot)$ denotes d-dimensional Lebesgue measure. Let $\acute{S}^*$ be the family of all subintervals of $\bar{I}^d$ of the form $\prod_{i=1}^{d}[0,u_i]$. Then the *star discrepancy* of $\{\mathbf{x}_j\}_{j=1}^{m}$ is

$$\mathrm{D}_m^*\left(\{\mathbf{x}_j\}_{j=1}^{m}\right) = \mathrm{D}_m\left(\acute{S}^*;\{\mathbf{x}_j\}_{j=1}^{m}\right).$$

The star discrepancy of $\{\mathbf{x}_j\}_{j=1}^m$ may be used to bound the error of approximation of $\int_{\bar{I}^d} \mathrm{f}(\mathbf{x})d\mathbf{x}$ by $m^{-1}\sum_{j=1}^m \mathrm{f}(\mathbf{x}_j)$. To do so, first define the *variation of* f *on* $\bar{I}^d$ *in the sense of Vitali,*

$$\mathrm{V}^{(d)}(\mathrm{f}) = \int_0^1 \mathrm{L} \int_0^1 \left| \frac{\partial^d \mathrm{f}}{\partial x_1 \mathrm{L}\ \partial x_d} \right| dx_1 \mathrm{K}\ dx_d$$

for functions f for which the individual partial derivatives are continu ous on $\bar{I}^d$. Next, let $\mathrm{V}^{(k)}(\mathrm{f}; i_1, \mathrm{K}, i_k)$ be the variation in the sense of Vitali of the restriction of f to the k-dimensional face $\{(x_1, \mathrm{K}, x_d) \in \bar{I}^d : x_j = 1 \text{ for } j \neq i_1, \mathrm{K}, i_k\}$. The *variation of* f *on* $\bar{I}^d$ *in the sense of Hardy and Krause* is

$$\mathrm{V}(\mathrm{f}) = \sum_{k=1}^d \sum_{1 \le i_1 \le \mathrm{K} \le i_k \le d} \mathrm{V}^{(k)}(\mathrm{f}; i_i, \mathrm{K}, i_k).$$

(See Niederreiter, 1992, Section 2.2, for an extension of this definiti on to functions f that are not d times continuously differentiable.) For any sequence $\{\mathbf{x}_j\}, \mathbf{x}_j \in \bar{I}^d$,

$$\left| m^{-1}\sum_{j=1}^m \mathrm{f}(\mathbf{x}_j) - \int_{\bar{\mathbf{I}}^d} \mathrm{f}(\mathbf{x})d\mathbf{x} \right| \le \mathrm{V}(\mathrm{f})\mathrm{D}_m^*(\mathbf{x}_1, \mathrm{K}, \mathbf{x}_m),$$

the Koksma-Hlawka inequality (Hlawka, 1961; Niederreiter, 1992, Theorem 2.11). The bound is strict (Niederreiter, 1992, Theorem 2.12).

Low discrepancy methods choose sequences $\{\mathbf{x}_j\}$ so as to minimize $\mathrm{D}_m^*\left(\{\mathbf{x}_j\}_{j=1}^m\right)$. Intuitively, the star discrepancy can be kept small by spacing the point s $\mathbf{x}_j$ evenly. A naive grid on $\bar{I}^d$ will achieve this, but requires an impractically large number of points for $d \ge 5$ in the same way as quadrature does. Low discrepancy methods substantial ly extend the range of practical d before succumbing to the curse of dimensionality. To describe two such sequences, begin with the unique base-b expansion of any integer n,

$$n = \sum_{j=0}^{\infty} \mathrm{a}_j(n) b^j,$$

where b is an integer exceeding 1 and $0 \le \mathrm{a}_j(n) < b$. The *radical-inverse function* ϕ_b *in base* b is defined by

$$\phi_b(n) = \sum_{j=0}^{\infty} \mathrm{a}_j(n) b^{-(j+1)}.$$

This function maps the integers $1, \mathrm{K}, m$ into m distinct points in the unit interval, maintaining a regular spacing between the points: if $m = b^k - 1$, k integer, then there are m evenly spaced points beginning with b^{-k} and ending with $1 - b^{-k}$. Let $\{b_j\}$ be a sequence of relatively prime integers all exceeding 1. (For example, $b_1 = 2, b_2 = 3, b_3 = 5, \mathrm{K}$.) The *Halton sequence in bases* b_1, K, b_d is

$$\{\mathbf{x}_j\}_{j=1}^{\infty}, \quad \mathbf{x}_j = \left[\phi_{b_1}(j), \mathrm{K}, \phi_{b_d}(j)\right]'$$

(Halton, 1960). The m-element *Hammersley sequence in bases* $b_1,\mathrm{K}\ ,b_d$ is

$$\left\{\mathbf{x}_j\right\}_{j=1}^{m},\quad \mathbf{x}_j=\left[j/m,\phi_{b_1}(j),\mathrm{K}\ ,\phi_{b_{d-1}}(j)\right]'$$

(Hammersley, 1960). (An even earlier, closely related sequence is th at of Richtmeyer, 1952, 1958, described in Hammersley and Handscomb, 1964.)

It may be shown (Niederreiter, 1992, Theorem 3.6) that for a Halton se quence in the pairwise relatively prime bases $b_1,\mathrm{K}\ ,b_d$,

$$\begin{aligned}\mathrm{D}_m^*\left(\left\{\mathbf{x}_j\right\}_{j=1}^{m}\right)&\le\frac{d}{m}+\frac{1}{m}\prod_{j=1}^{d}\left(\frac{b_{j-1}}{2\log b_j}\log m+\frac{b_j+1}{2}\right)\\&\le\left(\prod_{j=1}^{d}\frac{b_{j-1}}{2\log b_j}\right)m^{-1}(\log m)^d+0\left[m^{-1}(\log m)^{d-1}\right].\end{aligned}\tag{2.3.1}$$

For the corresponding Hammersley sequence, there is the somewhat better bound

$$\begin{aligned}\mathrm{D}_m^*\left(\left\{\mathbf{x}_j\right\}_{j=1}^{m}\right)&\le\frac{d}{m}+\frac{1}{m}\prod_{j=1}^{d-1}\left(\frac{b_{j-1}}{2\log b_j}\log m+\frac{b_j+1}{2}\right)\\&\le\left(\prod_{j=1}^{d-1}\frac{b_{j-1}}{2\log b_j}\right)m^{-1}(\log m)^{d-1}+0\left[m^{-1}(\log m)^{d-2}\right].\end{aligned}\tag{2.3.2}$$

The second inequalities in (2.3.1) and (2.3.2) imply that the optima l bases are the primes themselves, $b_1=2,\ b_2=3,\ b_3=5,\mathrm{K}$.

If the upper bounds in (2.3.1)-(2.3.2) are used to govern accuracy, then the number of function evaluations increases faster than geometrically with dimension, d, because of the presence of the term $\prod_{i=1}^{d}(b_i-1)/2\log b_i$ or $\prod_{i=1}^{d-1}(b_i-1)/2\log b_i$. Table 1 provides the number of evaluations required to assure that $\left|\sum_{j=1}^{m}\mathrm{f}\left(\mathbf{x}_j\right)-\int_{\bar{\mathbf{I}}^d}\mathrm{f}(\mathbf{x})d\mathbf{x}\right|\le c\ \ \left(c=10^{-2}\text{ or }10^{-5}\right)$ for a function for which the Hardy-Krause total variation is d. It also provides the actual number of evaluations required to guarantee an approximation error of c or less for the function $\mathrm{f}(\mathbf{x})=\sum_{j=1}^{d}x_j$. While the upper bound on the number of evaluations required increases faster than exponentially in the dimension d, the actual number required increases not much faster than linearly and is *much* smaller. In general, however, one will not know the value of the actual error of approximation. The difficulty of asses sing this error is a major disadvantage of low discrepancy and other deterministic algorithms for integration.

2.4 Other deterministic methods

In specialized settings integration in high dimensions can be made more tractable. The obvious limiting case is the one in which the entire problem may be solv ed analytically. But there are also classes of problems that cannot be solved analytically, w ith common features

that suggest specific approximations. An example is provided by Tierney and Kadane (1986) for a class of problems arising in Bayesian statistics and econ ometrics:

$$\mathrm{E}_n(\mathrm{g}) = \frac{\int_\Theta \mathrm{g}(\theta)\exp\left[l(\theta)\right]\pi(\theta)d\theta}{\int_\Theta \exp\left[l(\theta)\right]\pi(\theta)d\theta} = \frac{\int_\Theta \exp\left[n\,\mathrm{L}^*(\theta)\right]d\theta}{\int_\Theta \exp\left[n\,\mathrm{L}(\theta)\right]d\theta},$$

where $l(\theta)$ is a log-likelihood function; $\pi(\theta)$ is a prior density kernel; $\mathrm{g}(\theta)$ is a strictly positive function of interest; n is the number of observations entering the log-likelihood function; $\mathrm{L}(\theta) = \left[\log\pi(\theta) + l(\theta)\right]/n$; and $\mathrm{L}^*(\theta) = \left[\log \mathrm{g}(\theta) + \log\pi(\theta) + l(\theta)\right]/n$.

Let $\hat{\theta}$ denote the mode of L, and let $\Sigma = \partial^2 \mathrm{L}\left(\hat{\theta}\right)/\partial\theta\partial\theta'$. Laplace's approximation is

$$\int_\Theta \exp\left[n\,\mathrm{L}(\theta)\right]d\theta \approx \int_\Theta \exp\left[n\,\mathrm{L}\left(\hat{\theta}\right) - \tfrac{1}{2}n\left(\theta-\hat{\theta}\right)'\Sigma\left(\theta-\hat{\theta}\right)\right]d\theta = (2\pi)^{k/2}|\Sigma|^{1/2}\exp\left[n\,\mathrm{L}\left(\hat{\theta}\right)\right].$$

Similarly, if $\hat{\theta}^*$ is the mode of L^* and $\Sigma^* = \partial^2 \mathrm{L}^*\left(\hat{\theta}^*\right)/\partial\theta\partial\theta'$, then

$$\int_\Theta \exp\left[n\,\mathrm{L}^*(\theta)\right]d\theta \approx (2\pi)^{k/2}\left|\Sigma^*\right|^{1/2}\exp\left[n\,\mathrm{L}^*\left(\hat{\theta}^*\right)\right].$$

The error of approximation in each case is $\mathrm{O}\left(n^{-1/2}\right)$, but in the corresponding approximation

$$\hat{\mathrm{E}}_n(\mathrm{g}) = \left(\left|\Sigma^*\right|/|\Sigma|\right)^{1/2}\exp\left\{n\left[\mathrm{L}^*\left(\hat{\theta}^*\right) - \mathrm{L}\left(\hat{\theta}\right)\right]\right\},$$

the leading terms in the numerator and denominator cancel, and the resul ting error of approximation for $\hat{\mathrm{E}}_n(\mathrm{g})$ is $\mathrm{O}\left(n^{-1}\right)$ (Tierney and Kadane, 1986).

The approximate solution provided by this method is a substantial improv ement on previous approximations of this kind, which worked with a single expansi on about $\hat{\theta}$. The method exhibits two attractions shared by most specialized approximation s to integration in higher dimensions. First, it avoids the need for specific adaptive subr egion analysis required for quadrature, if indeed quadrature can be made to work at all . Second, once function-specific code has been written, the computations involve standa rd ascent algorithms to find $\hat{\theta}$ and $\hat{\theta}*$ and are usually extremely fast. This example also shares some limitations of this approach. First, reduction of approximation error t hrough higher order approximation is tedious at best, whereas in quadrature one can increase the number of points or subregions used and in Monte Carlo one can increase the number of iterations. Second, there is no way to evaluate the error of approximation; again, q uadrature and Monte Carlo will provide error estimates. Third, there is possibly time inten sive analytical work required for each problem in forming derivatives for different g as well as different l. And finally, the requirement that g be strictly positive is restrictive. Th e method may be extended to more general functions at the cost of some increase in complexity (T ierney, Kass, and Kadane, 1989).

3. Pseudorandom number generation

The analytical properties of virtually all Monte Carlo methods for numer ical integration, and more generally for simulation, are rooted in the assumption that it is possible to observe sequences of independent random variables, each distributed uniformly on the unit interval. Given this assumption, various methods, described in Section 3.2, may b e used to construct random variables and vectors with more complex distributions. Specific transformations from the uniform distribution on the unit interval to virtually all of t he classical distributions of mathematical statistics have been constructed using these methods. S ome examples are reviewed in Sections 3.3 and 3.4. These distributions, in turn, constit ute building blocks for the solutions of integration and simulation problems described subsequen tly in this chapter.

The assumption that it is possible to observe sequences of independent r andom variables, distributed uniformly or otherwise, constitutes a model or id ealization of what actually occurs. In this regard it plays the same role here with respec t to what follows as does the assumption of randomness in much of economic theory with respec t to the derived implications for optimizing behavior or does the assumption of randomnes s with respect to the development of methods of statistical inference in econometrics. In current methods for pseudorandom number generation, the observed sequences of numbers for wh ich the assumption of an i.i.d. uniform distribution on the unit interval is the model, are in fact deterministic. Since the algorithms that produce these observed sequenc es are known, the properties of the sequences may be studied analytically in a way that ev ents in the real world corresponding to assumptions of randomness in economic models may not. Thus, the adequacy or inadequacy of stochastic independence as a model for these s equences is on a surer footing than is this assumption as a model in economic or econom et ric theory. We begin this section with an overview of current methods of generating seq uences for which the independent uniform assumption should be an adequate model.

3.1 Uniform pseudorandom number generation

Virtually all pseudorandom number generators employed in practice are li near congruential generators and their elaborations. In the linear congruent ial generator a sequence of integers $\{J_i\}$ is determined by the recursion

$$J_i = (aJ_{i-1} + c) \bmod m. \tag{3.1.1}$$

The parameters a, c, and m determine the qualities of the generator. If $c = 0$, the resulting generator is a pure multiplicative congruential generator. For example, the multiplicative generator with $m = 2^{31} - 1 =$ 2147483647 (a prime) and $a =$ 16807, $a =$ 397204094, or $a =$ 950706376 is used in the IMSL scientific library (IMSL, 1994), and th e user may

choose between different values of c as well as set the seed J_0. The sequence $\{J_i\}$ is mapped into the pseudorandom uniform sequence $\{U_i\}$ by the transformation

$$U_i = J_i/m. \tag{3.1.2}$$

If m is prime, the sequence will cycle after producing exactly m distinct values; clearly one can do no better than $m = 2^{31} - 1$ for a sequence of positive integers with 32-bit arithmetic. There are many criteria for evaluating the i.i.d. uniform distribution o n the unit interval as a model for the resulting sequences $\{U_i\}$. Informal but useful discussions are provided by Press *et al.* (1986, pp. 192-194) and Bratley, Fox and Schrage (1987, pp. 216-220) . More technical and detailed evaluations, including discussion of the choice of c, may be found in Coveyou and McPherson (1967), Marsaglia (1972), Knuth (1981), and Fishman and Moore (1982, 1986).

There are many elaborations on pseudorandom number generation that build on the primitive of the linear or multiplicative congruential generator. In th e shuffled generator, a table is initialized with q seeds. The generator is then used in the obvious way to select a table entry pseudorandomly, and J_1 and U_1 are generated as described in the preceding paragraph. Then a new entry is selected pseudorandomly, U_2 is generated from that entry, and so on. If the congruential generator produced i.i.d. uniform random variables, so would the shuffled generator, and shuffled generators extend the upper bound o n cycle length to mq; this option is provided conveniently in IMSL. A shuffled generator de scribed by L'Ecuyer (1986) has cycle length over 10^{19}. However, the analytical properties of the shuffled generator are harder to evaluate. In another elaboration on th e basic approach, one may combine two pseudorandom sequences $\{J_i\}$ and $\{K_i\}$ from the congruential generator to produce a third sequence $\{L_i\}$ that is then mapped into U_i, $U_i = L_i/m$, in one of two ways: (a) Let $L_i = (J_i + K_i) \bmod m$, or (b) use $\{K_i\}$ to randomly shuffle $\{J_i\}$ and then set $\{L_i\}$ to the shuffled sequence. Both of these generators extend cycle length , but subtle issues arise in the combination of sequences. For a discussion of these issues and comparison of properties, consult Wichmann and Hill (1982) or L'Ecu yer (1986) for (a), Marsaglia and Bray (1968) or Knuth (1981, p. 32) for (b).

The add with carry generator (Marsaglia and Zaman, 1991) has a base b, lags r and s ($r > s$), and a seed vector $\mathbf{j}' = (j_1, \ldots, j_r, c)$ with integer elements $j_i: 0 \le j_i < b$ $(i = 1, \ldots, r)$ and carry bit $c = 0$ or 1. The generated sequence is $\mathbf{j}, \mathrm{f}(\mathbf{j}), \mathrm{f}[\mathrm{f}(\mathbf{j})], \ldots$ with

$$\mathrm{f}(j_1, \ldots, j_r, c) = \begin{cases} (j_2, \ldots, j_r, j_{r+1-s} + j_1 + c, 0) \text{ if } j_{r+1-s} + j_1 + c < b \\ (j_2, \ldots, j_r, j_{r+1-s} + j_1 + c - b, 1) \text{ if } j_{r+1-s} + j_1 + c \ge b \end{cases}.$$

With appropriately chosen base b, lags r and s, and seed vector $\mathbf{j}$, the generated sequence has period $b^r + b^s - 2$. Marsaglia and Zaman (1991) discuss appropriate choices of these

values. One example is $b = 2^{32} - 5, r = 43, s = 22,$ and seed vector consisting of any 43 integers in $\left[0, 2^{32} - 6\right]$. The sequence of vectors has a cycle exceeding 10^{414}, and all possible sequences of 43 integers appear within a cycle. (The add with carry generator is one of a family of closely related generators. Marsaglia and Zaman, 199 1, discuss the family.)

Since pseudorandom numbers are in fact deterministic, some consideration must be given to systematic differences between the two. One important quality is the cycle length. Most simulations on personal computers or workstations are unlikely to e xceed the cycle length of 2^{31} of typical good linear congruential generators. But a study carried ou t with vector or parallel processors could well exceed this length, and in such cases the shuffled or add with carry generator should be considered. Another quality is absen ce of serial correlation. This is easily tested but generally is not a problem. Gre enberger (1961) shows that the first order serial correlation coefficient of any linear congru ential generator is bounded above by $a^{-1}\left[1 - (6c/m) + 6(c/m)^2\right] + (a+6)/m$, and Knuth, 1981, p. 84, points out that for nearly all m the serial correlation coefficient is less than $1/\sqrt{m}$.

Evidence of pseudorandomness is usually exhibited in high dimensional sp aces. If one plots successive overlapping sequences of n pseduorandom numbers, then the sequences typically lie in a few hyperplanes of dimension $n-1$ each. For example, in the case of linear congruential generators the number of hyperplanes is no more than $(n!/m)^{1/n}$ (Marsaglia, 1968): e.g., if $m = 2^{31} - 1$, then sequences of length 6 lie on at most 108 distinct hyperplanes. In the add with carry generator, successive overlapping se quences of more than r values lie on hyperplanes with a separating distance is at least $1/\sqrt{3}$ (Tezuka *et al.*, 1993). One can determine the existence of such hyperplanes using the s pectral test first proposed in Coveyou and MacPherson (1967). Accessible descriptions of this test are provided in Knuth (1981) and Bratley, Fox, and Schrage (1987). Most simulation methods employ highly nonlinear transformations of $\{U_i\}$, as we shall see subsequently, so the distribution of sequences on hyperplanes does not carry over. (However , new problems can arise: see the discussion below of the Box and Muller transformation to construct normally distributed random variables.)

A few practical steps will avoid most problems. First, use only uniform pseudorandom number generators that are completely documented with references to the academic literature. Second, questions of execution time, often discussed in the academic literature, are irrelevant in computational economics: subsequent computations using pseudorandom uniform random sequences take much longer than the most elaborate varian ts on linear congruential generators, so that even if execution time for these genera tors could be driven

to zero, there would be no significant improvement in overall execution time. Third, one should ensure that cycle length is substantially greater than the length of the pseudorandom sequence to be generated. Finally, any publicly reported result based i n part on a sequence of pseudorandom numbers should be checked for sensitivity to the choice of generator. This does not imply numerical analysis that takes the investigator far f rom the problem of interest. A key advantage of Monte Carlo methods, to be discussed in Se ction 4, is that measures of accuracy are produced as a by-product based on the assumptio n that successive pseudorandom numbers are independently and identically distributed. Res ults obtained using variants of methods for producing these sequences should agree wit hin these measures of accuracy. For example, computations can be executed with di fferent seeds, with different values of c in (3.1.1), with or without shuffling, or using an add with carry or related generator. This requires only minor changes in code for most so ftware.

3.2 General methods for nonuniform distributions

Throughout this section, x will denote a random variable with cumulative distribution function (c.d.f.) F and support C, and u will denote a random variable with uniform distribution on the unit interval. If x is continuous, its probability density function (p.d.f.) will be denoted by f. We turn first to several general methods for mapp ing u into x.

Inverse c.d.f. Suppose x is continuous, and consequently the inverse c.d.f.

$$\mathrm{F}^{-1}(p) = \{c\colon \mathrm{P}(\mathrm{x} \le \mathrm{c}) = p\}$$

exists. Then x and $\mathrm{F}^{-1}(u)$ have the same distribution: $\mathrm{P}\left[\mathrm{F}^{-1}(u) \le d\right] = \mathrm{P}\left[u \le \mathrm{F}(d)\right] = \mathrm{F}(d)$. Hence pseudorandom drawings $\{x_i\}_{i=1}^{N}$ of x may be constructed as $\mathrm{F}^{-1}(u_i)$, where $\{u_i\}_{i=1}^{N}$ is a sequence of pseudorandom uniform numbers.

A simple example is provided by the exponential distribution with probab ility density $\mathrm{f}(x) = \lambda \exp(-\lambda x),\ x \ge 0$. Correspondingly, $\mathrm{F}(x) = 1 - \exp(-\lambda x),\ \mathrm{F}^{-1}(p) = -\log(1-p)/\lambda$, and consequently, $x = -\log(u)/\lambda$.

The inverse c.d.f. method is very easy to apply if an explicit, closed f orm expression for the inverse c.d.f. is available. Since most inverse c.d.f.'s req uire the evaluation of transcendental functions, the method may be inefficient relative to othe rs. (That is the case in the foregoing example; see von Neumann, 1951, or Forsythe, 1972, for a more efficient alternative.) In some cases, evaluation of the c.d.f. is superficially closed form to the user of a mathematical software library but in fact involves nontrivial numerica l integration of the kind discussed in Section 2. A leading example is provided by the stand ard normal distribution, for which specialized methods can be applied to the comput ation of F^{-1} (Hart

et al., 1968; Strecok, 1968), but for which acceptance and composition method s (discussed below) are more efficient.

Discrete distributions. Suppose that the random variable X takes on a finite number of values, without loss of generality the integers $1,\mathrm{K},n$ and $\mathrm{P}(X=i)=p_i$. The preferred methods will depend (among other things) on the number of draws to be made from the distribution. If only a few draws are to be made (as may be the case w ith the Markov chain Monte Carlo methods discussed in Section 6), then the obvious inverse m apping from the unit interval to the integers $1,\mathrm{K},n$ can be constructed and subsequently used to search for the appropriate integer corresponding to the drawn u. The disadvantage of this method is that the search time can be substantial. If many draws are to be made, then the alias method due to Walker (1974) and refined by Walker (1977) and Kronmal and Pe terson (1979) is more efficient. The basic idea is to draw an integer i from an equiprobable distribution on the first n integers, and choose i with probability r_i and its alias a_i with probability $1-r_i$. If the values of a_i and r_i are chosen correctly, then the resulting choice probabilities are p_i for i $(i=1,\mathrm{K},n)$. Setting up the table of r_i and a_i requires $\mathrm{O}(n)$ time (see Bratley, Fox, and Schrage, 1987, pp. 158-160, for an accessible discussion); whether this overhead is worthwhile depends on the value of n and the number of draws to be made from the discrete distribution. The aliasing algorithm is implemented in many ma thematical software libraries.

Acceptance methods. Suppose that x is continuous with p.d.f. $\mathrm{f}(x)$ and support C. Let g be the p.d.f. of a different continuous random variable z with p.d.f. $\mathrm{g}(z)$ which has a distribution from which it is possible to draw i.i.d. random variables a nd for which

$$\sup_{x\in C}\left[\mathrm{f}(x)/\mathrm{g}(x)\right]=a<\infty.$$

The function g is known as an *envelope* or *majorizing density* of f, and the distribution with p.d.f. g is known as the *source distributio*n. To generate x_i,

(a) Generate u;
(b) Generate z;
(c) If $u>\mathrm{f}(z)/\left[a\,\mathrm{g}(z)\right]$, go to (a);
(d) $x_i=z$.

The unconditional probability of proceeding from step (c) to step (d) in any pass is

$$\int_{-\infty}^{\infty}\left\{\mathrm{f}(z)/\left[a\,\mathrm{g}(z)\right]\right\}\mathrm{g}(z)dz=a^{-1},$$

and the unconditional probability of reaching step (d) with value at m ost c in any pass is

$$\int_{-\infty}^{c}\left\{\mathrm{f}(z)/\left[a\,\mathrm{g}(z)\right]\right\}\mathrm{g}(z)dz=a^{-1}\mathrm{F}(c).$$

Hence the probability that x_i is at most c at step (d) is $\mathrm{F}(c)$.

The principle of acceptance sampling is illustrated in Figure 2. The tw o essentials of applying this procedure are the ability to generate z and the finite upper bound on $\mathrm{f}(x)/\mathrm{g}(x)$. The efficiency of the method depends on the efficiency of generating z and the unconditional probability of acceptance, which is just the inverse of th e upper bound on $\mathrm{f}(x)/\mathrm{g}(x)$. (In this respect, acceptance sampling is closely related to importan ce sampling discussed in Section 4.3.) The great advantage of acceptance sampling is its ability to cope with arbitrary probability density functions as long as the two essentia l conditions are met and efficiency is acceptable for the purposes at hand. Notice that the method will work in exactly the same way if $\mathrm{f}(x)$ is merely the kernel of the p.d.f. of x (i.e., proportional to the p.d.f.) as long as $a = \sup_{x \in C}\left[\mathrm{f}(x)/\mathrm{g}(x)\right]$ (although in this case a^{-1} no longer provides the unconditional acceptance probability). This property can be exploited to advantage to avoid numerical approximation of unknown constants of integration.

Specific examples providing insight into the method may be found in the family of truncated univariate normal distributions. As a first example, consider the standard normal probability distribution truncated to the interval $(0,.5)$:

$$\mathrm{f}(x) = (.19146)^{-1}(2\pi)^{-1/2}\exp\left(-x^2/2\right) = 2.0837\exp\left(-x^2/2\right),\ 0 < x \leq .5.$$

The standard normal distribution itself is a legitimate source distribut ion, but since $\sup_{0<x\leq.5}\left[\mathrm{f}(x)/\mathrm{g}(x)\right] = (.19146)^{-1}$, the efficiency of this method is low. However, for a source distribution uniform on (0, .5], $\sup_{0<x\leq.5}\left[\mathrm{f}(x)/\mathrm{g}(x)\right] = 2.0837/2.0 = 1.0418$: the unconditional probability of acceptance is $(1.0418)^{-1} = .95985$. As a second example, consider the same distribution truncated to the interval $(5, 8]$:

$$\mathrm{f}(x) = \left(2.8665\times10^{-7}\right)^{-1}(2\pi)^{-1/2}\exp\left(-x^2/2\right) = \left(1.3917\times10^{6}\right)\exp\left(-x^2/2\right),\ 5 < x \leq 8.$$

The standard normal fails as a source distribution since the acceptance probability is 2.8665×10^{-7}. A uniform source density yields an acceptance probability of only .064 271. An exponential distribution translated to the truncation point is for ma ny purposes an excellent approximation to a severely truncated normal distribution (Ma rsaglia, 1964; Geweke, 1986), and for the exponential source density, setting the para meter equal to the truncation point is an optimal or near optimal choice (Geweke, 1991). One can readily verify that the acceptance probability for the source density

$$\mathrm{g}(x) = 5\exp\left[-5(x-5)\right],\ 5 < x \leq 8,$$

is .96406.

Optimizing acceptance sampling. Acceptance methods may readily be extended to multivariate distributions. This topic is taken up in detail in Section 4.2. We turn now to

the question of finding an optimal source distribution for a specified p roblem and develop results for the general case of univariate or multivariate distributions .

In general, suppose that it is desired to draw i.i.d. variables from a d istribution with target density kernel $f(\mathbf{x};\theta), \theta \in \Theta$, having support $C(\theta) \subseteq \Re^m$; the parameter vector θ indexes a family of density kernels $f(\cdot)$. Suppose that a family of source distributions with densities $g(\mathbf{x};\alpha), \alpha \in A \subseteq \Re^p$, having support $D(\alpha)$, has been identified, with the property that for all $\theta \in \Theta$, there exists at least one α for which $\sup_{\mathbf{x}\in C(\theta)} f(\mathbf{x};\theta)/g(\mathbf{x};\alpha) < \infty$. To accomplish the goal of i.i.d. sampling from $f(\mathbf{x};\theta)$, draws from $g(\mathbf{x};\alpha)$ are retained with probability $q(\alpha,\theta) f(\mathbf{x};\theta)/g(\mathbf{x};\alpha)$, where

$$q(\alpha,\theta) \equiv \left[\sup_{\mathbf{x}\in C(\theta)} f(\mathbf{x};\theta)/g(\mathbf{x};\alpha)\right]^{-1}.$$

Suppose the family of source densities $g(\cdot;\cdot)$ has been fixed, but not the value of α, and that the objective is to maximize the unconditional probability of accep ting the draw from the source distribution. Just as in the foregoing examples, this uncond itional probability is

$$\int_{D(\alpha)} \left[q(\alpha,\theta) f(\mathbf{x};\theta)/g(\mathbf{x};\alpha)\right] g(\mathbf{x};\alpha) d\mathbf{x} = q(\alpha,\theta).$$

Hence the problem is to determine the saddle point

$$\min_{\alpha\in A}\left\{\max_{\mathbf{x}\in C(\theta)}\left[\log f(\mathbf{x};\theta) - \log g(\mathbf{x};\alpha)\right]\right\}.$$

Given the usual regularity conditions, a necessary condition is that α be part of a solution of the $(m+p)$-equation system

$$\partial\left[\log f(\mathbf{x};\theta) - \log g(\mathbf{x};\alpha)\right]/\partial\mathbf{x} = \mathbf{0}$$

$$\partial \log g(\mathbf{x};\alpha)/\partial\alpha = \mathbf{0}.$$

As an example, consider the target density kernel

$$f(x;T,\eta) = (x/2)^{Tx/2}\left[\Gamma(x/2)\right]^{-T} \exp(-\eta x),$$

which arises as a conditional posterior density kernel for the degrees-o f-freedom parameter in a Student-*t* distribution (Geweke, 1992b, Appendix B). For the exponential family of source densities $g(x;\alpha) = \alpha \exp(-\alpha x)$, the regular necessary conditions are that

$$(T/2)\left[\log(x/2) + 1 - \psi(x/2)\right] + (\alpha - \eta) = 0,$$

$$\alpha^{-1} - x = 0,$$

where $\psi(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$ is the digamma function. The desired value of α is the solution of

$$(T/2)\left[-\log(2\alpha) + 1 - \psi(1/2\alpha)\right] + (\alpha - \eta) = 0,$$

which may be found using standard root-finding algorithms. Acceptance r ates of about .15 are reported in Geweke (1992b).

Adaptive methods. It may be possible to improve upon a source distribution, using information about the target distribution acquired in the sampling proce ss itself. A very useful application of this idea has been made to the problem of sampling from distributions with log-concave probability density functions. It is especially attrac tive when it is costly to evaluate the target density kernel at a point or when known source densi ties are inefficient or nonexistent. The exposition here closely follows Gilks and Wild (1992) , who build on some earlier work by Devroye (1986); see Wild and Gilks (1993) for a published algorithm. An application of this algorithm is discussed in Section 7.1 .

Let $\mathrm{h}(x) = \log \mathrm{f}(x)$. The support D of $\mathrm{f}(x)$ is connected, and $\mathrm{h}(x)$ is differentiable and weakly concave everywhere in D; i.e., $\mathrm{h}'(x)$ is monotonically nonincreasing in x on D.

Suppose that $\mathrm{h}(x)$ and $\mathrm{h}'(x)$ have been evaluated at k points in D, $x_1 \leq \mathbb{K} \leq x_k$, $k \geq 2$. We assume that if D is unbounded below, then $\mathrm{h}'(x_1) > 0$ and that if D is unbounded above, then $\mathrm{h}'(x_k) < 0$. Let the piecewise linear upper hull $\mathrm{u}(x)$ of $\mathrm{h}(x)$ be formed from the tangents to $\mathrm{h}(\cdot)$ at the x_j, as shown in Figure 3. For $j = 1, \mathbb{K}, k-1$ the tangents at x_j and x_{j+1} intersect at

$$w_j = \frac{\mathrm{h}(x_{j+1}) - \mathrm{h}(x_j) - x_{j+1}\,\mathrm{h}'(x_{j+1}) + x_j\,\mathrm{h}'(x_j)}{\mathrm{h}'(x_j) - \mathrm{h}'(x_{j+1})}.$$

Further let w_0 denote the lower bound of D (possibly $-\infty$) and w_k the upper bound of D (possibly $+\infty$). Then

$$\mathrm{u}(x) = \mathrm{h}(x_j) + (x - x_j)\,\mathrm{h}'(x_j), \quad x \in (w_{j-1}, w_j].$$

Similarly the piecewise linear lower hull $\mathrm{l}(x)$ of $\mathrm{h}(x)$ is formed from the chords between the x_j,

$$\mathrm{l}(x) = \frac{(x_{j+1} - x)\,\mathrm{h}(x_j) + (x - x_j)\,\mathrm{h}(x_{j+1})}{x_{j+1} - x_j}, \quad x \in (x_j, x_{j+1}].$$

For subsequent purposes it is useful to extend the definition to include

$$\mathrm{l}(x) = -\infty, \quad x < x_1 \text{ or } x > x_k.$$

At the start of an acceptance/rejection iteration, the function $\exp[u(x)]$ forms a source density kernel, and $\exp[\mathrm{l}(x)]$ is a squeezing density kernel. The iteration begins by drawing a value z from the distribution with kernel density function $\exp[u(x)]$. This may be done in two steps:

(a) Compute $p_j = \mathrm{P}(w_{j-1} < x \leq w_j) = I_j / I$ $(j = 1, \mathbb{K}, k)$, where

$$I_j = \begin{cases} \exp[\mathrm{h}(x_j) - x_j\,\mathrm{h}'(x_j)]\exp[\mathrm{h}'(x_j)(w_j - w_{j-1})]/\mathrm{h}'(x_j) & \text{if } \mathrm{h}'(x_j) \neq 0 \\ \mathrm{h}(x_j)(w_j - w_{j-1}) & \text{if } \mathrm{h}'(x_j) = 0 \end{cases}$$

and $I = \sum_{j=1}^{k} I_j$. Choose an interval $\left(w_{j-1}, w_j\right]$ from this discrete distribution as described above.

(b) Conditional on the choice of interval the source distribution is exponen tial. Draw z from this distribution as previously discussed.

The draw z is accepted or rejected by means of the acceptance sampling algorithm described above, but using the following shortcut. Having drawn u, we know that z will be accepted if $u \le \exp\left[l(z) - u(z)\right]$, and in this case no further computations are required. If $u > \exp\left[l(z) - u(z)\right]$, then evaluate $h(z)$ and $h'(z)$ and accept z if and only if $u \le \exp\left[h(z) - u(z)\right]$. In the latter case add z to the set of points $\left(x_1, \ldots, x_k\right)$, reordering the x_j's, and update $u(\cdot)$ and $l(\cdot)$, unless z is accepted and no more draws from the target distribution are needed. This completes the acceptance iteration.

Notice that this algorithm is more likely to update the source and squee zing densities the more discordant are these functions at a point. As the algorithm pr oceeds, the probability of acceptance of any draw increases toward 1, and the probab ility that an evaluation of h will be required for any draw falls to 0.

Composition algorithms. Formally, composition arises from a p.d.f. representation

$$f(x) = \int_{-\infty}^{\infty} g_y(x)\, dH(y).$$

A random variable Y from distribution H is generated, followed by a random variable X with p.d.f. g_y. This method goes back at least to Marsaglia (1961), who used it to generate normal random variables. It is also the natural method to use for mixtu re distributions. For example, suppose that x is drawn from a $N(0,.1^2)$ distribution with probability .95 and a $N(0,10^2)$ distribution with probability .05. The probability density

$$.95(2\pi)^{-1/2}(.1)^{-1}\exp\left(-x^2/.02\right) + .05(2\pi)^{-1/2}(10)^{-1}\exp\left(-x^2/200\right)$$

is strongly leptokurtic and not well suited to acceptance sampling. But the construction of the random variable in fact corresponds to a composition with

$$P(Y = 0) = .95,\ P(Y = 1) = .05,$$
$$g_{Y=0}(x) = (2\pi)^{-1/2}(.1)^{-1}\exp\left(-x^2/.02\right),\ g_{Y=1}(x) = .05(2\pi)^{-1/2}(10)^{-1}\exp\left(-x^2/200\right).$$

3.3 Selected univariate distributions

In most cases there is associated with each of the classical univariate distributions a substantial literature on the generation of corresponding pseudorandom v ariables. Good mathematical and statistical software libraries have drawn on this liter ature and are widely available. In many cases the most efficient and accurate routines are n ot simply implementations of the constructions that appear in the mathematical sta tistics literature, and

the user is well-advised to take advantage of the capital embodied in go od libraries. The discussion here is limited to illustrating how the techniques discussed in Section 3.2 are used in specific cases. More thorough surveys in the literature are p rovided by Bratley, Fox, and Schrage (1987, pp.164-189) and Devroye (1986). All of the methods discussed here are implemented in good software libraries, which should always be used. This discussion is not intended to form the basis of reliable code.

Binomial distribution. The binomial distribution indicates the probability of k successes in n independent trials if p is the probability of success in any given trial:

$$\mathrm{p}(k) = \binom{n}{k} p^k (1-p)^{(n-k)}.$$

The definition provides a direct method for generating the random variab le k, but is acceptably rapid only if n is small. For small values of np, the inverse c.d.f. method is practical since $\mathrm{p}(k)$ will typically require evaluation for only a few values of k. In all other cases, however, composition algorithms with acceptance methods are more efficient. Examples are given by Ahrens and Dieter (1980) and Kachitvichyanukul (1982).

Univariate normal distributions. Inverse c.d.f methods for the standard normal have already been mentioned. Acceptance sampling methods are not hard to des ign, especially if one exploits the exponential source distribution as first noted by Marsa glia (1964). Related and succeeding work by Marsaglia and Bray (1964); Marsaglia, MacLaren, and Bray (1964); and Kinderman and Ramage (1976) combining acceptance samplin g and composition form the basis for the generation of standard normal variabl es in most software libraries.

Box and Muller (1958) showed that if U_1 and U_2 are mutually independent standard uniform random variables, then

$$X = \cos(2\pi U_1)\sqrt{-2\log U_2}, \quad Y = \sin(2\pi U_1)\sqrt{-2\log U_2}$$

are independent standard normal random variables. (The key to the demo nstration lies in a transformation to polar coordinates.) The combination of this method w ith the linear congruential random number generator produces a pathology, however. If U_i and U_{i+1} are successive realizations of (3.1.1)-(3.1.2), then

$$U_{i+1} = \left[(amU_i + c) \bmod m\right] / m \Rightarrow$$

$$\cos(2\pi U_{i+1}) = \cos\left[2\pi(aU_i + c/m)\right], \quad \sin(2\pi U_{i+1}) = \sin\left[2\pi(aU_i + c/m)\right]$$

and hence

$$X_i = \cos\left[2\pi(aU_i + c/m)\right]\sqrt{-2\log U_i}, \quad Y_i = \sin\left[2\pi(aU_i + c/m)\right]\sqrt{-2\log U_i}.$$

All possible values of (X_i, Y_i) fall on a spiral. As an approximation to a *pair* of independent variables the distribution of (X_i, Y_i) could hardly be worse. However, if one discards Y_i, the sequence $\{X_i\}$ suffers from no known problems of this kind. This is one of the reasons that acceptance sampling and composition rather than the Box-Muller transformation is used in statistical libraries. It illustrates the ris ks involved in seemingly straightforward combinations of distribution theory with pseudorandom un iform variables.

Given a sequence of standard normal random variables $\{z_i\}$, a sequence from the general univariate normal distribution $\mathrm{N}(\mu, \sigma^2)$ can be generated through the familiar transformation $x_i = \mu + \sigma z_i$.

Gamma distributions. The gamma distribution is important in its own right, for included special cases like the chi-squared, and as a building block for other distributions like the beta. The gamma distribution with scale parameter λ and shape parameter a has probability density

$$\mathrm{f}(x) = \lambda \exp(-\lambda x)(\lambda x)^{a-1} / \Gamma(a),\ x \geq 0.$$

In general, random variables from this distribution may be generated eff iciently using composition algorithms and acceptance methods. Fast and accurate method s are complicated but readily available in statistical software libraries. Fo r example, IMSL uses the composition-acceptance methods of Ahrens and Dieter (1974) and Sch meiser and Lal (1980). A few special cases are worth note.

(a) If $a = 1$, then the distribution is exponential with parameter λ and the inverse c.d.f. method discussed above is much more efficient.

(b) If $a = 0.5$, then $x = z^2/2,\ z \sim \mathrm{N}(0, \lambda^2)$.

(c) If $\lambda = 0.5$, then $x \sim \chi^2(\nu), \nu = 2a$. If a is an integer, then x is the sum of a independent exponentially distributed random variables each with paramet er $\lambda = 0.5$. If ν is an odd integer, then x is the sum of $[\nu/2]$ independent exponentially distributed random variables plus the square of an indepen dent standard normal. For integers up to $\nu = 17$, these representations provide the basis for more efficient generation from the chi-squared distribution, b ut for larger integers it is more efficient to use the more general composition-accept ance methods.

3.4 Selected multivariate distributions

Generation of random vectors typically builds upon the ability to genera te univariate random variables. Just how this should be done is not always obvious, h owever, and

sometimes the obvious method is not the most efficient. The examples th at follow are intended only to illustrate this fact. Statistical software libraries s hould be consulted for implementation of these methods.

Multinomial distribution. The multinomial distribution indicates the probability of k_j realizations of outcome j, from m possible outcomes, in n independent trials. If p_j is the probability of outcome j in any given trial, then

$$\mathrm{p}(k_j) = \frac{n!}{\prod_{j=1}^{m} k_j!} \prod_{j=1}^{m} p_j^{k_j}, \quad k_j \geq 0 \text{ and } \sum_{j=1}^{m} k_j = n.$$

The decomposition of this distribution into its full conditionals, $\mathrm{p}(k_1)$, $\mathrm{p}(k_2|k_1)$,K , $\mathrm{p}(k_j|k_1,\mathrm{K}\ ,k_{j-1})$,K , $\mathrm{p}(k_m|k_1,\mathrm{K}\ ,k_{m-1})$, may be used to generate the k_j. We have

$$\mathrm{p}(k_1) = \binom{n}{k_1} p_1^{k_1} (1-p_1)^{(n-k_1)}, \; 0 \leq k_1 \leq n,$$

$$\mathrm{p}(k_j|k_1,\mathrm{K}\ ,k_{j-1}) = \binom{\tilde{n}_j}{k_j} \tilde{p}_j^{k_j} (1-\tilde{p}_j)^{(\tilde{n}_j - k_j)}, \; 0 \leq k_j \leq \tilde{n}_j,$$

$$\text{where } \tilde{n}_j \equiv n - \sum_{i=1}^{j-1} k_i, \; \tilde{p}_j \equiv p_j \Big/ \left(1 - \sum_{i=1}^{j-1} p_i\right).$$

These distributions are all binomial.

Multivariate normal distribution. The generation of a multivariate normal random vector $\underset{m\times 1}{\mathbf{x}}$ from the distribution $\mathrm{N}(\mu, \Sigma)$ is based on the familiar decomposition

$$\mathbf{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_m), \quad \mathbf{x} = \mu + \mathbf{A}\mathbf{z} \text{ with } \mathbf{A}\mathbf{A}' = \Sigma.$$

While any factorization $\mathbf{A}$ of Σ will suffice, it is most efficient to make $\mathbf{A}$ upper or lower triangular so that $m(m+1)/2$ rather than m^2 products are required in the transformation from z to x. The Cholesky decomposition, in which the diagonal elements of the upp er or lower triangular $\mathbf{A}$ are positive, is typically used.

Wishart distribution. If $\underset{m\times 1}{\mathbf{x}_i} \overset{IID}{\sim} \mathrm{N}(\mathbf{0}, \Sigma)$, the distribution of $\mathbf{A} = \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ is Wishart, with p.d.f.

$$\mathrm{f}(\mathbf{A}) = \frac{|\mathbf{A}|^{\frac{1}{2}(n-m)} \exp\left(-\frac{1}{2} \operatorname{tr} \Sigma^{-1}\mathbf{A}\right)}{2^{\frac{1}{2}(n-1)m} \pi^{m(m-1)/4} |\Sigma|^{\frac{1}{2}(n-1)} \prod_{i=1}^{m} \Gamma\left[\frac{1}{2}(n-i)\right]};$$

for brevity, $\mathbf{A} \sim \mathrm{W}(\Sigma, n-1)$. (For obvious reasons this distribution arises frequently in simulations. It is also important in Bayesian inference, where the post erior distribution of the inverse of the variance matrix for a normal population often has thi s form.) Direct

construction of $\mathbf{A}$ through generation of $\{\mathbf{x}_i\}_{i=1}^{n}$ becomes impractical for large n. A more efficient indirect method follows Anderson (1984). Let Σ have lower triangular Choleski decomposition $\Sigma = \mathbf{L}\mathbf{L}'$, and suppose $\mathbf{Q} \sim \mathrm{W}(\mathbf{I}_m, n-1)$. Then $\mathbf{L}\mathbf{Q}\mathbf{L}' \sim \mathrm{W}(\Sigma, n-1)$ (Anderson, 1984, pp. 254-255). Furthermore $\mathbf{Q}$ has representation

$$\mathbf{Q} = \mathbf{U}\mathbf{U}' \qquad u_{ij} = 0 \ (i < j < m)$$

$$u_{ij} \sim \mathrm{N}(0,1) \qquad u_{ii} \sim \chi^2(n-i)$$

$(i = 1, \mathrm{K}\ , m)$, with the u_{ij} mutually independent for $i \geq j$ (Anderson, 1984, p. 247). Even if n is quite small, this indirect construction is much more efficient than the direct construction.

4. Independence Monte Carlo

Building on the ability to produce sequences of vectors that are well de scribed as i.i.d. random variables, we return to the integration problem with particular a ttention to high dimensions. There are two distinct but closely related problems that ar ise in economics and econometrics.

Problem I is to evaluate

$$I = \int_D \mathrm{f}(\mathbf{x}) d\mathbf{x}.$$

Problem E is to evaluate

$$E = \mathrm{E}\big[\mathrm{g}(\mathbf{x})\big],$$

where $\mathbf{x}$ is a random vector with c.d.f. $\mathrm{P}(\mathbf{x})$. To simplify notation, assume that P is absolutely continuous and that $\mathbf{x}$ has a probability density function $\mathrm{p}(\mathbf{x})$. It is implicit in Problem E that $\int_D \mathrm{g}(\mathbf{x})\,\mathrm{p}(\mathbf{x}) d\mathbf{x}$ is absolutely convergent in its domain D. Detailed examples of Problems E and I are provided in Section 7.

If a random vector $\mathbf{z}$ has p.d.f. $\mathrm{p}(\mathbf{z})$, then any function $\mathrm{r}(\mathbf{z}) = a \cdot \mathrm{p}(\mathbf{z})$, $a > 0$, is said to be a kernel density function for $\mathbf{z}$. In order to express some key moments compactly, let $\mathrm{E}_\mathrm{r}\big[\mathrm{g}(\mathbf{z})\big]$ denote the expectation of $\mathrm{g}(\mathbf{z})$ if $\mathbf{z}$ has kernel density function $\mathrm{r}(\mathbf{z})$; similarly $\mathrm{var}_\mathrm{r}\big[\mathrm{g}(\mathbf{z})\big]$ for variance.

Many of the procedures discussed in this section are straightforward app lications of two results in basic mathematical statistics. Let $\{y_i\}$ be an i.i.d. sequence from a population, and let $\bar{y}_N = \frac{1}{N}\sum_{i=1}^N y_i$ and $s_N^2 = \frac{1}{N-1}\sum_{i=1}^N \big(y_i - \bar{y}_N\big)^2$. If the population has finite first moment, then $\mathrm{E}\big(\bar{y}_N\big) = \mathrm{E}(y)$ and the strong law of large numbers states that

$$\bar{y}_N \xrightarrow{a.s.} \mathrm{E}(y);$$

i.e., $\mathrm{P}\big[\lim_{N\to\infty} \bar{y}_N = \mathrm{E}(y)\big] = 1$. If the same population also has a finite variance σ^2, then the central limit theorem establishes that

$$\sqrt{N}\big[\bar{y}_N - \mathrm{E}(\mathrm{y})\big] \xrightarrow{d} \mathrm{N}\big(0, \sigma^2\big);$$

i.e., $\lim_{N\to\infty} \mathrm{P}\Big\{\sqrt{\mathrm{N}}\big[\bar{y}_N - \mathrm{E}(\mathrm{y})\big] \le c\sigma\Big\} = \Phi(c)$, where $\Phi(\cdot)$ is the c.d.f. of the N(0,1) distribution. In this case $\mathrm{E}\big(s_N^2\big) = \sigma^2$, and from the strong law of large numbers,

$$s_N^2 \xrightarrow{a.s.} \sigma^2.$$

4.1 Simple Monte Carlo

In the case of Problem I, suppose that

$$\mathrm{f}(\mathbf{x}) = \mathrm{g}(\mathbf{x})\,\mathrm{p}(\mathbf{x}),$$

with $\mathrm{p}(\mathbf{x}) \geq 0$ and $\int_D \mathrm{p}(\mathbf{x})d\mathbf{x} = p^*$, where p^* is a known positive constant. Then $\mathrm{p}(\mathbf{x})$ is a kernel density function. Suppose further that it is possible to draw ps eudorandom vectors $\{\mathbf{x}_i\}$ from the distribution with probability density function $\mathrm{p}(\mathbf{x})/p^*$, as described in Section 3. Since

$$I = \int_D \mathrm{f}(\mathbf{x})d\mathbf{x} = \int_D p^* \mathrm{g}(\mathbf{x})\left[\mathrm{p}(\mathbf{x})/p^*\right]d\mathbf{x} = \mathrm{E}_p\left[p^* \mathrm{g}(\mathbf{x})\right],$$

it follows that

$$I_N = N^{-1}p^* \sum_{i=1}^N \mathrm{g}(\mathbf{x}_i) \xrightarrow{a.s.} I. \tag{4.1.1}$$

The requirement that p^* is known may be weakened by replacing p^* with a sequence $p_N^* \xrightarrow{a.s.} p^*$ in the last expression. (Some practical methods of producing p_N^* at essentially no incremental cost are taken up in Section 4.2.) If p^* is known, then $\mathrm{E}(I_N) = I$, but if p^* must be replaced by a consistent estimator, then in general $\mathrm{E}(I_N) \neq I$ but (4.1.1) is still true.

If in addition $\int_D \mathrm{g}^2(\mathbf{x})\,\mathrm{p}(\mathbf{x})d\mathbf{x}$ is absolutely convergent, this result can be extended to provide a measure of the accuracy of I_N. Let

$$\sigma^2 = \mathrm{var}_p\left[p^* \mathrm{g}(\mathbf{x})\right] = p^{*-1}\int_D \left[p^* \mathrm{g}(\mathbf{x}) - I\right]^2 \mathrm{p}(\mathbf{x})d\mathbf{x}.$$

Then

$$\sqrt{N}(I_N - I) \xrightarrow{d} \mathrm{N}(0, \sigma^2), \quad N^{-1}\sum_{i=1}^N \left[p^* \mathrm{g}(\mathbf{x}_i) - I_N\right]^2 \xrightarrow{a.s.} \sigma^2.$$

(The result may be extended to include cases in which p^* is approximated by a sequence of p_N^*, but some changes are required; see Section 4.2.) This result makes e xact the intuitive notion that $\mathrm{p}(\cdot)$ should be chosen to mimic the shape of $\mathrm{f}(\cdot)$.

The solution of Problem E by simple Monte Carlo is even simpler, as long as it is possible to construct an i.i.d. sequence from the probability distributi on of $\mathbf{x}$ in $\mathrm{E}[\mathrm{g}(\mathbf{x})]$, for then $E_N = \frac{1}{N}\sum_{i=1}^N \mathrm{g}(\mathbf{x}_i) \xrightarrow{a.s.} E$ and $\mathrm{E}(E_N) = E\,\forall N$. It is not necessary to know the integrating constant of the kernel probability density for $\mathbf{x}$. If $\sigma^2 = \mathrm{var}[\mathrm{g}(\mathbf{x})]$ exists, then $\sqrt{N}(E_N - E) \xrightarrow{d} \mathrm{N}(0, \sigma^2)$ as well.

As an example, consider the problem

$$I = \int_{\Re^k} \mathrm{f}(\mathbf{x})d\mathbf{x} = \int_{\Re^k} \mathrm{g}(\mathbf{x})\,\mathrm{p}(\mathbf{x})d\mathbf{x} = \int_{\Re^k} \mathrm{g}(\mathbf{x}) \exp\left[-\tfrac{1}{2}(\mathbf{x}-\mu)'\,\mathbf{H}(\mathbf{x}-\mu)\right]d\mathbf{x},$$

where $\mathbf{H}$ is positive definite. Since $\mathrm{p}(\mathbf{x})$ is a multivariate normal kernel density function,

$$I_N = (2\pi)^{k/2}|\mathbf{H}|^{1/2} N^{-1}\sum_{i=1}^N \mathrm{g}(\mathbf{x}_i), \quad \mathbf{x}_i \overset{IID}{\sim} \mathrm{N}(\mu, \mathbf{H}^{-1}).$$

Because $\mathrm{p}(\mathbf{x}) \geq 0 \forall \mathbf{x} \in \Re^k$, $I_N \xrightarrow{a.s.} I$ regardless of the form of $\mathrm{f}(\mathbf{x})$. However, convergence will be impractically slow if $\mathrm{g}(\mathbf{x})$ is ill conditioned or (equivalently) μ and $\mathbf{H}$ are chosen so that $\mathrm{p}(\cdot)$ poorly mimics $\mathrm{f}(\cdot)$. If $\mathrm{var}_{\mathrm{p}}[g(\mathbf{x})]$ exists, then

$$\sigma^2 = (2\pi)^k |\mathbf{H}| \mathrm{var}_{\mathrm{p}}[\mathrm{g}(\mathbf{x})]$$

provides the pertinent measure of the adequacy of I_N as an approximation of I. Only this expression -- not the dimensionality k -- matters.

4.2 Acceptance methods

Acceptance methods may be used to evaluate integrals in much the same wa y as they are used to produce pseudorandom numbers. In Problem I, suppose that $0 \le \mathrm{g}(\mathbf{x}) \le a < \infty$ $\forall \mathbf{x} \in D$. Suppose further that p^* is known or equivalently that $\mathrm{p}(\mathbf{x})$ is a probability density function and not merely a kernel. Let $\{\mathbf{x}_i\}$ be an i.i.d. sequence drawn from a distribution function with p.d.f. $\mathrm{p}(\mathbf{x})$, and let u_i be a corresponding Bernoulli random variable,

$$u_i = 0 \text{ or } 1, \ \mathrm{P}(u_i = 1) = a^{-1}\mathrm{g}(\mathbf{x}_i).$$

Then

$$I_N = N^{-1} a \sum_{i=1}^N u_i \xrightarrow{a.s.} a\,\mathrm{E}_{\mathrm{p}}(u_i) = a\int_D a^{-1}\mathrm{g}(\mathbf{x})\,\mathrm{p}(\mathbf{x})d\mathbf{x} = I,$$
$$\mathrm{E}(I_N) = I \,\forall N, \quad \sqrt{N}(I_N - I) \xrightarrow{d} \mathrm{N}(0, \sigma^2),$$
$$\sigma^2 = aI - I^2, \quad aI_N - I_N^2 \xrightarrow{a.s.} \sigma^2. \tag{4.2.1}$$

This method may be extended to $\mathrm{g}(\mathbf{x})$ for which $-\infty < l \le \mathrm{g}(\mathbf{x}) \le u < \infty$, by defining $\mathrm{g}^+(\mathbf{x}) = \sup[0, \mathrm{g}(\mathbf{x})]$, $\mathrm{g}^-(\mathbf{x}) = -\inf[0, \mathrm{g}(\mathbf{x})]$, and approximating $\int_D \mathrm{g}^+(\mathbf{x})d\mathbf{x}$ and $\int_D \mathrm{g}^-(\mathbf{x})d\mathbf{x}$ separately. Observe that σ^2 is an increasing function of a and the unconditional probability of acceptance $\mathrm{P}(u_i = 1) = a^{-1}I$ is a decreasing function of a. If $\mathrm{p}(\mathbf{x}) \propto \mathrm{g}(\mathbf{x})$, then $\mathrm{P}(u_i) = 1$ and $\sigma^2 = 0$, but this is tantamount to being able to integrate $\mathrm{f}(\mathbf{x})$ analytically. In general one seeks to minimize a. If a is too large, then very few u_i will be accepted, and the method will be impractical.

In Problem E, acceptance methods may be applied to draw from the distrib ution with probability density $\mathrm{p}(\mathbf{x})$. If $\mathrm{h}(\mathbf{x})$ is a source density as described in Section 3.2, $0 \le \mathrm{p}(\mathbf{x})/\mathrm{h}(\mathbf{x}) \le a < \infty \,\forall\, \mathbf{x} \in D$, then a sequence of i.i.d. draws from the distribution with p.d.f. $\mathrm{p}(\mathbf{x})$ may be constructed. If we take $\{\mathbf{x}_i\}_{i=1}^N$ to be the accepted draws, then

$$E_N = N^{-1}\sum_{i=1}^N \mathrm{g}(\mathbf{x}_i) \xrightarrow{a.s.} E, \quad \mathrm{E}(E_N) = E \,\forall N, \quad \sqrt{N}(E_N - E) \xrightarrow{d} \mathrm{N}(0, \sigma^2),$$
$$\sigma^2 = \mathrm{var}_p[\mathrm{g}(\mathbf{x})], \quad s_N^2 = \sum_{i=1}^N [\mathrm{g}(\mathbf{x}_i) - E_N]^2 / N \xrightarrow{a.s.} \sigma^2. \tag{4.2.2}$$

If we take $\{\mathbf{z}_i\}_{i=1}^N$ to be draws from the source density, and $u_i = 1$ if $\mathbf{z}_i$ is accepted and $u_i = 0$ if not, then

$$E_N = \sum_{i=1}^N u_i \, \mathrm{g}(\mathbf{z}_i) \Big/ \sum_{i=1}^N u_i \xrightarrow{a.s.} E, \quad \sqrt{N}\left(E_N - E\right) \xrightarrow{d} \mathrm{N}\left(0, \sigma^2\right),$$
$$\sigma^2 = a \int_D \left[\mathrm{g}(\mathbf{z}) - E\right]^2 \mathrm{p}(\mathbf{z}) d\mathbf{z}, \quad a \sum_{i=1}^N u_i \left[\mathrm{g}(\mathbf{z}_i) - E\right]^2 \Big/ \sum_{i=1}^N u_i \xrightarrow{a.s.} \sigma^2. \quad (4.2.3)$$

(In this case one again seeks to choose $\mathrm{h}(\mathbf{x})$ so as to minimize a.) Which expression is more relevant depends on the particulars of the problem. We shall retur n to this topic in Section 4.4.

The acceptance method just described assumes that the probability densit y is known, including its constant of integration -- i.e., $\int_D \mathrm{p}(\mathbf{x}) dx = 1$. This assumption may be strong in practice. In Problem I, one may recognize $\mathrm{p}(\mathbf{x})$ as a probability density kernel, not knowing the constant of integration. Acceptance or adaptive methods mig ht be applied to draw from the distribution with kernel density $\mathrm{p}(\mathbf{x})$; these methods do not require that one know the constant of integration for $\mathrm{p}(\mathbf{x})$. If $\mathrm{p}(\mathbf{x})$ is the kernel and $p^* = \int_D \mathrm{p}(\mathbf{x}) d\mathbf{x}$, it is then the case for acceptance methods in Problem I that

$$I_N = N^{-1} a p^* \sum_{i=1}^N u_i \xrightarrow{a.s.} I.$$

Whether or not consistent evaluation of p^* is possible depends on the method used to draw variables from the distribution with kernel $\mathrm{p}(\mathbf{x})$. If the method is acceptance sampling or a variant on acceptance sampling (e.g., the adaptive method for log- concave densities described in Section 3.2), one can approximate p^* using the methods just described as long as the actual probability density (not just the kernel) of the source distribution for the target kernel $\mathrm{p}(\mathbf{x})$ is known. This produces a sequence p_i^* with the property $\bar{p}_N^* \equiv N^{-1} \sum_{i=1}^N p_i^* \xrightarrow{a.s.} p^*$. In this case clearly

$$I_N = N^{-1} a p_N^* \sum_{i=1}^N u_i \xrightarrow{a.s.} I, \quad \sqrt{N}\left(I_N - I\right) \xrightarrow{d} \mathrm{N}\left(0, \sigma^2\right),$$

but σ^2 is affected by the substitution of $\bar{p}_n^*$ for p^*.

One may work out expressions for σ^2 and a corresponding consistent (in N) approximation of σ^2, as has been done already in several cases. Such expressions are quite useful in the analytical comparison of approximation methods. But if th e goal is simply to assess approximation error, straightforward asymptotic expansion is much simpler. To illustrate the method, return to the case of simple Monte Carlo integrat ion with p^* unknown, (4.1.1). Let M be the number of i.i.d. draws from source density $\mathrm{h}(\mathbf{z})$ for target density $\mathrm{p}(\mathbf{z})$, define $a = \sup_D \left[\mathrm{p}(\mathbf{z}) / \mathrm{h}(\mathbf{z})\right]$, and let

$$y_i = \mathrm{p}(\mathbf{z}_i) / \mathrm{h}(\mathbf{z}_i)$$
$$u_i = \begin{cases} 1 \text{ with probability } \mathrm{p}(\mathbf{z}_i) / a\,\mathrm{h}(\mathbf{z}_i), \\ 0 \text{ otherwise} \end{cases}$$
$$w_i = u_i \, \mathrm{g}(\mathbf{z}_i).$$

Defining $\bar{y}_M = M^{-1}\sum_{i=1}^{M} y_i$, $\bar{u}_M = M^{-1}\sum_{i=1}^{M} u_i$, $\bar{w}_M = M^{-1}\sum_{i=1}^{M} w_i$,

$$I_M = \bar{y}_M \bar{w}_M / \bar{u}_M \xrightarrow{a.s.} I .$$

As long as $\int_D \mathrm{g}^2(x)\,\mathrm{p}(x)dx$ is absolutely convergent, $\sqrt{M}\left(I_M - I\right) \xrightarrow{d} \mathrm{N}\left(0, \sigma^2\right)$, and

$$I_M^2\left[\frac{\hat{\mathrm{var}}(y_i)}{\bar{y}_M^2} + \frac{\hat{\mathrm{var}}(w_i)}{\bar{w}_M^2} + \frac{\hat{\mathrm{var}}(u_i)}{\bar{u}_M^2} + \frac{2\,\hat{\mathrm{cov}}(y_i, w_i)}{\bar{y}_M \bar{w}_M} - \frac{2\,\hat{\mathrm{cov}}(y_i, u_i)}{\bar{y}_M \bar{u}_M} - \frac{2\,\hat{\mathrm{cov}}(w_i, u_i)}{\bar{w}_M \bar{u}_M}\right] \xrightarrow{a.s.} \sigma^2 .$$

(This expression may be derived by the delta method, i.e., by linearizi ng I_M in $\bar{y}_M$, $\bar{u}_M$ and $\bar{w}_M$. The terms $\hat{\mathrm{var}}(y_i)$, $\hat{\mathrm{cov}}(y_i, w_i)$, etc., are computed in the usual way from $\{y_i, w_i, u_i\}_{i=1}^{M}$.)

4.3 Importance sampling

The method of importance sampling may be used to solve Problem I or Prob lem E, under similar circumstances: one has available a probability distributio n with p.d.f. somewhat similar to the integrand $\mathrm{f}(\mathbf{x})$ in Problem I or the probability density function $\mathrm{p}(\mathbf{x})$ in Problem E and wishes to use an independent, identically distributed sample from this distribution to approximate I or E. Rather than use acceptance to generate an i.i.d. sample from the distribution with p.d.f. $\mathrm{p}(\mathbf{x})$, importance sampling uses all of the draws from the source probability distribution but weights that sample to obta in a convergent approximation. In this method the probability density function of the s ource distribution is called the importance sampling density, a term due to Hammersly and Hand scomb (1964), who were among the first to proposed the method. It appears to have bee n introduced to the economics literature by Kloek and van Dijk (1978). We shall denote th e importance sampling density $\mathrm{j}(\mathbf{x})$.

Suppose that for Problem I one can draw an i.i.d. sequence of random ve ctors $\{\mathbf{x}_i\}$ from the importance distribution and that the support of this distributi on includes D. Then

$$\mathrm{E}_{\mathrm{j}}\left[\mathrm{f}(\mathbf{x}_i)/\mathrm{j}(\mathbf{x}_i)\right] = \int_D \left[\mathrm{f}(\mathbf{x})/\mathrm{j}(\mathbf{x})\right]\mathrm{j}(\mathbf{x})d\mathbf{x} = \int_D \mathrm{f}(\mathbf{x})d\mathbf{x} = I .$$

Since $\mathrm{f}(\mathbf{x}_i)/\mathrm{j}(\mathbf{x}_i)$ is also an i.i.d sequence,

$$I_N \equiv N^{-1}\sum_{i=1}^{N}\left[\mathrm{f}(\mathbf{x}_i)/\mathrm{j}(\mathbf{x}_i)\right] \xrightarrow{a.s.} I$$

by the strong law of large numbers. Furthermore, $\mathrm{E}(I_N) = I\,\forall N$. This result is remarkable for its weakness: no upper bound on $\mathrm{f}(\mathbf{x})/\mathrm{j}(\mathbf{x})$ is required as is the case for $\mathrm{f}(\mathbf{x})/\mathrm{h}(\mathbf{x})$ in acceptance sampling. The requirement that the support of $\mathrm{j}(\mathbf{x})$ include D is necessary and usually trivial to verify.

In Problem E importance sampling may be attractive if there is no simple method of constructing pseudorandom numbers drawn from the distribution $\mathrm{P}(\cdot)$ underlying the expectation operator. If the constant of integration for the probabilit y density is known, then

$$E_N = N^{-1}\sum_{i=1}^{N}\left[\mathrm{g}(\mathbf{x}_i)\,\mathrm{p}(\mathbf{x}_i)/\mathrm{j}(\mathbf{x}_i)\right] \xrightarrow{a.s.} E \text{ and } \mathrm{E}(E_N) = E\ \forall N$$

as long as the support of the importance sampling distribution includes that of $\mathrm{P}(\cdot)$. If the constant of integration is not known and $\mathrm{p}(\mathbf{x})$ is merely the kernel of the probability density function, $\int_D \mathrm{p}(\mathbf{x})d\mathbf{x} = p^*$, then

$$N^{-1}\sum_{i=1}^{N}\left[\mathrm{g}(\mathbf{x}_i)\,\mathrm{p}(\mathbf{x}_i)/\mathrm{j}(\mathbf{x}_i)\right] \xrightarrow{a.s.} p^*E, \quad N^{-1}\sum_{i=1}^{N}\left[\mathrm{p}(\mathbf{x}_i)/\mathrm{j}(\mathbf{x}_i)\right] \xrightarrow{a.s.} p^*,$$

and hence

$$E_N \equiv \frac{\sum_{i=1}^{N}\left[\mathrm{g}(\mathbf{x}_i)\,\mathrm{p}(\mathbf{x}_i)/\mathrm{j}(\mathbf{x}_i)\right]}{\sum_{i=1}^{N}\left[\mathrm{p}(\mathbf{x}_i)/\mathrm{j}(\mathbf{x}_i)\right]} \xrightarrow{a.s.} E, \tag{4.3.1}$$

but of course $\mathrm{E}(E_N) \neq E$ in general. In either case $\mathrm{w}(\mathbf{x}) = \mathrm{p}(\mathbf{x})/\mathrm{j}(\mathbf{x})$ may be regarded as a weight function, large weights being assigned to those $\mathrm{g}(\mathbf{x}_i)$ for which the importance sampling distribution assigns smaller probability than does the probabil ity distribution $\mathrm{P}(\cdot)$.

To assess the accuracy of importance sampling approximations using a cen tral limit theorem, more is required. In the case of Problem I, suppose that $\int_D\left[\mathrm{f}^2(\mathbf{x})/\mathrm{j}(\mathbf{x})\right]d\mathbf{x}$ is absolutely convergent. Then $\mathrm{f}(\mathbf{x}_i)/\mathrm{j}(\mathbf{x}_i)$ is an i.i.d. sequence and

$$I_N \xrightarrow{a.s.} I, \quad \sqrt{N}(I_N - I) \xrightarrow{d} \mathrm{N}(0,\sigma^2),$$

$$\sigma^2 = \int_D\left[\frac{\mathrm{f}^2(\mathbf{x})}{\mathrm{j}(\mathbf{x})}\right]d\mathbf{x} - I^2 = \mathrm{E}_j\left[\frac{\mathrm{f}(\mathbf{x})}{\mathrm{j}(\mathbf{x})} - I\right]^2, \quad s_N^2 = N^{-1}\sum_{i=1}^{N}\frac{\mathrm{f}^2(\mathbf{x}_i)}{\mathrm{j}^2(\mathbf{x}_i)} - I_N^2 \xrightarrow{a.s.} \sigma^2. \tag{4.3.2}$$

It is therefore practical to assess the accuracy of I_N as an approximation of I. The convergence of $\int_D\left[\mathrm{f}^2(\mathbf{x})/\mathrm{j}(\mathbf{x})\right]d\mathbf{x}$ must be established analytically, however. If $\left|\mathrm{f}(\mathbf{x})/\mathrm{j}(\mathbf{x})\right|$ is bounded above on D or if D is compact and $\mathrm{f}^2(\mathbf{x})/\mathrm{j}(\mathbf{x})$ is bounded above, then convergence obtains. If neither of these conditions is satisfied, then verifying convergence may be difficult. In choosing an importance sampling density, it is esp ecially important to insure that the tails of $\mathrm{j}(\mathbf{x})$ decline no faster than those of $\mathrm{f}(\mathbf{x})$. If these conditions are not met, but one still proceeds with the approximation, then convergence is usually quite slow. Violation of the central limit theorem convergence condition then may be evidenced by values of s_N^2 that increase with N.

Assessing the accuracy of E_N as an approximation of E is complicated by the ratio of terms in (4.3.1). If both

$$\mathrm{E}_\mathrm{p}\left[w(\mathbf{x})\right] = \int_D\left[\mathrm{p}^2(\mathbf{x})/\mathrm{j}(\mathbf{x})\right]d\mathbf{x} \text{ and } \mathrm{E}_\mathrm{p}\left[\mathrm{g}^2(\mathbf{x})\,\mathrm{w}(\mathbf{x})\right] = \int_D\left[\mathrm{g}^2(\mathbf{x})\,\mathrm{p}(\mathbf{x})\right]d\mathbf{x} \tag{4.3.3}$$

are absolutely convergent, then

$$E_N \xrightarrow{a.s.} E, \quad \sqrt{N}(E_N - E) \xrightarrow{d} \mathrm{N}(0,\sigma^2),$$

$$\sigma^2 = \mathrm{E}_\mathrm{p}\left\{\left[\mathrm{g}(\mathbf{x}) - E\right]^2\mathrm{w}(\mathbf{x})\right\} = p^{*-1}\int_D\left\{\left[\mathrm{g}(\mathbf{x}) - E\right]^2\mathrm{w}(\mathbf{x})\,\mathrm{p}(\mathbf{x})\right\}d\mathbf{x}, \tag{4.3.4}$$

$$s_N^2 = \frac{N\sum_{i=1}^{N}\left[\mathrm{g}(\mathbf{x}_i) - E\right]^2 \mathrm{w}(\mathbf{x}_i)}{\left[\sum_{i=1}^{N}\mathrm{w}(\mathbf{x}_i)\right]^2} \xrightarrow{a.s.} \sigma^2 .$$

(Derivations are given in Geweke, 1989.) This result provides a practical way to assess approximation error and also indicates conditions in which the method of importance sampling will work well for Problem E. A small value of $\mathrm{E}_p[\mathrm{w}(\mathbf{x})]$, perhaps as reflected in a small upper bound on $\mathrm{w}(\mathbf{x})$, combined with small $\mathrm{var}_p[\mathrm{g}(\mathbf{x})]$, will lead to small values of σ^2. As in the case of Problem I, central limit theorem convergence conditions must be verified analytically.

There has been little practical work to date on the optimal choice of importance sampling distributions. Using a result of Rubinstein (1981, Theorem 4.3.1) one can show that the importance sampling density with kernel $|\mathrm{g}(\mathbf{x}) - E|\mathrm{p}(\mathbf{x})$ provides the smallest possible value of σ^2. This is not very useful, since drawing pseudorandom vectors from this distribution is likely to be awkward at best. There has been some attention to optimization within families of importance sampling densities (Geweke, 1989), but optimization procedures themselves generally involve integrals that in turn require numerical approximation. Adaptive methods use previously drawn $\mathbf{x}_i$ to identify large values of $\mathrm{f}(\mathbf{x})/\mathrm{j}(\mathbf{x})$, $\mathrm{w}(\mathbf{x})$, or $\mathrm{g}^2(\mathbf{x})\mathrm{w}(\mathbf{x})$ and modify $\mathrm{j}(\mathbf{x})$ accordingly (Evans, 1991). Such procedures can be convenient but are limited by the fact that $\mathbf{x}_i$ is least likely to be drawn where $\mathrm{j}(\mathbf{x})$ is small. Informal, deterministic methods for tailoring $\mathrm{j}(\mathbf{x})$ have worked well in some problems in Bayesian econometrics (Geweke, 1989).

In Problem I the objective in choosing the importance sampling density is to find $\mathrm{j}(\mathbf{x})$ that mimics the shape of $\mathrm{f}(\mathbf{x})$ as closely as possible; the relevant metric is (4.3.2). Finding $\mathrm{j}(\mathbf{x}) \propto \mathrm{f}(\mathbf{x})$ will drive σ^2 to zero, but this amounts to analytical solution of the problem since $\int_D \mathrm{j}(\mathrm{x})d\mathbf{x} = 1$. In Problem E the relevant metric (4.3.4) is more complicated, involving both the variance of $\mathrm{g}(\mathbf{x})$ and the closeness of $\mathrm{j}(\mathbf{x})$ to $\mathrm{p}(\mathbf{x})$ as reflected in $\mathrm{w}(\mathbf{x}) = \mathrm{p}(\mathbf{x})/\mathrm{j}(\mathbf{x})$. As long as $\mathrm{var}_p[\mathrm{g}(\mathbf{x})] > 0$, no choice of $\mathrm{j}(\mathbf{x})$ will drive σ^2 to zero, and if $\mathrm{var}_p[\mathrm{g}(\mathbf{x})] = 0$, then Problem E reduces to Problem I. If $\mathrm{j}(\mathbf{x}) \propto \mathrm{p}(\mathbf{x})$, then $\sigma^2 = \mathrm{var}_p[\mathrm{g}(\mathbf{x})]$, which can serve as a benchmark in evaluating the adequacy of $\mathrm{j}(\mathbf{x})$. The ratio $\sigma^2/\mathrm{var}_p[\mathrm{g}(\mathbf{x})]$ has been termed the relative numerical efficiency of $\mathrm{j}(\mathbf{x})$ (Geweke, 1989): it indicates the ratio of iterations using $\mathrm{p}(\mathbf{x})$ itself as the importance sampling density, to the number using $\mathrm{j}(\mathbf{x})$, required to achieve the same accuracy of approximation of E. Relative numerical efficiency much less than 1.0 (less than 0.1, certainly less than 0.01) indicates poor imitation of $\mathrm{p}(\mathbf{x})$ by $\mathrm{j}(\mathbf{x})$ in the metric (4.3.4), possibly the existence

of a better importance sampling distribution or the failure of the under lying convergence conditions (4.3.3).

4.4 A note on the choice of method

There is considerable scope for combining the methods discussed in Secti ons 3 and 4. For example, the pseudorandom number generation in making draws from the population with probability density $\mathrm{h}(\mathbf{x})$, in the case of acceptance sampling, or $\mathrm{j}(\mathbf{x})$, in the case of importance sampling, generally will involve several of the methods discu ssed in Section 3.2. In even moderately complex problems, the investigator needs to tailor th ese methods, balancing computational efficiency against demands for the development a nd checking of reliable code.

Acceptance sampling and importance sampling are clearly similar. In fac t, given a candidate source density, one has the choice of undertaking either accep tance or importance sampling. A straightforward comparison of approximation errors indicate s the issues involved in the choice. In Problem I, the variance in acceptance sampli ng is

$$\sigma_1^2 = \int_D \left[\mathrm{g}(\mathbf{x}) - I\right]^2 \mathrm{p}(\mathbf{x}) d\mathbf{x} = \int_D \mathrm{g}^2(\mathbf{x})\,\mathrm{p}(\mathbf{x}) d\mathbf{x} - I^2$$

if by *draw* we mean *accepted* draw. But if instead we mean every draw from the source distribution, the variance is

$$\sigma_2^2 = aI - I^2, \quad a = \sup_D \mathrm{g}(\mathbf{x}),$$

from (4.2.1). In importance sampling, where all draws are used but di fferentially weighted, the variance is

$$\sigma_3^2 = \int_D \mathrm{g}^2(\mathbf{x})\,\mathrm{p}(\mathbf{x}) d\mathbf{x} - I^2,$$

from (4.3.2). Hence given a choice between acceptance and importance sampling in Problem I, importance sampling is clearly preferred: it conserves inform ation from all draws, whereas the rejected draws in acceptance sampling require executi on time but do not further improve the accuracy of the approximation.

For Problem E the situation is different. The variance is

$$\sigma_4^2 = \int_D \left[\mathrm{g}(\mathbf{x}) - E\right]^2 \mathrm{p}(\mathbf{x}) d\mathbf{x}$$

for acceptance sampling (see (4.2.2)) if we count only accepted draw s and

$$\sigma_5^2 = a\int_D \left[\mathrm{g}(\mathbf{z}) - E\right]^2 \mathrm{p}(\mathbf{z}) d\mathbf{z}, \quad a = \sup_D \left[\mathrm{p}(\mathbf{z})/\mathrm{h}(\mathbf{z})\right]$$

if we count all draws (see (4.2.3)). For importance sampling, expre ssing (4.3.4) in the notation of acceptance sampling, we have

$$\sigma_6^2 = \int_D \left[\mathrm{g}(\mathbf{x}) - E\right]^2 \mathrm{w}(\mathbf{x})\,\mathrm{p}(\mathrm{x}) d\mathbf{x}, \quad \mathrm{w}(\mathbf{x}) = \mathrm{p}(\mathrm{x})/\mathrm{h}(\mathrm{x}).$$

Since $\sigma_4^2 \le \sigma_6^2 \le \sigma_5^2$, a choice between acceptance and importance sampling on grounds of computational efficiency rests on the particulars of the problem. If ev aluation of $g(\mathbf{x})$ is sufficiently expensive relative to evaluation of $p(\mathbf{x})/h(\mathbf{x})$, acceptance sampling will be more efficient; otherwise, importance sampling will be the choice.

In fact one may combine acceptance and importance sampling. Let c be any positive constant, and define

$$w(\mathbf{z}_i) = \begin{cases} p(\mathbf{z}_i)/c\,h(\mathbf{z}_i) \text{ if } p(\mathbf{z}_i)/h(\mathbf{z}_i) \ge c \\ 1 \text{ with probability } p(\mathbf{z}_i)/c\,h(\mathbf{z}_i) \text{ if } p(\mathbf{z}_i)/h(\mathbf{z}_i) < c \\ 0 \text{ otherwise} \end{cases}.$$

Then $\sum_{i=1}^n w(\mathbf{z}_i)\mathbf{g}(\mathbf{z}_i) \Big/ \sum_{i=1}^n w(\mathbf{z}_i) \xrightarrow{a.s.} E$. For any given problem there will be a value of c that minimizes the variance of approximation error relative to required computing time. This may be found experimentally; or for some analytical methods, see Mü ller (1991, Chapter 2). The hybrid method can result in dramatic increases in effi ciency when computation of $g(\mathbf{x})$ is relatively expensive (or there are many such functions to be evaluated) and the weight function $w(\mathbf{x})$ is small with high probability.

A more fundamental choice is that between the simulation methods discuss ed in this and the previous section and the deterministic algorithms outlined in Se ction 2. Many problems in economics require integration in very high dimensions. (Two examples are presented in Section 7.) For such problems the most practical determin istic procedures are the low discrepancy methods of Section 2.3. Tables 2 and 3 provide some specific comparisons for dimensions as high as $d = 100$. (Execution time for quadrature methods in these problems is approximately $8 \times 4^{d-10}$ seconds on a Sun 10/51 workstation: .01 seconds for $d = 5$, 8 seconds for $d = 10$, 3 months for $d = 20$, about 10^4 times the estimated age of the universe for $d = 40, \ldots$.)

Table 2 extends the analysis of the same problem taken up in Section 2.3 . As noted there, the bounds in (2.3.1) and (2.3.2) are useless for this proble m and most others. The actual Halton errors presented in Table 2 were found by direct computati on, using the first d primes as the bases. The Monte Carlo errors were found analytically. Two error bounds are presented, one based on a 95% confidence interval ($\pm 1.96\sigma$) and a second based on a $100(1\text{-}10^{-12})$% confidence interval ($\pm 7.13\sigma$). For lower dimensions the comparison is dominated by the convergence of the Halton sequence at rate $\log m/m$ compared with Monte Carlo at rate $m^{-1/2}$: the Halton sequence is much more accurate. But for any reasonable fixed value of m, the comparison in higher dimensions is dominated by an approximately exponential rate of error increase in d for the Halton sequence, contrasted with the rate $d^{1/2}$ for Monte Carlo. For $m = 1{,}000$ iterations, Monte Carlo is more efficient

for d exceeding about 25 if one applies the $p = .05$ standard and for d exceeding about 45 for the $p = 10^{-12}$ standard. For $m = 50{,}000$ the breakpoints occur around $d = 35$ and $d = 110$, respectively. (The Halton error is not monotone decreasing in m because of the systematic way in which points are selected.)

Table 3 provides a comparison of these methods for an example of Problem E. The Halton sequence is first mapped into the normal distribution applying th e inverse-c.d.f. transformation in each dimension. Each of the five panels provides appr oximations to successively higher moments, p, of the multivariate normal distribution. Within each panel, the comparison is dominated by the same features noted for Table 2. Com parisons across panels are dominated by important characteristics of each method. Monte Carlo errors are proportional to $\mathrm{E}\left[(z)^{2p}\right] = \left[(2p-1)\cdot(2p-3)\cdot\ldots\cdot 3\cdot 1\right]^{1/2}$, where $z \sim \mathrm{N}(0,1)$. Halton errors reflect an interaction between the ordering of the points and the charac teristics of x_i^p. When p is odd, x_i^p is an odd monotone increasing function of x_i, whereas the standard normal probability density function is even. For any fixed m, the Halton points systematically exclude positive x_i values for which the corresponding $-x_i$ value has been included. Hence the error is always negative (as it was in Table 2 for the same reason). When p is even, this is not the case and the size of the error is smaller as w ell. The tendency of the Halton sequence to systematically exclude larger x_i has more severe consequences for evaluation of the integral the higher the value of odd p. Thus, for $p = 5$ independence Monte Carlo becomes dominant for values of d exceeding a fairly small threshold.

The largest problems worked for Table 3 ($d = 100, m = 50{,}000$) required about 75 seconds on a Sun 10/51 when solved using a Halton sequence. Independenc e Monte Carlo was about 15 times faster in every case. The difference reflects the in herent speed of linear congruential generators, contrasted with the floating point operations r equired to generate a Halton sequence. For more complex and realistic problems the relative s peed of independence Monte Carlo is less important, since computation time typic ally will be dominated by subsequent computations involving the sequences produced by either method.

These comparisons illustrate the general rule that simulation methods ar e preferred for higher dimensional problems. If the dimension is very low, then quadrat ure methods are much faster and more accurate. For intermediate dimensions, quadrature is impractical and low discrepancy methods are more accurate than simulation methods. Just where the breakpoints occur is problem-specific, and the situation is complicated by the fact that there are no useful independent assessments of approximation error for low dis crepancy methods. Simulation methods always provide an assessment of numerical e rror as a by-product, for square-integrable functions. Combined with the checks for robustness of

results with respect to alternative uniform random number generators and seed values, these methods are practical and reliable for a much wider range of problems th an is any deterministic algorithm. As we shall see, their application in complex problems can be very natural.

5. Variance reduction

In any of the independence Monte Carlo methods a single draw can be repl aced by the mean of M identically but not independently distributed draws. For example, in s imple Monte Carlo for Problem I,

$$I_{N,M} = N^{-1}\sum_{i=1}^{N}\left\{M^{-1}\left[\sum_{j=1}^{M} \mathrm{g}\left(\mathbf{x}_{ij}\right)\right]\right\}.$$

For any $i \neq k$ $\mathbf{x}_{ij}$ and $\mathbf{x}_{k1}$ are independent, whereas $\mathbf{x}_{ij}$ and $\mathbf{x}_{i1}$ are dependent. Since all $\mathbf{x}_{ij}$ are drawn from the distribution with probability density $\mathrm{p}(\mathbf{x})$,

$$I_{N,M} \xrightarrow{a.s.} I, \quad \sqrt{N}\left(I_{N,M} - I\right) \xrightarrow{d} \mathrm{N}\left(0, \sigma^{*2}\right),$$

$$\sigma^{*2} = \mathrm{var}\left[M^{-1}\sum_{j=1}^{M} \mathrm{g}\left(\mathbf{x}_{ij}\right)\right], \quad s_N^{*2} = N^{-1}\sum_{i=1}^{N}\left[M^{-1}\sum_{j=1}^{M} \mathrm{g}\left(\mathbf{x}_{ij}\right) - I_{N,M}\right]^2 \xrightarrow{a.s.} \sigma^{*2}.$$

The idea is to set up the relation among $\mathbf{x}_{i1}, \mathrm{K}, \mathbf{x}_{iM}$ in such a way that $\sigma^{*2} < M^{-1}\,\mathrm{var}_p\left[\mathrm{g}\left(\mathbf{x}_{ij}\right)\right]$. If in addition the cost of generating the M-tuple is insignificantly greater than the cost of generating M independent variables from $\mathrm{p}(\mathbf{x})$, then $I_{N,M}$ provides a computationally more efficient approximation of I than does I_N.

There are numerous variants on this technique. This section takes up fo ur that account for most use of the method: antithetic variables, systematic sampling, c onditional expectations, and control variables. The scope for combining these vari ance reduction techniques with the methods of Section 4 or Section 6 is enormous. Rath er than list all the possibilities, the purpose here is to provide some appreciation of the c ircumstances in which each variant may be practical and productive.

5.1 Antithetic Monte Carlo

This technique is due to Hammersly and Morton (1956) and has been wide ly used in statistics, experimental design, and simulation (e.g., Mikhail, 1972; M itchell, 1973; Geweke, 1988). In antithetic simple Monte Carlo integration $M = 2$ correlated variables are drawn in each of N replications. Then,

$$\sigma^{*2} = \tfrac{1}{2}\left\{\mathrm{var}\left[\mathrm{g}\left(\mathbf{x}_{i1}\right)\right] + \mathrm{cov}\left[\mathrm{g}\left(\mathbf{x}_{i1}\right), \mathrm{g}\left(\mathbf{x}_{i2}\right)\right]\right\}.$$

As long as $\mathrm{cov}\left[\mathrm{g}\left(\mathbf{x}_{i1}\right), \mathrm{g}\left(\mathbf{x}_{i2}\right)\right] < 0$, antithetic simple Monte Carlo integration with $N/2$ replications will have smaller error variance than simple Monte Carlo it eration with N replications, and the computational requirements will be about the same.

To focus on the main ideas, consider the situation in which $\mathrm{p}(\mathbf{x})$ is symmetric about a point μ in Problem I set out in Section 4. In this case $\mathbf{x}_{i1} = \mu + \mathbf{w}_i$, $\mathbf{x}_{i2} = \mu - \mathbf{w}_i$ describes a pair of variables drawn from the distribution with p.d.f. $\mathrm{p}(\mathbf{x})$ with correlation matrix $-\mathbf{I}$. If $\mathrm{g}(\mathbf{x})$ were a linear function, then $\mathrm{var}\left\{\tfrac{1}{2}\left[\mathrm{g}\left(\mathbf{x}_{i1}\right) + \mathrm{g}\left(\mathbf{x}_{i2}\right)\right]\right\} = 0$, and variance reduction

would be complete. (Clearly $I = g(\mu)$; this case is of interest only as a limit for numerical integration problems.) At the other extreme, if $g(\mathbf{x})$ is also symmetric about μ, then $\mathrm{var}\left\{\frac{1}{2}\left[g(\mathbf{x}_{i1}) + g(\mathbf{x}_{i2})\right]\right\} = \mathrm{var}\left[g(\mathbf{x})\right]$: N replications of antithetic simple Monte Carlo integration will yield as much information as N replications of simple Monte Carlo, but will usually require about double the number of computations. As an intermed iate case, suppose that $d(y) = g(\mathbf{x}y)$ is either monotone nondecreasing or monotone nonincreasing for all $\mathbf{x}$. Then $g(\mathbf{x}_{i1}) - I$ and $g(\mathbf{x}_{i2}) - I$ must be of opposite sign if they are nonzero. This implies $\mathrm{cov}\left[g(\mathbf{x}_{i1}), g(\mathbf{x}_{i2})\right] < 0$, whence $\sigma^{*2} \leq \frac{1}{2}\mathrm{var}\left[g(\mathbf{x})\right] = \sigma^2/2$, and so antithetic simple Monte Carlo integration produces gains in efficiency.

The use of antithetic Monte Carlo integration is especially powerful in an important class of Bayesian learning and inference problems. In these problems $\mathbf{x}$ typically represents a vector of parameters unknown to an economic agent or an eco nometrician, and $p(\mathbf{x})$ is the probability density of that vector conditional on information av ailable. The integral I could correspond to an expected utility or a posterior probability. If the available information is based on an i.i.d. sample of size T, then it is natural to write $p_T(\mathbf{x})$ for $p(\mathbf{x})$. As T increases, the distribution $p_T(\mathbf{x})$ generally becomes increasingly symmetric and concentrated about the true value of the vector of unknown parameters, r eflecting the operation of a central limit theorem. In these circumstances $g(\mathbf{x})$ is increasingly well described by a linear approximation of itself over most of the support o f $p_T(\mathbf{x})$, as T increases. Suppose that the agent or econometrician approximates I using simple Monte Carlo with accuracy indicated by σ_T^2 or by antithetic simple Monte Carlo with accuracy indicated by σ_T^{*2}. Given some side conditions, mainly continuous differentiability of $g(\mathbf{x})$ in a neighborhood of the true value of the parameter vector $\mathbf{x}$ and a nonzero derivative of $g(\mathbf{x})$ at this point, it may be shown that $\sigma_T^{*2}/\sigma_T^2 \to 0$ (Geweke, 1988). Given additional side conditions, mainly twice continuous differentiability of $g(\mathbf{x})$ in a neighborhood of the true value of the parameter vector $\mathbf{x}$, it may be shown that $T\sigma_T^{*2}/\sigma_T^2$ converges to a constant. The constant is inversely related to the magnitude of $\partial g(\mathbf{x})/\partial \mathbf{x}$ and directly related to the magnitude of $\partial^2 g(\mathbf{x})/\partial \mathbf{x} \partial \mathbf{x}'$, each evaluated at the true value of the parameter vector $\mathbf{x}$ (Geweke, 1988). This result is an example of acceleration, because it indicates an interesting sequence of conditions under which the relative advantage of a variance reduction method increases without bound.

Application of the method of antithetic variables with techniques more c omplicated than simple Monte Carlo is generally straightforward. In the case of importa nce sampling, $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$ are drawn from the importance sampling density $j(\mathbf{x})$. In Problem I the term

$\frac{1}{2}\left[\mathrm{f}(\mathbf{x}_{i1})/\mathrm{j}(\mathbf{x}_{i1}) + \mathrm{f}(\mathbf{x}_{i2})/\mathrm{j}(\mathbf{x}_{i2})\right]$ replaces $\mathrm{f}(\mathbf{x}_i)/\mathrm{j}(\mathbf{x}_i)$. In Problem E, define $\mathrm{w}(\mathbf{x}) = \mathrm{p}(\mathbf{x})/\mathrm{j}(\mathbf{x})$ as before. Then

$$E_N \equiv \frac{\sum_{i=1}^{N}\left[\mathrm{g}(\mathbf{x}_{i1})\mathrm{w}(\mathbf{x}_{i1}) + \mathrm{g}(\mathbf{x}_{i2})\mathrm{w}(\mathbf{x}_{i2})\right]}{\sum_{i=1}^{N}\left[\mathrm{w}(\mathbf{x}_{i1}) + \mathrm{w}(\mathbf{x}_{i2})\right]} \xrightarrow{a.s.} E, \quad \sqrt{N}(E_N - E) \xrightarrow{d} \mathrm{N}(0, \sigma^{*2}),$$

$$\sigma^{*2} = \mathrm{E}_{\mathrm{p}}\left\{\left[\frac{\mathrm{g}(\mathbf{x}_{i1})\mathrm{w}(\mathbf{x}_{i1}) + \mathrm{g}(\mathbf{x}_{i2})\mathrm{w}(\mathbf{x}_{i2})}{\mathrm{w}(\mathbf{x}_{i1}) + \mathrm{w}(\mathbf{x}_{i2})} - E\right]^2 \frac{\mathrm{w}(\mathbf{x}_{i1}) + \mathrm{w}(\mathbf{x}_{i2})}{2}\right\},$$

$$s_N^2 = \frac{N\sum_{i=1}^{N}\left[\frac{\mathrm{g}(\mathbf{x}_{i1})\mathrm{w}(\mathbf{x}_{i1}) + \mathrm{g}(\mathbf{x}_{i2})\mathrm{w}(\mathbf{x}_{i2})}{\mathrm{w}(\mathbf{x}_{i1}) + \mathrm{w}(\mathbf{x}_{i2})} - E_N\right]^2 \left[\mathrm{w}(\mathbf{x}_{i1}) + \mathrm{w}(\mathbf{x}_{i2})\right]}{4\left\{\sum_{i=1}^{N}\left[\mathrm{w}(\mathbf{x}_{i1}) + \mathrm{w}(\mathbf{x}_{i2})\right]\right\}^2} \xrightarrow{a.s.} \sigma^{*2}.$$

These results are valid for any antithetic variables algorithm, even if $\mathrm{j}(\mathbf{x})$ is not symmetric and even if the variance of the approximation error σ^2 is increased rather than decreased in moving to the use of antithetic variables. The essential requirements a re that the $\mathbf{x}_{ij}$'s be drawn from the importance sampling distribution and that $\mathbf{x}_{ij}$ and $\mathbf{x}_{k1}$ be independent for $i \neq k$.

In complex problems involving multivariate $\mathbf{x}$, pseudorandom variables often may be generated by use of successive conditionals for $\mathbf{x}' = \left(\mathbf{x}'_{(1)}, \mathrm{K}\ , \mathbf{x}'_{(m)}\right)$,

$$\mathrm{p}(\mathbf{x}) = \mathrm{p}\left(\mathbf{x}_{(1)}\right)\mathrm{p}\left(\mathbf{x}_{(2)}\middle|\mathbf{x}_{(1)}\right)\mathrm{K}\ \mathrm{p}\left(\mathbf{x}_{(m)}\middle|\mathbf{x}_{(1)}\mathrm{K}\ \mathbf{x}_{(m-1)}\right).$$

In such cases a pair of antithetic variables $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$ may be created by constructing a pair for a single, convenient subvector $\mathbf{x}_{(j)}$. Especially if $\mathrm{g}(\mathbf{x}) = \mathrm{g}\left(\mathbf{x}_{(j)}\right)$, the benefits of antithetic Monte Carlo will then be realized in both Problem I and Problem E. An e xample of this use of antithetic variables is taken up in Section 7.2.

5.2 Systematic sampling

Systematic sampling (McGrath, 1970) combines certain advantages of det erministic and Monte Carlo methods. The former achieve great efficiency by systema tically choosing points for evaluation in specific low-dimensional problems; the latter p roduce indications of accuracy as a byproduct and are amenable to high-dimensional problems. Systematic sampling specifies an m-tuple of points as a deterministic function of a random vector $\mathbf{u}$,

$\mathbf{x}_j = \mathrm{f}_j(\mathbf{u}) \quad (j = 1, \mathrm{K}\ , m)$,

with the property that the induced distribution of every $\mathbf{x}_j$ is that of the probability density function $\mathrm{p}(\mathbf{x})$.

As a leading example consider the case of univariate x, with pseudorandom variables from the distribution of x constructed using the inverse c.d.f method (Section 3.2). Denote

$\mathrm{F}(c) = \mathrm{P}[x \le c]$, suppose u_i $(i = 1,\ldots,N)$ are independently and uniformly distributed on the unit interval, and take

$$x_{ij} = \mathrm{F}^{-1}\left(\left[u_i + j/m\right]\right) \quad (j = 1,\ldots,m),$$

where "[·]" denotes greatest fractional part. Clearly the method need not be limited to evenly spaced grids; e.g., Richtmeyer's method (Section 2.3) could just as easily be applied. Extension to higher dimensions is straightforward, but is subject to all of the problems of deterministic methods there. The advantage of systematic methods is tha t approximation error is generally $\mathrm{O}\left(m^{-1}\right)$ whereas that in Monte Carlo is $\mathrm{O}_p\left(N^{-1/2}\right)$.

In high-dimensional problems systematic sampling can be advantageous whe n confined to a subset of the vector $\mathbf{x}$ that is especially troublesome for Monte Carlo and/or is an important source of variation in the function $\mathrm{g}(\mathbf{x})$. As an example of the former condition, suppose it is difficult to find an importance sampling densit y that mimics $\mathrm{p}(\mathbf{x})$, but $\mathbf{x}' = \left(\underset{1\times m_1}{\mathbf{x}'_{(1)}}, \underset{1\times m_2}{\mathbf{x}'_{(2)}}\right)$, a good importance sampling density for the marginal p.d.f. $\mathrm{p}\left(\mathbf{x}_{(1)}\right)$ is available, and the inverse c.d.f. $\mathrm{F}^{-1}\left(p\middle|\mathbf{x}_{(1)}\right)$ of the conditional distribution of $\mathbf{x}_{(2)}$ can be evaluated. One may generate $\mathbf{x}_{1(i)}$ together with corresponding importance sampling weight w_i; draw $\left(u_1,\ldots,u_{m_2}\right)$ independently distributed on the unit interval; create the systematic sample

$$\mathbf{x}_{(2)ij_1\ldots j_{m_2}} = \mathrm{F}^{-1}\left(\left[u_1 + j_1/l_1\right],\ldots,\left[u_{m_2} + j_{m_2}/l_{m_2}\right]\right) \quad (j_k = 1,\ldots,l_k; k = 1,\ldots,m_2).$$

Then record

$$g_i = \left[\prod_{k=1}^{m_2} l_k\right]^{-1} \sum_{j_1=1}^{l_1} \cdots \sum_{j_{m2}=1}^{l_{m2}} \mathrm{g}\left(\mathbf{x}_{(1)i}, \mathbf{x}_{(2)ij_1\ldots j_{m2}}\right)$$

along with each weight w_i. Previous expressions in Section 4.3 for I_N, σ^2, and s_N^2 are then valid with g_i in place of $\mathrm{g}\left(\mathbf{x}_i\right)$. In particular (4.3.2) is still true, and s_N^2 may be used to assess the increase in accuracy yielded by systematic sampling with high er values of the l_k.

5.3 The use of conditional expectations

Suppose there is a partition of $\mathbf{x}$, $\mathbf{x}' = \left(\mathbf{x}'_{(1)}, \mathbf{x}'_{(2)}\right)$, such that

$$\mathrm{g}(\mathbf{x}) = \mathrm{g}\left(\mathbf{x}_{(1)}, \mathbf{x}_{(2)}\right) = \mathrm{g}^*\left(\mathbf{x}_{(1)}\right) l\left(\mathbf{x}_{(2)}\right),$$

where $l(\cdot)$ is linear; $\mathrm{p}(\mathbf{x}) = \mathrm{p}\left(\mathbf{x}_{(1)}, \mathbf{x}_{(2)}\right) = \mathrm{p}\left(\mathbf{x}_{(1)}\right)\mathrm{p}\left(\mathbf{x}_{(2)}\middle|\mathbf{x}_{(1)}\right)$; it is possible to draw pseudorandom vectors from the marginal distribution for $\mathbf{x}_{(1)}$ with p.d.f. $\mathrm{p}\left(\mathbf{x}_{(1)}\right)$; and $\mathrm{E}\left[\mathbf{x}_{(2)}\middle|\mathbf{x}_{(1)}\right]$ is known analytically. Then

$$\int_D \mathrm{g}(\mathbf{x})\,\mathrm{p}(\mathbf{x})\,d\mathbf{x} = \int_D \mathrm{g}^*\left(\mathbf{x}_{(1)}\right)\mathrm{p}\left(\mathbf{x}_{(1)}\right) l\left[\mathrm{E}\left(\mathbf{x}_{(2)}\middle|\mathbf{x}_{(1)}\right)\right] d\mathbf{x}_{(1)}, \qquad (5.3.1)$$

and

$$\mathrm{var}_{\mathrm{p}(\mathbf{x}_{(1)})}\left\{\mathrm{g}^{*}(\mathbf{x}_{(1)})\mathrm{I}\left[\mathrm{E}(\mathbf{x}_{(2)}|\mathbf{x}_{(1)})\right]\right\} \leq \mathrm{var}_{\mathrm{p}(\mathbf{x})}[\mathrm{g}(\mathbf{x})].$$

Consequently, application of Monte Carlo methods directly in (5.3.1) w ill produce an approximation error with smaller variance than would Monte Carlo in the general framework set forth in Section 4.

The use of conditional expectations in fact bears a close relationship t o antithetic Monte Carlo integration. In particular, if one could draw antithetic variable s $\mathbf{x}_{(2)i1}$ and $\mathbf{x}_{(2)i2}$ from the distribution with p.d.f. $\mathrm{p}(\mathbf{x}_{(2)}|\mathbf{x}_{(1)})$ with perfectly negative correlation, then $\frac{1}{2}(\mathbf{x}_{(2)i1}+\mathbf{x}_{(2)i2}) = \mathrm{E}(\mathbf{x}_{(2)}|\mathbf{x}_{(1)})$, and exactly the same result would be obtained.

More generally, whenever $\mathrm{g}(\mathbf{x})$ is a function of $\mathbf{x}_{(1)}$ only, it is usually worth noting whether $\mathrm{E}[\mathrm{g}(\mathbf{x}_{(1)})|\mathbf{x}_{(2)}]$ can be evaluated analytically. If so, then the variance of approximation error can be reduced by using the function of interest $\mathrm{E}[\mathrm{g}(\mathbf{x}_{(1)})|\mathbf{x}_{(2)i}]$ rather than $\mathrm{g}(x_{(1)i})$. Since $\mathrm{g}(x_{(1)}) = \mathrm{E}[\mathrm{g}(\mathrm{x}_{(1)})|x_{(2)}] + \eta$ with $\mathrm{cov}\{\eta, \mathrm{E}[\mathrm{g}(\mathrm{x}_{(1)})|x_{(2)}]\} = 0$,

$$\mathrm{var}_{\mathrm{p}(\mathbf{x}_{(2)})}\left\{\mathrm{E}\left[\mathrm{g}(\mathbf{x}_{(1)})|\mathbf{x}_{(2)}\right]\right\} \leq \mathrm{var}_{\mathrm{p}(\mathbf{x}_{(1)})}\left[\mathrm{g}(\mathbf{x}_{(1)})\right].$$

Against this improvement should be balanced the time required for the ad ditional computations, which are generally of no further use in generation of the $\mathbf{x}_i$; this time is usually small.

5.4 Control variables

It is often the case that one is able to solve approximations to Problem I or Problem E analytically. For example, if the mean μ of the distribution with p.d.f $\mathrm{p}(\mathbf{x})$ is known and one has available a linear approximation $\mathrm{g}^{(1)}(\mathbf{x})$ of the function $\mathrm{g}(\mathbf{x})$, then the mean of $\mathrm{g}^{(1)}(\mathbf{x})$ is $\mathrm{g}^{(1)}(\mu)$. Moreover if $\{\mathbf{x}_i\}_{i=1}^{N}$ is a pseudorandom sample drawn from the distribution with p.d.f. $\mathrm{p}(\mathbf{x})$, then $\mathrm{g}(\mathbf{x}_i)$ and $\mathrm{g}^{(1)}(\mathbf{x}_i)$ will be positively correlated if the linear approximation is good for most $\mathbf{x}_i$. In this situation the method of control variables, introduced by Kahn and Marshall (1953) and Hammersly and Handscomb (1 964), can be used to reduce the variance of the approximation error in I_N or E_N.

We develop the specific method for simple Monte Carlo integration in Pro blem I; extension to more involved methods is straightforward. Let $J_N = N^{-1}\sum_{i=1}^{N}\mathrm{h}(\mathbf{x}_i)$ have known mean J. (In the example given $\mathrm{h}(\mathbf{x}_i) = \mathrm{g}^{(1)}(\mathbf{x}_i)$, $J_N = N^{-1}\sum_{i=1}^{N}\mathrm{g}^{(1)}(\mathbf{x}_i)$ and $J = \mathrm{g}^{(1)}(\mu)$.) Consider approximations of the form

$$I'_N = I_N + \beta(J_N - J),$$

where I_N is computed as before. It is the case that $I'_N \xrightarrow{a.s.} I$, and as long as $\mathrm{var}_{\mathrm{p}}\left[\mathrm{h}(\mathbf{x}_i)\right]$ exists, a central limit theorem may still be used to evaluate numerical accuracy. One can easily verify that $\mathrm{var}(I'_N)$ is minimized by $\beta = -\mathrm{cov}(J_N, I_N)/\mathrm{var}(J_N)$, and in this case

$$\mathrm{var}(I'_N) = \mathrm{var}(I_N) - \frac{\mathrm{cov}^2(J_N, I_N)}{\mathrm{var}(J_N)} = \mathrm{var}(I_N)\left[1 - \mathrm{corr}^2(J_N, I_N)\right].$$

Usually the parameter β is unknown. It may be estimated in the obvious way from the replications.

This method is easily extended to the case in which a vector of estimate s $\mathbf{J}_N = \left(J_N^{(1)}, \mathrm{K}\ , J_N^{(q)}\right)'$ with known mean $\mathbf{J} = \left(J^{(1)}, \mathrm{K}\ , J^{(q)}\right)'$ is available. If we denote

$$\underset{q\times q}{\Sigma} = \mathrm{var}(\mathbf{J}_N), \quad \underset{q\times 1}{\mathbf{c}} = \mathrm{cov}(\mathbf{J}_N, I_N),$$

then the variance of the approximation

$$I'_N = I_N + \beta'(\mathbf{J}_N - \mathbf{J})$$

is minimized by $\beta = \Sigma^{-1}\mathbf{c}$, and in this case

$$\mathrm{var}(I'_N) = \mathrm{var}(I_N) - \mathbf{c}'\Sigma^{-1}\mathbf{c} = \mathrm{var}(I_N)\left[1 - \frac{\mathbf{c}'\Sigma^{-1}\mathbf{c}}{\mathrm{var}(I_N)}\right].$$

6. Markov chain Monte Carlo methods

All of the independence Monte Carlo methods for integration assume the a bility to efficiently generate pseudorandom variables from a distribution with spe cified probability density function $\mathrm{p}(\mathbf{x})$. But in many economic problems it is difficult or impossible to find a generation algorithm that is sufficiently efficient to be practical. An instructive limiting case is the one in which the constituents of $\mathbf{x}$ are independently distributed,

$$\mathrm{p}(\mathbf{x}) = \prod_{i=1}^{m} \mathrm{p}_i(x_i).$$

One could construct an acceptance sampling algorithm with a source densi ty $\mathrm{h}_i(z_i)$ corresponding to each $\mathrm{p}_i(z_i)$, and accept the draw with probability $\mathrm{p}(\mathbf{z})/a\,\mathrm{h}(\mathbf{z})$, where

$$a = \sup_z\left[\mathrm{p}(z)/\mathrm{h}(z)\right] = \prod_{i=1}^{m} a_i, \quad a_i = \sup_{z_i}\left[\mathrm{p}(z_i)/\mathrm{h}(z_i)\right] (i = 1,\mathrm{K}\ ,m).$$

Since a is directly proportional to the time required to obtain an accepted dra w (see Section 3.2) this expression makes clear that acceptance sampling can be subjec t to its own curse of dimensionality if the source density is constructed element-by-element. Essentially the same difficulty can arise in importance sampling, where it is manifested in only a few weights $\mathrm{w}(\mathbf{x}_i)$ accounting for the sum.

This example is of interest only as a limiting case. If the x_i really were independent, one could employ acceptance sampling element-by-element, and computation time would then be proportional to $\sum_{i=1}^{m} a_i$. An obvious extension of this idea to the general case is to write

$$\mathrm{p}(\mathbf{x}) = \mathrm{p}(x_1)\prod_{i=1}^{m} \mathrm{p}_{i|1,\mathrm{K}\ ,i-1}(x_i|x_1,\mathrm{K}\ ,x_{i-1})$$

and employ acceptance or importance sampling for each conditional. The difficulty here is that construction of probability density kernels for the marginal in x_1 and all but the last conditional require analytic integration. Notable simple cases aside, t his is not possible, and it remains impossible for subvectors as well as individual components.

This section takes up a recently developed generalization of independenc e Monte Carlo that has become known as *Markov chain Monte Carlo*. The idea is to construct a Markov chain with state space D and invariant distribution with p.d.f. $\mathrm{p}(\mathbf{x})$. Following an initial transient or *burn-in* phase, simulated values from the chain form a basis for approximating $\mathrm{E}_{\mathrm{p}}[g(\mathbf{x})]$, thus solving Problem E. If the p.d.f. $\mathrm{p}(\mathbf{x})$ does not contain an unknown factor of proportionality p^*, then Problem I is solved as well. What is required is to construct an appropriate algorithm and verify that its invariant distribution is uniq ue, with p.d.f. $\mathrm{p}(\mathbf{x})$.

Markov chain methods have a history in mathematical physics dating back to the algorithm of Metropolis *et al.* (1953). This method, which is described in Hammersly and Handscomb (1964, Section 9.3) and Ripley (1987, Section 4.7), was ge neralized by

Hastings (1970), who focused on statistical problems, and was further explored by Peskun (1973). A version particularly suited to image reconstruction and pro blems in spatial statistics was introduced by Geman and Geman (1984). This was subsequ ently shown to have great potential for Bayesian computation by Gelfand and Smith (199 0). Their work, combined with data augmentation methods (Tanner and Wong, 1987), has p roven very successful in the treatment of latent variables and other unobservables in economic models. (An example is given in Section 7.1.) Since 1990 application of Marko v chain Monte Carlo methods has grown rapidly; new refinements, extensions, and applications appear almost continuously.

This section concentrates on developing the methods, deferring serious e xamples to Section 7. We begin with a heuristic introduction to two widely used va riants of these methods, the Gibbs sampler and the Metropolis-Hastings algorithm (Secti on 6.1). Some theory of continuous state Markov chains required to demonstrate converg ence is given in Section 6.2. Easily verified sufficient conditions for convergence of t he Gibbs sampler are set forth in Section 6.3 and for convergence of the Metropolis-Hastings algorithm in Section 6.4. Some practical issues in assessing the error of approximat ion are treated in Section 6.5. Much of the treatment here draws heavily on the work of Ti erney (1991a, 1991b), who first used the theory of general state space Markov chains to demonstrate convergence, and Roberts and Smith (1992), who elucidated sufficient c onditions for convergence that turn out to be applicable in a wide variety of problems in economics.

6.1 Two Markov chain Monte Carlo algorithms

Motivated by the role of $p(\mathbf{x})$ in Problem I or Problem E, discussion here proceeds assuming that $\mathbf{x}$ is continuously distributed. However, there is no harm in regarding $\mathbf{x}$ as discrete on a first reading. A full development covering both the conti nuous and discrete cases is given in Section 6.2.

The Gibbs sampler begins with a partition, or *blocking*, of $\underset{m\times 1}{\mathbf{x}}$, $\mathbf{x}' = \left(\mathbf{x}'_{(1)}, \mathrm{K}\ , \mathbf{x}'_{(k)}\right)$. For $i = 1, \mathrm{K}\ , k$, $\mathbf{x}'_{(i)} = \left(\mathbf{x}_{i1}, \mathrm{K}\ , \mathbf{x}_{im(i)}\right)$ and $m(i) \geq 1$; $\sum_{i=1}^{k} m(i) = m$; and the x_{ij} are the components of $\mathbf{x}$. Let $p\left(\mathbf{x}_{(i)} \middle| \mathbf{x}_{(-i)}\right)$ denote the conditional p.d.f.'s induced by $p(\mathbf{x})$, where $\mathbf{x}_{(-i)} = \left\{\mathbf{x}_{(j)}, j \neq i\right\}$.

Suppose we were given a single drawing $\mathbf{x}^0$, $\mathbf{x}'^0 = \left(\mathbf{x}'^0_{(1)}, \mathrm{K}\ , \mathbf{x}'^0_{(k)}\right)$, from the distribution with p.d.f. $p(\mathbf{x})$. Successively make drawings from the conditional distribution as follo ws:

$$
\begin{aligned}
\mathbf{x}^1_{(1)} &\sim \mathrm{p}\left(\cdot \,\middle|\, \mathbf{x}^0_{(-1)}\right) \\
\mathbf{x}^1_{(2)} &\sim \mathrm{p}\left(\cdot \,\middle|\, \mathbf{x}^1_{(1)}, \mathbf{x}^0_{(3)}, \mathrm{K}\ , \mathbf{x}^0_{(k)}\right) \\
&\ \ \mathrm{M} \\
\mathbf{x}^1_{(j)} &\sim \mathrm{p}\left(\cdot \,\middle|\, \mathbf{x}^1_{(1)}, \mathrm{K}\ , \mathbf{x}^1_{(j-1)}, \mathbf{x}^0_{(j+1)}, \mathrm{K}\ , \mathbf{x}^0_{(k)}\right) \\
&\ \ \mathrm{M} \\
\mathbf{x}^1_{(k)} &\sim \mathrm{p}\left(\cdot \,\middle|\, \mathbf{x}^1_{(-k)}\right).
\end{aligned}
\tag{6.1.1}
$$

This defines a transition process from $\mathbf{x}^0$ to $\mathbf{x}^1 = \left(\mathbf{x}'^1_{(1)}, \mathrm{K}\ , \mathbf{x}'^1_{(k)}\right)$. The Gibbs sampler is defined by the choice of blocking and the forms of the conditional densi ties induced by $\mathrm{p}(\mathbf{x})$ and the blocking. Since $\mathbf{x}^0 \sim \mathrm{p}(\mathbf{x})$, $\left(\mathbf{x}^1_{(1)}, \mathrm{K}\ , \mathbf{x}^1_{(j-1)}, \mathbf{x}^1_{(j)}, \mathbf{x}^1_{(j+1)}, \mathrm{K}\ , \mathbf{x}^1_{(k)}\right) \sim \mathrm{p}(\mathbf{x})$ at each step in (6.1.1) by definition of the conditional density. In particul ar, $\mathbf{x}^1 \sim \mathrm{p}(\mathbf{x})$.

Iteration of the algorithm produces a sequence $\mathbf{x}^0, \mathbf{x}^1, \mathrm{K}\ , \mathbf{x}^t, \mathrm{K}$ which is a realization of a Markov chain with probability density function kernel for the transit ion from point $\mathbf{x}$ to point $\mathbf{y}$ given by

$$
\mathrm{K}_G(\mathbf{x}, \mathbf{y}) = \prod\nolimits_{1=1}^{k} \mathrm{p}\left[\mathbf{y}_{(1)} \,\middle|\, \mathbf{x}_{(j)}(j > 1), \mathbf{y}_{(j)}(j < 1)\right].
$$

Any single iterate $\mathbf{x}^t$ retains the property that it is drawn from the distribution with p.d.f. $\mathrm{p}(\mathbf{x})$.

For the Gibbs sampler to be practical, it is essential that the blocking be chosen in such a way that one can make the drawings (6.1.1) in an efficient manner. For many problems in economics, the blocking is natural and the conditional distributions are familiar; Section 7.1 provides an example. In making the drawings (6.1.1) all the methods o f Sections 3 and 4 are at our disposal. Observe that in this context acceptance sampling i s attractive relative to importance sampling, since the former produces independent, identically distributed, unweighted drawings from the conditional distribution.

Of course, it is generally difficult or impossible to make even one init ial draw from the distribution with p.d.f. $\mathrm{p}(\mathbf{x})$. The purpose of that assumption here is to marshal an informal argument that $\mathrm{p}(\mathbf{x})$ is the p.d.f. of the invariant distribution of the Markov chain. A leading practical problem is to elucidate conditions in which the dist ribution of $\mathbf{x}^t$ will converge to that corresponding to $\mathrm{p}(\mathbf{x})$ for any choice of $\mathbf{x}^0$ in the domain D, and we turn to this in Section 6.3.

The Metropolis-Hastings algorithm begins with an arbitrary transition pr obability density function $\mathrm{q}(\mathbf{x}, \mathbf{y})$ and a starting value $\mathbf{x}^0$. If $\mathbf{x}^t = \mathbf{x}$, the random vector generated

from $q(\mathbf{x},\mathbf{y})$ is considered as a candidate value for $\mathbf{x}^{t+1}$. The algorithm actually sets $\mathbf{x}^{t+1} = \mathbf{y}$ with probability

$$\alpha(\mathbf{x},\mathbf{y}) = \min\left\{\frac{p(\mathbf{y})q(\mathbf{y},\mathbf{x})}{p(\mathbf{x})q(\mathbf{x},\mathbf{y})}, 1\right\};$$

otherwise, the algorithm sets $\mathbf{x}^{t+1} = \mathbf{x} = \mathbf{x}^t$. This defines a Markov chain with a generally mixed continuous-discrete transition probability from $\mathbf{x}$ to $\mathbf{y}$ given by

$$K(\mathbf{x},\mathbf{y}) = \begin{cases} q(\mathbf{x},\mathbf{y})\alpha(\mathbf{x},\mathbf{y}) \text{ if } \mathbf{y} \neq \mathbf{x} \\ 1 - \int_D q(\mathbf{x},\mathbf{z})\alpha(\mathbf{x},\mathbf{z})d\mathbf{z} \text{ if } \mathbf{y} = \mathbf{x} \end{cases}.$$

This form of the algorithm is due to Hastings (1970). The Metropolis *et al.* (1953) form takes $q(\mathbf{x},\mathbf{y}) = q(\mathbf{y},\mathbf{x})$. A simple variant that is often useful is the independence chain (Tierney, 1991a, 1991b), $q(\mathbf{x},\mathbf{y}) = j(\mathbf{y})$. Then

$$\alpha(\mathbf{x},\mathbf{y}) = \min\left\{\frac{p(\mathbf{y})j(\mathbf{x})}{p(\mathbf{x})j(\mathbf{y})}, 1\right\} = \min\left\{\frac{w(\mathbf{y})}{w(\mathbf{x})}, 1\right\},$$

where $w(\mathbf{x}) = p(\mathbf{x})/j(\mathbf{x})$. The independence chain is closely related to acceptance sampling (Section 4.2) and importance sampling (Section 4.3). But rather tha n place a low (high) probability of acceptance or a low (high) weight on a draw that is too likely (unlikely) relative to $p(\mathbf{x})$, the independence chain assigns a high (low) probability of accepting the candidate for the next draw.

There is a simple two-step argument that motivates the convergence of th e sequence $\{\mathbf{x}^t\}$ generated by the Metropolis-Hastings algorithm to $p(\cdot)$. (This approach is due to Chib and Greenberg, 1994.) First, observe that if any transition proba bility function $p(\mathbf{x},\mathbf{y})$ satisfies the reversibility condition

$$p(\mathbf{x})p(\mathbf{x},\mathbf{y}) = p(\mathbf{y})p(\mathbf{y},\mathbf{x}),$$

then it has $p(\cdot)$ as its invariant distribution. To see this, note that

$$\int p(\mathbf{x})p(\mathbf{x},\mathbf{y})d\mathbf{x} = \int p(\mathbf{y})p(\mathbf{y},\mathbf{x})d\mathbf{x} = p(\mathbf{y})\int p(\mathbf{y},\mathbf{x})d\mathbf{x} = p(\mathbf{y}).$$

The second step is to consider the implications of the requirement that $K(\mathbf{x},\mathbf{y})$ be reversible: $p(\mathbf{x})K(\mathbf{x},\mathbf{y}) = p(\mathbf{y})K(\mathbf{y},\mathbf{x})$. For $\mathbf{y} \neq \mathbf{x}$ it implies that

$$p(\mathbf{x})q(\mathbf{x},\mathbf{y})\alpha(\mathbf{x},\mathbf{y}) = p(\mathbf{y})q(\mathbf{y},\mathbf{x})\alpha(\mathbf{y},\mathbf{x}).$$

Suppose (without loss of generality) that $p(\mathbf{x})q(\mathbf{x},\mathbf{y}) \geq p(\mathbf{y})q(\mathbf{y},\mathbf{x})$. If we take $\alpha(\mathbf{y},\mathbf{x}) = 1$ and $\alpha(\mathbf{x},\mathbf{y}) = p(\mathbf{y})q(\mathbf{y},\mathbf{x})/p(\mathbf{x})q(\mathbf{x},\mathbf{y})$, this equality is satisfied.

In implementing the Metropolis-Hastings algorithm, the transition probab ility density function must share two important properties. First, it must be possibl e to generate $\mathbf{y}$ efficiently from $q(\mathbf{x},\mathbf{y})$. All the methods of Sections 3 and 4 are potential tools for these drawings. (Once again, acceptance sampling is attractive relative to i mportance sampling.) A second key characteristic of a satisfactory transition process is that the unconditional

acceptance rate not be so low that the time required to generate a suffi cient number of distinct $\mathbf{x}^t$ is too great.

6.2 Mathematical background

Let $\{\mathbf{x}^t\}_{t=0}^{\infty}$ be a Markov chain defined on $D \subseteq \Re^m$ with transition kernel $\mathrm{K}: D \times D \to \Re^+$ such that, with respect to a σ-finite measure ν on the Borel σ-field of $\Re^m$, for ν-measurable A,

$$\mathrm{P}\left(\mathbf{x}^t \in A \middle| \mathbf{x}^{t-1} = \mathbf{x}\right) = \int_A \mathrm{K}(\mathbf{x}, \mathbf{y}) d\nu(\mathbf{y}) + \mathrm{r}(\mathbf{x}) \chi_A(\mathbf{x}),$$

$$\text{where } \mathrm{r}(\mathbf{x}) = 1 - \int_D \mathrm{K}(\mathbf{x}, \mathbf{y}) d\nu(\mathbf{y}) \text{ and } \chi_A(\mathbf{x}) = \begin{cases} 1 \text{ if } \mathbf{x} \in A \\ 0 \text{ if } \mathbf{x} \notin A \end{cases}.$$

The measure ν will be Lebesgue for continuous distributions and discrete for discrete distributions.

The transition kernel K is substochastic: it defines only the distributi on of accepted candidates. Assume that K has no absorbing states, so that $\mathrm{r}(\mathbf{x}) < 1 \; \forall \, \mathbf{x} \in D$. The corresponding substochastic kernel over t steps is then defined iteratively,

$$\mathrm{K}^{(t)}(\mathbf{x}, \mathbf{y}) = \int \mathrm{K}^{(t-1)}(\mathbf{x}, \mathbf{z}) \mathrm{K}(\mathbf{z}, \mathbf{y}) d\nu(\mathbf{z}) + \mathbf{K}^{(t-1)}(\mathbf{x}, \mathbf{y}) \mathrm{r}(\mathbf{y}) + \left[\mathrm{r}(\mathbf{x})\right]^{t-1} \mathrm{K}(\mathbf{x}, \mathbf{y}).$$

This describes all t-step transitions that involve at least one accepted move. As a functio n of $\mathbf{y}$ it is the p.d.f. with respect to ν of $\mathbf{x}^t$, given $\mathbf{x}^0 = \mathbf{x}$, excluding realizations with $\mathbf{x}^t = \mathbf{x} \; \forall \, j = 1, \ldots, t$.

An invariant distribution for the Markov chain is a function $\mathrm{p}(\mathbf{x})$ that satisfies

$$\mathrm{P}(A) = \int_A \mathrm{p}(\mathbf{x}) d\nu(\mathbf{x}) = \int_D \left\{ \int_A \mathrm{K}(\mathbf{x}, \mathbf{y}) d\nu(\mathbf{y}) + \mathrm{r}(\mathbf{x}) \chi_A(\mathbf{x}) \right\} \mathrm{p}(\mathbf{x}) d\nu(\mathbf{x})$$

$$= \int_D \mathrm{P}\left(\mathbf{x}^t \in A \middle| \mathbf{x}^{t-1} = \mathbf{x}\right) \mathrm{p}(\mathbf{x}) d\nu(\mathbf{x})$$

for all ν-measurable A. Let $D^* = \{\mathbf{x} \in D : \mathrm{p}(\mathbf{x}) > 0\}$. The kernel K is p-*irreducible* if for all $\mathbf{x} \in D^*$, $\mathrm{P}(A) > 0$ implies that $\mathrm{P}\left(\mathbf{x}^t \in A \middle| \mathbf{x}^0 = \mathbf{x}\right) > 0$ for some $t \geq 1$. It is *aperiodic* if there exists no ν-measurable partition $D = \bigcup_{s=0}^{r-1} B_s$ $(r \geq 2)$ such that

$$\mathrm{P}\left(\mathbf{x}^t \in B_{t \bmod (r)} \middle| \mathbf{x}^0 = \mathbf{x} \in B_0\right) = 1 \quad \forall \, t.$$

Define $|\mathrm{f}| = \int_D |\mathrm{f}(\mathbf{x})| d\nu(\mathbf{x})$ for all ν-measurable functions f defined on D. If K is p-irreducible and aperiodic, then

(A) For all $\mathbf{x}^0 \in D$, $\lim_{t \to \infty} \left| \mathrm{K}^{(t)} - \mathrm{p} \right| = 0$;

(B) If g is p-integrable, then for all $\mathbf{x}^0 \in D$,

$$N^{-1} \sum_{t=1}^{N} \mathrm{g}\left(\mathbf{x}^t\right) \xrightarrow{a.s.} \int_D \mathrm{g}(\mathbf{x}) \mathrm{p}(\mathbf{x}) d\nu(\mathbf{x})$$

(Tierney, 1991b, based on Numelin, 1984).

The kernel K is Harris recurrent if $\mathrm{P}\left[\mathbf{x}^t \in B \text{ i.o.}\right] = 1$ for all ν-measurable B with $\int_B \mathrm{p}(\mathbf{x}) d\nu(\mathbf{x}) > 0$ and all $\mathbf{x}^0 \in D$. (A general discussion of recurrence is provided by Numelin (1984, Chapter 3).) If K is p-irreducible and Harris recurre nt, then

(C) The invariant probability distribution $\mathrm{p}(\mathbf{x})$ is unique.

(Numelin, 1984, Corollary 5.2; Tierney, 1991b, Section 3.1). Harris r ecurrence eliminates situations like the one shown in Figure 4, where the support is disconne cted and the Markov chain is the Gibbs sampler. Note that if $\mathbf{x}^0 \in D_i$, it is impossible that $\mathbf{x}^t \in D_j$ $(j \neq i,$ any $t > 0)$. In the situation portrayed in Figure 4, there are two invariant distributions, one for D_1 (reached if $\mathbf{x}^0 \in D_1$) and one for D_2 (reached if $\mathbf{x}^0 \in D_2$).

6.3 Convergence of the Gibbs sampler

The Gibbs sampler requires that the conditional probability density func tions

$$\mathrm{p}\left[\mathbf{x}_{(i)} \middle| \mathbf{x}_{(-i)}\right] = \mathrm{p}(\mathbf{x}) \Big/ \int_{x_{(i)}} \mathrm{p}(\mathbf{x}) d\nu_i\left(x_{(i)}\right) \quad (i = 1, \ldots, k)$$

be well-defined on their supports. In this case the transition kernel d ensity is

$$\mathrm{K}_G(\mathbf{x}, \mathbf{y}) = \prod_{1=1}^{k} \mathrm{p}\left[\mathbf{y}_{(1)} \middle| \mathbf{x}_{(j)}(j > 1), \mathbf{y}_{(j)}(j < 1)\right].$$

If $\mathbf{x}^0 \in D$, then $\mathrm{p}(\mathbf{x})$ is the density of an invariant distribution of the chain defined by K_G:

$$\int_D \mathrm{K}_G(\mathbf{x}, \mathbf{y})\, \mathrm{p}(\mathbf{x}) d\nu(\mathbf{x})$$

$$= \mathrm{p}\left(\mathbf{y}_{(k)} \middle| \mathbf{y}_{(-k)}\right) \int \mathrm{p}\left[\mathbf{y}_{(k-1)} \middle| \mathbf{x}_{(k)}, \mathbf{y}_{(j)}(j < k-1)\right] \int \mathrm{p}\left[\mathbf{y}_{(k-2)} \middle| \mathbf{x}_{(k)}, \mathbf{x}_{(k-1)}, \mathbf{y}_{(j)}(j < k-2)\right]$$

$$\llcorner \int \mathrm{p}\left[\mathbf{y}_{(2)} \middle| \mathbf{y}_{(1)}, \mathbf{x}_{(j)}(j > 2)\right] \int \mathrm{p}\left[\mathbf{y}_{(1)} \middle| \mathbf{x}_{(j)}(j > 1)\right] \int \mathrm{p}\left[\mathbf{x}_{(1)} \middle| \mathbf{x}_{(j)}(j > 1)\right] d\nu_1\left(\mathbf{x}_{(1)}\right)$$

$$\mathrm{p}\left[\mathbf{x}_{(2)} \middle| \mathbf{x}_{(j)}(j > 2)\right] d\nu_2\left(\mathbf{x}_{(2)}\right) \mathrm{p}\left[\mathbf{x}_{(3)} \middle| \mathbf{x}_{(j)}(j > 3)\right] d\nu_3\left(\mathbf{x}_{(3)}\right)$$

$$\llcorner \quad \mathrm{p}\left[\mathbf{x}_{(k-1)} \middle| \mathbf{x}_{(k)}\right] d\nu_{k-1}\left(\mathbf{x}_{(k-1)}\right) \mathrm{p}\left[\mathbf{x}_{(k)}\right] d\nu_k\left(\mathbf{x}_{(k)}\right)$$

$$= \mathrm{p}\left(\mathbf{y}_{(k)} \middle| \mathbf{y}_{(-k)}\right) \int \mathrm{p}\left[\mathbf{y}_{(k-1)} \middle| \mathbf{x}_{(k)}, \mathbf{y}_{(j)}(j < k-1)\right] \int \mathrm{p}\left[\mathbf{y}_{(k-2)} \middle| \mathbf{x}_{(k)}, \mathbf{x}_{(k-1)}, \mathbf{y}_{(j)}(j < k-2)\right]$$

$$\llcorner \int \mathrm{p}\left[\mathbf{y}_{(2)} \middle| \mathbf{y}_{(1)}, \mathbf{x}_{(j)}(j > 2)\right] \int \mathrm{p}\left[\mathbf{y}_{(1)} \middle| \mathbf{x}_{(j)}(j > 2)\right] \mathrm{p}\left[\mathbf{x}_{(3)} \middle| \mathbf{x}_{(j)}(j > 3)\right] d\nu_3\left(\mathbf{x}_{(3)}\right)$$

$$\llcorner \quad \mathrm{p}\left[\mathbf{x}_{(k-1)} \middle| \mathbf{x}_{(k)}\right] d\nu_{k-1}\left(\mathbf{x}_{(k-1)}\right) \mathrm{p}\left[\mathbf{x}_{(k)}\right] d\nu_k\left(\mathbf{x}_{(k)}\right)$$

$$= \mathrm{p}\left(\mathbf{y}_{(k)} \middle| \mathbf{y}_{(-k)}\right) \int \mathrm{p}\left[\mathbf{y}_{(k-1)} \middle| \mathbf{x}_{(k)}, \mathbf{y}_{(j)} (j < k-1)\right] \int \mathrm{p}\left[\mathbf{y}_{(k-2)} \middle| \mathbf{x}_{(k)}, \mathbf{x}_{(k-1)}, \mathbf{y}_{(j)} (j < k-2)\right]$$

$$\mathrm{L} \quad \int \mathrm{p}\left[\mathbf{y}_{(1)}, \mathbf{y}_{(2)} \middle| \mathbf{x}_{(j)} (j > 3)\right] \quad \mathrm{L} \quad \mathrm{p}\left[\mathbf{x}_{(k-1)} \middle| \mathbf{x}_{(k)}\right] d\nu_{k-1}\left(\mathbf{x}_{(k-1)}\right) \mathrm{p}\left[\mathbf{x}_{(k)}\right] d\nu_k\left(\mathbf{x}_{(k)}\right)$$

$$= \quad \mathrm{L}$$

$$= \mathrm{p}\left(\mathbf{y}_{(k)} \middle| \mathbf{y}_{(-k)}\right) \int \mathrm{p}\left[\mathbf{y}_{(k-1)} \middle| \mathbf{x}_{(k)}, \mathbf{y}_{(j)} (j < k-1)\right] \int \mathrm{p}\left[\mathbf{y}_{(k-2)} \middle| \mathbf{x}_{(k)}, \mathbf{x}_{(k-1)}, \mathbf{y}_{(j)} (j < k-2)\right]$$
$$\cdot \mathrm{p}\left[\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \mathrm{K}\ , \mathbf{y}_{(k-3)} \middle| \mathbf{x}_{(k-1)}, \mathbf{x}_{(k)}\right] \mathrm{p}\left[\mathbf{x}_{(k-1)} \middle| \mathbf{x}_{(k)}\right] d\nu_{k-1}\left(\mathbf{x}_{(k-1)}\right) \mathrm{p}\left[\mathbf{x}_{(k)}\right] d\nu_k\left(\mathbf{x}_{(k)}\right)$$

$$= \mathrm{p}\left(\mathbf{y}_{(k)} \middle| \mathbf{y}_{(-k)}\right) \int \mathrm{p}\left[\mathbf{y}_{(k-1)} \middle| \mathbf{x}_{(k)}, \mathbf{y}_{(j)} (j < k-1)\right] \mathrm{p}\left[\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \mathrm{K}\ , \mathbf{y}_{(k-2)} \middle| \mathbf{x}_{(k)}\right] \mathrm{p}\left[\mathbf{x}_{(k)}\right] d\nu_k\left(\mathbf{x}_{(k)}\right)$$

$$= \mathrm{p}\left(\mathbf{y}_{(k)} \middle| \mathbf{y}_{(-k)}\right) \mathrm{p}\left[\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \mathrm{K}\ , \mathbf{y}_{(k-1)}\right] \quad = \quad \mathrm{p}(\mathbf{y}).$$

If ν is discrete, p-irreducibility of K_G is sufficient for results (A), (B), and (C) in Section 6.2 (Tierney, 1991b). The continuous (Lebesgue measure) cas e is technically more difficult, but it may be shown that three simple conditions are jointly sufficient for results (A), (B), and (C) (Roberts and Smith, 1992):

(1) $\mathrm{p}(\mathbf{x})$ is lower semicontinuous at 0;

(2) $\int \mathrm{p}(\mathbf{x}) dx_i$ is locally bounded $(i = 1, \mathrm{K}\ , k)$;

(3) D^* is connected.

A function $\mathrm{h}(\mathbf{x})$ is lower semicontinuous at 0 if, for all $\mathbf{x}$ with $\mathrm{h}(\mathbf{x}) > 0$, there exists an open neighborhood $N_{\mathbf{x}} \supset \mathbf{x}$ and $\varepsilon > 0$ such that for all $\mathbf{y} \in N_{\mathbf{x}}, \mathrm{h}(\mathbf{y}) \geq \varepsilon > 0$. This condition rules out situations like the one shown in Figure 5, where the probability density is uniform on a closed set. For any point $\mathbf{x}$ on the boundary there is no open neighborhood $N_{\mathbf{x}} \supset \mathbf{x}$ such that for all $\mathbf{y} \in N_{\mathbf{x}}$, $\mathrm{h}(\mathbf{y})$ is bounded away from 0. The point A is absorbing.

The local boundedness condition, together with lower semicontinuity at 0 , ensures that the Markov chain is aperiodic. It does so by guaranteeing that for the sequence of support sets $B^t(\mathbf{x}) = \left\{\mathbf{y} \in D^* : \mathrm{K}_G^{(t)}(\mathbf{x}, \mathbf{y}) > 0\right\}$, $B^t(\mathbf{x}) \subseteq B^{t+1}(\mathbf{x})$ for all $t \geq 1$ and all $\mathbf{x} \in D^*$ (Roberts and Smith, 1992, Lemma 3).

Connectedness of D^*, together with conditions (1) and (2), implies that the Gibbs sampler is p-irreducible (Roberts and Smith, 1992, Theorem 2). Condit ions (2) and (3) further imply that the probability measure P corresponding to $\mathrm{p}(\mathbf{x})$ is absolutely

continuous, and consequently (Tierney, 1991b, Corollary 1) the Gibbs s ampler is Harris recurrent. Therefore $\mathrm{p}(\mathbf{x})$ is the unique invariant probability density of the Gibbs sampler.

These conditions are by no means necessary for convergence of the Gibbs sampler; Tierney (1991b) provides substantially weaker conditions. However, th e conditions stated here are satisfied for a very wide range of problems in economics and ar e much easier to verify than the weaker conditions.

6.4 Convergence of the Metropolis-Hastings algorithm

Take the transition probability density function $\mathrm{q}(\mathbf{x},\mathbf{y})$ of Section 6.1 to be a Markov chain kernel with respect to ν, $\mathrm{q}: D^* \times D^* \to \Re^+$. Defining $\alpha: D^* \times D^* \to [0,1]$ as before, define $\mathrm{K}_H: D^* \times D^* \to \Re^+$ by

$$\mathrm{K}_H(\mathbf{x},\mathbf{y}) = \mathrm{q}(\mathbf{x},\mathbf{y})\alpha(\mathbf{x},\mathbf{y}).$$

This is the substochastic kernel governing transitions of the chain from $\mathbf{x}$ to $\mathbf{y}$ that are accepted according to the probability $\alpha(\mathbf{x},\mathbf{y})$. The distribution $\mathrm{p}(\mathbf{x})d\nu(\mathbf{x})$ is invariant if for all ν-measurable sets A,

$$\mathrm{P}(A) = \int_A \mathrm{p}(\mathbf{x})d\nu(\mathbf{x}) = \int_D \mathrm{P}[\mathbf{y} \in A|\mathbf{x}]\,\mathrm{p}(\mathbf{x})d\nu(\mathbf{x}).$$

Recalling that

$$\mathrm{P}[\mathbf{y} \in A|\mathbf{x}] = \int_A \mathrm{K}_H(\mathbf{x},\mathbf{y})d\nu(\mathbf{y}) + \left[1 - \int_D \mathrm{K}_H(\mathbf{x},\mathbf{z})d\nu(\mathbf{z})\right]\chi_A(\mathbf{x}),$$

$$\int_D \mathrm{P}[\mathbf{y} \in A|\mathbf{x}]\,\mathrm{p}(\mathbf{x})d\nu(\mathbf{x})$$

$$= \int_D \int_A \mathrm{K}_H(\mathbf{x},\mathbf{y})d\nu(\mathbf{y})\,\mathrm{p}(\mathbf{x})d\nu(\mathbf{x}) + \int_D \chi_A(\mathbf{x})\,\mathrm{p}(\mathbf{x})d\nu(\mathbf{x}) - \int_D \int_D \mathrm{K}_H(\mathbf{x},\mathbf{y})d\nu(\mathbf{y})\chi_A(\mathbf{x})\,\mathrm{p}(\mathbf{x})d\nu(\mathbf{x})$$

$$= \int_D \int_A \mathrm{K}_H(\mathbf{x},\mathbf{y})d\nu(\mathbf{y})\,\mathrm{p}(\mathbf{x})d\nu(\mathbf{x}) + \int_A \mathrm{p}(\mathbf{x})d\nu(\mathbf{x}) - \int_A \int_D \mathrm{K}_H(\mathbf{x},\mathbf{y})d\nu(\mathbf{y})\,\mathrm{p}(\mathbf{x})d\nu(\mathbf{x}).$$

Since $\mathrm{p}(\mathbf{x})\mathrm{K}_H(\mathbf{x},\mathbf{y}) = \min[\mathrm{p}(\mathbf{y})\mathrm{q}(\mathbf{y},\mathbf{x}), \mathrm{p}(\mathbf{x})\mathrm{q}(\mathbf{x},\mathbf{y})]$ is symmetric in $\mathbf{x}$ and $\mathbf{y}$, the last expression reduces to $\int_A \mathrm{p}(\mathbf{x})d\nu(\mathbf{x}) = \mathrm{P}(\mathbf{x} \in A)$.

From this derivation it is clear that invariance is unaffected by an arb itrary scaling of $\mathrm{K}_H(\mathbf{x},\mathbf{y})$ by a constant c. The choice of c affects the properties of the Metropolis-Hastings algorithm in important practical ways. Larger values of c result in fewer rejected

draws but slower convergence to $p(\mathbf{x})$, whereas smaller values of c increase the proportion of rejected candidates but accelerate the rate of convergence to $p(\mathbf{x})$.

Roberts and Smith (1992) show that the convergence properties of the Hastings-Metropolis algorithm are inherited from those of $q(\mathbf{x},\mathbf{y})$: if q is aperiodic and p-irreducible, then so is the Hastings-Metropolis algorithm. If $q(\mathbf{x},\mathbf{y})$ is constructed as a Gibbs sampler (as is often the case), then the conditions set forth in Section 6.3 may be used to verify aperiodicity and p-irreducibility. A Hastings-Metropolis chain is always Harris recurrent, and therefore the invariant distribution p is unique.

6.5 Assessing convergence and numerical accuracy

In any practical application one is concerned with the discrepancy between $\mathrm{E}[g(\mathbf{x})] = \int_D g(\mathbf{x})p(\mathbf{x})d\mathbf{x}$ and its numerical approximation $N^{-1}\sum_{i=1}^{N} g(\mathbf{x}_i)$. Consider the decomposition

$$N^{-1}\sum_{t=1}^{N} g(\mathbf{x}^t) - \mathrm{E}[g(\mathbf{x})] = \left\{\mathrm{E}\left[N^{-1}\sum_{t=1}^{N} g(\mathbf{x}^t)\middle|\mathbf{x}^0\right] - \mathrm{E}[g(\mathbf{x})]\right\}$$
$$+\left\{N^{-1}\sum_{t=1}^{N} g(\mathbf{x}^t) - \mathrm{E}\left[N^{-1}\sum_{t=1}^{N} g(\mathbf{x}^t)\middle|\mathbf{x}^0\right]\right\} = \mathrm{A}_N(\mathbf{x}^0) + \mathrm{B}_N(\mathbf{x}^0).$$

The term $\mathrm{A}_N(\mathbf{x}^0)$ is nonstochastic and in general nonzero, but $\lim_{N\to\infty}\mathrm{A}_N(\mathbf{x}^0) = 0$ if conditions set forth earlier in this section are satisfied. The purpose of a transient or burn-in phase is to reduce $\mathrm{A}_N(\mathbf{x}^0)$, but for any finite transient period it will still be the case in general that $\mathrm{A}_N(\mathbf{x}^0) \neq 0$. This difficulty is termed the convergence or sensitivity to initial conditions problem. The term $\mathbf{B}_N(\mathbf{x}_0)$ is stochastic and is the analog of $E_N - E$ or $I_N - I$ for acceptance or importance sampling. This term vanishes as $N \to \infty$, but assessing its size is complicated by the fact that $\{\mathbf{x}^t\}$ is neither independently nor identically distributed. This difficulty may be termed the numerical accuracy problem.

A leading cause of slow convergence is multimodality of the probability distribution, for example, as shown in Figure 6 for a Gibbs sampler. In the limit multimodality approaches disconnectedness of the support, and increasingly large values of N are required for $\mathrm{A}_N(\mathbf{x}^0)$ to be close to 0. This difficulty is essentially undetectable given a single Markov chain: for a chain of any fixed length, one can imagine multimodal distributions for which the probability of leaving the neighborhood of a single mode is arbitrarily small. This sort of convergence problem is precisely the same as the multimodality problem in optimization, where iteration from a single starting value can by itself never guarantee the determination of a global optimum. Multimodal disturbances are difficult to manage by any method, including those discussed in Section 4. In the context of the Markov chain Monte Carlo algorithms, the question may be recast as one of sensitivity

to initial conditions: $\mathbf{x}_A^0$, $\mathbf{x}_B^0$, and $\mathbf{x}_C^0$ will lead to quite different chains, in Figure 6, unless the simulations are sufficiently long.

A Markov chain Monte Carlo algorithm can be made fully robust against se nsitivity to initial conditions by constructing many very long chains. Just how one should trade off the number of chains against their length for a given budget of computation time is problem specific and as a practical matter not yet full understood. Many of the issues involved are discussed by Gelman and Rubin (1992), Geyer (1992), and their discus sants and cited works. In an extreme variant of the multiple chains approach, the chain is restarted many times, with initial values chosen independently and identically distribu ted from an appropriate distribution. But finding an appropriate distribution may b e difficult: one that is too concentrated reintroduces the difficulties exemplified by Figure 6; one that is too diffuse may require excessively long chains for convergence. These problems asi de, proper use of the output of Markov chain Monte Carlo in a situation of multimodality r equires specialized diagnostics; Zellner and Min (1992) have obtained some interesting res ults of this kind. At the other extreme a single starting value is used. This approach provid es the largest number of iterations toward convergence, but diagnostics of the type of problem illustrated in Figure 6 will not be as clear.

In specific circumstances a central limit theorem applies to $\mathrm{B}_N(\mathbf{x}^0)$, which may therefore be used to assess the numerical accuracy problem. To develop one set of such circumstances, suppose that the Markov chain is stationary. This could be guaranteed by drawing $\mathbf{x}^0$ from the stationary distribution. Such a drawing would be time consumi ng (if not, i.i.d. sampling from $\mathrm{p}(\mathbf{x})$ is possible), but only one is required. Alternatively, one could iterate the chain many times beginning from an arbitrary initial v alue, discard all but the last iteration, and take this value as drawn from the stationary dis tribution to begin a new chain. Suppose $\mathrm{var}_{\mathrm{p}}[\mathrm{g}(\mathbf{x})]$ is finite and denote $\gamma_i = \mathrm{cov}_{\mathrm{K}}\left[\mathrm{g}(\mathbf{x}^t), \mathrm{g}(\mathbf{x}^{t+i})\right]$. A Markov chain with kernel K is *reversible* if $\mathrm{K}(\mathbf{x},\mathbf{y}) = \mathrm{K}(\mathbf{y},\mathbf{x})$ for all $\mathbf{x},\mathbf{y} \in D$. Hastings-Metropolis chains are always reversible; Gibbs sampling chains are not (Geyer, 199 2, Section 2). If the Markov chain is stationary, p-irreducible, and reversible, then

$$N \operatorname{var}(g_N) \xrightarrow{a.s.} \sigma^2 = \sum_{i=-\infty}^{\infty} \gamma_i ,$$

and if $\sigma^2 < \infty$, then

$$\sqrt{N}(g_N - G) \xrightarrow{d} \mathrm{N}(0, \sigma^2)$$

(Kipnis and Varadhan, 1986).

In the absence of reversibility, known sufficient conditions for central limit theorems are strong and difficult to verify. For example, if for some $m < \infty$ $\mathrm{P}(\mathbf{x}^{t+m} \in A | \mathbf{x}^t = \mathbf{x}) \Big/ \int_{\mathbf{A}} \mathrm{p}(\mathbf{x}) dv(\mathbf{x})$ is bounded below uniformly in $\mathbf{x}$, then D is a small state

space and $\{\mathbf{x}^t\}$ is uniformly ergodic (Tierney, 1991b, Proposition 2). Then if $\mathrm{var}_p[g(\mathbf{x})]$ is finite, there exists $\sigma^2 < \infty$ such that $\sqrt{N}(g_N - G) \xrightarrow{d} \mathrm{N}(0, \sigma^2)$. The boundedness condition, however, is generally difficult to establish.

In neither circumstance is there a known sufficient condition for approx imation of the variance term σ^2 of the central limit theorem. The problem is formally quite similar to estimating the variance of the sample mean $\bar{z}_N = N^{-1}\sum_{t=1}^{N} z_t$ of a stationary time series $\{z_t\}$. In the time series problem, well-established mixing conditions (rates o f decay for $\mathrm{cov}(z_t, z_{t+i})$) are sufficient for consistent estimation of $\mathrm{var}(\bar{z}_N)$ (e.g., Hannan, 1970, pp. 207-210). In time series applications these conditions remain assumpti ons. The difficulty in applying these conditions to Markov chain Monte Carlo is that they ca nnot be established from verifiable fundamentals.

Nevertheless, applications of the time series procedures as if sufficien t mixing conditions obtain appear to give quite reliable results for real problem s in economics. That is, applying a central limit theorem as if the output of the Markov chai n Monte Carlo algorithm were a stationary process satisfying the mixing conditions yie lds accurate probability statements about the output of the same algorithm applied to the same problem with a new starting value and initial seed for the random number generat or (Geweke, 1992a; Geyer, 1992). This leads to a conservative but practical procedure for assessing the accuracy and reliability of Markov chain Monte Carlo. First, execute se veral short runs -- a burn-in of 50 to 100 iterations followed by a chain of length $N = 500$ or $N = 1000$ is sufficient for many problems. Examine the g_N and their standard errors as assessed by conventional time series procedures for a single time series to see whet her the scatter of each g_N across the short runs is consistent with these standard errors. If nec essary, increase the length of the short runs until this consistency is achieved . Second, choose the last value of one of the short runs, and use it as the starting value of a long run of from $N = 10^4$ to $N = 10^6$ iterations. As a final check, compare the g_N from the single long run with the confidence intervals implied by the short runs. Report the fin al value of g_N, together with its numerical standard error as computed by time series me thods for a single series.

7. Some examples

The usefulness of all of these methods lies as much in their appropriate combination as in the application of any one individually. We turn now to some example s that illustrate some useful combinations and in the process treat a few topics closely r elated to integration and simulation.

7.1 Stochastic volatility

Models in which the volatility of asset returns varies smoothly over tim e have received considerable attention in recent years. (For a survey of several appro aches, see Bollerslev, Chou, and Kroner, 1992.) Persistent but changing volatility is an evid ent characteristic of returns data. Since the conditional distribution of returns is relevant in the theory of portfolio allocation, proper treatment of volatility is important. Time -varying volatility also affects the properties of real growth and business cycle models.

A simple model of time-varying volatility is the stochastic volatility m odel, the descriptive properties of which have been examined by a series of invest igators beginning with Taylor (1986). The approach here closely follows that of Jacquie r, Polson, and Rossi (1994). Let r_t denote the one-period return of a single asset, and let $\mathbf{x}_t$ be a vector of deterministic time series such as indicators for day of the week, holida ys, etc. A simple stochastic volatility model is

$$r_t = \beta' \mathbf{x}_t + \varepsilon_t, \quad \varepsilon_t = h_t^{1/2} u_t \tag{7.1.1}$$

$$\log h_t = \alpha + \delta \log h_{t-1} + \sigma_v v_t \tag{7.1.2}$$

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} \overset{IID}{\sim} \mathrm{N}(0, \mathbf{I}_2). \tag{7.1.3}$$

At time T an economic agent is concerned with future returns $r_{T+1}, \mathrm{K}, r_{T+q}$ through an expected utility function

$$\mathrm{E}\left[\mathrm{V}\left(r_{T+1}, \mathrm{K}, r_{T+q}; \mathbf{z}\right) \middle| \Phi_T\right] = \mathrm{E}\left[\mathrm{V}\left(\mathbf{r}_q; \mathrm{K}\right) \middle| \Phi_T\right], \tag{7.1.4}$$

where $\mathbf{z}$ is a generic vector of other arguments which may be known or unknown at time T.

Evaluation of this expected utility function requires the solution of an integration problem. We will consider this problem for three different specificatio ns of the information set Φ_T in turn. Denoting $\mathbf{r}_T = \left(r_1, \mathrm{K}, r_T\right)'$, $\mathbf{x}_{T+q} = \left(x_1, \mathrm{K}, \mathbf{x}_{T+q}\right)'$, $\theta' = \left(\beta', \alpha, \delta, \sigma_v\right)$, and $\mathbf{h}_T = \left(h_1, \mathrm{K}, h_T\right)'$, these are

$$\Phi_T^{(1)} = \left\{\mathbf{r}_T, \mathbf{x}_{T+q}, \theta, \mathbf{h}_T\right\}; \quad \Phi_T^{(2)} = \left\{\mathbf{r}_T, \mathbf{x}_{T+q}, \theta\right\}; \quad \Phi_T^{(3)} = \left\{\mathbf{r}_T, \mathbf{x}_{T+q}\right\}.$$

As one may readily verify, deterministic approximations of the type disc ussed in Section 2 are inconvenient for this problem. Even explicit expressions of the int egrals in closed form

are awkward and unrevealing. Simulation methods are much more direct an d have the added advantage that one set of simulations can suffice for several alte rnative values of the other arguments $\mathbf{z}$ in (7.1.4). These arguments might include taste parameters or the va lues of decision variables which themselves do not affect $\mathbf{r}_q$. (Section 7.2 provides an example involving explicit optimization.)

The solution for the problem for $\Phi_T^{(1)}$ is simple. In the notation of Section 4, repeated period-by-period simulation of $\mathbf{x} = \mathbf{r}_q$ provides an independent identically distributed sample $\left\{\tilde{\mathbf{r}}_q^{(i)}\right\}_{i=1}^N$ with a probability density $\mathrm{p}(\mathbf{x}) = \mathrm{p}\left(\mathbf{r}_q \middle| \Phi_T^{(1)}\right)$ that we have not even expressed. Then

$$\mathrm{E}\left[\mathrm{V}\left(r_{T+1}, \mathrm{K}, r_{T+q}; \mathrm{K}\right) \middle| \Phi_T^{(1)}\right] = \int \mathrm{g}(\mathbf{x})\,\mathrm{p}(\mathbf{x})\,d\mathbf{x},$$

where $\mathrm{g}(\mathbf{x}) = \mathrm{V}(\mathbf{x}; \mathrm{K}) = \mathrm{V}\left(r_{T+1}, \mathrm{K}, r_{T+q}; \mathrm{K}\right)$. Consequently,

$$\mathrm{E}\left[\mathrm{V}\left(r_{T+1}, \mathrm{K}, r_{T+q}; \mathrm{K}\right) \middle| \Phi_T^{(1)}\right] \approx N^{-1} \sum_{i=1}^N \mathrm{V}\left(\tilde{\mathbf{r}}_q^{(i)}\right).$$

The problem for $\Phi_T^{(2)}$ is more difficult. Rather than $\mathbf{h}_T$ itself, the agent has available only

$$\begin{aligned}
\mathrm{p}\left(\mathbf{h}_T \middle| \mathbf{r}_T, \mathbf{x}_T, \theta\right) &= \mathrm{p}\left(\mathbf{h}_T, \mathbf{r}_T \middle| \mathbf{x}_T, \theta\right) \Big/ \mathrm{p}\left(\mathbf{r}_T\right) \\
&= \mathrm{p}\left(\mathbf{r}_T \middle| \mathbf{h}_T, \mathbf{x}_T, \theta\right) \mathrm{p}\left(\mathbf{h}_T \middle| \mathbf{x}_T, \theta\right) \Big/ \mathrm{p}\left(\mathbf{r}_T\right) \propto \mathrm{p}\left(\mathbf{r}_T \middle| \mathbf{h}_T, \mathbf{x}_T, \theta\right) \mathrm{p}\left(\mathbf{h}_T \middle| \mathbf{x}_T, \theta\right) \\
&= (2\pi)^{-T/2} \prod_{t=1}^T h_t^{-1/2} \exp\left[-\sum_{t=1}^T \varepsilon_t^2 \Big/ 2h_t\right] \\
&\quad \cdot (2\pi)^{-T/2} \sigma_v^{-T} h_t^{-1} \exp\left[-\sum_{t=1}^T \left(\log h_t - \alpha - \delta \log h_{t-1}\right)^2 \Big/ 2\sigma_v^2\right] \\
&\propto \prod_{t=1}^T h_t^{-3/2} \exp\left[-\sum_{t=1}^T \varepsilon_t^2 \Big/ 2h_t\right] \exp\left[-\sum_{t=1}^T \left(\log h_t - \alpha - \delta \log h_{t-1}\right)^2 \Big/ 2\sigma_v^2\right], \qquad (7.1.5)
\end{aligned}$$

where $\varepsilon_t = r_t - \beta' \mathbf{x}_t$. The simple Monte Carlo solution of the previous problem could be extended to this one if one could draw an i.i.d. sample $\left\{\tilde{\mathbf{h}}_T^{(i)}\right\}_{i=1}^N$ from the distribution implied by the last kernel. This is clearly not possible, nor are there obvious source or importance sampling distributions for the methods of Sections 4.2 or 4.3.

This problem can be solved in a number of ways, and a comparison of thre e alternatives is instructive. All begin with the kernels of the conditional probabili ty densities for individual h_t implied by (7.1.5). For $t = 2, \mathrm{K}, T-1$ the kernel is

$$\mathrm{p}\left[h_t \middle| h_s (t \neq s), \theta, \varepsilon_t\right] \propto h_t^{-3/2} \exp\left(-\varepsilon_t^2 / 2h_t\right) \exp\left[-\left(\log h_t - \mu_t\right)^2 \Big/ 2\sigma^2\right], \qquad (7.1.6)$$

where

$$\mu_t = \frac{\alpha(1-\delta) + \delta\left(\log h_{t-1} + \log h_{t+1}\right)}{1+\delta^2}, \quad \sigma^2 = \frac{\sigma_v^2}{1+\delta^2}.$$

(Similar expressions for h_1 and h_T may be constructed.)

The first two approaches construct a Gibbs sampler for the h_t, drawing and successively replacing $h_1, h_2, \mathrm{K}, h_T$. Each cycle of drawing and replacement produces the next realization of $\tilde{\mathbf{h}}_T^{(i)}$ in the Markov chain. Note from (7.1.5) that $\lim_{h_t \to 0} \mathrm{p}\left(h_t \middle| \mathbf{r}_t, \mathbf{x}_t, \theta\right) = 0$ for any $t = 1, \mathrm{K}, T$, and since the support of $\mathbf{h}_T$ is the positive orthant of $\Re^T$ the probability density function of $\mathbf{h}_T$ is lower semicontinuous at 0. The remaining sufficient conditions for convergence of the Gibbs sampler are clearly satisfied. Conditional on each $\tilde{\mathbf{h}}_T^{(i)}$ in the chain, draw a single $\tilde{\mathbf{r}}_q^{(i)}$ as in the problem for $\Phi_T^{(1)}$. Since $\left|\mathrm{p}\left(\tilde{\mathbf{h}}_T^{(i)}\right) - \mathrm{p}\left(\mathbf{h}_T \middle| \mathbf{r}_T, \mathbf{x}_T, \theta\right)\right| \to 0$, it follows that $\left|\mathrm{p}\left(\tilde{\mathbf{r}}_q^{(i)}\right) - \mathrm{p}\left(\mathbf{r}_q \middle| \mathbf{r}_T, \mathbf{x}_{T+q}, \theta\right)\right| \to 0$. Both approaches work directly with the conditional distribution of $H_t = \log h_t$, which from (7.1.6) is given by

$$\log \mathrm{p}\left(H_t \middle| H_s (s \neq t), \theta, \varepsilon_t\right) = -\exp\left(-\varepsilon_t^2/2\right)\exp\left(-H_t\right) - \left(H_t - \mu_t^*\right)^2 \big/ 2\sigma^2 \qquad (7.1.7)$$

(up to an additive constant), where $\mu_t^* = \mu_t - .5\sigma^2$, but differ in the method for obtaining H_t.

The first approach is to use acceptance sampling. A reasonable source d istribution is $\mathrm{N}\left(\mu_t^*, \sigma^2\right)$, for which the acceptance probability is

$$\exp\left[-\left(\varepsilon_t^2/2\right)\exp\left(-H_t\right)\right] = \exp\left(-\varepsilon_t^2/2h_t\right).$$

The acceptance probability falls below .01 if and only if ε_t^2/h_t exceeds 9.2, which is highly unlikely if the model reasonably well describes the distribution of the returns r_t. The acceptance probability could be improved somewhat using the optimizing p rocedures set out in Section 3.2, but given the favorable acceptance probabilities for the $\mathrm{N}\left(\mu_t^*, \sigma^2\right)$ source distribution, the additional overhead might not be warranted.

The second approach is to note that the log-conditional kernel densities (7.1.7) are strictly concave and apply the adaptive method of Gilks and Wild (1992) . Their algorithm (described in Section 3.2) may be initialized by noting that $H_t = \mu_t^*$ lies to the left of the mode of the log-conditional and a solution of $\left(1 - H_t + H_t^2/2\right)\exp\left(-\varepsilon_t^2/2\right) - \left(H_t - \mu_t^*\right)/\sigma^2$ lies to the right of the mode. Except for the method of drawing H_t, the solution of the problem proceeds as in the first approach.

The third approach is to construct a Metropolis-Hastings independence ch ain. This is done by forming a Metropolis step M_t for each h_t and then combining all T steps into a single transition $M = M_1 M_2 \mathrm{K}\ M_T$. At each M_t either a candidate new value is accepted or the old value of h_t is retained. Thus, when M operates on the old $\mathbf{h}_T$ it generally produces a mixture of old and new h_t in the new $\mathbf{h}_T$. The transition kernel M is p-irreducible and aperiodic, and an argument like the one in Section 6.4 shows that $\mathrm{p}\left(\mathbf{h}_T \middle| \mathbf{r}_T, \mathbf{x}_T, \theta\right)$ is the invariant distribution of M (Jacquier, Polson, and Rossi, 1994, Section 2). A useful distribution for the Metropolis-Hastings independence chain is the gamma distribution for h_t^{-1} with shape parameter $a = \left[1 - 2\exp\left(\sigma^2\right)\right] \big/ \left[1 - \exp\left(\sigma^2\right)\right] + .5$ and scale parameter

$\lambda = (a-1)\exp\left(\mu_t + .5\sigma^2\right) + .5\varepsilon_t^2$. Combined with an appropriate scaling of the transition kernel, as discussed in Section 6.4, this chain produces convergence at a practical rate (see Jacquier, Polson, and Rossi, 1994, Section 2.4, for details).

The solution of the problem for $\Phi_T^{(2)}$ is directly usable in the solution of the problem for $\Phi_T^{(3)}$, in the context of the Gibbs sampler. From the form of (7.1.1)-(7.1 .3) the probability density kernel for θ and $\mathbf{h}_T$ underlying the expectations operator in (7.1.4) is

$$\begin{aligned} &\prod_{t=1}^{T} h_t^{-1/2} \exp\left[-\sum_{t=1}^{T}\left(r_t - \beta'\mathbf{x}_t\right)^2 / 2h_t\right] \\ &\quad \cdot \sigma_v^{-T} \exp\left[-\sum_{t=1}^{T}\left(\log h_t - \alpha - \delta \log h_{t-1}\right)^2 / 2\sigma_v^2\right] \mathrm{p}\left(\beta, \alpha, \delta, \sigma_{\mathrm{v}}\right), \end{aligned} \tag{7.1.8}$$

where $\mathrm{p}\left(\beta, \alpha, \delta, \sigma_{\mathrm{v}}\right)$ is the prior probability density function of $\theta' = \left(\beta', \alpha, \delta, \sigma_v\right)$. A Gibbs sampler with blocking $\left(\mathbf{h}_T, \theta\right)$ will alternate drawing and substitution for $\mathbf{h}_T \left| \mathbf{r}_T, \mathbf{x}_T, \theta \right.$ and $\theta \left| \mathbf{r}_T, \mathbf{x}_T, \mathbf{h}_T \right.$. The drawing for $\mathbf{h}_T$ is the same one constructed to solve the problem for $\Phi_T^{(2)}$. The second drawing is facilitated by noting that the kernel of (7.1.8) in θ may be expressed

$$\begin{aligned} &\propto \prod_{t=1}^{T} \exp\left[-\sum_{t=1}^{T}\left(r_t - \beta'\mathbf{x}_t\right)_t^2 / 2h_t\right] \\ &\quad \cdot \sigma_v^{-(T+1)} \exp\left[-\sum_{t=1}^{T}\left(\log h_t - \alpha - \delta \log h_{t-1}\right)^2 / 2\sigma_v^2\right] \end{aligned}$$

if the prior probability distribution has the conventional improper kern el $\mathrm{p}\left(\beta, \alpha, \delta, \sigma_{\mathrm{v}}\right) \propto \sigma_v^{-1}$. Thus, β and $\left(\alpha, \delta, \sigma_v\right)$ are conditionally independent. In each case the distribution follows from standard treatments of Bayesian learning a bout a linear model (e.g., Poirier, 1995, Section 9.9):

$$\beta \sim \mathrm{N}\left(\mathbf{b}, \mathbf{Q}^{-1}\right), \text{ where } \mathbf{Q} = \sum_{t=1}^{T} h_t^{-1}\mathbf{x}_t\mathbf{x}_t' \text{ and } \mathbf{b} = \mathbf{Q}^{-1}\sum_{t=1}^{T} h_t^{-1}\mathbf{x}_t\mathbf{r}_t,$$

for β and

$$S^2/\sigma_v^2 \sim \chi^2(T-2), \quad (\alpha, \delta)' \left| \sigma_v \right. \sim \mathrm{N}\left(\mathbf{c}, \sigma_v^2\mathbf{P}^{-1}\right), \text{ where}$$

$$\mathbf{P} = \begin{bmatrix} \mathbf{T} & \sum_{t=1}^{T} \log h_{t-1} \\ \sum_{t=1}^{T} \log h_{t-1} & \sum_{t=1}^{T} \log^2 h_{t-1} \end{bmatrix}, \quad \mathbf{c} = \mathbf{P}^{-1}\begin{bmatrix} \sum_{t=1}^{T} \log h_{t-1} \\ \sum_{t=1}^{T} \log h_t \log h_{t-1} \end{bmatrix},$$

$$\text{and } S^2 = \sum_{t=1}^{T}\left(\log h_t - c_1 - c_2 \log h_{t-1}\right),$$

for $\left(\alpha, \delta, \sigma_v\right)$.

7.2 Integration and optimization

The solution of all but the simplest dynamic optimization problems canno t be expressed in closed form. Since the objective function in these problem s is expected utility, integration is required to evaluate a candidate solution. Finding a goo d numerical approximation to the solution therefore requires optimizations of a func tion which can be evaluated only inexactly. Moreover this evaluation must in general be r epeated many times

in the process of approximating the solution. Several approaches to thi s important problem have been proposed: a good introduction is provided by Taylor and Uhlig (1990) and the papers following that article; more recent work includes McGrattan (199 3). Here we discuss a widely applicable procedure that uses Monte Carlo integration to solve dynamic optimization problems subject to an imposed parameterization of the deci sion rule and then loosens the parametric restrictions so as to approach the optimum. The description here closely follows Smith (1991) who invented the method. The notation an d assumptions are largely those of Stokey and Lucas (1989, Chapter 9).

The problem. Many dynamic optimization problems can be expressed

$$\max_{\{\mathbf{x}_t\}_{t=1}^{\infty}} \mathrm{E}_0 \sum_{t=0}^{\infty} \beta^t \mathrm{r}\left(\underset{p\times 1}{\mathbf{x}_t}, \mathbf{x}_{t+1}, \underset{1\times 1}{\mathbf{z}_t} \right) \tag{7.2.1}$$

$$\text{given } \mathbf{x}_0, \mathbf{z}_0 \text{ and subject to } \mathbf{x}_{t+1} \in \Gamma(\mathbf{x}_t, \mathbf{z}_t) \ \forall\ t.$$

The sequence of state vectors $\{\mathbf{z}_t\}_{t=1}^{\infty}$ is a Markov process with transition density

$$\mathrm{v}(\mathbf{z}_{t+1} | \mathbf{z}_t);\ \mathbf{z}_t \in Z \subseteq \Re^1 \ \forall\ t; \tag{7.2.2}$$

and Z is either compact or countable. The decision vector $\mathbf{x}_t \in X \subseteq \Re^p$; X is closed and convex. The agent observes the state vector $\mathbf{s}_t' = (\mathbf{x}_t', \mathbf{z}_t') \in S = X \times Z$ prior to choosing $\mathbf{x}_{t+1}$. The operator E_0 denotes expectations conditional on the period 0 information set $\mathbf{s}_0$. The return function r is bounded, continuous in $(\mathbf{x}_t, \mathbf{x}_{t+1}, \mathbf{z}_t)$, and concave in $(\mathbf{x}_t, \mathbf{x}_{t+1}) \ \forall\ \mathbf{z}_t \in Z$. The correspondence Γ is nonempty, compact- and convex-valued, and continuous. The convexity of Γ precludes problems with discrete choice sets; for a treatment of discre te choice similar to the one here for continuous choice, see Geweke, Slonim , and Zarkin (1992).

These assumptions imply the existence of a unique, time-invariant contin uous decision rule $w: S \to X$ that expresses optimal $\mathbf{x}_{t+1} = \mathrm{w}(\mathbf{x}_t, \mathbf{z}_t)$ (Stokey and Lucas, 1989, Chapter 9). The optimization problem is to determine the decision rule. The approac h taken here is to replace w with a rule of thumb characterized by a vector of parameters $\underset{k\times 1}{\psi}$:

$$\mathbf{x}_{t+1} = \mathrm{h}(\mathbf{x}_t, \mathbf{z}_t; \psi),\ \psi \in C \subseteq \Re^k, C \text{ compact.} \tag{7.2.3}$$

This rule closes the model. Given $\mathbf{s}_0$, $\mathbf{z} = \{\mathbf{z}_t\}_{t=1}^{T}$, and ψ, (7.2.2)-(7.2.3) determines $\mathbf{x} = \{\mathbf{x}_t\}_{t=1}^{T+1} = \mathrm{q}(\mathbf{z}; \psi, \mathbf{s}_0)$ through the obvious iterations.

Let $\mathrm{b}(\mathbf{x}, \mathbf{z}; \mathbf{s}_0) = \sum_{t=0}^{T} \beta^t \mathrm{r}(\mathbf{x}_t, \mathbf{x}_{t+1}, \mathbf{z}_t)$ denote the utility delivered by the sequences $\mathbf{x}$ and $\mathbf{z}$ given $\mathbf{s}_0$ for the dynamic optimization problem with horizon truncated at T. Repressing $\mathbf{s}_0$ to maintain notational simplicity, $\mathrm{g}(\mathbf{z}, \psi) = \mathrm{b}[\mathrm{q}(\mathbf{z}; \psi, \mathbf{s}_0), \mathbf{z}; \mathbf{s}_0]$ is delivered utility for decision rule h with parameterization ψ. Given h the agent chooses the best possible ψ, which we shall denote

$$\psi_0 = \arg\max_{\psi} \mathrm{E}_0\left[\mathrm{g}(\mathbf{z}, \psi)\right]. \tag{7.2.4}$$

Problem (7.2.4) is a simplification of Problem (7.2.1), but it still cannot be solved analytically. The chief complication is the evaluation of the integral associated with E_0 in (7.2.4). The key idea in the solution described here is to simulate t he behavior of $\mathbf{s}$ for different values of ψ, thereby providing approximations to $\mathrm{E}_0\left[\mathrm{g}(\mathbf{z}, \psi)\right]$. As we shall see, arbitrarily good approximations to ψ_0 may be obtained in this way. By increasing T and employing a sequence of functions h that are increasingly flexible throu gh a longer parameter vector ψ, the solution of (7.2.4) may be made to approximate that of (7.2.1) (Smith, 1991).

The algorithm. Generate n i.i.d. sequences $\tilde{\mathbf{z}}^{(i)} = \left\{\tilde{\mathbf{z}}_t^{(i)}\right\}_{t=1}^{T}$ according to (7.2.2), and take $\Theta = \left\{\tilde{\mathbf{z}}^{(i)}\right\}_{i=1}^{n}$ to be the collection of these sequences. If we let $\mathrm{Q}_n(\Theta, \psi)$ $= \sum_{i=1}^{n} \mathrm{g}\left(\tilde{\mathbf{z}}^{(i)}, \psi\right)$, then $n^{-1}\mathrm{Q}_n(\Theta, \psi) \xrightarrow{a.s.} \mathrm{E}_0\left[\mathrm{g}(\mathbf{z}, \psi)\right]$. Since the set of sequences Θ is fixed,

$$\hat{\psi}_n = \arg\max_{\psi} n^{-1}\mathrm{Q}_n(\Theta, \psi)$$

is a well-defined, deterministic optimization problem that can be solved using standard hill climbing methods. These methods will be more efficient to the extent th at $\partial \mathrm{r}/\partial \mathrm{h}$ and $\partial \mathrm{h}/\partial\psi$ (better yet, $\partial^2 r/\partial \mathrm{h}^2$ and $\partial^2 \mathrm{h}/\partial\psi\partial\psi'$ in addition) can be evaluated analytically.

Asymptotic properties. Given four further assumptions, $\hat{\psi}_n \xrightarrow{a.s.} \psi_0$ and central limit theorems may be used to assess the accuracy of the approximation o f ψ_0 by $\hat{\psi}_n$ and of $\mathrm{E}_0\left[\mathrm{g}(\mathbf{z}, \psi)\right]$ by $n^{-1}\mathrm{Q}_n(\Theta, \psi)$.

(1) $\mathrm{g}(\mathbf{z}, \psi)$ is twice continuously differentiable in ψ for all $\mathbf{z}$.

(2) The following functions are regular:

(a) $\mathrm{g}(\mathbf{z}, \psi)$, $\partial \mathrm{g}(\mathbf{z}, \psi)/\partial\psi$, $\partial^2 \mathrm{g}(\mathbf{z}, \psi)/\partial\psi\partial\psi'$;

(b) $\left[\partial \mathrm{g}(\mathbf{z}, \psi)/\partial\psi\right]\left[\partial \mathrm{g}(\mathbf{z}, \psi)/\partial\psi'\right]$;

(c) $\mathrm{g}^2(\mathbf{z}, \psi)$.

Regular is used in the sense of Tauchen (1985). Denoting the probability den sity function of $\mathbf{z}$ by $\mathrm{f}(\mathbf{z})$, $\mathrm{d}(\mathbf{z}, \psi)$ is regular if

(i) $\mathrm{d}(\mathbf{z}, \psi)$ is measurable in $\mathbf{z} \forall\, \psi \in C$;

(ii) d is separable (Huber, 1967);

(iii) d is dominated -- i.e., $\exists \mathrm{b} \ni \int \mathrm{b}(\mathbf{z}) d\mathbf{z} < \infty$ and $\left|\mathrm{d}(\mathbf{z}, \psi)\right| < \mathrm{b}(\mathbf{z}) \forall\, \psi \in C$;

(iv) $\mathrm{d}(\mathbf{z}, \psi)$ is continuous in $\psi \forall\, \mathbf{z}$.

(3) $\mathrm{E}\left[\mathrm{g}(\mathbf{z},\psi)\right]$ (the existence of which is guaranteed by Assumption 2(a)) is unique ly maximized at ψ_0, an interior point of C.

(4) $\mathrm{E}\left[\partial^2 \mathrm{g}(\mathbf{z},\psi_0)/\partial\psi\partial\psi'\right]$ (the existence of which is also guaranteed by Assumption 2(a)) is nonsingular.

Given these four further assumptions, one can usefully approximate ψ_0:

$$\hat{\psi}_n \xrightarrow{p} \psi_0, \quad n^{1/2}(\hat{\psi}_n - \psi_0) \xrightarrow{d} \mathrm{N}(\mathbf{0},\mathbf{V});$$

$$\mathbf{V} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}, \text{ with } \mathbf{A} = \mathrm{E}\left[\frac{\partial^2 \mathrm{g}(\mathbf{z},\psi_0)}{\partial\psi\partial\psi'}\right], \ \mathbf{B} = \mathrm{E}\left[\frac{\partial \mathrm{g}(\mathbf{z},\psi_0)}{\partial\psi}\frac{\partial \mathrm{g}(\mathbf{z},\psi_0)}{\partial\psi'}\right];$$

$$\hat{\mathbf{A}}_n = n^{-1}\frac{\partial^2 \mathrm{Q}_n(\Theta,\hat{\psi}_n)}{\partial\psi\partial\psi'} \xrightarrow{p} \mathbf{A}, \quad \hat{\mathbf{B}}_n = n^{-1}\sum_{i=1}^{n}\frac{\partial \mathrm{g}(\tilde{\mathbf{z}}^{(i)},\hat{\psi}_n)}{\partial\psi}\frac{\partial \mathrm{g}(\tilde{\mathbf{z}}^{(i)},\hat{\psi}_n)}{\partial\psi'} \xrightarrow{p} \mathbf{B}.$$

Under exactly the same conditions, one can also usefully approximate $\mathrm{E}\left[\mathrm{g}(\mathbf{z},\psi)\right]$:

$$n^{-1}\mathrm{Q}_n(\Theta,\hat{\psi}_n) \xrightarrow{a.s.} \mathrm{E}\left[\mathrm{g}(\mathbf{z},\psi)\right], \quad n^{1/2}\left\{n^{-1}\mathrm{Q}_n(\Theta,\hat{\psi}_n) - \mathrm{E}\left[\mathrm{g}(\mathbf{z},\psi)\right]\right\} \xrightarrow{d} \mathrm{N}(0,\sigma^2);$$

$$\hat{\sigma}_n^2 = n^{-1}\sum_{i=1}^{n}\mathrm{g}^2(\tilde{\mathbf{z}}^{(i)},\hat{\psi}_n) - \left[n^{-1}\mathrm{Q}_n(\Theta,\hat{\psi}_n)\right]^2 \xrightarrow{p} \sigma^2 = \mathrm{var}\left[\mathrm{g}(\mathbf{z},\psi)\right].$$

Proofs are given by Smith (1991) who uses asymptotic theory developed by Amemiya (1985) and Tauchen (1985). The second result is especially useful i n valuing the approximation error: see Smith (1991, Section 5).

Antithetic variables. In many applications the conditional distribution of the exogenous state vector $\mathbf{z}_t$, with probability density function $\mathrm{v}(\mathbf{z}_t|\mathbf{z}_{t-1})$, is smooth and symmetric or nearly symmetric. The return function r is commonly monoto ne increasing or decreasing in each element of $\mathbf{z}_t$ and may be nearly linear over most of the support of the distribution of $\mathbf{z}_t$. In such circumstances there are substantial gains in the use of antit hetic variables as described in Section 5.1. Let $\tilde{\mathbf{z}}^{(i1)}$ and $\tilde{\mathbf{z}}^{(i2)}$ denote such an antithetic pair. (Exactly how the pair is drawn will depend on the particulars of the pr oblem. What is essential, as discussed in Section 5.1, is that $\tilde{\mathbf{z}}^{(i1)}$ and $\tilde{\mathbf{z}}^{(i2)}$ be identically distributed.) Consider $n/2$ replications of $\tilde{\mathbf{z}}^{(i1)}$ and $\tilde{\mathbf{z}}^{(i2)}$ in lieu of n replications of $\tilde{\mathbf{z}}^{(i)}$. Redefine

$$\mathrm{Q}_n(\Theta,\psi) = \sum_{i=1}^{n/2}\left[\mathrm{g}(\tilde{\mathbf{z}}^{(i1)},\psi) + \mathrm{g}(\tilde{\mathbf{z}}^{(i2)},\psi)\right]$$

with $\Theta = \left\{\tilde{\mathbf{z}}^{(i1)},\tilde{\mathbf{z}}^{(i2)}\right\}_{i=1}^{n/2}$ and take $\hat{\psi}_n = \arg\max_{\psi} n^{-1}\mathrm{Q}_n(\Theta,\psi)$. Then $\hat{\psi}_n$ and $n^{-1}\mathrm{Q}_n(\theta,\psi)$ are consistent for ψ and $\mathrm{E}\left[\mathrm{g}(\mathbf{z},\psi)\right]$ as before. There are again central limit theorems, but now

$$\mathbf{V} = \mathbf{A}^{-1}\mathbf{B}^*\mathbf{A}^{-1}, \text{ with } \mathbf{B}^* = \mathbf{B} + \tfrac{1}{2}(\mathbf{C}+\mathbf{C}'), \text{ where}$$

$$\mathbf{C} = \mathrm{E}\left[\frac{\partial \mathrm{g}(\tilde{\mathbf{z}}^{(i1)},\psi_0)}{\partial\psi}\frac{\partial \mathrm{g}(\tilde{\mathbf{z}}^{(i2)},\psi_0)}{\partial\psi'}\right], \quad \hat{\mathbf{C}}_n = \left(\frac{n}{2}\right)^{-1}\sum_{i=1}^{n/2}\frac{\partial \mathrm{g}(\tilde{\mathbf{z}}^{(i1)},\psi_0)}{\partial\psi}\frac{\partial \mathrm{g}(\tilde{\mathbf{z}}^{(i2)},\psi_0)}{\partial\psi'} \xrightarrow{p} \mathbf{C}$$

and

$$\sigma^2 = \text{var}\left[\text{g}\left(\mathbf{z}, \psi_{\mathbf{0}}\right)\right] + \text{cov}\left[\text{g}\left(\tilde{\mathbf{z}}^{(i1)}, \psi_{\mathbf{0}}\right), \text{g}\left(\tilde{\mathbf{z}}^{(i2)}, \psi_{\mathbf{0}}\right)\right],$$

$$n^{-1}\sum_{i=1}^{n/2}\left[\text{g}^2\left(\tilde{\mathbf{z}}^{(i1)}, \hat{\psi}_n\right) + \text{g}^2\left(\tilde{\mathbf{z}}^{(i2)}, \hat{\psi}_n\right)\right] + \left(\frac{n}{2}\right)^{-1}\sum_{i=1}^{n}\text{g}\left(\tilde{\mathbf{z}}^{(i1)}, \hat{\psi}_n\right)\text{g}\left(\tilde{\mathbf{z}}^{(i2)}, \hat{\psi}_n\right)$$

$$-2\left[n^{-1}\text{Q}_n\left(\Theta, \hat{\psi}_n\right)\right]^2 \xrightarrow{p} \sigma^2.$$

Smith (1991) applies this method to a variant of the Brock and Mirman (1972) growth model. The characteristic of the model that is important for the succes s of the use of antithetic variables is that the exogenous state variables move smoothly over time and the return function is only modestly nonlinear over most of the support of $\mathbf{z}$. Using only 100 antithetic pairs and $T = 800$, Smith determines ψ up to four significant figures. The suboptimality of the resulting decision rules turns out to be equivalent to a per-period decrease in consumption of 2×10^{-5}%.

References

Ahrens, J.H., and U. Dieter, 1974, "Computer Methods for Sampling from Gamma, Beta, Poisson, and Binomial Distributions," *Computing* **12**:223-246.

Ahrens, J.H., and U. Dieter, 1980, "Sampling from Binomial and Poisson Distributions: A Method with Bounded Computation Times," *Computing* **25**: 193-208.

Amemiya, T., 1985, *Advanced Econometrics*. Cambridge: Harvard University Press.

Anderson, T.W., 1984, *An Introduction to Multivariate Statistical Analysis*. New York: Wiley. (Second edition)

Bollerslev, T., R. Chou, and K.F. Kroner, 1992, "ARCH Modelling in Finance," *Journal of Econometrics* **52**: 5-59.

Box, G.E.P., and M.E. Muller, 1958, "A Note on the Generation of Random Normal Deviates," *Annals of Mathematical Statistics* **29**: 610-611.

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

Brock, W.A., and L.J. Mirman, 1972, "Optimal Economic Growth and Uncertainty: the Discounted Case," *Journal of Economic Theory* **4**: 497-513.

Chib, S., and E. Greenberg, 1994, "Understanding the Metropolis-Hastings Algorithm," Washington University John M. Olin School of Business working paper.

Coveyou, R.R., and R.D. MacPherson, 1967, "Fourier Analysis of Uniform Random Number Generators," *Journal of the ACM* **14**: 100-119.

Davis, P.J., and P. Rabinowitz, 1984, *Methods of Numerical Integration*. Orlando: Academic Press. (Second edition)

Devroye, L., 1986, *Non-uniform Random Variate Generation*. New York: Springer-Verlag.

Evans, M., 1991, "Adaptive Importance Sampling and Chaining," *Contemporary Mathematics* **115** (Statistical Multiple Integration): 137-142. Providence: American Mathematical Society.

Fishman, G.F., and L.R. Moore, III, 1982, "A Statistical Evaluation of Multiplicative Random Number Generators with Modulus 2^{31}-1," *Journal of the American Statistical Association* **77**: 129-136.

Fishman, G.F., and L.R. Moore, III, 1986, "An Exhaustive Analysis of Multiplicative Congruential Random Number Generators with Modulus 2^{31}-1," *SIAM Journal on Scientific and Statistical Computing* **7**: 24-45.

Forsythe, G.E., 1972, "Von Neumann's Comparison Method for Random Sampling from the Normal and Other Distributions," *Mathematical Computation* **26**: 817-826.

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

Gelman, A., and D.B. Rubin, 1992, "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science* **7**: 457-472.

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

Genz, A., 1991, "Subregion Adaptive Algorithms for Multiple Integrals," in N. Flournoy and R.K. Tsutakawa (eds.), *Contemporary Mathematics* **115** (Statistical Multiple Integration): 23-31. Providence: American Mathematical Society.

Genz, A., 1993, "Subregion Adaptive Integration of Functions Having a Dominant Peak," University of Washington working paper.

Genz, A., and A. Malik, 1980, "An Adaptive Algorithm for Numerical Integration over an N-Dimensional Rectangular Region," *Journal of Computational and Applied Mathematics* **6**: 295-302.

Genz, A., and A.A. Malik, 1983, "An Imbedded Family of Fully Symmetric Numerical Integration Rules," *SIAM Journal of Numerical Analysis* **20**: 580-588.

Geweke, J., 1986, "Exact Inference in the Inequality Constrained Normal Linear Regression Model," *Journal of Applied Econometrics* **1**: 127-141.

Geweke, J., 1988, "Antithetic Acceleration of Monte Carlo Integration in Bayesian Inference," *Journal of Econometrics* **38**: 73-89.

Geweke, J., 1989, "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica* **57**: 1317-1340.

Geweke, J., 1991, "Efficient Simulation from the Multivariate Normal and Student-*t* Distributions Subject to Linear Constraints," in E.M. Keramidas (ed.), *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, 571-578.

Geweke, J., 1992a, "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments," in J.M. Bernardo *et al.* (eds.), *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting.* Oxford: Clarendon Press.

Geweke, J., 1992b, "Priors for Macroeconomic Time Series," Federal Reserve Bank of Minneapolis Institute for Empirical Macroeconomics Discussion Paper No. 64.

Geweke, J., R. Slonim, and G. Zarkin, 1992, "Econometric Solution Methods for Dynamic Discrete Choice Problems," University of Minnesota Department of Economics working paper.

Geyer, C.J., 1992, "Practical Markov Chain Monte Carlo," *Statistical Science* **7**: 473-481.

Gilks, W.R., and P. Wild, 1992, "Adaptive Rejection Sampling for Gibbs Sampling," *Applied Statistics (JRSS Series C)* **41**: 337-348.

Golub, G.H., and J.H. Welsch, 1969, "Calculation of Gaussian Quadrature Rules," *Mathematics of Computation* **23**: 221-230.

Gradshteyn, I.S., and I.M. Ryzhik, 1965, *Tables of Integrals, Series, and Products*. New York: Academic Press, 1965.

Greenberger, M., 1961, "Notes on a New Pseudo-Random Number Generator," *Journal of the ACM* **8**: 163-167.

Halton, J.M., 1960, "On the Efficiency of Evaluating Certain Quasi-random Sequences of Points in Evaluating Multi-dimensional Integrals," *Numerische Mathematik* **2**: 84-90.

Hammersley, J.M., 1960, "Monte Carlo Methods for Solving Multivariable Problems," *Annals of the New York Academy of Sciences* **86**: 844-874.

Hammersly, J.M., and D.C. Handscomb, 1964, *Monte Carlo Methods*. London: Methuen and Company.

Hammersly, J.M., and K.W. Morton, 1956, "A New Monte Carlo Technique: Antithetic Variates," *Proceedings of the Cambridge Philosophical Society* **52**: 449-474.

Hannan, E.J., 1970, *Multiple Time Series*. New York: Wiley.

Hart, H.F., E.W. Cheney, C.L. Lawson, H.J. Maehly, C.K. Mesztenyi, J.R. Rice, H.G. Thacher, Jr., and C. Witzgall, 1968, *Computer Approximations*. New York: Wiley.

Hastings, W.K., 1970, "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika* **57**: 97-109.

Hlawka, E., 1961, "Funktionen von Beschrankter Variation in der Theorie der Gleichverteilung," *Annali di Matematica Pura Ed Applicata* **54**: 325-333.

Huber, P.J., 1967, "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions," in L.M. LeCam and J. Neyman (eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 221-234. Berkeley: University of California Press.

IMSL, 1994, *IMSL Stat/Library*. Houston: Visual Numerics, Inc.

Jacquier, E., N.G. Polson, and P.E. Rossi, 1994, "Bayesian Analysis of Stochastic Volatility Models," *Journal of Business and Economic Statistics*, forthcoming.

Judd, K.L., 1991, "Numerical Methods in Economics," Hoover Institution Stanford University manuscript.

Kachitvichyanukul, V., 1982, "Computer Generation of Poisson, Binomial, and Hypergeometric Random Variates." Unpublished Ph.D. dissertation, Purdue University.

Kahn, M., and A.W. Marshall, 1953, "Methods of Reducing Sample Size in Monte Carlo Computations," *Operations Research* **1**: 263-278.

Kinderman, A.J., and J.G. Ramage, 1976, "Computer Generation of Normal Random Variables," *Journal of the American Statistical Association* **71**: 893-896.

Kipnis, C., and S.R.S. Varadhan, 1986, "Central Limit Theorem for Additive Functionals of Reversible Markov Processes and Applications to Simple Exclusions," *Communications in Mathematical Physics* **104**: 1-19.

Kloek, T., and H.K. van Dijk, 1978, "Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo," *Econometrica* **46**: 1-20.

Knuth, D.E., 1981, *The Art of Computer Programming, Volume 2: Seminumerical Algorithms*. Reading: Addison-Wesley. (Second edition)

Kronmal, R.A., and A.V. Peterson, 1979, "On the Alias Method for Generating Random Variables from a Discrete Distribution," *American Statistician* **33**: 214-218.

L'Ecuyer, P., 1986, "Efficient and Portable Combined Pseudorandom Number Generators," *Communications of the ACM* **29**: 304-313.

Marsaglia, G., 1961, "Expressing a Random Variable in Terms of Uniform Random Variables," *Annals of Mathematical Statistics* **32**: 894-899.

Marsaglia, G., 1964, "Generating a Variable from the Tail of a Normal Distribution," *Technometrics* **6**: 101-102.

Marsaglia, G., 1968, "Random Numbers Fall Mainly in the Planes," *Proceedings of the National Academy of Sciences* **60**: 25-28.

Marsaglia, G., 1972, "The Structure of Linear Congruential Sequences," in S.K. Zarema (ed.), *Applications of Number Theory to Numerical Analysis*. New York: Academic Press.

Marsaglia, G., and T.A. Bray, 1964, "A Convenient Method for Generating Normal Variables," *SIAM Review* **6**: 260-264.

Marsaglia, G., and T.A. Bray, 1968, "On-line Random Number Generators and Their Use in Combinations," *Communications of the ACM* **11**:757-759.

Marsaglia, G., M.D. MacLaren, and T.A. Bray, 1964, "A Fast Procedure for Generating Normal Random Variables," *Communications of the ACM* **7**: 4-10.

Marsaglia, G. and A. Zaman, 1991, "A New Class of Random Number Generators," *The Annals of Applied Probability* **1**: 462-480.

McGrath, E.I., 1970, *Fundamentals of Operations Research*. San Francisco: West Coast University Press.

McGrattan, E., 1993, "Solving the Stochastic Growth Model with a Finite Element Method," Federal Reserve Bank of Minneapolis working paper.

McNamee, J., and F. Stenger, 1967, "Construction of Fully Symmetric Numerical Integration Formulas," *Numerical Mathematics* **10**: 327-344.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H.Teller, and E. Teller, 1953, "Equation of State Calculations by Fast Computing Machines," *The Journal of Chemical Physics* **21**: 1087-1092.

Mikhail, W.M., 1972, "Simulating the Small-sample Properties of Econometric Estimators," *Journal of the American Statistical Association* **67**: 620-624.

Mitchell, B., 1973, "Variance Reduction by Antithetic Variates in G1/G/1 Queueing Simulation," Operations Research **21**: 988-997.

Müller, P., 1991, "Numerical Integration in Bayesian Analysis," Purdue University unpublished Ph.D. dissertation.

Niederreiter, H., 1992, *Random Number Generation and Quasi-Monte Carlo Methods.* Philadelphia: Society for Industrial and Applied Mathematics.

Numelin, E., 1984, *General Irreducible Markov Chains and Non-negative Operators.* Cambridge: Cambridge University Press.

Peskun, P.H., 1973, "Optimum Monte-Carlo Sampling using Markov Chains," *Biometrika* **60**: 607-612

Piessens, R., E. deDoncker-Kapenga, C.W. Überhuber, and D.K. Kahander, 1983, *QUADPACK.* New York: Springer-Verlag.

Poirier, D., 1995, *Intermediate Statistics and Econometrics: A Comparative Approach.* Cambridge: MIT Press.

Press, W.H., B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling, 1986, *Numerical Recipes: The Art of Scientific Computing.* Cambridge: Cambridge University Press.

Richtmeyer, R.D., 1952, "On the Evaluation of Definite Integrals and a quasi-Monte Carlo Method Based on Properties of Algebraic Numbers," Report LA-1342. Los Alamos: Los Alamos Scientific Laboratories.

Richtmeyer, R.D., 1958, "A Non-random Sampling Method, Based on Congruences for Monte Carlo Problems, Report NYO-8674. New York: Institute of Mathematical Sciences, New York University.

Ripley, R.D., 1987, *Stochastic Simulation.* New York: Wiley.

Roberts, G.O., and A.F.M. Smith, 1992, "Simple Conditions for the Convergence of the Gibbs Sampler and Metropolis-Hastings Algorithms," University of Cambridge Statistical Laboratory Research Report No. 92-30.

Rubinstein, R.Y., 1981, *Simulation and the Monte Carlo Method.* New York: Wiley.

Schmeiser, B.W., and R. Lal, 1980, "Squeeze Methods for Generating Gamma Variates," *Journal of the American Statistical Association* **75**: 679-682.

Smith, A.A., 1991, "Solving Stochastic Dynamic Programming Problems using Rules of Thumb," Queen's University Department of Economics Discussion Paper No. 816.

Stokey, N.L., and R.E. Lucas, Jr., 1989, *Recursive Methods in Economic Dynamics.* Cambridge: Harvard University Press.

Strecok, A.J., 1968, "On the Calculation of the Inverse of the Error Function," *Mathematics of Computation* **22**: 144-158.

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

Tauchen, G., 1985, "Diagnostic Testing and Evaluation of Maximum Likelihood Models," *Journal of Econometrics* **30**: 415-443.

Taylor, S., 1986, *Modelling Financial Time Series*. New York: Wiley.

Taylor, J., and H. Uhlig, 1990, "Solving Nonlinear Stochastic Growth Models: A Comparison of Alternative Solution Methods," *Journal of Business and Economic Statistics* **8**: 1-17.

Tezuka, S., P. L'Ecuyer, and R. Couture, 1993, "On the Lattice Structure of the Add-with-carry and Subtract-with-borrow Random Number Generators," *ACM Transactions on Modeling and Computer Simulation* **3**: 315-331.

Tierney, L., 1991a, "Exploring Posterior Distributions Using Markov Chains," in E.M. Keramaidas (ed.), *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, 563-570. Fairfax: Interface Foundation of North America, Inc.

Tierney, L., 1991b, "Markov Chains for Exploring Posterior Distributions," Technical Report No. 560, University of Minnesota School of Statistics. Forthcoming, *Annals of Statistics*.

Tierney, L., and J.B. Kadane, 1986, "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association* **81**: 82-86.

Tierney, L., R.E. Kass, and J.B. Kadane, 1989, "Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions," *Journal of the American Statistical Association* **84**: 710-716.

van Dooren, P., and L. de Ridder, 1976, "An Adaptive Algorithm for Numerical Integration over an N-Dimensional Rectangular Region," *Journal of Computational and Applied Mathematics* **2**: 207-217.

von Neumann, J., 1951, "Various Techniques Used in Connection with Random Digits," *National Bureau of Standards Applied Mathematics*, Series 12, pp. 36-38.

Walker, A.J., 1974, "New Fast Method for Generating Discrete Random Numbers with Arbitrary Frequency Distributions," *Electronics Letters* **10**: 127-128.

Walker, A.J., 1977, "An Efficient Method for Generating Discrete Random Variables with General Distributions," *ACM Transactions on Mathematical Software* **3**: 253-256.

Wichmann, B.A., and I.D. Hill, 1982, "An Efficient and Portable Pseudo-random Number Generator," *Applied Statistics* **31**: 188-190.

Wild, P., and W.R. Gilks, 1993, "Adaptive Rejection Sampling from Log-concave Density Functions," *Applied Statistics (JRSS Series C)* **42**: 701-708.

Zellner, A., and C. Min, 1992, "Gibbs Sampler Convergence Criteria," University of Chicago Graduate School of Business working paper.

.

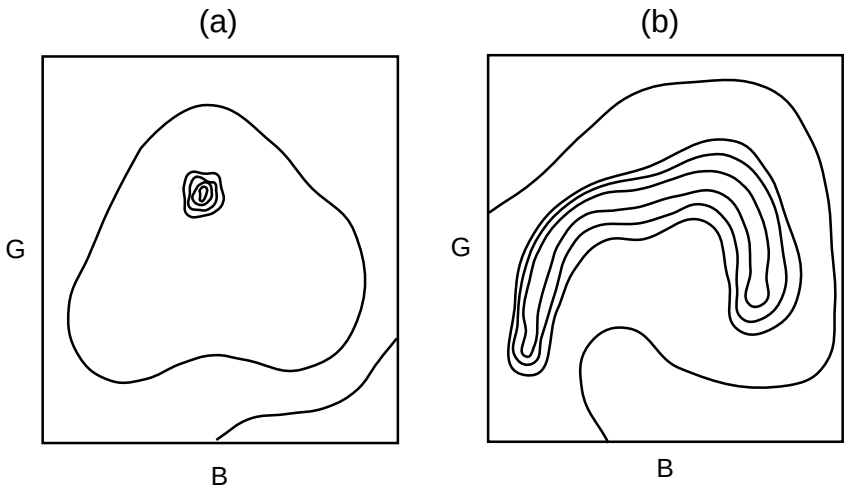

Figure 1. Contours of the function to be integrated are shown.

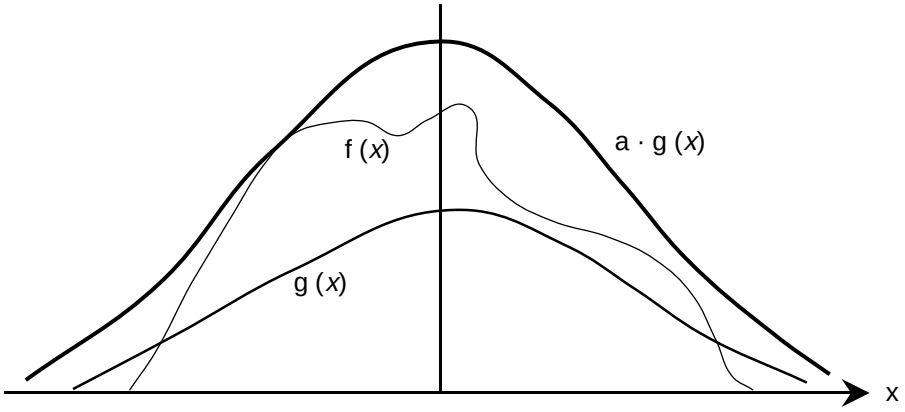

Figure 2. The target density is f(*x*), the source density is g(*x*), and $a = \sup[f(x)/ g(x)]$.

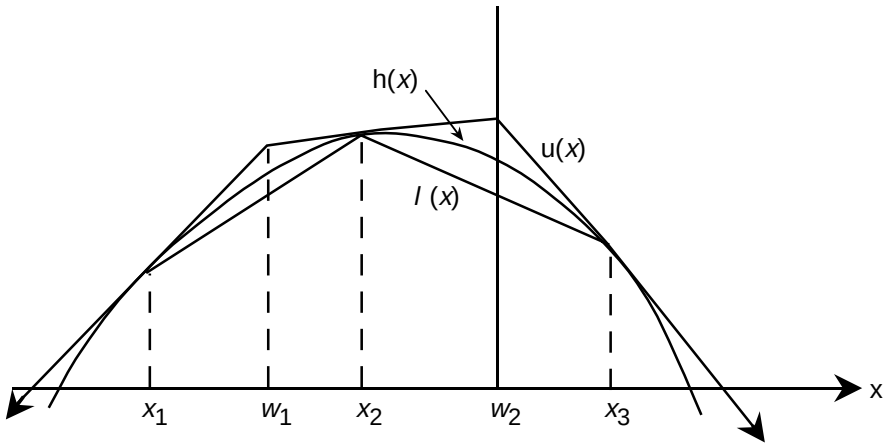

Figure 3. The function h(x) = log f(x), where f(x) is a log-concave p.d.f. The lower hull l (x) is formed by the chords joined at the x_j, and the upper hull u(x) is formed by the tangents at the x_j which are joined at the w_j.

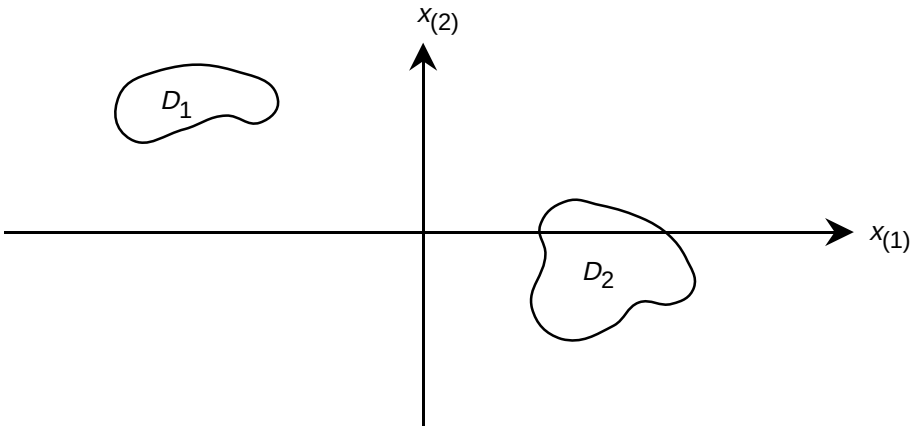

Figure 4. The disconnected support $D = D_1 \cup D_2$ for the probability distribution implies that a Gibbs sampler with blocking $(x_{(1)}, x_{(2)})$ will not be Harris recurrent. In the example shown it cannot converge from any star ting value.

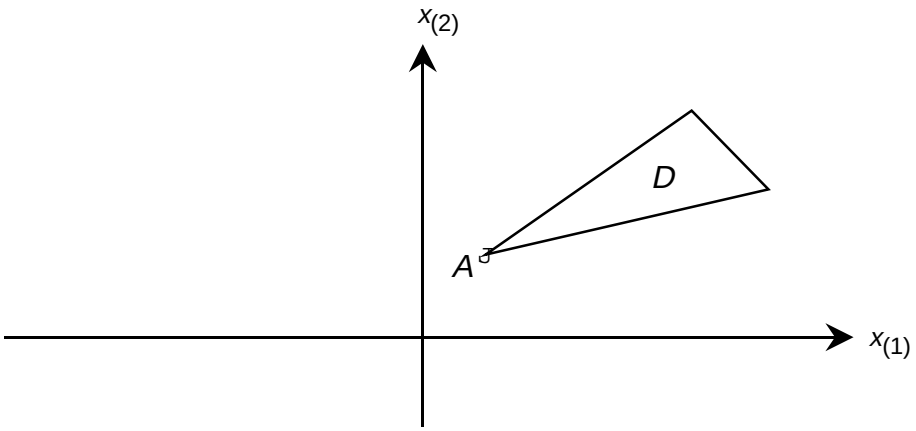

Figure 5. The probability density p($\boldsymbol{x}$) is uniform on the closed set D and consequently is not lower semicontinuous at 0. The point A is absorbing for the Gibbs sampler with blocking $(x_{(1)}, x_{(2)})$, so if $\boldsymbol{x}^0 = A$ convergence will not occur.

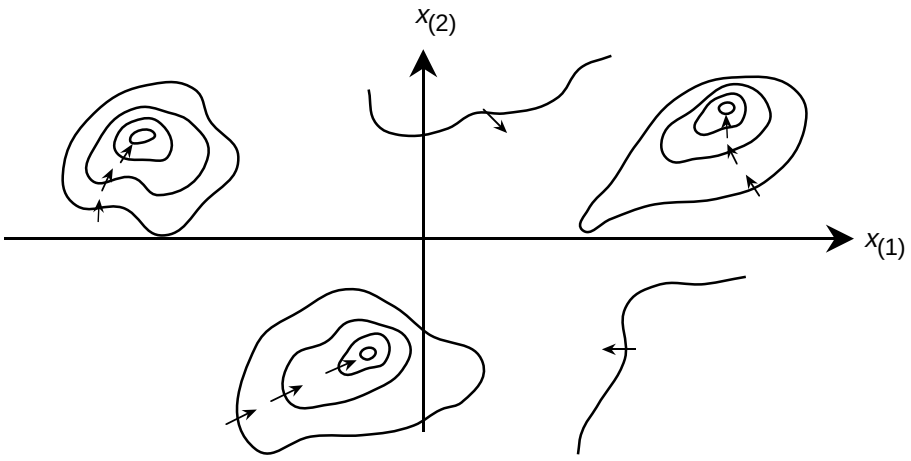

Figure 6. Iso-probability density contours of a multimodal bivariate distribution are shown. (Arrows indicate directions of increased densi ty.) Given sufficiently steep gradients the Gibbs sampler will converge very slowly.

Table 1

Evaluations required to approximate $\int_{\bar{I}^d} \mathrm{f}(\mathbf{x})d\mathbf{x}$, $\mathrm{f}(\mathbf{x}) = \sum_{j=1}^{d} \mathrm{f}\left(x_j\right)$,
with maximum error c: Actual number and upper bound

d	2	3	4	5
$c = .01$:				
Actual	228	442	661	1060
Bound	19,335	1,014,825	9.154×10^{7}	1.522×10^{10}
$c = 10^{-5}$:				
Actual	640,426	1,039,188	1,523,433	2,379,162
Bound	52,477,915	3.469×10^{9}	3.513×10^{11}	6.114×10^{13}

Table 2

Error comparison for Halton sequence and independence Monte Carlo

$$\int_{\bar{I}^d} \mathrm{f}(\mathbf{x})d\mathbf{x},\ \ \mathrm{f}(\mathbf{x}) = \sum_{i=1}^{d} x_i$$

$$m = 1{,}000$$

d	Halton error	Halton bound	MC error($p = .05$)	MC error($p = 10^{-12}$)
5	-7.526×10^{-3}	9.302×10^{2}	.04000	.1455
10	-.02807	6.053×10^{19}	.05658	.2058
20	-.1097	2.616×10^{29}	.08002	.2911
40	-.3824	8.225×10^{72}	.1132	.4117
60	-.8202	2.467×10^{121}	.1386	.5042
80	-1.476	1.250×10^{173}	.1600	.5822
100	-2.062	1.447×10^{227}	.1789	.6509

$$m = 50{,}000$$

d	Halton error	Halton bound	MC error($p = .05$)	MC error($p = 10^{-12}$)
5	-2.786×10^{-4}	1.071×10^{2}	5.658×10^{-3}	.02058
10	-8.861×10^{-4}	3.533×10^{10}	8.002×10^{-3}	.02911
20	-3.537×10^{-3}	3.225×10^{30}	.01132	.04117
40	-.02216	3.356×10^{75}	.01600	.05822
60	-.02768	2.2990×10^{127}	.01960	.07131
80	-.05681	2.4186×10^{181}	.02263	.08234
100	-.08779	5.235×10^{237}	.02530	.09205

Table 3

Error comparison for Halton sequence and Monte Carlo

$$f(\mathbf{x}) = \sum_{i=1}^{d} x_i, \quad \mathbf{x} \sim N(\mathbf{0}, \mathbf{I}_d); \text{ evaluate } E[f(\mathbf{x})]$$

	$m = 1{,}000$			$m = 50{,}000$		
		Monte Carlo error			Monte Carlo error	
d	Halton error	($p = .05$)	($p = 10^{-12}$)	Halton error	($p = .05$)	($p = 10^{-12}$)
5	-.04190	.1386	.5042	-1.808×10^{-3}	.01960	.07131
10	-.1411	.1960	.7131	-5.552×10^{-3}	.02772	.1008
20	-.5497	.2772	1.008	-.02076	.03920	.1426
40	-1.7306	.3920	1.426	-.06548	.05544	.2017
60	-3.3617	.4801	1.747	-.1461	.06790	.2470
80	-5.6578	.5544	2.017	-.2573	.07840	.2852
100	-7.8073	.6198	2.255	-.2336	.08765	.3189

$$f(\mathbf{x}) = \sum_{i=1}^{d} x_i^2, \quad \mathbf{x} \sim N(\mathbf{0}, \mathbf{I}_d); \text{ evaluate } E[f(\mathbf{x})]$$

	$m = 1{,}000$			$m = 50{,}000$		
		Monte Carlo error			Monte Carlo error	
d	Halton error	($p = .05$)	($p = 10^{-12}$)	Halton error	($p = .05$)	($p = 10^{-12}$)
5	-.0496	.2400	.8733	-1.664×10^{-3}	.03395	.1235
10	-.0941	.3395	1.2350	-2.418×10^{-3}	.04801	.1747
20	-.0864	.4801	1.746	-4.611×10^{-3}	.06790	.2470
40	.2436	.6790	2.470	-6.367×10^{-3}	.0962	.3493
60	.5680	.8316	3.0252	-3.662×10^{-3}	.1176	.4278
80	.4982	.9602	3.4932	.0243	.1358	.4940
100	1.449	1.074	3.906	-.04932	.1518	.5523

Table 3 (continued)

$$f(\mathbf{x}) = \sum_{i=1}^{d} x_i^3, \quad \mathbf{x} \sim N(\mathbf{0}, \mathbf{I}_d); \text{ evaluate } E[f(\mathbf{x})]$$

	$m = 1{,}000$			$m = 50{,}000$		
		Monte Carlo error			Monte Carlo error	
d	Halton error	($p = .05$)	($p = 10^{-12}$)	Halton error	($p = .05$)	($p = 10^{-12}$)
5	-.3500	.5368	1.953	-.02286	.07591	.2761
10	-1.083	.7591	2.762	-.06800	.1073	.3906
20	-4.072	1.074	3.906	-.2386	.1518	.5523
40	-11.865	1.518	5.523	-.6821	.2174	.7811
60	-19.564	1.859	6.765	-1.411	.2630	.9567
80	-27.78	2.147	7.811	-2.641	.3036	1.104
100	-36.18	2.400	8.733	-2.218	.3395	1.235

$$f(\mathbf{x}) = \sum_{i=1}^{d} x_i^4, \quad \mathbf{x} \sim N(\mathbf{0}, \mathbf{I}_d); \text{ evaluate } E[f(\mathbf{x})]$$

	$m = 1{,}000$			$m = 50{,}000$		
		Monte Carlo error			Monte Carlo error	
d	Halton error	($p = .05$)	($p = 10^{-12}$)	Halton error	($p = .05$)	($p = 10^{-12}$)
5	-.7442	1.420	5.167	-.03612	.2008	.7307
10	-1.046	2.008	7.307	-.04667	.2840	1.0333
20	-.8494	2.840	10.33	-.07076	.4017	1.461
40	7.504	4.016	14.61	.03523	.5681	2.067
60	16.88	4.919	17.90	.1150	.0957	2.521
80	23.48	5.681	20.66	-.1898	.8034	2.923
100	32.94	6.351	23.105	-.7909	.8982	3.268

Table 3 (continued)

$$f(\mathbf{x}) = \sum_{i=1}^{d} x_i^5, \quad \mathbf{x} \sim N(\mathbf{0}, \mathbf{I}_d); \text{ evaluate } E[f(\mathbf{x})]$$

	$m = 1{,}000$			$m = 50{,}000$		
		Monte Carlo error			Monte Carlo error	
d	Halton error	($p = .05$)	($p = 10^{-12}$)	Halton error	($p = .05$)	($p = 10^{-12}$)
5	-3.216	4.260	15.50	-.3365	.6026	2.192
10	-1.043	6.025	21.92	-1.006	.8521	3.100
20	-36.44	8.521	31.00	-3.433	1.205	4.384
40	-118.9	12.05	43.84	-9.549	1.704	6.200
60	-202.8	14.76	53.69	-14.50	2.087	7.593
80	-281.6	17.04	62.00	-13.11	2.410	8.760
100	-359.7	19.05	69.32	-23.97	2.695	9.803