SIAM J. SCI. COMPUT.
Vol. 15, No. 6, pp. 1251–1279, November 1994

QUASI-RANDOM SEQUENCES AND THEIR DISCREPANCIES*

WILLIAM J. MOROKOFF† AND RUSSEL E. CAFLISCH†

Abstract. Quasi-random (also called low discrepancy) sequences are a deterministic alternative to random sequences for use in Monte Carlo methods, such as integration and particle simulations of transport processes. The error in uniformity for such a sequence of N points in the s-dimensional unit cube is measured by its discrepancy, which is of size $(\log N)^s N^{-1}$ for large N, as opposed to discrepancy of size $(\log\log N)^{1/2} N^{-1/2}$ for a random sequence (i.e., for almost any randomly chosen sequence). Several types of discrepancies, one of which is new, are defined and analyzed. A critical discussion of the theoretical bounds on these discrepancies is presented. Computations of discrepancies are presented for a wide choice of dimension s, number of points N, and different quasi-random sequences. In particular for moderate or large s, there is an intermediate regime in which the discrepancy of a quasi-random sequence is almost exactly the same as that of a randomly chosen sequence. A simplified proof is given for Woźniakowski's result relating discrepancy and average integration error, and this result is generalized to other measures on function space.

Key words. Monte Carlo, quasi-random, discrepancy, Brownian sheet

AMS subject classifications. 65C05, 10K30

1. Introduction.

Since the beginning of the computer age, Monte Carlo methods have been used to evaluate integrals, solve integral equations, and simulate physical processes [7]. These methods use a sequence of points, usually a deterministic pseudo-random approximation to a randomly chosen sequence, to sample the values of the integrand function or the possible steps in a process. Over the years a number of techniques, such as variance reduction through stratification, have been developed to improve the accuracy of these methods. An alternative technique is to replace the pseudo-random sequence with a deterministic sequence having better uniformity properties.

Uniformity of a sequence is measured by its discrepancy, which is the error in representation of the volume of subsets of the unit cube by the fraction of points in the subsets. Several different definitions of discrepancy can be formulated [8], [14], [17], including a sup over rectangles or an L_1 or L_2 integral over rectangles using either all rectangles or only those with one vertex at the origin. Integration error can be related to discrepancy either through the Koksma–Hlawka inequality [8], [14], [17] or Woźniakowksi's identity [26], which states that the discrepancy is equal to the average integration error with respect to the Brownian sheet measure. We present a critical discussion of the various definitions of discrepancy, one of which is new, and a simplified proof of Woźniakowksi's result is given, which allows the result to be generalized to other measures on function space.

A quasi-random (or low discrepancy) sequence in the s-dimensional cube is a sequence for which the discrepancy is roughly of size $(\log N)^s N^{-1}$ for large N, which is the minimum size possible. These sequences are more uniform than random sequences because randomly chosen points tend to clump, leading to discrepancy of size $(\log\log N)^{1/2} N^{-1/2}$. Evidence of this clumping is shown in a planar projection of a pseudo-random sequence in Fig. 12; while the top graph of Fig. 13 shows the uniformity that can be achieved with quasi-random points. At the other extreme, regular lattices of points work well in low dimension, but in high dimension they are not very useful. Points cannot be added to a lattice incrementally. Instead, a given s-dimensional lattice can only be refined by increasing the number of points by a factor 2^s; i.e., the discrepancy of a lattice is of size $O(1)$, except at special values of N

*Received by the editors July 17, 1991; accepted for publication (in revised form) August 30, 1993. Research was supported in part by the Air Force Office of Scientific Research grant AFOSR 90-0003.

†Mathematics Department, University of California Los Angeles, Los Angeles, California 90024-1555 (morokoff@math.ucla.edu, caflisch@math.ucla.edu).

at which the lattice is completely refined. Moreover for large s it is usually impossible to put down enough lattice points to get good resolution.

Quasi-random sequences combine the advantage of a random sequence (points can be added incrementally) with the advantage of a lattice (no clumping of points). Examples of such sequences that will be considered below include the Halton sequence, Sobol' sequence, and Faure sequence. Bounds on the discrepancy of these sequences, as well as other analytic properties, have been previously derived using number theoretic techniques [8], [14]–[17]. An alternative method for generating quasi-random sequences and bounds on integration error using a dynamical systems approach is presented in [19].

The main portion of this paper consists of computations and critical discussion of the discrepancy for these sequences over a large range of values of N and s. In particular for large dimension s, the theoretical bound $(\log N)^s N^{-1}$ is only meaningful for extremely large values of N; i.e., $N = O(e^s)$.

In an attempt to directly understand the uniformity properties of quasi-random sequences in high dimension, two-dimensional projections are presented for a variety of quasi-random sequences. These can show considerable clumping in high dimension. Finally we present timing results for the generation of the different quasi-random sequences.

Previous computational studies of quasi-random sequences (as well as scrambled quasi-random sequences) and their discrepancy by Braaten and Weller [1], Bratley and Fox [2], Bratley, Fox, and Niederreiter [3], Levitan [9], and Pages and Xiao [20] presented some useful, but less complete, results. Some of their results were due to transient effects, such as those described below for the T_N^* discrepancy, and some of the multidimensional integration tests were only performed for product functions. Sarkar and Prasad [22] have given a comparison of Halton, scrambled Halton, and Faure sequences as applied to an absorption problem. In a cogent article, Press and Teukolsky [21] discussed the Sobol' sequences and computational methods for generating them and showed how discontinuous integrand functions decrease the effectiveness of Monte Carlo integration with quasi-random sequences. In two companion papers [12], [13] we present computational studies and some analysis for quasi-Monte Carlo methods applied to integration and simulation of some simple transport processes.

2. Discrepancy and integration error for quasi-random sequences. Given that pseudo-random sequences work well in Monte Carlo integration, it seems reasonable to ask if other deterministic sequences might also work. More precisely, it seems that the independence of random numbers plays a secondary role to their uniformity in Monte Carlo calculations; so sequences with better uniformity properties may lead to smaller errors. In order to develop this idea it is necessary to define a uniform sequence and some measure of its uniformity. The following is based on Niederreiter's development of the topic in [14].

Let I^s denote the s-dimensional unit cube. An infinite sequence $\{x_n\}$ in I^s is called uniformly distributed if for all Jordan measurable subsets J of I^s,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathcal{X}_J(\boldsymbol{x}_n) = m(J)$$

holds, where $\mathcal{X}_J$ is the characteristic function of J, and $m(J)$ is the volume of J. Thus in the limit of an infinite number of points, every region in I^s has proportionally the right number of points. From this definition it follows that a sequence $\{\boldsymbol{x}_n\}$ is uniformly distributed if for all Riemann integrable functions f defined on I^s it holds that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(\boldsymbol{x}_n) = \int_{I^s} f(\boldsymbol{x})\, d\boldsymbol{x}\,.$$

It follows from the Central Limit Theorem that a sequence of independent, random points chosen from the interval I^s with probability density one is indeed a uniformly distributed sequence with probability one.

Practically, it is only possible to deal with a finite number of integration nodes, so it is necessary to define some measure of uniformity for finite point sets. Such a quantity is known as discrepancy. For a set $J \subseteq I^s$ and a sequence of N points $\{\boldsymbol{x}_n\}$ in I^s, define

$$R_N(J) = \frac{1}{N}\sum_{n=1}^{N} \chi_J(\boldsymbol{x}_n) - m(J)\,.$$

Various kinds of discrepancies can be defined then by restricting J to a certain class of sets and taking a norm of R_N over this class. If E is the set of all subrectangles of I^s, then the L_∞ and L_2 norms are defined as

$$D_N = \sup_{J\in E} \mid R_N(J) \mid, \tag{1}$$

$$T_N = \left[\int_{(\boldsymbol{x},\boldsymbol{y})\in I^{2s},\, x_i<y_i} (R_N(J(\boldsymbol{x},\boldsymbol{y})))^2\, d\boldsymbol{x}\, d\boldsymbol{y}\right]^{\frac{1}{2}}. \tag{2}$$

Here $J(\boldsymbol{x},\boldsymbol{y})$ indicates the rectangle with opposite corners at $(\boldsymbol{x},\boldsymbol{y})$. If E^* is the set of subrectangles with one corner at $\mathbf{0}$, then the star discrepancies are defined as

$$D_N^* = \sup_{J\in E^*} \mid R_N(J) \mid, \tag{3}$$

$$T_N^* = \left[\int_{I^s} (R_N(J(\boldsymbol{x})))^2\, d\boldsymbol{x}\right]^{\frac{1}{2}}. \tag{4}$$

Here $J(\boldsymbol{x})$ is the rectangle with a corner at $\mathbf{0}$ and a corner at $\boldsymbol{x}$. It should be noted that this is not the standard notation as used by Niederreiter [14] and others. In the past, T_N has denoted what is here referred to as T_N^*. The new notation is necessary because the L_2 discrepancy over all rectangles T_N had not been previously defined nor used. To be consistent with the sup discrepancy, it makes sense to relabel the original L_2 discrepancy over rectangles with a corner at zero as T_N^*, and call the new L_2 discrepancy T_N. Another kind of discrepancy, J_N, is obtained by taking the sup over all convex sets. No L_2 analog exists for this class. The infinite sequence $\{\boldsymbol{x}_n\}$ being uniformly distributed is equivalent to $\lim_{N\to\infty} D_N = 0$, where D_N refers to the discrepancy of the first N terms of the sequence. The statement is true for all of the above discrepancies.

The importance of discrepancy as an error bound for Monte Carlo integration can be seen from the Koksma–Hlawka inequality, which in one dimension for smooth functions f of bounded variation reads

$$\varepsilon(f) = \left|\int_0^1 f(x)\, dx - \frac{1}{N}\sum_{n=1}^{N} f(x_n)\right| \le V(f)\, D_N^*, \tag{5}$$

where D_N^* is the discrepancy of the sequence $\{x_n\}$ and $V(f) = \int_0^1 |df|$ is the variation of f.

Inequality (5) can be extended to higher dimensions; however, the definition of variation must be modified. Assume for the moment that f is sufficiently smooth on I^s so that

$$V^{(s)}(f) = \int_{I^s} \left|\frac{\partial^s f}{\partial t_1 \ldots \partial t_s}\right| dt$$

exists. Define for all $k \leq s$ and all sets of k integers $1 \leq i_1 < i_2 < \cdots < i_k \leq s$ the quantity

$$V^{(k)}(f; i_1, \ldots, i_k) = \int_{I^k} \left| \frac{\partial^k f}{\partial t_{i_1} \ldots \partial t_{i_k}} \right|_{t_j = 1, j \neq i_1, \ldots, i_k} dt_{i_1} \ldots dt_{i_k} .$$

Similarly define $D_N^*(i_1, \ldots, i_k)$ to be the star discrepancy of the orthogonal projection of the sequence $\{\boldsymbol{x}_n\}$ on to the appropriate k-dimensional subspace of I^s. Then

$$\left| \int_{I^s} f(\boldsymbol{x})\, d\boldsymbol{x} - \frac{1}{N} \sum_{n=1}^{N} f(\boldsymbol{x}_n) \right| \leq \sum_{k=1}^{s} \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq s} V^{(k)}(f; i_1, \ldots, i_k)\, D_N^*(i_1, \ldots, i_k) .$$

It is possible to obtain a similar bound involving $T_N^*(i_1, \ldots, i_k)$ and an L_2 version of $V^{(k)}$ defined by

$$W^{(k)}(f; i_1, \ldots, i_k) = \left[\int_{I^k} \left(\frac{\partial^k f}{\partial t_{i_1} \ldots \partial t_{i_k}} \Big|_{t_j = 1, j \neq i_1, \ldots, i_k} \right)^2 dt_{i_1} \ldots dt_{i_k} \right]^{\frac{1}{2}} .$$

This requires a stronger condition on the integrand f than just being of L_1 bounded variation. The resulting bound is

$$\left| \int_{I^s} f(\boldsymbol{x})\, d\boldsymbol{x} - \frac{1}{N} \sum_{n=1}^{N} f(\boldsymbol{x}_n) \right| \leq \sum_{k=1}^{s} \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq s} W^{(k)}(f; i_1, \ldots, i_k)\, T_N^*(i_1, \ldots, i_k) .$$

This argument may be made rigorous for a broader class of f known as functions of bounded variation in the sense of Hardy and Krause. See [14] for details.

It is easy to see that $D_N^* \geq D_N^*(i_1, \ldots, i_k)$ for all $k \leq s$. If the variation of f in the sense of Hardy and Krause is defined as

$$V(f) = \sum_{k=1}^{s} \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq s} V^{(k)}(f; i_1, \ldots, i_k) ,$$

then inequality (5) follows immediately. This relationship does not follow when the T_N^* discrepancy is used; in fact, in our computational tests, $T_N^* \leq T_N^*(i_1, \ldots, i_k)$.

The inequality (5) shows that if sequences exist with lower discrepancy than random sequences, better than random convergence may be possible. Several such low discrepancy sequences are discussed in the next section. However, first it is helpful to examine some of the basic properties of discrepancy.

For a random sequence it can be shown from the law of iterated logarithms that

$$D_N = \mathcal{O}\left(\left(\frac{\log \log N}{N} \right)^{\frac{1}{2}} \right)$$

with probability one (see [14, p. 971]). This is true for any dimension s. Calculations carried out in §6 show that $E(T_N^2) = C_s/N$, where $E(\cdot)$ is the expectation, taken to be an integral over the space I^{sN} of possible values of the s coordinates of the N random points. C_s is a constant depending on dimension. Both of these estimates show $N^{-.5}$ type convergence, which corresponds to the standard Monte Carlo error behavior for integration with random nodes.

Various relationships exist among the different notions of discrepancy, which allow a sequence to be termed low discrepancy without specifying the measure. Perhaps the simplest relationship is

$$D_N^* \leq D_N \leq 2^s D_N^* .$$

This is clear from the fact that E^* is a subset of E, while any set in E can be written as a combination of 2^s sets in E^*. From the basic fact that L_∞ norms are larger than L_2 norms, it follows that

$$T_N^* \leq D_N^* .$$

The relationship between T_N^* and T_N is discussed in §5; in general, the star discrepancy is larger. In [16] Niederreiter establishes the relationship

$$C_s (D_N)^{\frac{s+2}{2}} \leq T_N^* .$$

In his book with Kuipers [8] he also shows that the isotropic discrepancy over convex sets satisfies

$$J_N \leq 4s(D_N)^{\frac{1}{s}} .$$

This bound is improved in [13] under the assumption suggested by Press and Teukolsky [21].

Other properties of discrepancy include the lower bound established by Roth and discussed in [8];

$$T_N^* > C_s \frac{(\log N)^{\frac{s-1}{2}}}{N} .$$

Halton [6] showed the existence of infinite sequences in any dimension which satisfy

$$D_N = \mathcal{O}\left(\frac{(\log N)^s}{N}\right) .$$

This bound is regarded as the best possible. This is an important result because it offers hope that the standard Monte Carlo $N^{-1/2}$ convergence can be improved considerably. Sequences with this property are the topic of the §4.

3. Average integration error. A direct relation between the integration error $\varepsilon(f)$ and the L^2 discrepancy has been derived by Woźniakowski [26]. He showed that $(T_N^*)^2$ is equal to the average integration error, i.e.,

$$(T_N^*)^2 = E(\varepsilon(f)^2) \tag{6}$$

in which $\varepsilon(f)$ denotes the integration error on the left side of (5) above. The average is taken with respect to the "Brownian sheet" measure, which is a generalization of Brownian motion with s-dimensional "time." In particular the measure is concentrated on functions that are roughly "half-differentiable" (i.e., they have Hölder exponent nearly equal to $\frac{1}{2}$), so that they have infinite variation. This shows that the Koksma–Hlawka inequality (5) is a vast overestimate, at least for this class of functions.

The Brownian sheet measure is a measure on function space. It is a natural generalization of the simple Brownian motion $b(x)$ to multidimensional "time" $\boldsymbol{x}$. Denote $\boldsymbol{x}' = (x_i')_{i=1}^s$ in which $x_i' = 1 - x_i$ and (with the usual abuse of notation) denote $f(\boldsymbol{x}') = f(\boldsymbol{x})$; also

denote the finite difference operator $D_i f(\boldsymbol{x}') = f(\boldsymbol{x}' + \Delta_i \hat{e}_i) - f(\boldsymbol{x}')$ in which $\hat{e}_i$ is the ith coordinate vector. The Brownian sheet is based at the point $\boldsymbol{x}' = 0$, i.e., $\boldsymbol{x} = (1, \ldots, 1)$, and has $f(\boldsymbol{x}' = 0) = f(1, \ldots, 1) = 0$. For any point $\boldsymbol{x}'$ in I^s and any set of positive lengths Δ_i (with $x_i' + \Delta_i < 1$), the multidimensional difference $D_1 \ldots D_s f(\boldsymbol{x})$ is a normally distributed random variable with mean zero and variance

$$E((D_1 \ldots D_s f(\boldsymbol{x}))^2) = \Delta_1 \ldots \Delta_s. \tag{7}$$

This implies that

$$E(df(\boldsymbol{x})df(\boldsymbol{y})) = \delta(\boldsymbol{x} - \boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y} \tag{8}$$

in which df is understood in the sense of the Ito calculus [10]. Moreover $f(\boldsymbol{x}') = 0$ if $x_i' = 0$ for any i, and for any $\boldsymbol{x}$ in I^s, $f(\boldsymbol{x})$ is normally distributed with mean zero and variance

$$E(f(\boldsymbol{x})^2) = \prod_{i=1}^{s} x_i'. \tag{9}$$

The Brownian sheet has the same covariance properties as the product of independent Brownian motions $\tilde{f}(\boldsymbol{x}) = \prod_{i=1}^{s} x_i b_i(x_i)$, but $f \neq \tilde{f}$ since the product is not normally distributed.

The derivation of (6) in [26] was simply a calculation of each side of the equation. Here we present a new derivation that follows naturally the properties of the Brownian sheet measure. First rewrite the integration error $E(f)$ using integration by parts, following the steps of the proof of the Koksma–Hlawka inequality [14]. Note that

$$dR(\boldsymbol{x}) = \left\{ \frac{1}{N} \sum_{n=1}^{N} \delta(\boldsymbol{x} - \boldsymbol{x}_n) - 1 \right\} d\boldsymbol{x} \tag{10}$$

in which $R(\boldsymbol{x}) = R_N(J(\boldsymbol{x}))$ as defined in §2. Also $R(\boldsymbol{x}) = 0$ if $x_i = 0$ and $f(\boldsymbol{x}) = 0$ if $x_i = 1$ for any i, which implies that the boundary terms all disappear in the following integration by parts:

$$\begin{aligned} \varepsilon(f) &= \left| \int_{I^s} f(\boldsymbol{x})d\boldsymbol{x} - \frac{1}{N} \sum_{n=1}^{N} f(\boldsymbol{x}_n) \right| \\ &= \left| \int_{I^s} \left\{ 1 - \frac{1}{N} \sum_{n=1}^{N} \delta(\boldsymbol{x} - \boldsymbol{x}_n) \right\} f(\boldsymbol{x})d\boldsymbol{x} \right| \\ &= \left| \int_{I^s} R(\boldsymbol{x})df(\boldsymbol{x}) \right|. \end{aligned}$$

The quantity df in this identity is defined here through the Ito calculus, even though $V(f) = \infty$ with probability one.

It follows from (8) that the average square error is

$$\begin{aligned} E(\varepsilon(f)^2) &= E\left(\left(\int_{I^s} R(\boldsymbol{x})df(\boldsymbol{x}) \right)^2 \right) \\ &= \int_{I^s \times I^s} R(\boldsymbol{x})R(\boldsymbol{y})E(df(\boldsymbol{x})df(\boldsymbol{y})) \\ &= \int_{I^s} R(\boldsymbol{x})^2 d\boldsymbol{x} \\ &= (T_N^*)^2. \end{aligned} \tag{11}$$

One unnatural feature of the Brownian sheet measure used above is that the functions f are all required to vanish on the boundaries $x_i' = 0$ (i.e., $x_i = 1$) for all i. This restriction can be removed by a generalization of the Brownian sheet that puts the values on the boundaries $x_i' = 0$ to be generalized Brownian sheets.

First set $f(\boldsymbol{x}' = 0) = f(1, \ldots, 1) = 0$. Then on the s coordinate lines emanating from the origin, let f be given by s independent Brownian motions. Next on the two-dimensional boundaries, where all but two of the x_i' are zero, define f by the Brownian sheet property with the given boundary conditions on the two sides. Continue this procedure until all boundaries and finally the interior have been defined. The resulting measure still satisfies the equation (8) and has the following covariance:

$$(E(f(\boldsymbol{x}')f(\boldsymbol{y}')) = \prod_{i=1}^{s} \min(x_i', y_i') + \sum_{j=1}^{s} \prod_{i\neq j, i=1}^{s} \min(x_i', y_i') + \cdots + \sum_{i=1}^{s} \min(x_i', y_i'). \tag{12}$$

In particular note that

$$E(df(\boldsymbol{x}')f(\boldsymbol{y}')) = 0 \tag{13}$$

if $\boldsymbol{x}'$ is an interior points of I^s and $y_i' = 0$ for some i.

Now compute the average integration error as before. The boundary terms do not vanish but are all independent of each other. Use integration by parts to find that the integration error is

$$\begin{aligned} \varepsilon &= \left| \int_{I^s} f(\boldsymbol{x}) dR(\boldsymbol{x}) \right| \\ &= \left| \int_{I^s} R(\boldsymbol{x}) df(\boldsymbol{x}) + \sum_{i=1}^{s} \int_{I^s \cap x_i = 1} f(\boldsymbol{x}) dR(\boldsymbol{x}) \right| \end{aligned}$$

since $R(\boldsymbol{x}) = 0$ if $x_i = 0$ for any i. The first term and the sum are independent according to (13); also integration by parts can be performed on each term in the sum. The result is similar to the covariance equation (12); i.e.,

$$\begin{aligned} E(\varepsilon^2) &= \int_{I^s}\int_{I^s} R(\boldsymbol{x})R(\boldsymbol{y})E(df(\boldsymbol{x})df(\boldsymbol{y})) \\ &\quad + \sum_{i=1}^{s} \int_{I^s \cap x_i = 1} \int_{I^s \cap y_i = 1} R(\boldsymbol{x})R(\boldsymbol{y})E(df(\boldsymbol{x})df(\boldsymbol{y})) \\ &\quad + \cdots + \sum_{i=1}^{s} \int_{0<x_i<1, x_j=1 \text{ if } j\neq i} \int_{0<x_i<1, y_j=1 \text{ if } j\neq i} R(\boldsymbol{x})R(\boldsymbol{y})E(df(\boldsymbol{x})df(\boldsymbol{y})) \qquad (14) \\ &= \int_{I^s} R(\boldsymbol{x})^2 d\boldsymbol{x} + \sum_{i=1}^{s} \int_{I^s \cap x_i = 1} R(\boldsymbol{x})^2 d\boldsymbol{x} + \cdots + \sum_{i=1}^{s} \int_{0<x_i<1, x_j=1 \text{ if } j\neq i} R(\boldsymbol{x})^2 d\boldsymbol{x} \\ &= T_n^* + \sum_{i=1}^{s} T_n^*(i) + \cdots + \sum_{i=1}^{s} T_n^*(1, \ldots, i-1, i+1, \ldots, s) \end{aligned}$$

in which $T_n^*(i_1, \ldots, i_k)$ is the L_2^* discrepancy for the sequence projected onto the boundary $x_j = 1$ for $j = i_1, \ldots, i_k$.

The point 0 still plays a special role in this measure. A still more uniform measure would be to center the Brownian sheet at a random point $\boldsymbol{y}$ inside I^s. In this case the first term in $E(\varepsilon^2)$ is just the L^2 discrepancy T_N (without *), but there are correlations between the boundary terms and the interior, as well as between different boundary terms, so that no simple equation results.

4. Low discrepancy sequences. Numerous sequences have been shown to have $\mathcal{O}((\log N)^s/N)$ behavior for their discrepancies. Included among these are sequences of the form

$$\boldsymbol{x}_n = ([n\alpha_1], \ldots, [n\alpha_s])$$

where the α_i are irrational numbers, which are linearly independent over the rationals, and $[\cdot]$ denotes the fractional portion of the number. Another sequence that has been suggested involves using a pseudo-random generator. If $\boldsymbol{x}_1 = (\rho_1, \ldots, \rho_s)$ is a random point, then the next term in the quasi-random sequence is given by $\boldsymbol{x}_2 = (\rho_2, \ldots, \rho_s, \rho_{s+1})$, where ρ_{s+1} is the next number produced by the generator. Various techniques have been applied to determine the discrepancy bounds for these sequences.

The sequences that have generated the most attention, and which have been studied in the current research, are those based on the p-adic expansion of the integers. For any integer n, let $(n)_p = a_k a_{k-1} \ldots a_0$ be the base p expansion of n with $0 \le a_i < p$. Define

$$S_p(n) = \frac{a_0}{p} + \frac{a_1}{p^2} + \cdots + \frac{a_k}{p^{k+1}}.$$

Then $0 < S_p(n) < 1$ for all n, and the sequence $S_p(n)$ is a one-dimensional uniformly distributed sequence. For $p = 2$ this is known as the van der Corput sequence. An s-dimensional generalization of this sequence $\{\boldsymbol{x}_n\}$, known as the Halton sequence, is given by

$$\boldsymbol{x}_n = (S_{p_1}(n), \ldots, S_{p_s}(n)),$$

where $(p_1, \ldots, p_s)$ are relatively prime integers, usually taken to be the first s primes. Several authors have derived bounds for the discrepancy of the Halton sequence. Meijer [11] shows that

$$D_N^* \le C_s^H \frac{(\log N)^s}{N} + \mathcal{O}\left(\frac{(\log N)^{s-1}}{N}\right) \tag{15}$$

where

$$C_s^H = \prod_{k=1}^{s} \frac{p_k - 1}{2 \log p_k}. \tag{16}$$

An unfortunate aspect of this bound is that the constant in the leading term grows super-exponentially with dimension. The difficulties of the Halton sequence in high dimension are discussed further in §6 and in [13].

The Sobol' [23] and the Faure [4] sequences are also based, at least indirectly, on p-adic expansions of the integers. Niederreiter [17], [18] has developed and expanded a general theory of (t, s)-sequences, which encompasses the theory behind both of these sequences. The Sobol' sequence is an example of a (t, s)-sequence with $p = 2$ independent of s, and with t growing with s. The Faure sequence is another example, but it requires that $t = 0$ by setting $p = p(s)$, where $p(s)$ is the smallest prime greater than or equal to s. In each case there is a discrepancy bound equivalent to (15). For Sobol' the coefficient of the $(\log N)^s N^{-1}$ term takes the form

$$C_s^S = \frac{2^t}{s!(\log 2)^s}.$$

Sobol' [23] gives the bound for $t = t(s)$ in the bound for the Sobol' sequence as

$$K \frac{s \log s}{\log \log s} \le t(s) \le \frac{s \log s}{\log 2} + \mathcal{O}(s \log \log s),$$

which shows that $t(s)$ grows super-linearly. Like the Halton bound, this constant grows super-exponentially with s, although it is not nearly as large as the Halton constant. For the Faure sequence, the coefficient can be written

$$C_s^F = \frac{1}{s!}\left(\frac{p(s)-1}{2\log p(s)}\right)^s .$$

This has the desirable property that $\lim_{s\to\infty} C_s = 0$. As Faure's calculations show [4], C_s^F is smaller than both C_s^S and C_s^H, and it goes to zero as dimension increases while the others go to infinity. Because of this smaller bound, it has been claimed that the Faure sequence is superior. A comparison of these sequences in actual computation is made in §6 below.

The actual construction of these sequences is rather complicated, and it is best to check the papers of Sobol', Faure, and Niederreiter for a complete description. Press and Teukolsky [21] and Bratley and Fox [2] give detailed descriptions of the implementation of the Sobol' sequence. For periodic integrands, the method of good lattice points, described in [14], also offers promise, although this has not been studied in the current work.

5. Theoretical bounds on discrepancy. The Halton, Sobol', and Faure sequences discussed in §4 are now studied in detail. First the nature of the error bounds for discrepancy of these sequences is examined. Then actual calculations of the L_2 discrepancies are presented as a means of comparison and as a method of predicting performance in integration. This is followed by a discussion of certain properties of the sequences that are revealed through studying the two-dimensional orthogonal projections. Finally some computational aspects of the sequences are examined, and recommendations are made for their use.

As described in §4, the Halton, Sobol', and Faure sequences all have discrepancy bounds of the form

$$D_N^* \le C_s \frac{(\log N)^s}{N} + \mathcal{O}\left(\frac{(\log N)^{s-1}}{N}\right) .$$

The difficulty with basing any conclusions on this bound for discrepancy is illustrated through the following considerations. Only the bound for the Faure sequence will be considered, as it is the smallest. Let $b_F^s(N) = C_s^F(\log N)^s N^{-1}$ denote the leading term of the bound on the Faure sequence including the constant given above. The best way to examine the behavior of discrepancy with respect to N is to consider a log log plot. Thus let $x = \log N$, which gives

$$\log(b_F^s(N)) = \log(C_s^F) + s\,\log x - x .$$

This function has a maximum, which can be found by setting the derivative

$$\frac{d\log(b_F^s(N))}{dx} = \frac{s}{x} - 1$$

equal to zero. Then the maximum occurs at $x = s$, or when $N = e^s$. Because the general trend of discrepancy for a uniform sequence should be to decrease with increasing N, it follows that the bound cannot be a useful measure of performance until after its maximum has been attained. Thus in high dimensions, the bound gives no information until a very large number of points is used. Moreover, in order to get the same rate of decay of error as with random numbers, $N = e^{2s}$ points are required. The bound has a rate of convergence of $N^{-.95}$ only when $N = e^{20s}$. Even in low dimensions, an extraordinary number of points is required for the bound to indicate near $\frac{1}{N}$ performance.

Not only is the convergence rate predicted by the bound somewhat questionable, but the actual value of the bound is rather large, and grows with dimension, despite the fact that

C_s^F goes to zero. Figure 1 shows the growth as a function of dimension of $\log(b_F^s(N))$ at $N = e^s$ where it attains its maximum. As discrepancy is bounded by one, and thus the log of discrepancy must be negative, a large positive value for the log of the bound is another indication that for N near the maximum, the bound is not accurate.

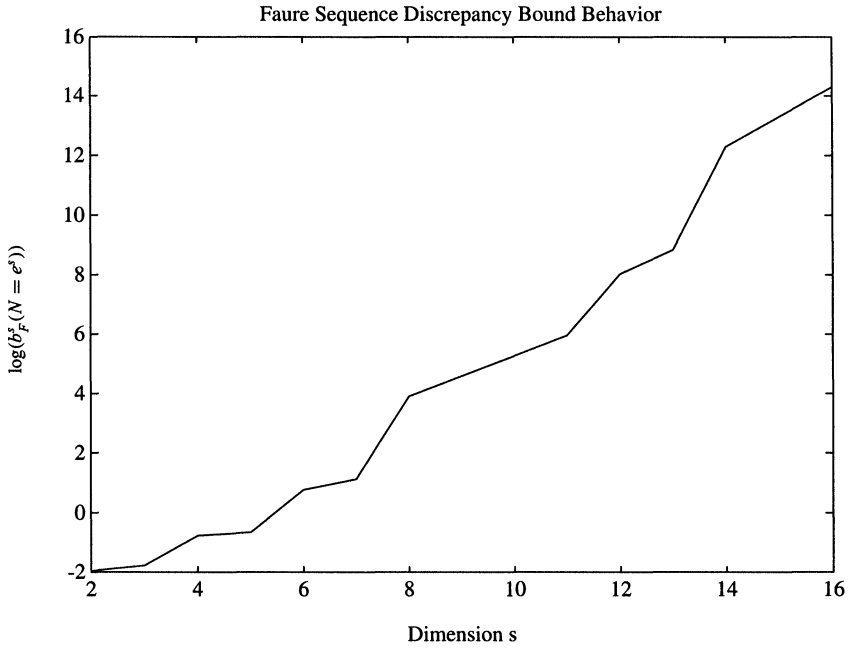

FIG. 1. *Maximum value of discrepancy bound.*

It should be noted that for fixed N, $b_F^s(N)$ is also an increasing function of s until it achieves its maximum at a value of s which is somewhat larger than N (asymptotically in N, the maximum occurs at $s = N^{e/2}$). Thus for fixed N the leading order term does go to zero as dimension increases, but only after passing through a large maximum whose value is super-exponential in N.

It should also be noted that the influence of the terms other than the leading order one has been neglected here. The effect of including these extra terms on the converge rate in N is minor, at most a factor of $\log(N)/(\log(N) - 1)$. However, the leading order constant C_s may be an underestimate. This further illustrates the inadequacies of the discrepancy bound.

Figure 2 shows a plot of $\log b_F^s(N)$ as a function of $\log N$ for dimensions 4 and 16. Also plotted is the actual value of $\log T_N$ for each sequence. The calculation of this quantity is described in the next section. These graphs show that even in the lower dimension, the bound does not accurately predict the behavior of the L_2 discrepancy in slope or in magnitude. Of course the bound was derived for D_N^* and not T_N, and the two measures of discrepancy do not necessarily have the same convergence properties. However, both measure the uniformity of a sequence and thus the convergence of T_N should be related to how well the sequence performs in practice. The difference becomes even greater as dimension increases, as illustrated by the plot for 16 dimensions. This may be somewhat deceptive because D_N^* increases with dimension, while T_N decreases; however, both are still bounded by one. No bound has been specifically derived for T_N, but it is worthwhile to consider this quantity, because as described next, it seems to indicate what one can expect from actual calculations.

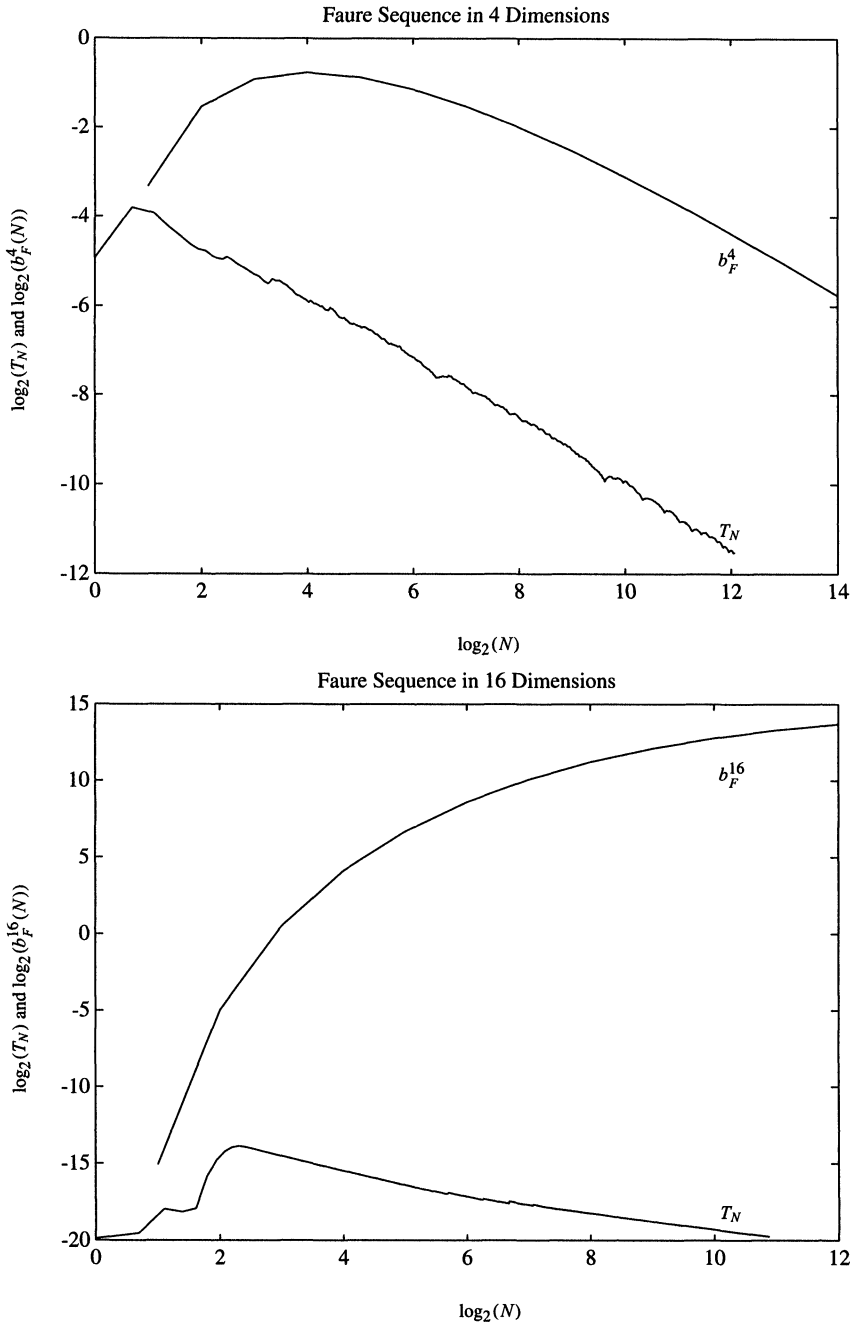

FIG. 2. *L_2 discrepancy T_N and leading term of bound for Faure in* 4 *and* 16 *dimensions.*

6. Calculation of L_2 discrepancy. In his review article on quasi-Monte Carlo methods and sequences [14], Niederreiter discusses only the L_2 star discrepancy T_N^* based on rectangles, which have one corner at the point **0**. An explicit formula for this quantity was first derived by Warnock [25] and subsequently used by Braaten and Weller [1] and Sarkar and Prasad [22].

The result obtained for a given sequence $\{x_n\}$ of N terms is

$$(T_N^*)^2 = \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\prod_{i=1}^{s}\left(1-\max(x_{n,i},x_{m,i})\right) - \frac{2^{-s+1}}{N}\sum_{n=1}^{N}\prod_{i=1}^{s}(1-x_{n,i}^2) + 3^{-s}$$

If the sequence is random, such that each coordinate of each term is an independent random number, then by integrating over the space I^{sN}, the expected value of $(T_N^*)^2$ for a random sequence can be found to be

$$E\left((T_N^*)^2\right) = \frac{2^{-s}-3^{-s}}{N}.$$

While useful in theoretical discussions due to its relationship with D_N^*, T_N^* suffers as a means of comparing sequences and predicting performance because of the strong emphasis it puts on points near $\mathbf{0}$. If $\boldsymbol{x}_j = (0,\ldots,0)$ is a point of the N term sequence, then the dominant term in the calculation of $(T_N^*)^2$ comes from the double sum when $n = m = j$. This term contributes $1/N^2$ to the sum, which tends to dominate all other terms in the sum. Thus $T_N^* \approx \frac{1}{N}$. A similar result is obtained if the sequence contains a point with all coordinates near zero. If this point is excluded from the sequence, however, there is no longer a dominant term, and T_N^* appears rather different. This can be seen by comparing the plots in Fig. 3. Of course this is a transient effect with diminishing influence as N increases. However, as dimension increases, so does the length of the transient region.

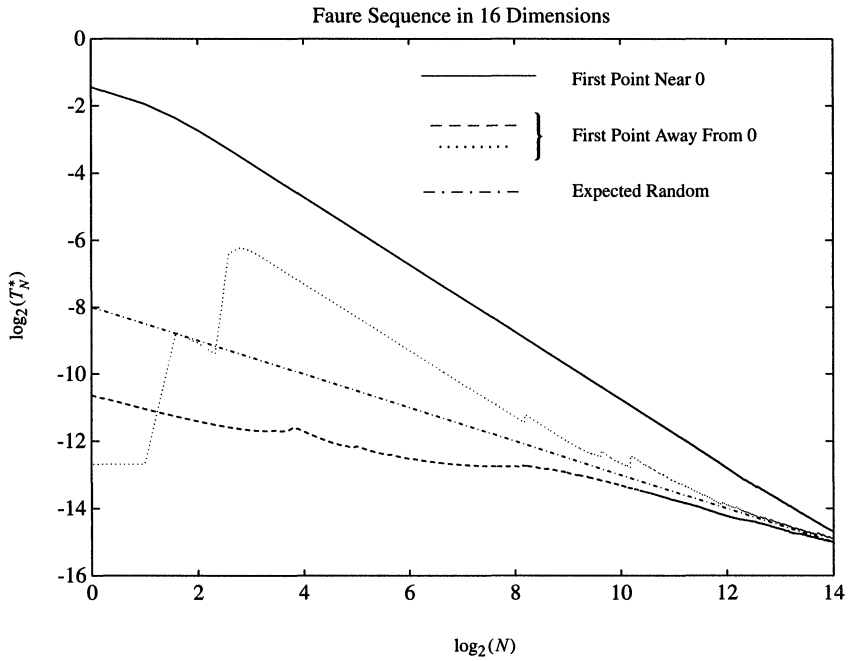

FIG. 3. *L_2^* discrepancy of Faure with and without point near zero.*

As an alternative to T_N^*, the modified L_2 discrepancy T_N was defined in §2. As with the L_2 star discrepancy, it is possible to derive an exact formula for T_N for any given sequence $\{\boldsymbol{a}_n\}$ of N terms (the notation for a sequence is changed here from $\boldsymbol{x}$ to $\boldsymbol{a}$ to help distinguish

between the terms of the sequence and the points defining the rectangles). Using the Heaviside function

$$\theta(y) = \begin{cases} 1, & y > 0, \\ 0, & y \le 0, \end{cases}$$

it is possible to rewrite R_N as

$$R_N(J(\mathbf{y}, \mathbf{z})) = \frac{1}{N} \sum_{n=1}^{N} \prod_{i=1}^{s} \theta(z_i - a_{n,i}) \cdot \theta(a_{n,i} - y_i) - \prod_{i=1}^{s} (z_i - y_i) .$$

Squaring this quantity and integrating over the domain described above leads to T_N^2, which can be expressed as $T_N^2 = A + B + C$, where

$$A = \frac{1}{N^2} \sum_{n=1}^{N} \sum_{m=1}^{N} \prod_{i=1}^{s} \int_{(y_i, z_i) \in I^2, y_i < z_i} \theta(z_i - a_{n,i}) \theta(a_{n,i} - y_i) \theta(z_i - a_{m,i}) \theta(a_{m,i} - y_i) dy_i dz_i,$$

$$B = \frac{-2}{N} \sum_{n=1}^{N} \prod_{i=1}^{s} \int_{(y_i, z_i) \in I^2, y_i < z_i} (z_i - y_i) \theta(z_i - a_{n,i}) \theta(a_{n,i} - y_i) dy_i dz_i,$$

$$C = \prod_{i=1}^{s} \int_{(y_i, z_i) \in I^2, y_i < z_i} (z_i - y_i)^2 dy_i dz_i .$$

These quantities can be evaluated as follows.

For A.

$$\begin{aligned} &\int_{(y_i, z_i) \in I^2, y_i < z_i} \theta(z_i - a_{n,i}) \theta(a_{n,i} - y_i) \theta(z_i - a_{m,i}) \theta(a_{m,i} - y_i) dy_i dz_i \\ &\quad = \int_0^1 \left[\int_{y_i}^1 \theta(z_i - a_{n,i}) \theta(z_i - a_{m,i}) dz_i \right] \theta(a_{n,i} - y_i) \theta(a_{m,i} - y_i) dy_i \\ &\quad = \int_0^1 [1 - \max(y_i, a_{n,i}, a_{m,i})] \theta(a_{n,i} - y_i) \theta(a_{m,i} - y_i) dy_i \\ &\quad = [1 - \max(a_{n,i}, a_{m,i})] \cdot \min(a_{n,i}, a_{m,i}). \end{aligned}$$

Thus

$$A = \frac{1}{N^2} \sum_{n=1}^{N} \sum_{m=1}^{N} \prod_{i=1}^{s} [1 - \max(a_{n,i}, a_{m,i})] \cdot \min(a_{n,i}, a_{m,i}).$$

For B.

$$\begin{aligned} &\int_{(y_i, z_i) \in I^2, y_i < z_i} (z_i - y_i) \theta(z_i - a_{n,i}) \theta(a_{n,i} - y_i) dy_i dz_i \\ &\quad = \int_0^1 \left[\int_{y_i}^1 (z_i - y_i) \theta(z_i - a_{n,i}) dz_i \right] \theta(a_{n,i} - y_i) dy_i \\ &\quad = \int_0^{a_{n,i}} \left[\int_{y_i}^1 (z_i - y_i) \theta(z_i - a_{n,i}) dz_i \right] dy_i. \end{aligned}$$

The inner integral is the area of a trapezoid with corners $(a_{n,i}, 0)$, $(1, 0)$, $(1, 1 - y_i)$, and $(a_{n,i}, a_{n,i} - y_i)$. Thus we have

$$\int_{y_i}^{1} (z_i - y_i)\theta(z_i - a_{n,i})dz_i = \frac{1}{2}(1 - a_{n,i})(1 + a_{n,i} - 2y_i) .$$

Substituting this in the previous equation, it follows that

$$\begin{aligned}\int_0^{a_{n,i}} \frac{1}{2}(1 - a_{n,i})(1 + a_{n,i} - 2y_i)dy_i &= \frac{1}{2}(1 - a_{n,i})\left[(1 + a_{n,i})y_i - y_i^2\right]_0^{a_{n,i}} \\ &= \frac{1}{2}a_{n,i}(1 - a_{n,i}).\end{aligned}$$

Thus

$$B = -\frac{2^{-s+1}}{N}\sum_{n=1}^{N}\prod_{i=1}^{s} a_{n,i}(1 - a_{n,i}) .$$

For C.

$$\begin{aligned}\int_{(y_i,z_i)\in I^2, y_i<z_i} (z_i - y_i)^2 dz_i dy_i &= \int_0^1 \left[\frac{1}{3}(z_i - y_i)^3\right]_{y_i}^1 dy_i \\ &= \int_0^1 \frac{1}{3}(1 - y_i)^3 dy_i \\ &= \frac{1}{12}.\end{aligned}$$

Thus

$$C = 12^{-s} .$$

Combining these elements leads to the formula

$$\begin{aligned}(T_N)^2 = \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\prod_{i=1}^{s}[1 - \max(a_{n,i}, a_{m,i})] \cdot \min(a_{n,i}, a_{m,i}) \\ -\frac{2^{-s+1}}{N}\sum_{n=1}^{N}\prod_{i=1}^{s} a_{n,i}(1 - a_{n,i}) + 12^{-s} .\end{aligned}$$

As with the star discrepancy, it is possible to compute the expected value of this quantity for a random sequence. This root mean square (rms) expectation of T_N is given by

$$\begin{aligned}E(T_N^2) &= \int_{I^{sN}} T_N^2 \prod_{n=1}^{N}\prod_{i=1}^{s} da_{n,i} \\ &= \frac{1}{N}6^{-s}(1 - 2^{-s}).\end{aligned}$$

Thus again the average L_2 discrepancy of a random sequence decreases like $N^{-1/2}$, corresponding nicely with the random Monte Carlo bound.

By comparing the formulas for T_N^* and T_N, it appears likely that $T_N^* > T_N$ for all sequences and all N, although this has not been proved. It is certainly true for the expected value of a random sequence, and it has been borne out in all computations. Figure 4 compares the two discrepancies for a couple of versions of the Faure sequence (created by starting at different places in the sequence). The qualitative behavior of the two discrepancies is similar for large N, but T_N is smoother and has a shorter, less extreme, transient region. This becomes even more important in higher dimensions, where the transient region is considerably longer.

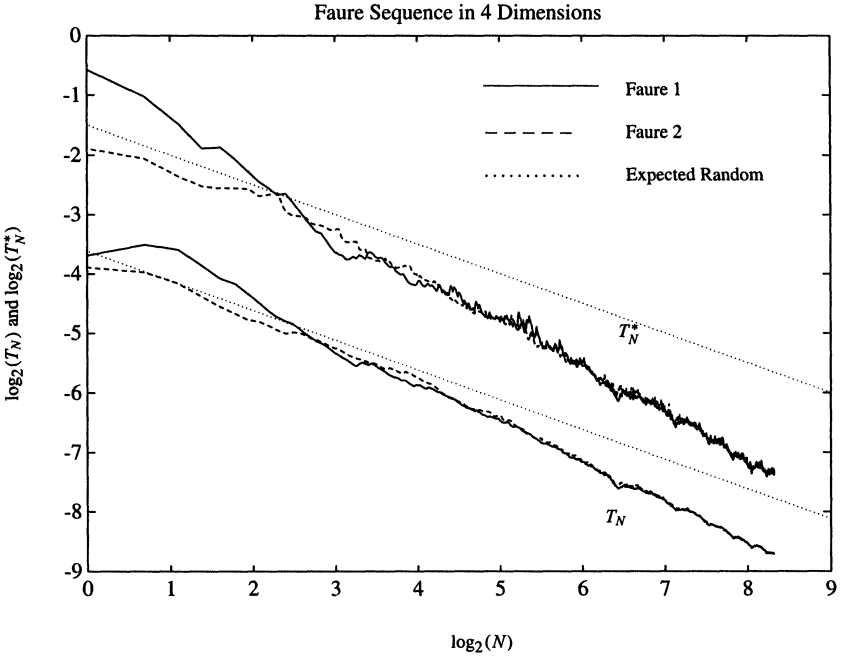

FIG. 4. *Comparison of T_N and T_N^*.*

A disadvantage of using T_N to measure discrepancy is that no direct connection has been established between it and integration error. Whereas the relationship $D_N^* < D_N$ allows the Koksma–Hlawka inequality to be modified to include D_N, the sup discrepancy taken over all rectangles, it is not possible to change the L_2 version from T_N^* to T_N. Nevertheless, actual computation of T_N indicates that it is a useful gauge of integration error convergence as a function of N. Figures 5–8 show plots of T_N (solid line) on a log log base 2 scale. The rms expectation of T_N for a random sequence (dashed line) is also plotted, along with the function $\frac{1}{N}$ (dotted line) for reference.

There are several interesting features of these plots, which should be noted. Sobol' [23] predicts that when N equals large enough powers of two, the value of discrepancy should have a local minimum. The plot for the Sobol' sequence in three dimensions shows this kind of behavior. After $N = 2^{10}$ there appears to be a cusp at the powers of 2. Closer examination of this phenomenon shows that a minimum actually occurs at a few points short of the power of two. Sobol' also predicts how large N must be before this occurs, with the cut-off value increasing with dimension. For three and four dimensions, he shows greater uniformity for $N = 2^6$ and above. However, the plots do not reveal any particularly noteworthy behavior until, as mentioned above, $N = 2^{10}$ for three dimensions and $N = 2^{13}$ for four dimensions. For $s = 8$ Sobol's formula for the cut off value predicts improved discrepancy for powers of

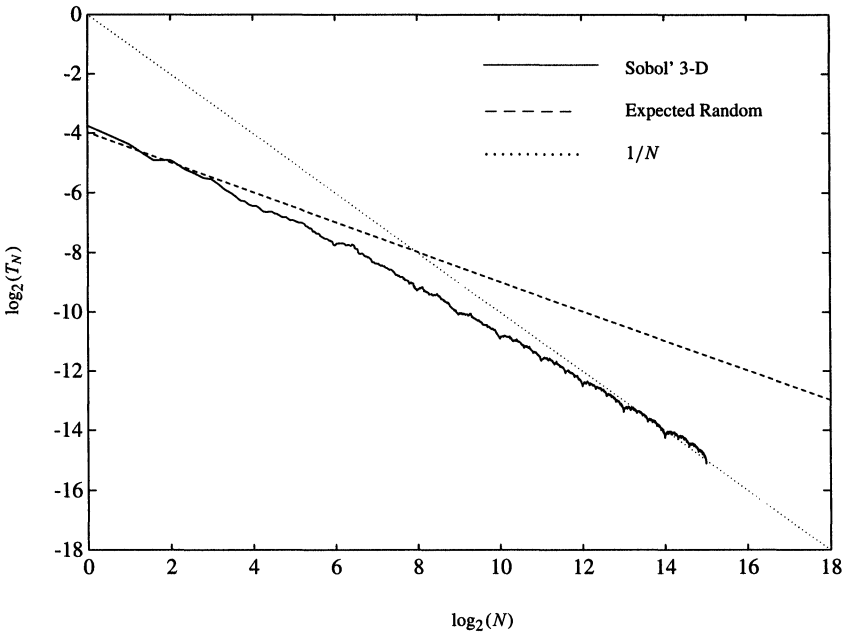

FIG. 5. *T_N for three-dimensional Sobol' sequence.*

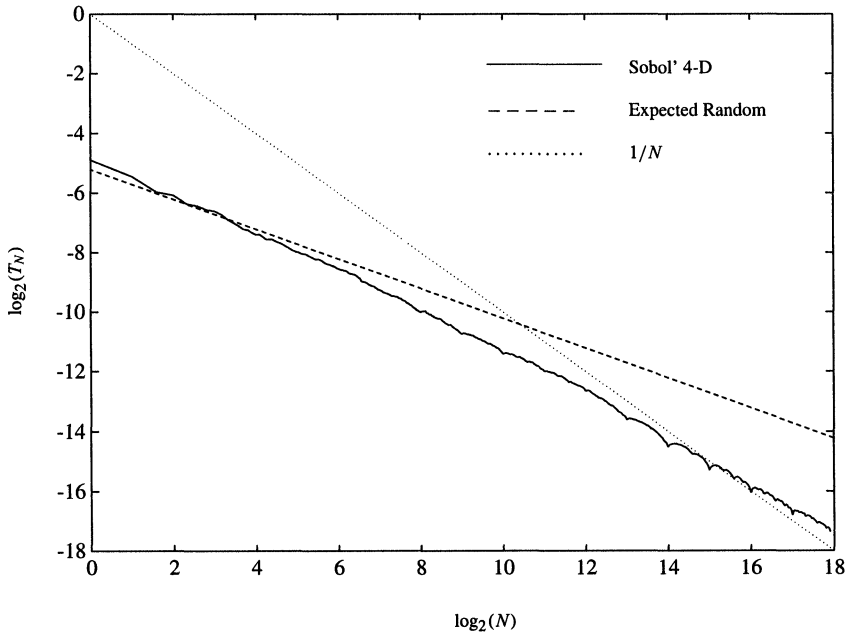

FIG. 6. *T_N for four-dimensional Sobol' sequence.*

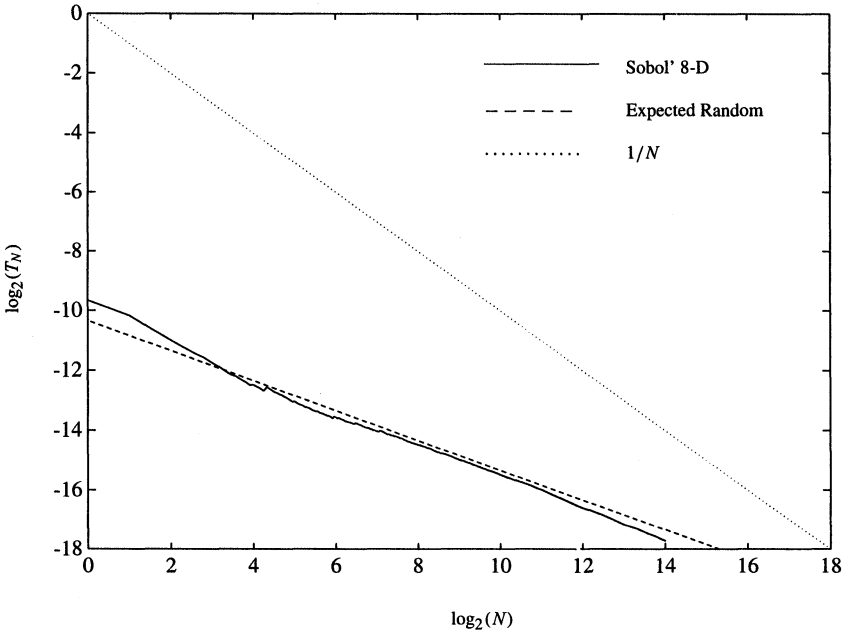

FIG. 7. *T_N for eight-dimensional Sobol' sequence.*

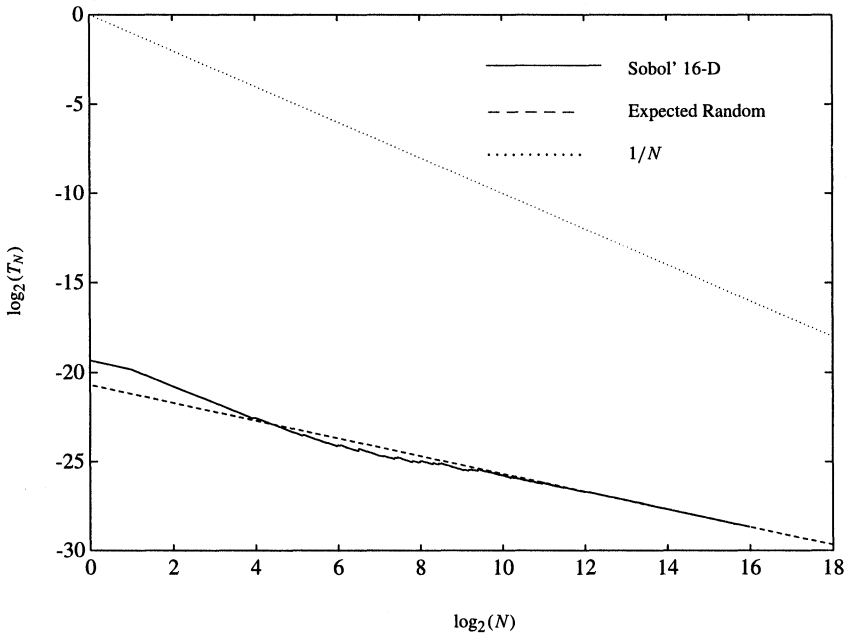

FIG. 8. *T_N for sixteen-dimensional Sobol' sequence.*

two greater than or equal to $N = 2^{22}$. Since this number is larger than four million, it is not surprising that nothing special is seen on the plot for eight dimensions, which only goes out to about 16,000.

A more important observation to be made from these plots is the transition of T_N from random-like behavior for low values of N, to perhaps eventual $\frac{1}{N}$-type convergence. For dimensions three and four, T_N starts near the rms expectation curve of T_N for a random sequence, but fairly quickly starts to decay at a faster rate than $N^{-1/2}$. The transition seems to occur around the point where the rms expectation curve and the $\frac{1}{N}$ curve intersect. This is easily shown to occur at around $N = 6^s$. This is a purely heuristic estimate, since the curve $\frac{1}{N}$ is used only as an approximation to the asymptotic behavior of T_N and does not mean much for smaller values of N. However, it does provide a rough estimate of the nature of T_N. In eight dimensions this predicts a transition at around $N = 2^{20}$. Figure 7 shows that T_N is just beginning to break away from the rms expected curve around $N = 2^{14}$. In 16 dimensions, after an initial transient region which is near the rms expectation of T_N for a random sequence, the value of T_N for the Sobol' sequence lies almost exactly on the rms expectation curve out to $N = 2^{16}$ and probably considerably farther. For this dimension the heuristic estimate predicts the transition at 2^{41}. It might be hoped that this is an overestimate; however, this kind of exponential growth of the transition point is similar to that of the maximum point for the theoretical bound on discrepancy. Figure 9 compares the L_2 discrepancies of various sequences in 16 dimensions. Except for within the initial transient, all of the sequences behave almost identically; that is, as if they were random. This indicates that in high dimensions, unless one uses a very large number of points, quasi-random sequences are no more uniform than random sequences.

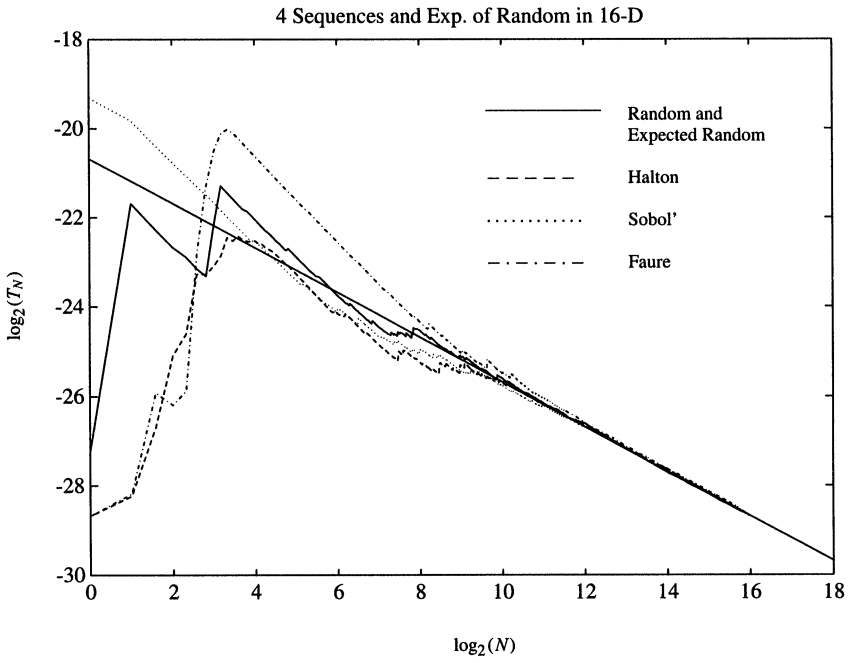

FIG. 9. *L_2 discrepancy in 16 dimensions for various sequences.*

It should also be noted that the value of T_N is insensitive to where the sequence begins. All the sequences considered are produced by mapping the sequence of integers $\{n = 1, 2, 3, \ldots\}$

to points in I^s; however, it is not mandatory to start with $n = 1$. Any number of the initial terms can be discarded without affecting T_N much, except in the transient region. The same is true for T_N^*, but the change in the transient region may be much more extreme if a point very close to zero is included. It is also possible that, when these initial terms are discarded, Sobol's improved bounds on discrepancy are no longer valid. For Sobol' these occur at powers of 2; for Faure they should occur at powers of $p(s)$. However, as pointed out above, this improvement is only of practical value for low dimensions; moreover, from the discrepancy plots, it is clear that the value of discrepancy at the special values of N is not all that much lower than otherwise. Orthogonal projections (discussed next), which are calculations of discrepancy and computations of integrals, show that there is not much difference between any two subsequences of equal length. Thus it does not matter what value of n corresponds to the beginning of the sequence.

For the T_N^* discrepancy, similar dependence on dimension s is observed. To optimize the sequence for a given dimension s, we consider "Hammersley"-type sequences in which the components in the first dimension are lattice points $\frac{n}{N}$, while the other components come from an $(s-1)$-dimensional quasi-random (or random) sequence. Figures 10 and 11 show the T_N^* discrepancy for a variety of sequences, a pure random sequence, a Hammersley random sequence, a Hammersley–Halton sequence, and a normal Halton sequence, in dimensions 2, 10, and 15. The random sequences have discrepancy of size $N^{-1/2}$ in all dimensions. In dimension 2 the quasi-random sequences have discrepancy of size N^{-1}, but in dimension 10 it is of size $N^{-1/2}$ for small N before beginning to drop off faster for larger N. In dimension 15 the quasi-random sequences have discrepancy almost exactly that of a random sequence (i.e., of size $N^{-1/2}$) for the values of N computed here, although it must eventually approach size N^{-1} for extremely large N.

7. Orthogonal projections. Another approach to understanding quasi-random sequences is to look at two-dimensional orthogonal projections of the points in I^s. The assumption made here is that if a sequence is uniformly distributed in I^s, then the two-dimensional sequences formed by pairing coordinates (i.e., the two-dimensional orthogonal projections) should also be uniformly distributed. Moreover the discrepancy of projections of a sequence occur explicitly in the average error identity (6). The appearance of nonuniformity in these projections is an indication of potential problems in using a quasi-random sequence for integration. However, a sequence with very nonuniform behavior in some projection may in fact be reasonably uniform in I^s. Of course, attempting to integrate a function that has strong dimensional dependence on just the two dimensions in question will lead to poor results, but for many functions a bad pairing of dimensions may not have much influence.

Here a catalog of potentially bad behavior is given for all the sequences under consideration, along with some insight into the source of these problems. First it is worthwhile to consider a pseudo-random sequence. Figure 12 shows the projection of 4096 points on the first and 16th dimensions of the sequence generated in Matlab (using seed zero). The points appear to be randomly distributed and fairly uniform. Any decent pseudo-random number generator should be able to produce this effect for orthogonal projections. Nothing particularly different was seen from examining other projections of this sequence.

Figure 13 shows the projection of 4096 points of the Halton sequence onto the first and second dimensions and the 28th and 29th dimensions. Compared with the random sequence, this low-dimensional projection appears to be considerably more uniform, and thus a better sequence. However, difficulties with high dimensions occur, as observed in [1]. If approximately 5900 points are used, then the projection onto the 28th and 29th dimensions would be almost perfectly uniform. However, this would not be true for any other dimensional pairings, and as more points are added the uniformity would disappear. The problem here arises from

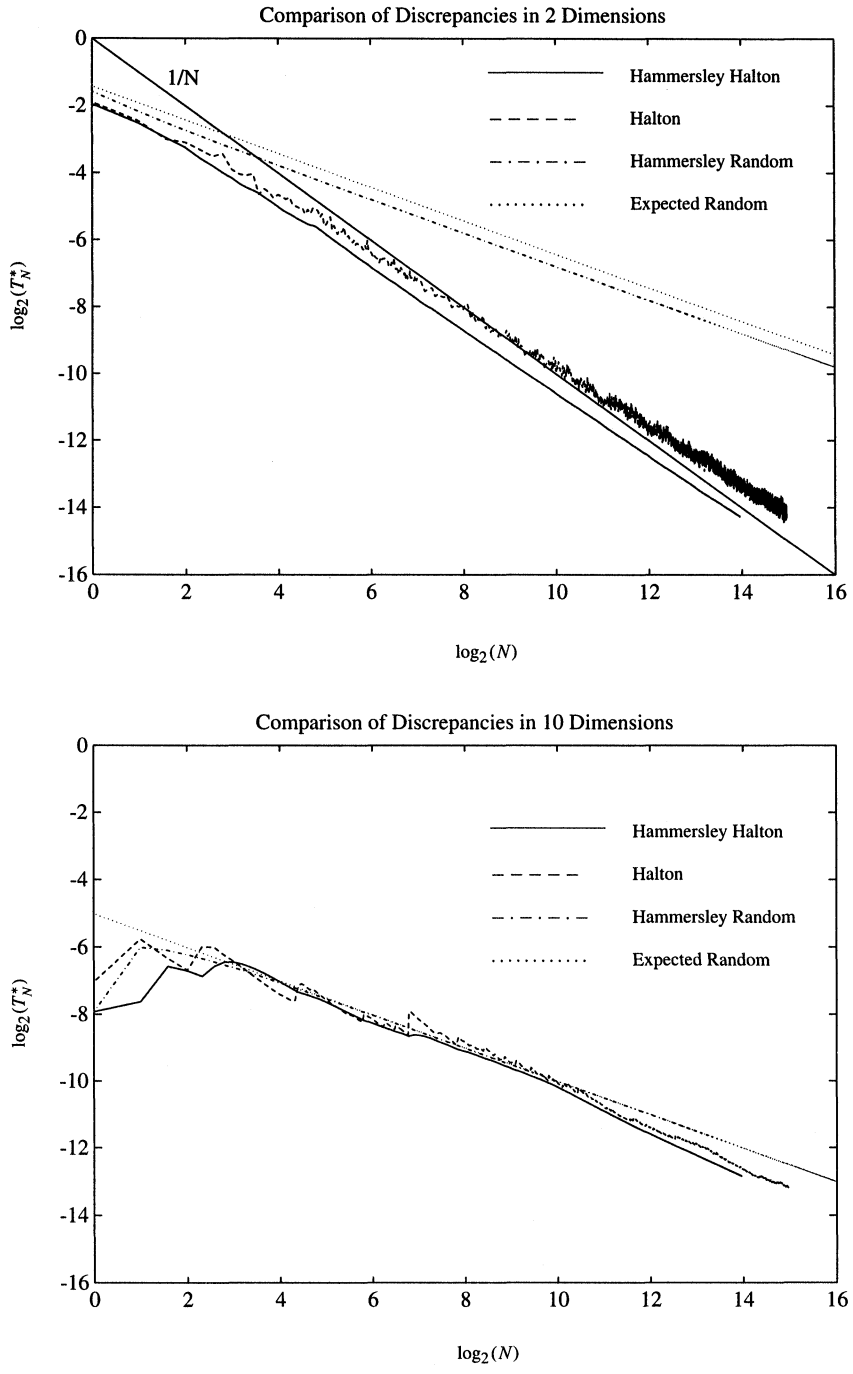

FIG. 10. L_2^* *discrepancy in* 2 *and* 10 *dimensions for various sequences.*

the use of large primes, in this case 107 and 109 for the 28th and 29th dimensions respectively. The 28th dimension of the Halton sequence consists of monotone increasing subsequences of length 107 terms. When this is paired with the monotone subsequences of length 109 for the 29th dimension, the lines seen in the plot occur.

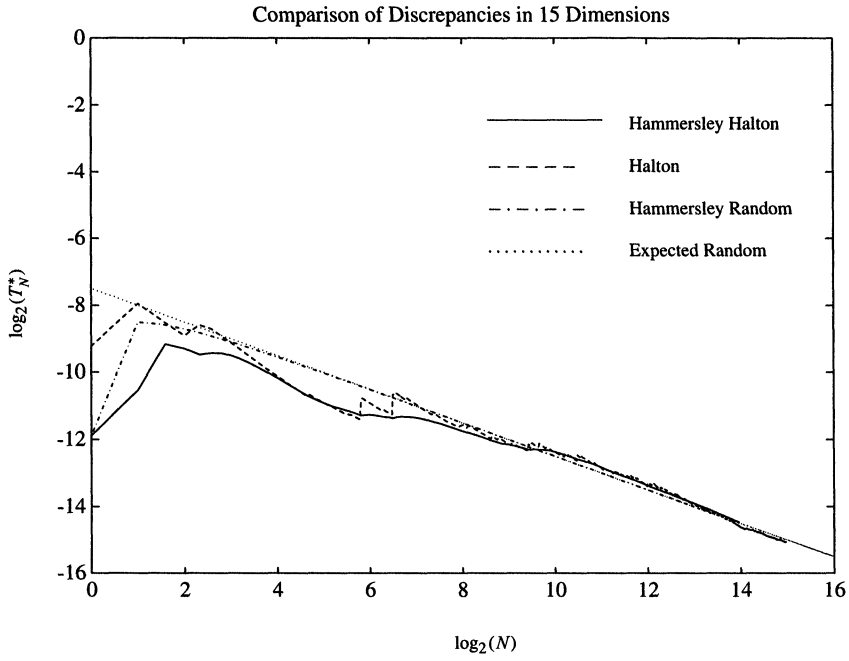

FIG. 11. *L_2^* discrepancy in 15 dimensions for various sequences.*

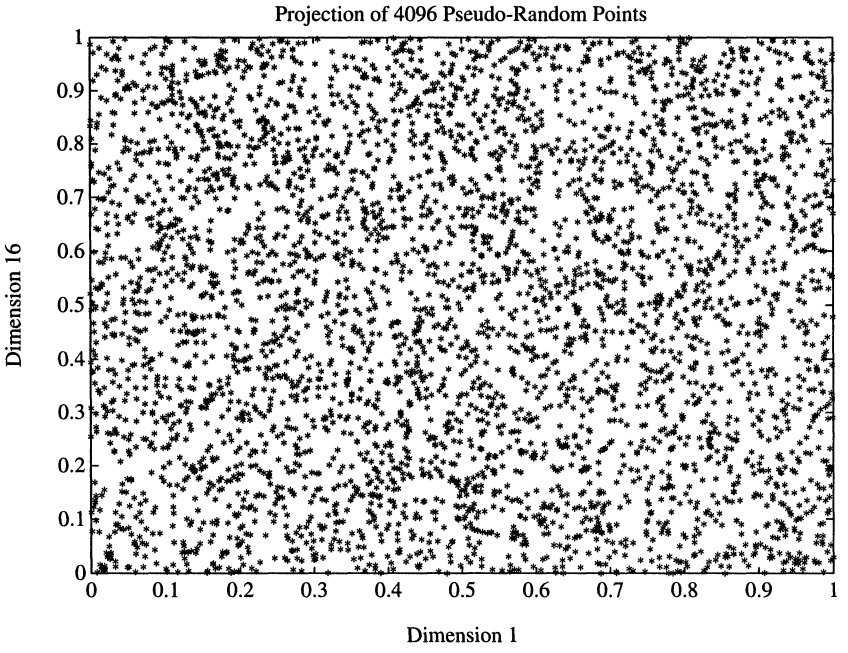

FIG. 12. *Two-dimensional projection of random sequence.*

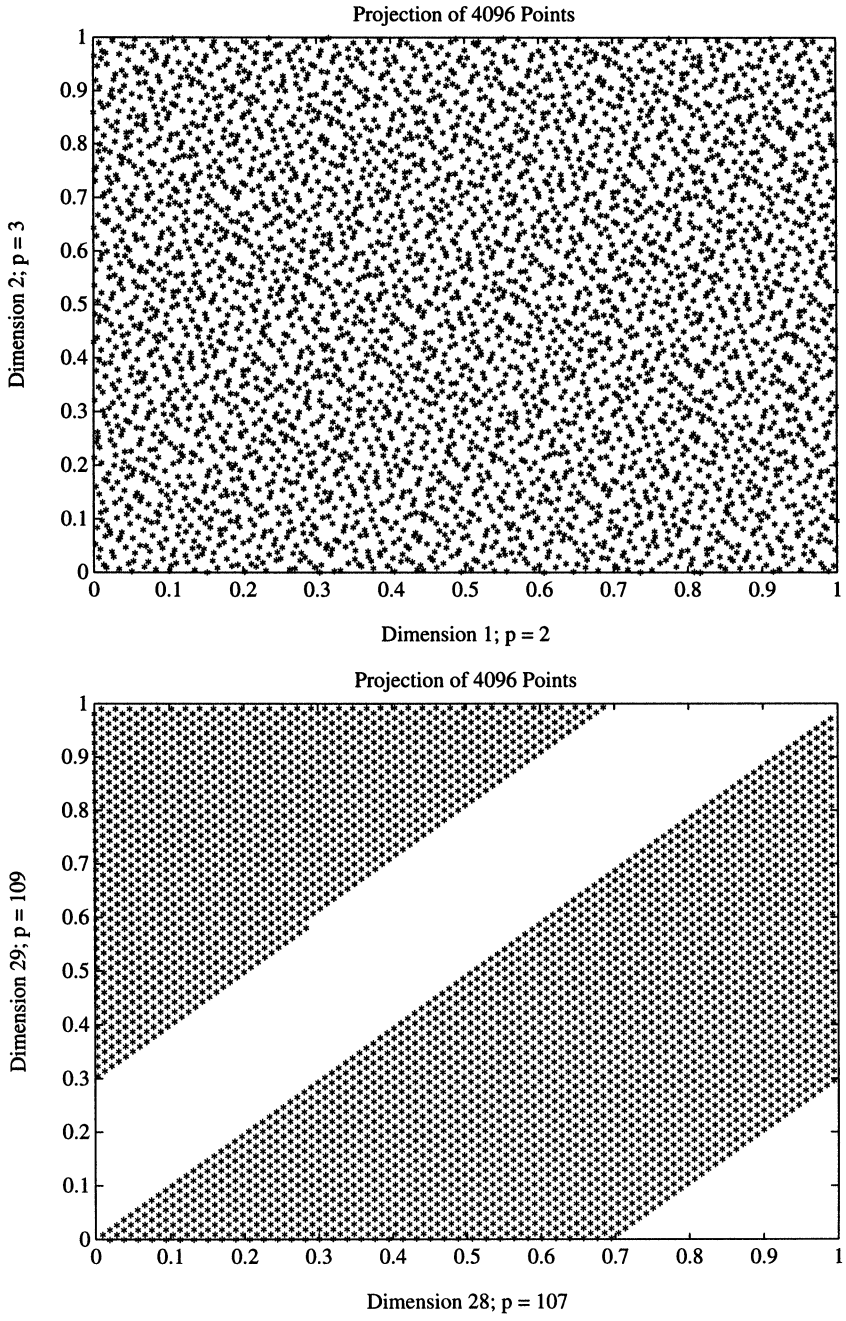

FIG. 13. *Two-dimensional projection of Halton sequence.*

To improve this situation Braaten and Weller [1] suggest a scrambling or permutation procedure, which preserves the traditional $(\log N)^s/N$-type bound for the discrepancy. A less elaborate, but easier to implement, scrambling technique was used in the current work. Here the sequence was simply (pseudo)randomly scrambled independently in each dimension.

For example, if N points in I^s were required, then s sequences of N random numbers were generated and sorted from smallest to largest. This mapping of original position in the sequence to final position was then used to permute the Halton sequence. Figure 14 shows a two-dimensional projection of the 29-dimensional randomly scrambled sequence. At least in terms of projections, scrambling seems to greatly improve the Halton sequence in high dimensions. In each dimension, this procedure does not change the one-dimensional discrepancy of the N points. As the pairings across dimensions are (pseudo-)random, this may lead to slower, more random-like, convergence, although the actual value of discrepancy for a given N will hopefully be smaller over a reasonable range of N.

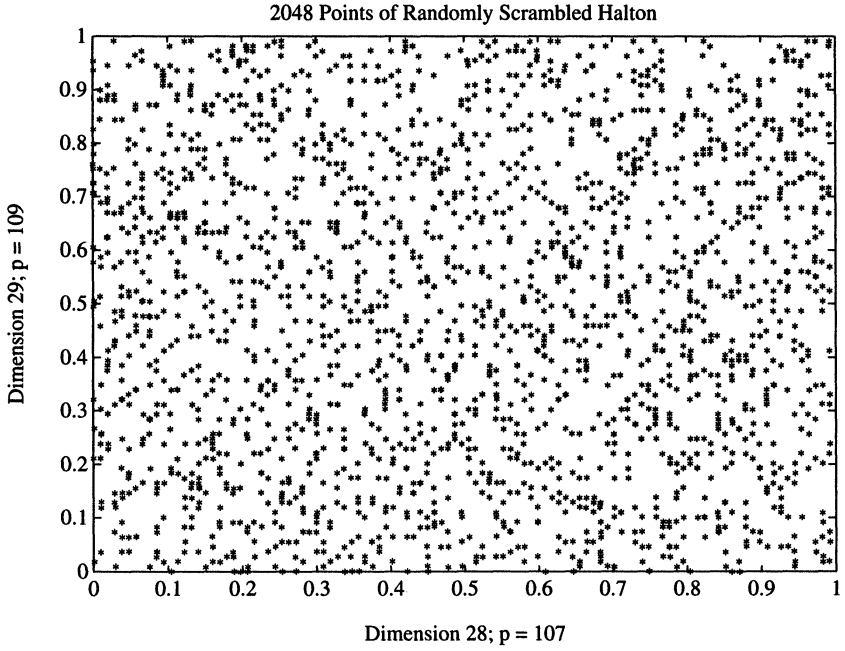

FIG. 14. *28th versus 29th dimensions of scrambled Halton sequence.*

To compare the standard Halton sequence to the scrambled version, Braaten and Weller [1] compute the discrepancy T_N^* of the first 1000 points of each sequence in 8, 12, and 16 dimensions. This result is somewhat misleading, though, because of the use of the T_N^*. This measure of uniformity weights the point zero and points near zero, such as the first terms of the Halton sequence in high dimensions, much greater than other points in the unit cube. The calculated value of T_N^* for the Halton sequence is almost entirely determined by the first point. If the sequences formed by deleting the first ten points of Halton and scrambled Halton are compared, the values of T_N^* are almost identical for this range of number of points and dimension. This is not to say that scrambling does not improve the sequence, but just that the improvement cannot be seen through calculation of T_N^*.

Press and Teukolsky [21] give some examples of the projections of the Sobol' sequence to illustrate how it fills out the unit square. They show how the first 256 points lay down a fairly uniform, but distinct, pattern in the square, and how the next 256 points fill in the gaps left by the first group. The points are put down to be uniform, and additional points "know" about the spaces left by the original points, so they are put down to make the whole sequence even more uniform.

To understand what can potentially go wrong here, it is necessary to have a feel for how the Sobol' sequence is generated. Each dimension of this sequence is just a permutation of the Halton sequence with prime base 2 (this is also known as the van der Corput sequence) whenever $N = 2^m$ for $m = 0, 1, 2, 3, \ldots$. These permutations are generated from irreducible polynomials over the field {0,1}. Ideally, polynomials of the lowest degree possible are used; however, as dimension increases, it is necessary to use polynomials of higher and higher degree. To generate a one-dimensional sequence from a polynomial of degree d, $d-1$ odd integers $j_1, \ldots, j_{d-1}$ must be chosen with the restriction that $j_i < 2^i$. Thus there are $2^d - 1$ possible ways of picking the starting values. Sobol' has given a list of good starting values for dimension up to 16 [24]. These are said to be better because they produce sequences that satisfy an additional uniformity property.

What can go wrong with the Sobol' sequence involves the pairing of dimensions. The fact that each dimension is a permutation of the same sequence allows for certain correlations to develop. In some cases this is good, because it allows for the phenomenon described above where points fill in the gaps left by previous points. However, these correlations can also produce regions in the unit square where no points fall until N becomes extremely large. Figure 15 shows a "good" pairing of dimensions using Sobol's second and third dimensions with his recommended starting values. A "bad" pairing of dimensions is also shown, representing what would be the 27th and 28th dimensions following Sobol's convention for associating dimension with generating polynomial. The polynomials used here are $x^7 + x^5 + x^4 + x^2 + x + 1$ and $x^7 + x^5 + x^4 + x^3 + x^2 + x + 1$ and the starting values are (1,3,5,11,3,3,35) for the 27th dimension and (1,1,7,5,11,59,113) for the 28th dimension. If one or two of these starting values were changed, then the problem illustrated in the graph would disappear. However, it does not seem possible to tell a priori that this is a bad pairing. Moreover, neither set of starting values is particularly at fault, because when they are paired with other dimensions there is no such pathological behavior.

For 29 dimensions, there are 406 pairings of dimensions that could be checked for such correlations; this probably should be done if the Sobol' sequence in this high a dimension is to be used. Sobol' may have checked this for the recommended first 16 dimensions; however, the uniformity property that these sequences satisfy does not exempt them from such bad behavior. There may also be higher-dimensional correlations, which would be difficult to detect.

The bad behavior seen in the second plot of Figure 15 can be explained in terms of the filling-in-holes idea. If 8192 (2^{13}) points are used, the plot looks almost identical to what is shown for 4096. However, the next 8192 points fall only where the gaps appear. Thus by $N = 16,384$, the projection plot is almost perfectly uniform. The problem is that the cycle for filling in holes is 2^{13}, which is too long.

The idea behind the Faure sequence is an extension of the theory of the Sobol' sequence. This theory, which has been somewhat extended by Niederreiter [15], is based on the idea of the elementary rectangle base p in I^s This is a rectangle that is a product of s intervals of the form $[ap^{-d}, (a+1)p^{-d})$, where a is an integer less than p^d and d is a nonnegative integer. For arbitrary integer m, the goal is to construct a sequence such that every subsequence of length p^m of the form $(k-1)p^m < n \leq kp^m$ (n is the index of the sequence) has the property that each elementary rectangle base p of volume p^{-m} contains exactly one point of the subsequence. Faure constructs such a sequence by taking p to be the smallest prime greater than or equal to s. Figure 16 shows the projections onto the first and second dimensions of 3125 points of the five-dimensional sequence ($p = 5$), and 2197 points of the 13-dimensional sequence ($p = 13$). The second plot of this figure shows some considerable difficulties, which may initially seem surprising given that the sequence was constructed so that every elementary rectangle base 13 with volume $\frac{1}{N}$ contains exactly one point if N is a power of 13. However, Figure 17 illustrates

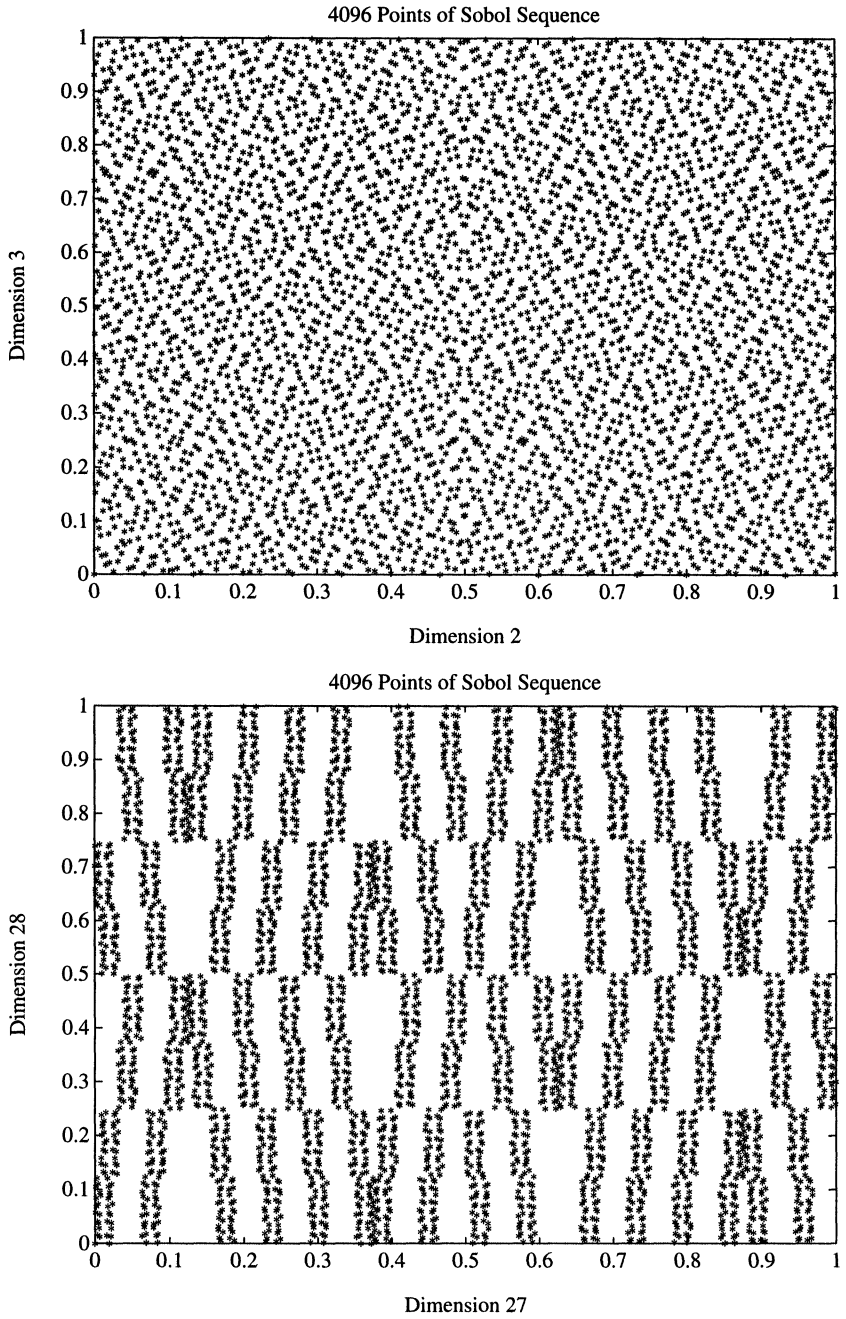

FIG. 15. *Two-dimensional projection of Sobol' sequence.*

how this can happen. The grid shown divides the unit square into elementary rectangles of volume 13^{-3}. It is clear that each rectangle does have exactly one point of the sequence in it; unfortunately, the distribution of the point inside the rectangle is not uniform. It will take approximately $13^4 = 28,561$ points before the square is more satisfactorily filled.

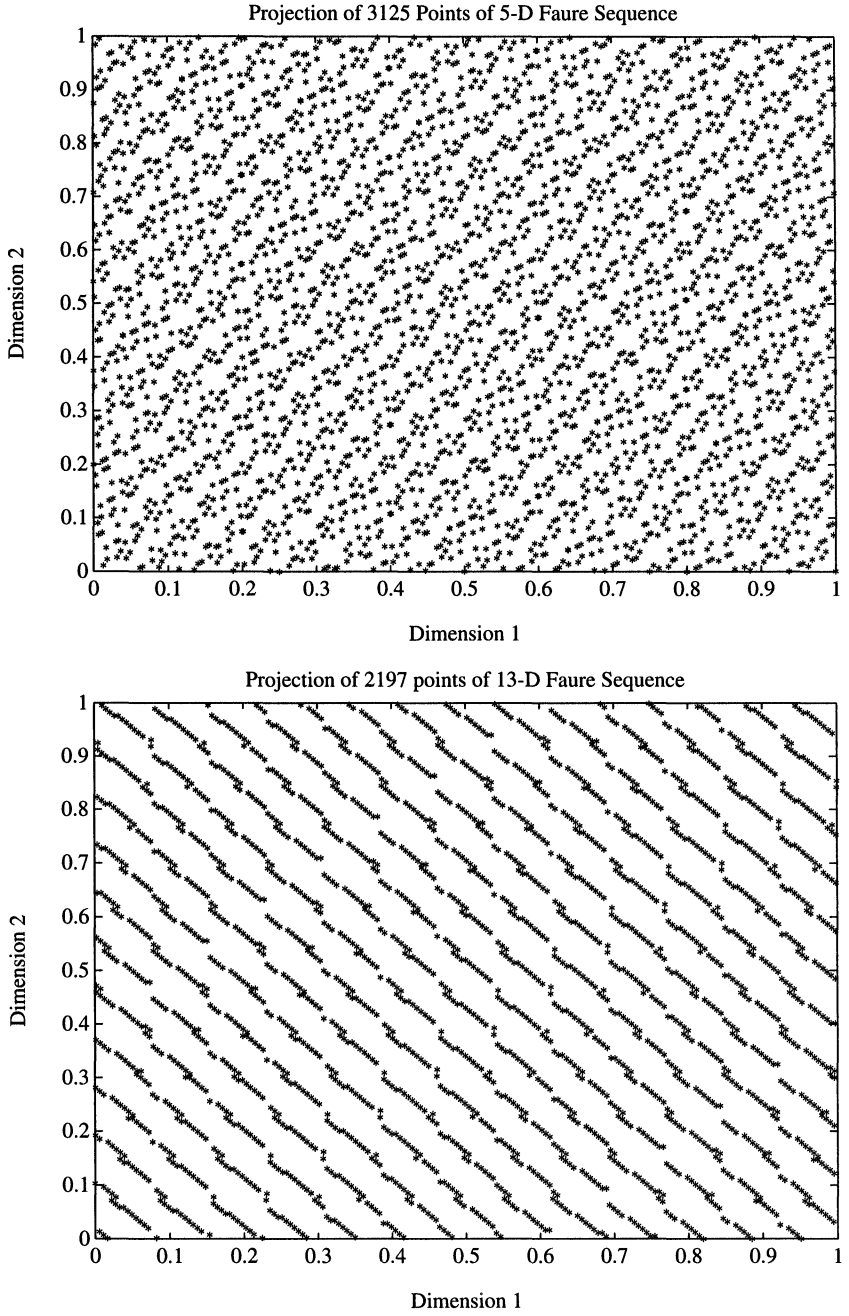

FIG. 16. *Two-dimensional projection of Faure sequence.*

As noted above, even if a sequence has poor two-dimensional projections, it may still be fairly uniform in I^s, and there are many functions which it may integrate quite well. However, it is important to be aware of the potential problems these sequences may have, and the orthogonal projections are a good means of identifying and assessing the difficulties.

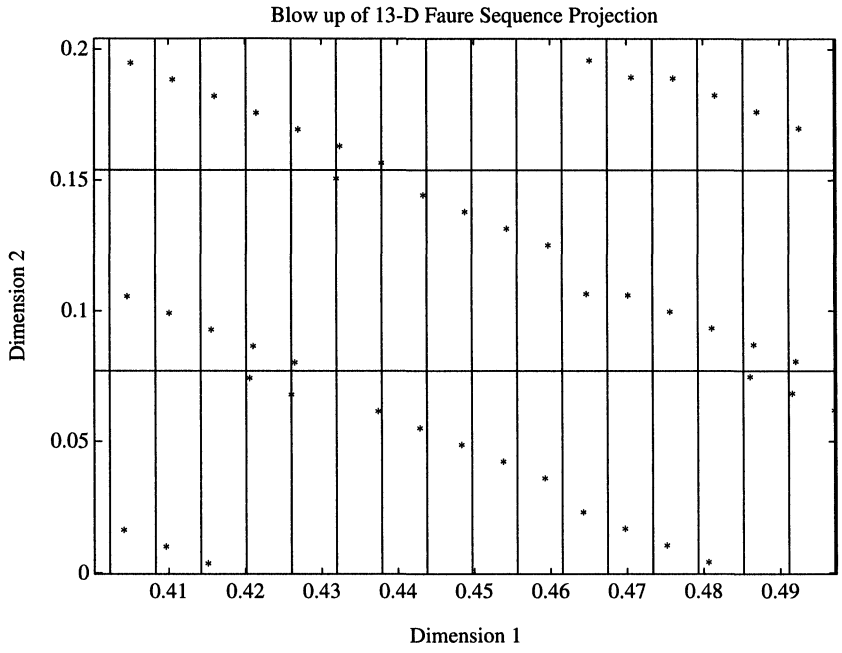

FIG. 17. *Two-dimensional projection of Faure sequence.*

8. Computational speed. Another aspect of quasi-random sequences worth considering is the computer time required to generate them. Fox [5] and Bratley and Fox [2] present such results for various values of N and dimensions and for a pseudo-random number generator along with the three quasi-random sequences under consideration here. Their calculations were done on a Cyber 855 computer and include calls to the initialization routine and a routine to evaluate a simple integral, as well as to the sequence generator. They conclude that the time spent on the initialization routine is negligible. In comparing the sequences, they find that Sobol' is 1 to 3 times faster than their random number generator and 3 to 5 times faster than Halton. They find Halton to be approximately 4 times faster than Faure. They also state that when run on a different computer, the ordering remained the same, but the ratios for computing times for the various sequences were much different. The results show that computation time is approximately proportional to dimension and to number of points used.

Similar timing experiments were run for this work on an Alliant FX/80. The pseudo-random number generator used was the routine lib_vdran supplied by Alliant and found in the common library. The results of these experiments are given in Table 1. These results are not definitive, since we have not made much effort to optimize our code. Nevertheless we expect that they will be of interest to the potential user. Here, the only thing timed was the sequence generating subroutine; the initialization routine was not included, nor was any integral evaluated. Here, the random number generator is the fastest, and its times are proportional to the number of points and the dimension. Sobol' again is faster than Halton, but only slightly so, around 1.2 times faster. Both Sobol' and Halton have timings that are proportional to N, but that grow slower than linearly with dimension. This is probably related to the vector and parallel aspects of the Alliant. Sobol' and Halton are approximately 4 to 5 times faster than Faure.

TABLE 1
Timings for QMC *sequences (in seconds).*

Generator	N	$s = 5$	$s = 10$	$s = 20$	$s = 40$
Random	1000	0.00363	0.00702	0.0138	0.0284
Halton	1000	0.0263	0.0335	0.0424	0.0568
Sobol'	1000	0.0225	0.0257	0.0321	0.0467
Faure	1000	0.102	0.0976	0.123	0.189
Random	10,000	0.0353	0.0707	0.142	0.285
Halton	10,000	0.265	0.342	0.434	0.535
Sobol'	10,000	0.231	0.269	0.367	0.517
Faure	10,000	1.14	1.08	1.28	1.99
Random	100,000	0.361	0.720	1.45	2.88
Halton	100,000	2.69	3.47	4.31	5.29
Sobol'	100,000	2.35	2.73	3.52	5.12
Faure	100,000	13.1	11.8	13.7	20.5

As Fox points out, sequence generation time is frequently only a fraction of what is required to evaluate a complicated integrand. Thus for many realistic problems, the question of which generator is fastest is not all that important.

9. Conclusions. The computations described above show strong dependence of the discrepancy on dimension s. While the theoretical bound N^{-1} is observed in any dimension for sufficiently large N, it appears that there is a transition value of N, below which the discrepancy is of size $N^{-1/2}$. For such values of N, random-like behavior of the sequence can be expected. This transition point grows exponentially with dimension.

Comparison between different quasi-random sequences have also been presented. While the discrepancy bound suggests that Faure is a superior sequence, the actual calculation of the discrepancy indicates that all the sequences are about the same. The orthogonal projections show that all of the sequences have potential problems as dimension increases; however, Halton is probably the worst, because all its high-dimensional pairings will be nonuniform for large ranges of N. To a certain extent, Faure has the same problem, but the degree of nonuniformity is not as severe. Sobol' may be able to avoid this problem if the starting values are carefully checked for two-dimensional correlations. Of course, this does not preclude three- (or higher-) dimensional projection problems. Although a direct connection has not been demonstrated, we expect that nonuniformity of projections will lead to poorer performance of Monte Carlo methods for many functions.

Finally, computational timings put Sobol' and Halton on about the same ground, while Faure is considerably slower.

The actual value of these sequences must be judged by their performance in Monte Carlo methods. In the companion papers [12], [13] we present computational experiments with quasi-Monte Carlo methods applied to multidimensional integration and to simulation of the heat equation. Again it is found that the performance of these methods degrades with increasing dimension. Nevertheless, quasi-Monte Carlo methods using quasi-random sequences consistently give significant, but limited, improvement over standard Monte Carlo methods using random or pseudo-random sequences.

REFERENCES

[1] E. BRAATEN AND G. WELLER, *An improved low-discrepancy sequence for multidimensional quasi-Monte Carlo integration*, J. Comput. Phys., 33 (1979), pp. 249–258.

[2] P. BRATLEY AND B. L. FOX, *Implementing Sobol's quasirandom sequence generator*, ACM Trans. Math. Software, 14 (1988), pp. 88–100.
[3] P. BRATLEY, B. L. FOX, AND H. NIEDERREITER, *Implementation and tests of low-discrepancy sequences*, ACM Trans. Modeling and Comp. Sim., 2 (1992), pp. 195–213.
[4] H. FAURE, *Discrépance de suites associées à un système de numération (en dimension s)*, Acta Arith., 41 (1982), pp. 337–351.
[5] B. L. FOX, *Implementation and relative efficiency of quasirandom sequence generators*, ACM Trans. Math. Software, 12 (1986), pp. 362–376.
[6] J. H. HALTON, *On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals*, Numer. Math., 2 (1960), pp. 84–90.
[7] M. H. KALOS AND P. A. WHITLOCK, *Monte Carlo Methods*, Vol. I, John Wiley and Son, New York, 1986.
[8] L. KUIPERS AND H. NIEDERREITER, *Uniform Distribution of Sequences*, John Wiley and Son, New York, 1974.
[9] Y. L. LEVITAN et al., *Short communications on quasi-random sequences for numerical computations*, U.S.S.R. Comput. Math. and Math. Phys., 28 (1988), pp. 88–92.
[10] H. MCKEAN, *Stochastic Integrals*, Academic Press, New York, 1969.
[11] H. G. MEIJER, *The discrepancy of a g-adic sequence*, Indag. Math. (N.S.), 30 (1968), pp. 54–66.
[12] W. J. MOROKOFF AND R. E. CAFLISCH, *A quasi-Monte Carlo approach to particle simulation of the heat equation*, SIAM J. Numer. Anal., 30 (1993), pp. 1558–1573.
[13] ———, *Quasi-Monte Carlo integration*, J. Comp. Phys., to appear.
[14] H. NIEDERREITER, *Quasi-Monte Carlo methods and pseudo-random numbers*, Bull. Amer. Math. Soc., 84 (1978), pp. 957–1041.
[15] ———, *Quasi-Monte Carlo methods for multidimensional numerical integration*, in Numerical Integration III, International Series of Numerical Math. 85, H. Brass and G. Hämmerlin eds., Birkhäuser Verlag, Basel, 1988.
[16] ———, *Application of diophantine approximations to numerical integration*, in Diophantine Approximation and Its Applications, C. F. Osgood, ed., Academic Press, New York, 1973, pp. 129–199.
[17] ———, *Point sets and sequences with small discrepancy*, Monatsh. Math., 104 (1987), pp. 273–337.
[18] ———, *Random Number Generators and Quasi-Monte Carlo Methods*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.
[19] G. PAGES, *Van der Corput sequences, Kakutani transforms and one-dimensional numerical integration*, 1992, preprint.
[20] G. PAGES AND Y. J. XIAO, *Sequences with low discrepancy and pseudo-random numbers: theoretical results and numerical tests*, 1992, preprint.
[21] W. H. PRESS AND S. A. TEUKOLSKY, *Quasi- (that is, sub-) random numbers*, Comput. Phys., 3 (1989), pp. 76–79.
[22] P. K. SARKAR AND M. A. PRASAD, *A comparative study of pseudo and quasi random sequences for the solution of integral equations*, J. Comput. Phys., 68 (1987), pp. 66–88.
[23] I. M. SOBOL', *The distribution of points in a cube and the approximate evaluation of integrals*, U.S.S.R. Comput. Math. and Math. Phys., 7 (1976), pp. 86–112.
[24] ———, *Uniformly distributed sequences with additional uniformity property*, U.S.S.R. Comput. Math. and Math. Phys., 16, (1976), pp. 1332–1337.
[25] T. T. WARNOCK, *Computational investigations of low discrepancy point sets*, in Applications of Number Theory to Numerical Analysis, S. K. Zaremba, ed., Academic Press, New York, 1972, pp. 319–343.
[26] H. WOŹNIAKOWSKI, *Average case complexity of multivariate integration*, Bull. Amer. Math. Soc., 24 (1991), pp. 185–194.