Journal of Computational and Applied Mathematics 259 (2014) 798–804

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Uniform point sets and the collision test

Ahmet Göncü [a,*], Giray Ökten [b]

[a] Department of Mathematical Sciences, Xian Jiaotong-Liverpool University, Suzhou 215123, China
[b] Department of Mathematics, Florida State University, Tallahassee, FL 32306-4510, USA

ARTICLE INFO

Article history:
Received 18 October 2012
Received in revised form 10 July 2013

Keywords:
Collision test
Uniform point sets
Quasi-Monte Carlo
Error bounds

ABSTRACT

Monte Carlo and quasi-Monte Carlo methods are popular numerical tools used in many applications. The quality of the pseudorandom sequence used in a Monte Carlo simulation is essential to the accuracy of its estimates. Likewise, the quality of the low-discrepancy sequence determines the accuracy of a quasi-Monte Carlo simulation. There is a vast literature on statistical tests that help us assess the quality of a pseudorandom sequence. However, for low-discrepancy sequences, assessing quality by estimating discrepancy is a very challenging problem, leaving us with no practical options in very high dimensions. In this paper, we will discuss how a certain interpretation of the well-known collision test for pseudorandom sequences can be used to obtain useful information about the quality of low-discrepancy sequences. Numerical examples will be used to illustrate the applications of the collision test.

1. Introduction

Monte Carlo (MC) and quasi-Monte Carlo (QMC) methods are two popular numerical tools used in a wide variety of applications. It is convenient to describe these methods in the context of numerical integration, where one estimates $\int_{(0,1)^s} f(x)dx$ by function averages $\frac{1}{N}\sum_{i=1}^{N} f(x_i)$. In MC, the points x_i come from a pseudorandom sequence, and in QMC, a low-discrepancy sequence. Convergence of the function averages to the integral, and error estimation, is done in fundamentally different ways: the theory of MC methods is based on probability, whereas QMC methods is based on number theory, in particular, the theory of uniform distribution modulo one. A comprehensive survey of QMC can be found in the monograph by Niederreiter [1].

The quality of the number sequence used in MC and QMC is essential to the accuracy of their estimates. In MC, a common method to assess the quality of a pseudorandom sequence is to subject the sequence to a battery of statistical tests. These tests check how well the sequence approximates the ideal *random number sequence*, since the theory of MC is based on the assumption that the numbers that we observe are i.i.d. realizations of a random variable. There is a vast literature on the theory of statistical tests for randomness; see [2] for a survey.

Measuring the quality of a low-discrepancy sequence brings additional challenges. There are two main approaches: theoretical, and numerical. The theoretical approach compares different sequences by computing the upper bounds for their discrepancy; the limitations of this approach are discussed in [3]. The most commonly used numerical approach is to compare different sequences by the exact error they produce when they are applied to some problems with known solutions. This approach is practical, but does not necessarily reveal why a sequence is better, or whether the advantages observed in one problem can be extrapolated to other problems.

* Corresponding author. Tel.: +86 15190015627; fax: +86 51281880442.
E-mail addresses: goncuahm@gmail.com, Ahmet.Goncu@xjtlu.edu.cn (A. Göncü), okten@math.fsu.edu (G. Ökten).

0377-0427/$ – see front matter
http://dx.doi.org/10.1016/j.cam.2013.07.019

The alternative numerical approach is to compute the star-discrepancy, or one of its variants. The star-discrepancy of vectors $x_1, \ldots, x_N \in [0, 1)^s$ is

$$D_N^*(x_i) = \sup_S |A_N(S)/N - \lambda(S)| \tag{1}$$

where S is an s-dimensional interval of the form $\prod_{i=1}^s [0, \alpha_i)$, $A_N(S)$ counts the number of vectors x_i, $1 \le i \le N$, that are in S, $\lambda(S)$ is the Lebesgue measure of S, and the supremum is taken over the family of all such intervals. If the supremum is taken over intervals of the form $\prod_{i=1}^s [\alpha_i, \beta_i)$, then we obtain the so-called (extreme) discrepancy. Computing the star-discrepancy is an NP-hard problem [4]. There are a few algorithms in the literature that compute estimates for star-discrepancy [5–11]. Because of the computational complexity, these algorithms are typically used when the dimension is in the low tens and the number of vectors is a few hundreds; values that are much smaller than what is used in practical applications in simulation. In the next section, we will discuss how a well known statistical test for randomness can be used as a measure of quality for low-discrepancy sequences. In the numerical examples we will show how the test can give useful information about the underlying sequences in dimensions up to 20, and the number of vectors up to 40,000.

2. Uniform point sets and the collision test

A popular type of low-discrepancy sequences used in QMC integration is (t, s)-sequences. Let $s \ge 1$, $b \ge 2$ be integers. An elementary interval in base b is a subinterval of $[0, 1)^s$ of the form $\prod_{j=1}^s \left[a_j/b^{d_j}, (a_j + 1)/b^{d_j}\right)$, where $d_j \ge 0$ and $0 \le a_j < b^{d_j}$. For $0 \le t \le m$, a finite sequence of b^m points in $[0, 1)^s$ 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 of the sequence. An infinite sequence of points $q_1, q_2, \ldots$ is called a (t, s)-sequence in base b if the finite sequence $q_{kb^m+1}, \ldots, q_{(k+1)b^m}$ is a (t, m, s)-net in base b for all $k \ge 0$ and $m \ge t$.

The crucial property of a net (or, sequence) is how elements of the net are distributed among the elementary intervals: the proportion of elements of the net that fall in an elementary interval is equal to the volume of the elementary interval. This property also appears in the notion of a *uniform point set*, which is a generalization of (t, m, s)-nets introduced by Niederreiter [12]. Consider a general probability space $(X, \mathcal{B}, \mu)$, where X is an arbitrary nonempty set, $\mathcal{B}$ is a σ-algebra of subsets of X, and μ is a probability measure defined on $\mathcal{B}$. Let $\mathcal{M}$ be a nonempty subset of $\mathcal{B}$. For a point set $\mathcal{P} = \{x_1, \ldots, x_N\}$ and $M \subseteq X$, define $A(M; \mathcal{P})$ as the number of elements in $\mathcal{P}$ that belong to M. A point set $\mathcal{P}$ of N elements of X is called $(\mathcal{M}, \mu)$-uniform if

$$A(M; \mathcal{P})/N = \mu(M) \tag{2}$$

for all $M \epsilon \mathcal{M}$. Clearly, a (t, m, s)-net in base b is a $(\mathcal{M}, \mu)$-uniform point set where the measure μ is the Lebesgue measure on $(0, 1)^s$, and $\mathcal{M}$ is the collection of all elementary intervals in base b of volume b^{t-m}.

Niederreiter [12] proves error bounds when uniform point sets are used in QMC integration. An interesting feature of these bounds is that they do not require the integrand to be smooth. For example, for the special case when $\mathcal{M} = \{M_1, \ldots, M_K\}$ is a partition of X, Niederreiter [12] proves (Theorem 2, p. 288)

$$\left| \frac{1}{N} \sum_{n=1}^{N} f(x_n) - \int_X f d\mu \right| \le \sum_{j=1}^{K} \mu(M_j)(G_j(f) - g_j(f)) \tag{3}$$

where $G_j(f) = \sup_{t \in M_j} f(t)$ and $g_j(f) = \inf_{t \in M_j} f(t)$, $1 \le j \le K$.

The above error bound suggests the following intuition for a good QMC integration rule: find an appropriate partition of the domain, and then construct a point set that allocates elements to each set in the partition in proportions that are equal to the measure of each set. If we use a sequence $\{x_n\}_{n=1}^{\infty}$, instead of a finite point set $\{x_n\}_{n=1}^{N}$, then we certainly cannot preserve the property of a uniform set for all N. Instead, we would strive to obtain sequences where blocks of finite elements are uniform point sets infinitely often (like the relationship between a (t, m, s)-net and (t, s)-sequence), and between blocks the proportion of elements do not deviate too much from the measure of each set. We can then measure the quality of a uniform *sequence* by its deviation from perfect proportions.

The notion of a partition of a domain, and points falling in each set of the partition in equal proportions, also appears in the collision test, which is a statistical test for pseudorandom sequences. Consider K urns, and imagine tossing (randomly) a smaller number, N, of balls into the urns. We toss balls one after another. A ball falls into a specific urn with probability $1/K$. We say there is a collision if a ball falls into an urn that already has a ball. The probability distribution of the number of collisions and the percentage points for the distribution can be computed exactly. Given a pseudorandom number generator, we can count the number of collisions it gives, and reject the generator if this number is too small or too big. See [2] for a detailed discussion of the collision test.

We adopt the following interpretation of the collision test where urns are sets in *some* partition $\mathcal{M}$ of the domain X, and balls are elements of a sequence $\{x_n\}_{n=1}^{\infty}$ in X. The partition $\mathcal{M}$ is not necessarily tied to the way the sequence is constructed, rather, it can be obtained using the specific features of the problem at hand. For example, Spanier [13] considers some problems from particle transport theory where one integrates functions over a space of random walks X. Spanier partitions the space as $X = (\cup \Lambda_l) \cup \Lambda_\infty$ where Λ_l consists of random walks that terminates in l steps. This is a natural partition to consider in these type of problems, and Spanier remarks that (p. 134, [13]) a crucial test for QMC (or, MC) simulation is