Statistics and Computing (1997) **7**, 217–228

Exploring quasi Monte Carlo for marginal density approximation

M. OSTLAND and B. YU

University of California at Berkeley, Berkeley CA 94720, USA

Received September 1996 and accepted June 1997

We first review quasi Monte Carlo (QMC) integration for approximating integrals, which we believe is a useful tool often overlooked by statistics researchers. We then present a manually-adaptive extension of QMC for approximating marginal densities when the joint density is known up to a normalization constant. Randomization and a batch-wise approach involving $(0, s)$-sequences are the cornerstones of our method. By incorporating a variety of graphical diagnostics the method allows the user to adaptively allocate points for joint density function evaluations. Through intelligent allocation of resources to different regions of the marginal space, the method can quickly produce reliable marginal density approximations in moderate dimensions. We demonstrate by examples that adaptive QMC can be a viable alternative to the Metropolis algorithm.

Keywords: adaptive, marginal distribution, Metropolis, algorithm, quasi Monte Carlo

1. Introduction

A common computational problem in statistics is the calculation of marginal densities. In some cases the joint density is too complicated to allow an analytical solution, so we must turn to approximations. Such a situation arises naturally in Bayesian statistics when a complicated posterior density is known only up to a normalization constant. We will address the general problem of approximating a $d \geq 1$ dimensional marginal density of an $s > d$ dimensional non-negative density f, where f is integrable and known up to a normalizing constant. We have two primary goals: to introduce quasi Monte Carlo (QMC) integration to researchers new to QMC, and to explore an adaptive variation of QMC integration. We demonstrate by examples that adaptive QMC can be a viable alternative to the commonly used Metropolis algorithm.

An obvious question is why do we need alternatives to Metropolis in the first place? Metropolis is easy to implement (relying mostly on being able to evaluate the unnormalized joint density), and has had its share of successes. The basic idea behind Metropolis is to create a Markov chain with stationary distribution equal to the target distribution. Thus, once the chain is run far enough to reach approximate stationarity, the next batch of iterations will be an approximate (dependent) sample from the desired distribution. However, many serious questions persist: How far is 'far enough' in order to reach stationarity? How does one best handle the dependency between samples in a 'sticky' chain? How does one choose a good transition kernel for a particular density? How does one choose a starting value? Which is better: one long chain or several shorter chains? Without a clear consensus on these important questions there is a strong need for alternative methods which can serve as external checks for Metropolis. Moreover, many Metropolis diagnostics suggested in the literature are unintuitive, difficult to implement, or require problem-specific code. Furthermore, Cowles and Carlin (1996) recently reviewed thirteen convergence diagnostics and concluded, 'all the methods can fail to detect the sorts of convergence failure that they were designed to identify.' So there is clearly room for alternative methods. Leonard, Hsu and Ritter (1994) and Tierney, Kass and Kadane (1989) demonstrate algebraic approximations which can be useful alternatives in some instances. Here we provide a numeric alternative for moderate dimensions based on QMC.

In the Bayesian framework, Shaw (1988) used quasi Monte Carlo (QMC) methods to approximate posterior moments and marginal density evaluations of a posterior f. Unlike Metropolis which strives to build up samples from those regions of the domain where f has greatest

0960-3174 © 1997 Chapman & Hall

mass, QMC attempts in some sense to blanket the domain of a function with an evenly spread net of points and use function evaluations at these points to approximate the function's integral. The differing nature of these methods suggests that agreement in the two results lends credibility to the analysis. In this paper we extend the simple implementation of QMC to a scheme that allocates points for joint density evaluations adaptively according to which regions seem to need more evaluations. To emphasize the adaptive nature of our QMC extension, we refer to our method as AQMC. In addition to adaptiveness, AQMC enables rough assessment of the accuracy of the approximations. These assessments are based on an introduction of randomness to the QMC nets and, as suggested by Owen (1996), utilization of a batch-wise approach based on $(0, s)$-sequences.

The organization of the paper is as follows: Section 2.1 introduces regular quasi Monte Carlo integration for readers unfamiliar with QMC. Section 2.2 is a brief description of the Metropolis algorithm for comparison. The useful QMC sequence known as the (t, s)-sequence is defined in Section 2.3. Technical details of how to construct such a sequence are provided in the Appendix. Section 3 develops the application of AQMC to the approximation of marginal densities. A variety of graphical diagnostics are proposed to help assess the quality of the approximations as well as answer the related question of which regions require more joint density evaluations. Section 4 demonstrates these techniques on a Bayesian posterior density example, comparing the results to that of Metropolis. The limits of AQMC are also probed as we apply the method to higher-dimension problems. Finally, a discussion of the results is found in Section 5.

2. Quasi Monte Carlo and Metropolis

2.1. *Quasi Monte Carlo*

Quasi Monte Carlo methods impact a variety of topics in statistics (see Fang and Wang, 1993), but we focus on the application of QMC to multi-dimensional numerical integration. In this context, QMC can be viewed as a quadrature technique, but it distinguishes itself from other more standard quadrature techniques because of the number theory behind the selection of points at which to evaluate the integrand. Such theory is beyond the scope of this paper, but to gain some insight into how QMC works, it is instructive to first take a quick look at Monte Carlo integration.

Let f be an integrable function on the s-dimensional unit cube $C^s = [0, 1]^s$. Let $I = \int_{C^s} f \mathrm{d}x$. If I cannot be calculated analytically it must be approximated. Much literature is devoted to special quadrature techniques for numerical solutions to this very problem (see Davis and Rabinowitz 1984). However, the standard techniques become prohibitively inefficient when s becomes large. When s is large, one route is the Monte Carlo (MC) method. In its simplest incarnation MC integration consists of taking a random sample $X_1, \ldots, X_n$ from the uniform distribution on C^s. Then the approximation

$$\hat{I}_n = \frac{1}{n}\sum_{j=1}^{n} f(X_j) \tag{1}$$

converges to I by the Law of Large Numbers. Additionally, if $f \in L_2$ then the Central Limit Theorem (CLT) guarantees that $\sqrt{n}(\hat{I}_n - I) \longrightarrow N(0, \sigma^2)$, where convergence is in distribution and $\sigma^2 = \int_{C^s}(f - I)^2 \mathrm{d}x$. MC is simple to implement and it is viable in high dimensions since the rate of convergence does not depend on the dimension. Also obtaining an estimate of the approximation error is simple using the sample variance. On the other hand, the rate of convergence is only of the order $n^{-\frac{1}{2}}$ and, since the CLT is in terms of the probability distributions, bounds on the error $(\hat{I}_n - I)$ are given with probability statements attached to them. Thus, the method performs well on average, but the particular sample path that one observes may lead to a terrible approximation. Furthermore, with an undersized sample it is possible to get a very small estimated standard error even when MC is far from the right answer. Hence, many researchers are rightfully wary about using Monte Carlo integration when high precision is demanded. There are variance reduction methods that improve on this simple version (see Rubinstein, 1981; Ripley, 1987), but neither the order nor type of convergence is altered.

In moderate dimensions quasi Monte Carlo integration succeeds on the two points where MC falters with a deterministic error bound at a better asymptotic rate. For thorough details on QMC, see Niederreiter (1992) and Hua and Wang (1981). The idea behind QMC integration is that the uniform random vectors $X_1, \ldots, X_n$ in (1) are not 'evenly spread' enough – a term we will formally define below. So, instead of using uniform random vectors, QMC employs a deterministic set of points $x_1, \ldots, x_n$ that are spread evenly over C^s. The resulting estimator is again (1), only this time using the deterministic set $x_1, \ldots, x_n$. We will refer to a set of points meeting certain criteria for being spread evenly as a QMC net (also called a Number Theoretic Net, or an NT-net for short). Figure 1 compares two sets of 25 points in the two-dimensional unit square. The first set is of (pseudo)random uniforms like those used in Monte Carlo integration, and the second is a QMC net known as a $(0, 2, 2)$-net in base 5.

The discrepancy is the most common measure of evenness of spread for a set of points in the unit cube. Let $\rho = \{x_j,\ j = 1, \ldots, n\}$ be a set of points in C^s. For fixed ρ define $U_n(y)$ as the function on R^s given by

$$U_n(y) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \leq y\}} \tag{2}$$

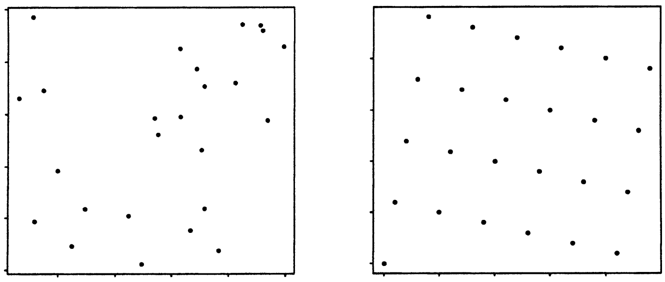

Fig. 1. *25 Random uniforms and a (0, 2, 2)-net in base 5.*

where 1_A is the indicator function of the set A, and $\{x_i \leq y\}$ is understood as being with respect to component-wise order. This is simply the empirical distribution function of ρ. Then, the *discrepancy* of ρ is defined as

$$D(n, \rho) = \sup_{y \in C^s} |U_n(y) - U(y)| \qquad (3)$$

where $U(y)$ is the cumulative distribution function of a uniform random vector on C^s.

It turns out that for a wide class of functions (those of bounded variation in the sense of Hardy and Krause, which is a smoothness requirement on the integrand – see Niederreiter (1978)) the absolute error of a QMC approximation is bounded by a constant times the discrepancy of the set of points used in (1). This is good news since it is possible to construct a set of points $\rho = \{x_1, \ldots, x_n\}$ such that $D(n, \rho)$ is order $(\log n)^{s-1}/n$, which is a superior asymptotic rate to the $n^{-\frac{1}{2}}$ given by Monte Carlo integration. Moreover, the bound on the error is deterministic, not just in probability.

Well-known methods for constructing sets of low discrepancy include the *good lattice point*, the *good point*, *Halton*, *scrambled Halton*, *Haber*, *Hammersley*, and (t, m, s)*-nets*. Fang and Wang (1993) and Shaw (1988) compare many of these methods on test functions. Our paper makes extensive use of the (t, m, s)-net and the associated infinite sequence. Because we can construct a nested sequence of increasingly large sets of points such that each set has excellent equi-distribution properties (see Section 2.3 for a description of these properties), we are able to adaptively allocate resources and get estimates of approximation errors. A construction of these sequences is found in the Appendix.

Regular QMC is inferior to Monte Carlo integration in one important area: in QMC the approximation error is difficult to estimate. The forementioned error bound is not very useful since it is neither sharp nor calculable in general. Cranley and Patterson (1976), Shaw (1988) and Owen (1995, 1996) among others have discussed introducing randomness into the nets in order to get estimates of the error. By re-introducing randomness these authors have created hybrid QMC–Monte Carlo methods enjoying some of the best properties of both QMC and MC. We will follow the lead of these ideas in assessing the quality of our approximations.

2.2. *Metropolis algorithm*

A vast literature on this topic has erupted in recent years. Since many readers are likely to be familiar with this algorithm, we present only a quick recipe for implementation. For a more complete discussion see Gelman and Rubin (1992), Chib and Greenberg (1994), Cowles and Carlin (1996), Geyer (1992) and Tierney (1994).

According to the Metropolis algorithm, an s-dimensional Markov chain is formed as follows:

1. Find a transition kernel $q(\cdot, \cdot)$ for a symmetric Markov Chain from which one can directly draw a sample.
2. Pick an initial value X_0.
3. Generate a candidate Y from the distribution $q(X_i, \cdot)$.
4. Calculate the ratio $r = \alpha(X_i, Y)$ where $\alpha(x, y) = \min\left(\frac{f(y)}{f(x)}, 1\right)$.
5. Set $X_{i+1} = Y$ with probability r, and $X_{i+1} = X_i$ otherwise.
6. Repeat steps 3–5 N times, for some large N.

Then X_n is a Markov Chain that has stationary distribution with density f (see Tierney, 1994). Convergence to the stationary distribution is guaranteed once the chain is shown to be irreducible and aperiodic – conditions that are satisfied by the usual choices for $q(x, \cdot)$. However the rate of convergence is generally not known, and, especially with a poor choice of transition kernel and/or starting value, convergence could take a very long time. There appears to be no fail-safe method for determining if a chain has reached stationarity based on the iterations up to N (see Gelman and Rubin (1992) and Cowles and Carlin (1996) for discussion on this point). Standard procedure is to discard ('burn in') the first n_0 iterations of the chain. At this point many researchers separate the remaining $N - n_0$ iterations into batches and estimate the mean and variance of the target distribution from these batches, taking care to consider auto-correlations. In the examples that follow we will be concerned with marginal density approximation and will use a slightly naive treatment of considering the $N - n_0$ iterations as i.i.d. samples from the joint distribution. Then density estimation techniques are employed with the samples.

2.3. (t, m, s)*-nets and* (t, s)*-sequences*

In this section we define (t, m, s)-nets and the associated sequences. This type of QMC sequence will be instrumental in the methods presented later, and a construction of such sequences is provided in the Appendix. For integers $s \geq 1$ and $b \geq 2$, an *elementary interval* in base b is an s-dimensional sub-rectangle of $C^s = [0, 1]^s$ of the form

$$E = \prod_{i=1}^{s} \left[\frac{a_i}{b^{k_i}}, \frac{a_i + 1}{b^{k_i}}\right] \qquad (4)$$

where k_i, a_i are integers with $k_i \geq 0$ and $0 \leq a_i < b^{k_i}$. The usual volume of such a rectangle is thus $V(E) = b^{-\delta}$ where $\delta = \sum_{i=1}^{s} k_i$. Let $m \geq 0$ be an integer and $t \leq m$ be a nonnegative integer, then the set of points $x_1, \ldots, x_{b^m}$ from C^s is a (t, m, s)-*net in base b* if every elementary interval in base b of volume b^{t-m} contains exactly b^t points from the set. So, smaller values of t imply that the property holds at a finer resolution, resulting in stronger equi-distributional statements. At the other extreme, taking $t = m$ makes the trivial statement that all points in the set lie in $[0, 1]^s$. The points in Fig. 2 are a $(0, 2, 2)$-net in base 3. The three frames demonstrate how each elementary interval of volume 3^{-2} contains exactly one point.

Finally, for an integer $t \geq 0$, an infinite sequence $\{x_i\}_{i \geq 1}$ of points from C^s is a (t, s)-*sequence in base b* if for every $m \geq t$ the blocks of length b^m are (t, m, s)-nets in base b. Specifically, for all $k \geq 0$ and $m \geq t$ the set of b^m points $x_{kb^m+1}, \ldots, x_{(k+1)b^m}$ is a (t, m, s)-net in base b. The useful property of (t, s)-sequences is that we do not need to specify the size of the net ahead of time, yet each finite subsequence of points up to the b^m-th point will have the strong equi-distributional properties of a (t, m, s)-net, $m = 1, 2, \ldots$. It is for precisely this reason (and the ease of construction) that we use (t, s)-sequences in the sequel. However, other QMC sequences possessing this important property could be substituted.

3. Approximating marginal densities

In this section we propose a method of adaptively approximating marginal densities using $(0, s)$-sequences. We begin by reviewing a simple version of QMC and gradually enhance this to full AQMC, which incrementally allocates points in a $(0, s)$-sequence for density evaluations, where the number and location of points is determined adaptively according to the specific problem. Diagnostics for guiding the adaptive allocation and estimating the approximation error are an integral part of AQMC. We want to approximate a particular d_1-dimensional marginal density of an s-dimensional random variable having density f, where $1 \leq d_1 < s$. Also, $f = g/c$ where the normalization constant c is possibly unknown. Without loss of generality, take the marginal space, denoted by T_1, as the first d_1 coordinates of R^s. The orthogonal complement T_2 is therefore spanned by the last $d_2 = s - d_1$ coordinates of R^s. Let $x = (x_1, \ldots, x_s)$ represent a point in s-dimensional space. Then the marginal density is

$$f_{T_1}(x_1, \ldots, x_{d_1}) = \int_{T_2} f(x)\, \mathrm{d}x_{d_1+1} \cdots \mathrm{d}x_s$$

$$= \frac{1}{c} \int_{T_2} g(x)\, \mathrm{d}x_{d_1+1} \cdots \mathrm{d}x_s \qquad (5)$$

$$= \frac{1}{c} g_{T_1}(x_1, \ldots, x_{d_1}). \qquad (6)$$

If we assume for now that the domain of f is in a known s-dimensional rectangle with finite volume, then changing variables enables us to consider only f with domain equal to the unit s-cube.

In non-adaptive QMC n density evaluations are made at each of m points in the marginal space, where $n \gg m$ are fixed. To implement, create a d_1-dimensional QMC net N_1 (which we refer to as the *main net*) of m points, and a d_2-dimensional QMC net N_2 (the *auxiliary net*) of n points. Then, for each point $x = (x_1, \ldots, x_{d_1}) \in N_1$ approximate $g_{T_1}(x)$ by

$$\hat{g}_{T_1}(x) = \frac{1}{n} \sum_{i=1}^{n} g(x, y_i) \qquad (7)$$

where $y_i = (y_{i_1}, \ldots, y_{id_2})$ is the ith point in the auxiliary net. Appealing to (1) the normalization constant can be approximated by

$$\hat{c} = \frac{1}{m} \sum_{j=1}^{m} g_{T_1}(x_i). \qquad (8)$$

If $d_1 = 1$ and the marginal density is known to be somewhat smooth, then a specialized quadrature technique such as Simpson's rule could be used to improve on (8). If c is known, it can be combined with (7) into (6) to approximate $f_{T_1}(x)$. Otherwise, $\hat{c}$ is substituted for c. In the case that c is known, the absolute error $|c - \hat{c}|$ and corresponding relative error can be used to help assess the quality of the approximations. This is essentially the QMC method described by Shaw (1988), among others.

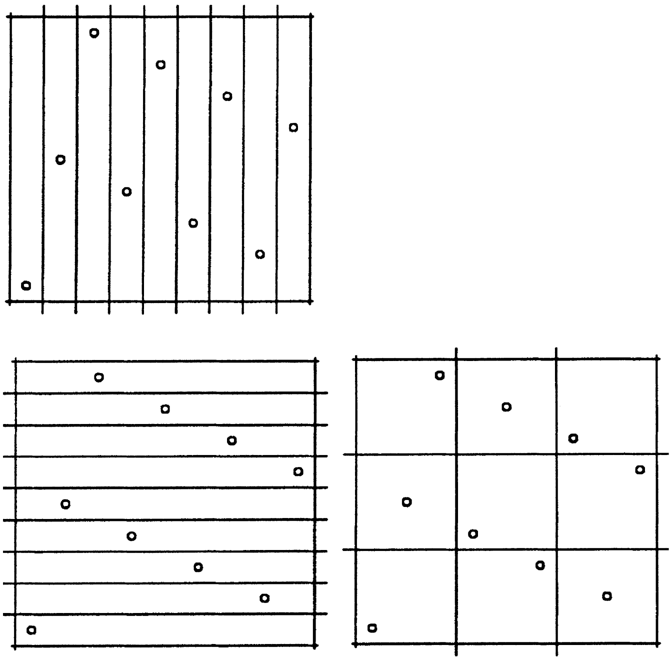

Fig. 2. *The elementary intervals for a (0, 2, 2)-net in base 3.*

Example. We illustrate regular QMC by example with a bimodal multivariate Gaussian. We will revisit this example later as we develop full AQMC. Let $\mathbf{I}$ be the 4-dimensional identity matrix, and let $\mathbf{J}$ be the 4-dimensional vector of ones. We consider the mixture of two $s = 4$-dimensional Gaussians with means $\mu_1 = \mathbf{0}$ and $\mu_2 = 4\mathbf{J}$ and covariance matrices $\Sigma_1 = \mathbf{I}$, and $\Sigma_2 = 0.25\mathbf{I}$. Take the mixing probability to be 0.5 on each of the two modes. We approximate the $d_1 = 1$-dimensional density of the first coordinate. To proceed, the function is truncated outside of $[-4, 6]^4$ and transformed into the unit cube. The two panels of Fig. 3 show the resulting density approximations for two choices of (m, n). Solid lines indicate the true marginal density which was calculated analytically. Clearly, both sets of approximations are very close to the truth. In each case, the usually unavailable relative error $|c - \hat{c}|/c$ is less than 1% . Each set of computations took only a few seconds on a Sparc 20.

3.1. *Assessing quality of marginal approximations*

Here we introduce diagnostics for assessing the quality of QMC approximations. Shaw (1988), who was interested in posterior moments, describes using complete replication based on independent random origin shifts on the QMC nets to assess variability. In other words, he takes a d-dimensional QMC net, M_1, and forms $r - 1$ new nets $M_2, \ldots, M_r$ where $M_j = \{(x_1 + U_{1j}, \ldots, x_d + U_{dj}) : (x_1, \ldots, x_d) \in M_1\}$ where the Us are independent standard uniform and the addition is modulus one. Using each net individually he gets r independent approximations for the posterior moment, which he then uses to assess variability. Although natural and easy to implement, this is computationally prohibitive if the net size is at all large. However, in the context of marginal density approximation one can proceed without resorting to complete replication provided the marginal density is fairly smooth and the main net is not too sparse. Specifically, one uses independent random origin-shifted auxiliary nets for each point in the main net. One then simply compares the approximations at neighbouring points in N_1. Wild fluctuations between neighbouring approximations suggest that the approximations are poor and indicates the need for increasing the size of the auxiliary net. In essence, this diagnostic assumes smoothness of the marginal density in order to use neighbouring approximations as a proxy for complete replication.

Next we present a diagnostic based on a non-replicated run which was suggested by Owen (1996). Unlike the random origin shifts, this diagnostic does not require the assumption of smoothness on f_{T_1}. We partition the evaluations into batches of 'pseudo-replicates' and examine the variability between approximations based on the batches. We work with the approximated g_{T_1} instead of the f_{T_1} which have an extra component of variability from $\hat{c}$ when c is unknown. Suppose $n = kb^j$ for some integers $k \geq 2$ and $j \geq 1$. From Section 2.3 we know each of the k batches of b^j points is a $(0, j, d_2)$-net in base b. For each $x \in N_1$ one can thus approximate $g_{T_1}(x)$ separately based on each of the k sub-nets. For fixed x, call these approximations $\hat{g}_l$, $l = 1, \ldots, k$. If these batches were true independent replicates, then

$$V_n(x) := \frac{1}{k}\sum_{l=1}^{k}(\hat{g}_l - \bar{g})^2 \qquad (9)$$

would be an estimate for the variance of $\bar{g}$, which is the approximation based on all n evaluations. Then, $\sqrt{V_n(x)}$ can be compared with c (or $\hat{c}$) to assess convergence. However, these are not true independent replicates since N_2 has such structure. But (9) can still be used as a rough estimate for variability. Owen (1996) discusses some conditions on f for which (9) is conservative. Since the batch approximations are not independent, the choice of k and j makes a difference, and in our experience smaller values of k seem to lead to better approximations of variability. If $d_1 \leq 2$ a plot of $\sqrt{V_n(x)}$ versus x is possible. We will refer to such a plot as an *SD plot.*

Example (continued). We continue the multivariate Gaussian example with $m = 40$ and $n = 1000$ and implement the 40 independent random origin shifts of the auxiliary net. We form batch variances $V_n(x)$ using $j = 5$, resulting in $k = 4$ full batches of $3^5 = 243$ points each.

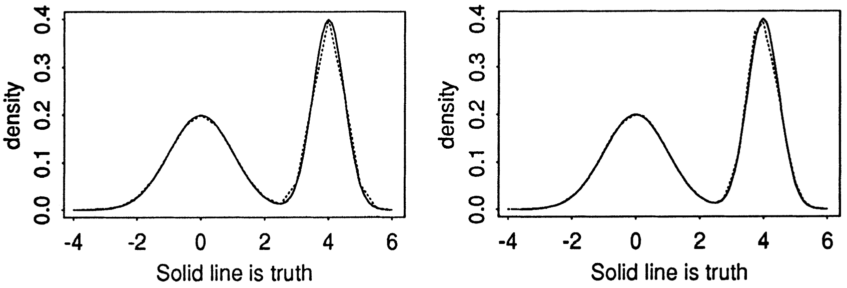

Fig. 3. *QMC Approximations: (m, n)= (20, 8000) and (m, n)= (40, 4000). Approximated marginal density evaluations are connected by straight lines. Solid lines show truth.*

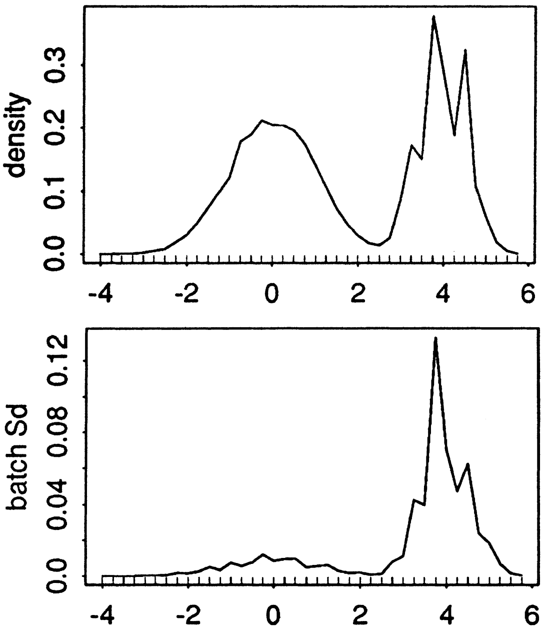

Fig. 4. *QMC approximation and SD plot for (m, n) = (40, 1000) with random origin shifts.*

Figure 4 shows the approximated marginal density evaluations and the SD plot.

Since Normality promises smooth marginal densities, the jagged lines connecting the approximations indicate that 1000 evaluations is not sufficient for points under the second mode. The large within batch SD's relative to c for these points adds support for the need for more points. Here the relative error is over 9%. In contrast, Fig. 5 displays similar plots for $m = 40$ and $n = 4000$, which are the same values used to form the right panel of Fig. 3.

3.2. *Adaptive nets*

In this section we make QMC adaptive in three ways. First, the size of the auxiliary net can be increased if the diagnostics suggest that approximation errors for the marginal density evaluations are unacceptably large. Second, the auxiliary net can vary in length from point to point in the marginal space. Thus, marginal evaluations that are easy to approximate need not consume as much computing time as those that are more difficult to approximate. Third, the main net can adapt to the needs of the problem by becoming more concentrated in those regions of the marginal space where the marginal density seems to be highly variable. We hope to capture important features of the marginal density without resorting to an increase in the concentration of the main net throughout the entire marginal space. We go through these steps in order.

1. Size of auxiliary net. We grow the auxiliary net in stages by taking increasingly large segments of a $(0, d_2)$-sequence. After the jth step, letting n_j denote the size of the auxiliary net, one examines the diagnostics from Section 3.1. By proceeding in steps, other useful diagnostics can be carried out. One can plot $\hat{c}_i$ versus n_i for $i = 1, \ldots, j$, where $\hat{c}_i$ is the estimate for c after the ith step. This plot will be referred to as the *normalization constant plot*. Achieving a plateau in the normalization constant plot is a necessary condition for convergence. However, care must be taken since the aggregate nature of the $\hat{c}$s means that $\hat{c}_j$ will not change much if $n_j - n_{j-1}$ is small relative to n_{j-1}. If $d_1 \leq 2$ then one can also plot the marginal approximations after steps j and $j - 1$ on the same axes. Lack of agreement establishes the need for more evaluations. The same warnings as with the normalization constant plot apply. Based on the results of the diagnostics, one decides if the number of points in the auxiliary net should be increased.

2. Focusing resources. After the jth step, the diagnostics may indicate regions of the marginal space for which g_{T_1} is well approximated, while other regions require more evaluations. One does not want to waste evaluations on the former of these regions. So, we 'turn off' these points, skipping over them when doing more evaluations at the rest of the points in the main net.

3. Concentration of the main net. Separate from the issue of whether or not g_{T_1} is well approximated for individual points in main net is whether or not the main net N_1 is sufficiently dense. Indeed, even if one knew $g_{T_1}(x)$ exactly for all $x \in N_1$, it is possible that the main net is too sparse to capture important features of the underlying marginal

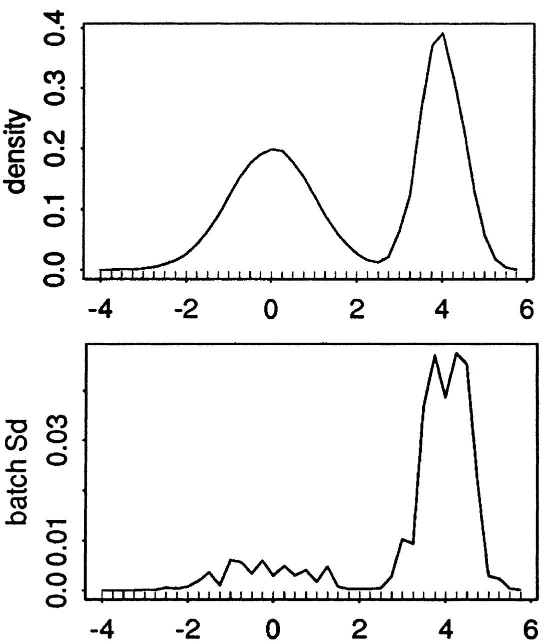

Fig. 5. *QMC approximation and SD plot for (m, n)= (40, 4000) with random origin shifts.*

density. One can never be completely certain that the main net is sufficiently dense for an unknown f, but there are diagnostics capable of suggesting when resolution is insufficient in a region. For example, if the marginal density is known to be smooth, then neighbouring estimates which do not change as n increases yet remain far apart from each other suggest the need for increased concentration in that neighbourhood. This manifests itself in the marginal density estimate appearing highly piecewise linear rather than smooth.

If the main net seems sparse in a region, we add points. First we turn back on points that were turned off during a previous step. If necessary we then add brand new points, although doing so in such a way as to maintain high local equi-distribution is no trivial task in more than one or two dimensions. But this is not a major problem since adding points to sub-regions destroys global equi-distribution and invalidates using (8) to approximate c. So, alternative quadrature methods allowing non-equally spaced points must used. This is not a problem provided d_1 is small.

Next, although many problems come with information about the support of f, the assumption that the domain of f lies in a known rectangle rarely holds. We proceed by guessing the support of f and acting as if it were the truth. Any guess should be spot checked by investigating f at points outside of the rectangle. This is quite an *ad hoc* approach, but it is a similar blend of art, science, and understanding of the physical problem behind the data that is used by Metropolis practitioners for finding starting values and transition kernels. The main net can be adapted, so the guess is less important in the corresponding coordinates. However, in the complement space too large a rectangle makes more evaluations necessary whereas too small a rectangle means we are missing support of the joint density. Clearly, estimating the wrong quantity is the greater evil, therefore the rectangle should be chosen somewhat conservatively. We do not pursue this here, but Fang and Wang (1993) discuss methods for generating QMC nets on non-rectangular regions, which may be more natural in some problems.

Example (continued). We now demonstrate a full AQMC treatment of our multivariate Gaussian example. We start with $m = 40$ points in the main net and an initial batch of $n_0 = 1000$ points in the auxiliary net. The first batch results in Fig. 4 as previously discussed. Based on the diagnostics, we 'turn off' the points under the first mode and increase the auxiliary net in steps of 1000 points. After six more batches, we end up with the marginal density approximation shown in the left panel of Fig. 6. The diagnostics (not all shown here) suggest that the approximations are close to the truth. In fact our relative error is about 1%. We note that since over half of the points in the main net only required 1000 evaluations each, the total number of evaluations is less than that required in Fig. 5, yet the accuracy is comparable. This demonstrates the advantage of adaptiveness.

For comparison, we also used Metropolis to approximate the same marginal density. A multivariate Gaussian transition kernel was used with covariance $K\mathbf{I}$. Four separate chains were run with K taken as 0.5, 1, 2, and 4. Starting values for the four runs were picked independently from a uniform distribution over the set used as support for the AQMC treatment. In each case the first 10 000 iterations were burned off and the remaining 75 000 were retained as the 'sample'. Density estimation using a Gaussian kernel was then applied to the samples to estimate the marginal density of the first coordinate. The results are found in Fig. 7. The variation between estimates demonstrates that the choice of transition kernel can have enormous impact on the results. Even the best of the four kernels explored here led to an unsatisfactory approximation when compared with the AQMC approximation. Also, there is extra variability introduced by density estimation when using any sampling based method.

4. Examples

The following (as well as the previous) AQMC examples were carried out using routines written in C by the authors. An interface with S-plus was used to create the graphics

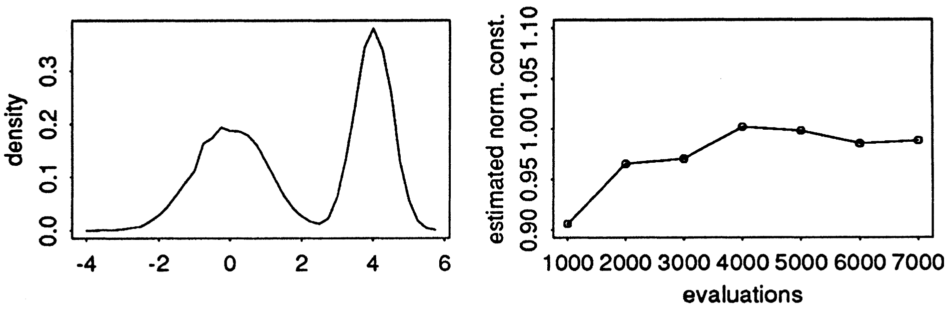

Fig. 6. *AQMC approximation and normalization constant plot. The approximations forming the first mode are based on 1000 evaluations, while those for the second are based on 7000 evaluations.*

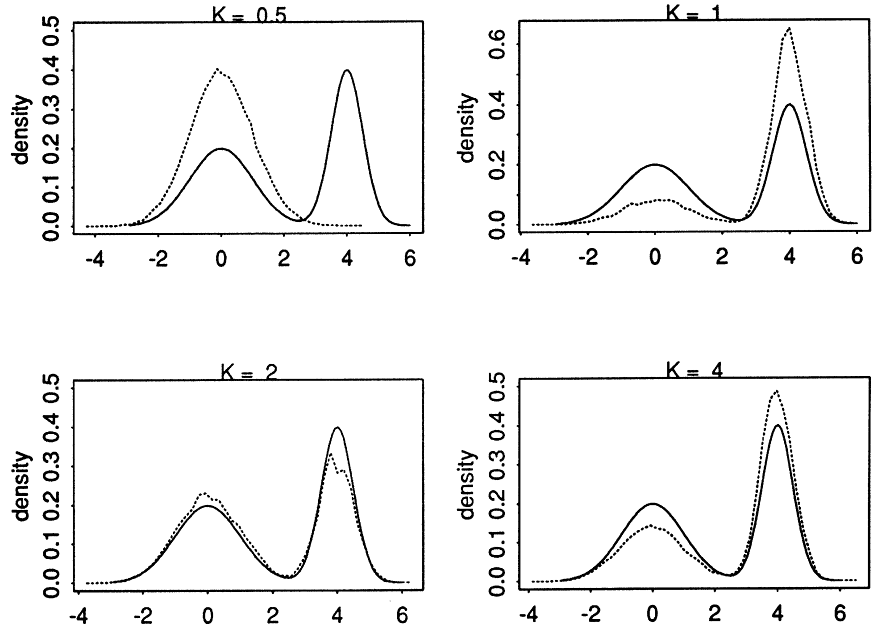

Fig. 7. *Metropolis approximations for a variety of transition kernels. The transition kernel is Gaussian with covariance K times the identity matrix. The first 10 000 iterations are burned off and the next 75 000 iterations form the sample. Density estimation used on samples to form dotted lines. The solid line is truth.*

and handle some of the small details. The functions apply the methods of the previous sections to the case of a one-dimensional marginal distribution. We are currently working towards making these functions available through Statlib.

4.1. *Bayesian posterior density*

Similar to the Metropolis–Hastings example of Chib and Greenberg (1995), we illustrate AQMC on a Bayesian analysis of an AR(2) model where our prior distribution is uniform over the region in which the series is stationary. In this situation the posterior distribution is intractable, but by focusing evaluations in the important areas AQMC quickly and accurately approximates the posterior marginal densities of the model parameters.

To begin with we simulated 100 observations from the model

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \epsilon_t, \quad t = 1,2,\ldots,100, \tag{10}$$

with $\alpha_1 = 1$, $\alpha_2 = -0.5$, $\sigma^2 = 1$, and ϵ_t iid $N(0,\sigma^2)$. For stationarity, $\theta = (\alpha_1,\alpha_2)'$ must lie in the region S where

$$S = \{(x,y) \in R^2 : x+y < 1; -x+y < 1; y > -1\}. \tag{11}$$

Following Box and Jenkins (1976), the likelihood function for this model given the $n = 100$ observations $(y_1,\ldots,y_n)$ is

$$l(\theta,\sigma^2) = \Psi(\theta,\sigma^2) \times (\sigma^2)^{-(n-2)/2} \times \exp\left[-\frac{1}{2\sigma^2}\sum_{s=3}^{n}(y_s - (y_{s-1},y_{s-2})\theta)^2\right] \tag{12}$$

where

$$\Psi(\theta,\sigma^2) = (\sigma^2)^{-1}|V^{-1}|^{1/2}\exp\left[-\frac{1}{2\sigma^2}(y_1,y_2)V^{-1}(y_1,y_2)'\right] \tag{13}$$

is the density of (y_1,y_2), and

$$V^{-1} = \begin{pmatrix} 1-\alpha_2{}^2 & -\alpha_1(1+\alpha_2) \\ -\alpha_1(1+\alpha_2) & 1-\alpha_2{}^2 \end{pmatrix}. \tag{14}$$

To do a Bayesian analysis where the prior is uniform on S, the posterior density is proportional to $g(\alpha_1,\alpha_2,\sigma^2) := l(\theta,\sigma^2)1_{\{\theta\in S\}}$. We use AQMC to approximate the marginal density for each of the three parameters separately. This is a particularly nice example for AQMC since the support of the first two coordinates of the posterior is known to lie in S. This suggests first approximating the marginal density of σ^2, making use of the restrictions on θ to ‘guess’ the support of f. Then, once the marginal of σ^2 is approximated, we incorporate this information to guess the rectangle for approximating the marginal density of α_1. Finally, we repeat this for α_2.

The solid lines in Fig. 8 show the AQMC marginal density approximation for each parameter. In each case $m = 20$ was the initial size of the main net and steps of 1000 evaluations in the auxiliary net were used. For each parameter, the adaptive techniques quickly indicated where more evaluations were needed, and just as importantly, where additional evaluations were not needed. The approximation for σ^2 was done first, and it used up to 10 000 evaluations at each point in order for all of the diagnostics to be satisfied. Using the results on σ^2 to narrow the domain of f, the approximation for α_1 required at most only 6000 evaluations. Similarly, α_2 required at most 2000 evaluations per point. In each case, additional points were added to the main net in active regions of the support in order to improve the picture of the marginal density. Judging by Fig. 8, the AQMC density approximations are centred on the true values of the parameters that generated the data. We note that it was necessary to scale up the unnormalized joint density in order to maintain numerical stability.

The dashed lines in Fig. 8 display the marginal density approximations based on a run of the Metropolis algorithm. A tri-variate normal jumping kernel was used with covariance 0.05**I**. An initial 5000 iterations were burned off, and 10 000 more iterations of the chain were executed to form the Metropolis sample. As is often the case, this is quite a bit fewer overall joint density evaluations than was necessary for AQMC. It should be noted however, that four chains were run examining the proportion of transitions (see Gelman *et al.*, 1995) before the value of 0.05 was finally accepted. Similar to before, kernel density estimation was used to estimate the marginal densities based on the respective coordinate of the Metropolis sample. Figure 8 shows that Metropolis and AQMC give very similar results.

4.2. *Some results in higher dimensions*

QMC must saturate a region with points in order to produce an accurate approximation. Since such saturation gets exponentially more difficult as dimension increases, there comes a point when QMC and AQMC are no longer feasible. Exactly how high a dimension can be handled depends on the particular f. For example, the one-dimensional marginal density of a unimodal, 10-dimensional Gaussian with identity covariance matrix was well approximated in less than 200 000 total joint density evaluations by AQMC (see Fig. 10). However, a bi-modal Gaussian as in our earlier 4-dimensional example was prohibitively difficult in only 8 dimensions. In particular, after about 5 million joint density evaluations, the diagnostics still indicate the need for more points. Figure 9 shows the AQMC approximations at this point.

Figure 10 shows discretized L_1 and supremum norm distances between approximations and truth for AQMC (solid) and Metropolis (dashed) as a function of joint density evaluations for the 10-dimensional Gaussian example of the previous paragraph. The well-tuned Metropolis chain was burned in 20 000 evaluations accounting for the right shift of the dotted line. Clearly, AQMC needs more evaluations to get started than Metropolis. But once the region starts to get saturated relative to the variability of f, AQMC appears just as accurate (in the L_1 and sup norm sense) as Metropolis. When using less well-tuned Metropolis chains, AQMC can eventually eclipse Met-

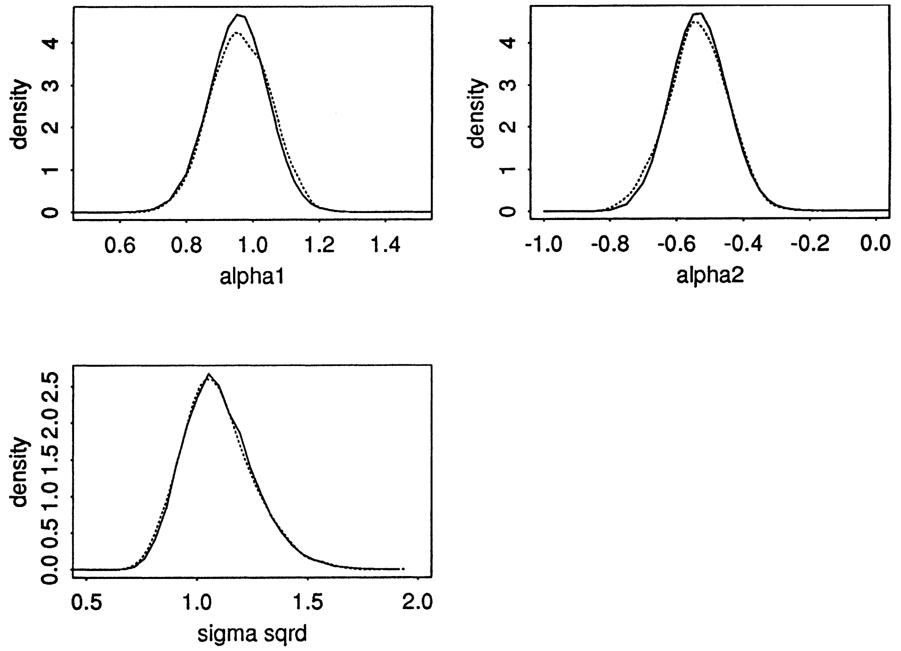

Fig. 8. *AQMC (solid) and Metropolis (dashed) approximations to marginal posterior densities in AR(2) model. For Metropolis: 5000 iterations burned, then density estimation done on next 10 000 iterations. Transition kernel is Gaussian with covariance 0.05 times identity matrix. For AQMC: (m, n)=(20, 1000) initially. Active points received 10 000, 6000, and 2000 evaluations respectively for σ^2, α_1, and α_2.*

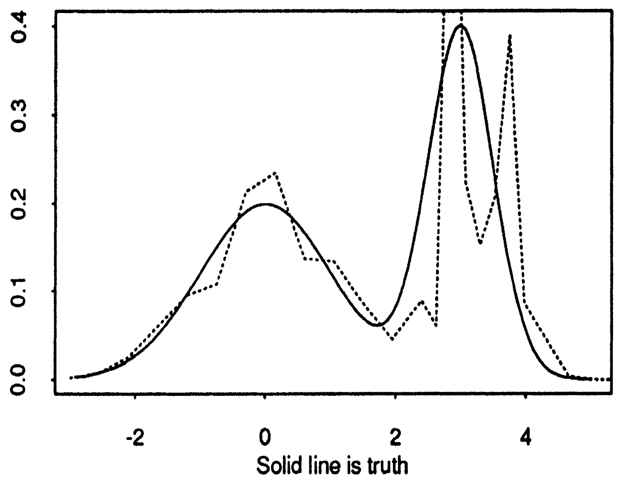

Fig. 9. *8-dimensional Gaussian mixture of the same form as the example in Section 3 (only with the modes a bit closer together). Solid line is truth. Note poor performance of AQMC despite nearly 5 million joint density evaluations in total.*

ropolis over the range of evaluations examined. In more difficult examples, such as that of Fig. 9, the point of saturation comes too late for AQMC to be of practical use.

5. Discussion

In this section we discuss some of the strengths and weaknesses of AQMC and Metropolis. Our view is that AQMC (and QMC in general) is worth consideration in a variety of situations. However, there are situations where QMC-based methods are of less value. First, if the dimension s of the space is large, or the estimated domain of f has too much volume relative to the variation of f, then saturation with points can be computationally prohibitive. How large is 'too large' depends on the particular f and how much computational muscle one has. Our examples show that a unimodal Gaussian with $s = 10$ was handled painlessly by AQMC, but more challenging bi-modal investigations suggest that AQMC is too slow for such an f around $s = 8$. Also, if marginal densities for several functions of the underlying random variable are desired, Metropolis handles this easily given a quality single sample from the marginal distribution. However, QMC requires analytical work to express the new marginal density in terms of the old or possibly even a brand new QMC calculation. A further strength of Metropolis was hinted at in the AR(2) example. If there is more than one margin of interest, QMC requires that each margin be treated sepa-

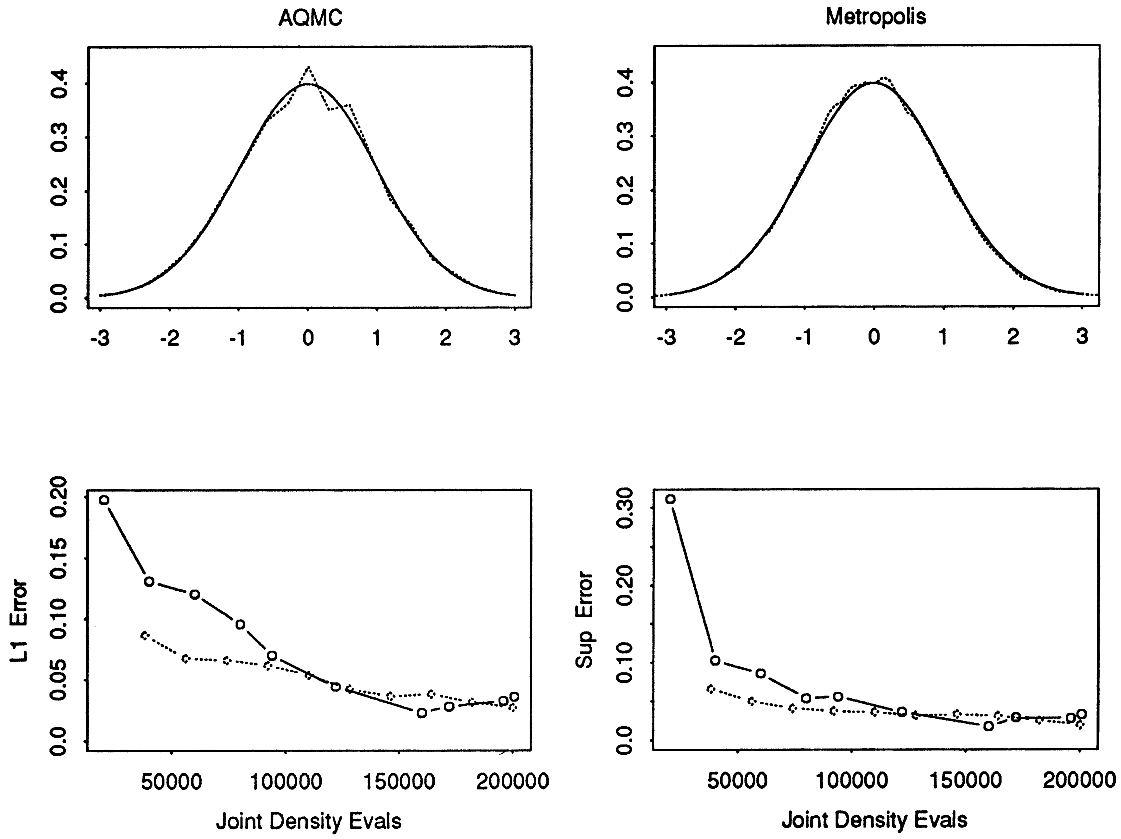

Fig. 10. *Marginal density approximations for 10-dimensional standard Normal. L_1 distances between approximations and truth as a function of joint density evaluations. Solid lines are for AQMC, dashed for Metropolis.*

rately, whereas Metropolis can simply examine the different margins of a single sample. In our example, three margins is not enough to cause major problems, especially since the marginal densities become successively easier to approximate.

For problems that do not fall into any of the categories listed in the previous paragraph, AQMC has several strengths. As our examples show AQMC can be very accurate. And by adaptively allocating resources AQMC can be computationally competitive with Metropolis. Further, approximating marginal densities with AQMC does not require any density estimation. In addition to being another source of subjectivity, density estimation can add substantial computational costs – especially if several binwidths or smoothing kernels are explored. Also, we agree with Cowles and Carlin (1996) who report (p. 902), 'many of the MCMC diagnostics proposed in the statistical literature to date are fairly difficult to use, requiring problem-specific coding and perhaps analytical work.' Such difficulty opens the door that much wider for human error. We believe that faulty computer code does not get nearly the credit it deserves for spoiling analyses. The possibility of coding error exists in AQMC too, but the differing approaches of AQMC and Metropolis ensure that beyond the joint density evaluations, no single bug can invalidate both results. Moreover, the simple intuition of adding more points where more points seem necessary means that complicated problem-specific code need not be written to use AQMC. In fact, we used the exact same code (aside from joint density evaluations, of course) for all the examples given.

Finally, we have mentioned the lack of consensus in applying Metropolis in terms of choosing the transition kernel, burn length, run length, number of chains, etc. and Fig. 7 demonstrates the importance of these decisions. Of course, similar complaints can be made of AQMC. There are no free lunches, and AQMC pays in having to guess the approximate support of the joint density and having to know when enough points have been laid down. But knowing the approximate support of the joint density requires a different sort of knowledge than knowing what is a good transition kernel for a particular joint density. The fact that these types of knowledge are different makes AQMC an excellent check for the results of Metropolis and vice versa. And in problems such as our AR(2) example where partial or complete knowledge of the support of f is available a priori, AQMC involves less guesswork or fine tuning to implement.

6. Conclusion

Our experience with AQMC for approximating marginal densities suggests that AQMC is highly accurate and computationally competitive with Metropolis when the overall dimension is modest and both the number and dimension of marginal distributions of interest are small. Furthermore, AQMC can be particularly effective if information about the support of the joint distribution is available, as in the case of our AR(2) example. Most importantly, once routines for generating points from $(0, s)$-sequences are written, AQMC is straightforward and intuitive to implement, and the batch-wise approach provides a variety of useful diagnostics that are suggestive of the magnitude of the approximation errors. Finally, by allowing the researcher to focus computational resources where needed, AQMC is able to handle problems beyond the computational feasibility limits of regular QMC.

In the hands of an experienced practitioner who has years of iterative sampling experience upon which to draw, we do not doubt that Metropolis and Metropolis–Hastings are powerful tools for exploring marginal distributions in a wide class of problems. And it is also true that for high-dimensional problems there is often no substitute for the iterative sampling methods. However, for the non-expert working in modest dimensions with only a few parameters of interest, AQMC is an excellent alternative to Metropolis – both as an independent check for results from a 'black box' implementation of Metropolis and as an easy-to-use, viable method in its own right. Furthermore, in moderate dimensions where the support of the joint density is well known, the intuitive AQMC may be preferable over a Metropolis that requires complicated fine tuning.

Acknowledgements

We would like to thank Art Owen for helpful discussions and access to his preprints. We also thank the two anonymous referees for their valuable comments and Rudy Beran for bringing our attention to QMC in the first place. Both authors gratefully acknowledge ARO grant DAAH04-94-G-0232 for support.

References

Box, G. E. P. and Jenkins, G. M. (1976) *Time Series Analysis Forecasting and Control*, San Francisco: Holden Day.

Chib, S. and Greenberg, E. (1994) Bayes inference for regression models with ARMA(p,q) errors. *Journal of Econometrics*, **64**, 183–206.

Chib, S. and Greenberg, E. (1995) Understanding the Metropolis–Hastings algorithm. *American Statistician*, **49**, 327–35.

Cowles, M. K. and Carlin, B. P. (1996) Markov chain monte carlo convergence diagnostics: a comparitive review. *Journal of the American Statistical Association*, **91**, 883–904.

Cranley, R. and Patterson, T. N. L. (1976) Randomization of number theoretic methods for multiple integration. *SIAM Journal of Numerical Analysis*, **13**, 904–14.

Davis, P. J. and Rabinowitz, P. (1984) *Methods of Numerical Integration, 2nd edn*, Orlando: Academic Press Inc.

Fang, K. T. and Wang, Y. (1993) *Number Theoretic Methods in Statistics*, London: Chapman & Hall.

Faure, H. (1982) Discrépance de suites assocités à un syteme de numération (en dimensions). *Acta Arithmetica*, **41**, 337–351.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457–72.

Gelman, A., Roberts, G. and Gilks, W. (1995) Efficient Metropolis jumping rules. In J.M. Bernardo, J.O. Berger, A. P. Dawid and A. F. M. Smith (eds) *Bayesian Statistics 5*. New York: Oxford Press.

Geyer, C. (1992) Practical Markov chain Monte Carlo (with discussion). *Statistical Science*, **7**, 473–511.

Hua, L. K. and Wang, Y. (1981) *Application of Number Theory to Numerical Analysis*, Berlin: Springer.

Leonard, T., Hsu, J. S. J. and Ritter, C. (1994) The Laplacian t-approximation in Bayesian inference. *Statistica Sinica*, **4**, 127–42.

Niederreiter, H. (1978) Quasi-Monte Carlo methods and pseudorandom numbers. *Bulletin of the American Mathematical Society*, **84**, 957–1042.

Niederreiter, H. (1992) *Random Number Generation and Quasi-Monte Carlo Methods*, Philadelphia: Siam.

Owen, A. (1995) Randomly permuted (t,m,s)-nets and (t,s)-sequences. In H. Niederreiter and J.S. Shiue (eds) *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing* NY: Springer.

Owen, A. (1996) Monte Carlo variance of scrambled equidistribution quadriture. *SIAM Journal of Numerical Analysis*, to appear.

Ripley, B. (1987) *Stochiastic Simulation*, New York: Wiley.

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

Shaw, J. E. H. (1988) A quasirandom approach to integration in Bayesian statistics. *Annals of Statistics*, **16**, 895–914.

Tierney, L. (1994) Markov chains for exploring posterior distributions. *Annals of Statistics*, **22**, 1701–62.

Tierney, L., Kass, R. E. and Kadane, J. B. (1989) Approximate marginal densities of nonlinear functions. *Biometrika*, **76**, 425–33.

Appendix

Construction of $(0, s)$-sequences for prime bases

Here we show how to construct points from a $(0, s)$-sequence in base b. This construction was introduced by Faure and requires that $b \geq 2$ is prime and the dimension satisfies $s \leq b$. We note that constructions for many non-prime bases are available. Here we simply outline the steps and refer interested readers to Niederreiter (1992) for thorough details. Our notation and development draw heavily from chapter four of Niederreiter (1992).

For $n = 1, 2, \ldots$, first find the base b expansion of $n - 1$. That is find integers $a_j \in \{0, \ldots, b-1\}$ such that

$$n - 1 = \sum_{r=0}^{m_n} a_r b^r. \tag{15}$$

Here, m_n is just the number of digits necessary to express $n - 1$ in base b. For $1 \leq i \leq s$, and $1 \leq j \leq m_n$ define

$$y_{nj}(i) = \left(\sum_{r=0}^{m_n} c_{jr}(i) a_r \right) \bmod b, \tag{16}$$

where

$$\begin{aligned} c_{ir}(i) &= 0 && \text{for } 0 \leq r < j-1 \\ &= \binom{r}{j-1} (i-1)^{r-j+1} && \text{for } j-1 \leq r \leq m_n \end{aligned}$$

and 0^0 is taken to be 1 by convention. Then, the nth point of a $(0, s)$-sequence in base b is given by $x_n = (x_n^{(1)}, \ldots, x_n^{(s)})$ where

$$x_n^{(i)} = \sum_{j=1}^{m_n} y_{nj}(i) b^{-j}. \tag{17}$$