INFORMATICA, 2016, Vol. 27, No. 2, 335–349

DOI: http://dx.doi.org/10.15388/Informatica.2016.88

Linear Time Estimators for Assessing Uniformity of Point Samples in Hypercubes

Günter RUDOLPH*, Simon WESSING
Fakultät für Informatik, Technische Universität Dortmund
Otto-Hahn-Str. 14, 44227 Dortmund, Germany
e-mail: guenter.rudolph@tu-dortmund.de, simon.wessing@tu-dortmund.de

Received: February 2016; accepted: May 2016

Abstract. We investigate the problem of detecting a point set's deviation from uniformity in the unit hypercube. High uniformity is for example desirable in Monte Carlo methods for numerical integration, but also for obtaining a good worst-case bound in global optimization. In high dimensions, many points are required to get reliable results, so the point sets are preferably generated by fast methods such as quasirandom sequences. Unfortunately, assessing their uniformity often requires quadratic time. So, we present several numerical summary characteristics of point sets that can be computed in linear time. They do not measure uniformity directly, but by comparing them to reference values for the uniform distribution, deviations from uniformity can be quickly detected. The necessary reference values are also derived here, if possible exactly, else approximately.

Key words: uniformity, measure, linear time, hypercube.

1. Introduction

Uniform sampling is an important tool in many application areas. For example, in Monte Carlo methods for numerical integration, a high uniformity of the used point sample is desirable to achieve a low approximation error. For this purpose, deterministic quasirandom sequences with *low discrepancy* (i.e. low deviation from uniformity) were developed. They provide a worst-case error bound for the integration error in numerical integration (Niederreiter, 1992). Corresponding randomized variants provide a variance reduction compared to random uniform sampling (L'Ecuyer, 2008).

In the following, n depicts the number of points and d the dimension. For quasirandom sequences, the asymptotic discrepancy for $n \to \infty$ is usually known. (Formally, we speak of low discrepancy sequences when their discrepancy has a convergence order of $O(n^{-1}\log^d n)$, Niederreiter, 1992, p. 32). However, for practical comparisons of (arbitrary) finite point sets this information is not sufficient. In this case we ideally want to calculate the discrepancy explicitly. In certain cases this is possible with a run time of $O(dn^2)$ (Morokoff and Caflisch, 1994). There also exists an asymptotically faster algorithm with run time $O(n\log^d n)$, which in practice is unfortunately only faster for $n > 2^{2d}$

*Corresponding author.

(Matoušek, 1998). Instead of developing a new algorithm for measuring discrepancy (or some other uniformity-related measure), our aim in this work is to investigate on measures that can be trivially calculated as Monte Carlo estimators in $O(dn)$ time, but which only tell us something about potential deviations from uniformity.

In the following, the *mean distance to the boundary* is identified as an important characteristic of point sets. Another potentially useful measure is the distance between the centre of mass of a point set and the centroid of the region of interest. Both are nothing more but conventional Monte Carlo estimates. As the expected values of these measures for random uniform points are useful reference values, they are derived analytically. Although neither characteristic is sufficient for measuring uniformity, they both can quickly detect certain deviations from it. This way, they can contribute to a richer description of experimental designs and point sets in general, for example as part of statistical tests for complete spatial randomness (Illian *et al.*, 2008, pp. 83–98).

Apart from the numerical integration mentioned above, global optimization is another application area where the worst-case error bound depends on a uniformity measure. This measure is defined as follows:

Definition 1. (Covering radius) (See Niederreiter, 1992, p. 148.) If $(\mathcal{X}, \rho)$ is a bounded metric space and the point set $\mathcal{P}$ consists of $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$, then the *covering radius* of $\mathcal{P}$ in $\mathcal{X}$ is defined by

$$d_n(\mathcal{P}, \mathcal{X}) = \sup_{\mathbf{x}\in\mathcal{X}} \left\{ \min_{1\leqslant i\leqslant n} \left\{\rho(\mathbf{x}, \mathbf{x}_i)\right\}\right\} = \sup_{\mathbf{x}\in\mathcal{X}} \left\{d_{\mathrm{nn}}(\mathbf{x}, \mathcal{P})\right\}.$$

Niederreiter coined the term *dispersion* for d_n, which "may be viewed as a measure for the deviation from denseness" (Niederreiter, 1992, p. 149). However, the name did not become widely accepted, because it does not reflect the intuitive meaning of d_n and by now it is also used differently in other diversity-related research (see, e.g. Erhan, 1990; Lunacek and Whitley, 2006). Meinl *et al.* (2011) call this indicator the n-centre measure. We will use the name *covering radius*, which is used for example by Damelin *et al.* (2010), because d_n is the smallest radius for which closed balls around the points of $\mathcal{P}$ completely cover $\mathcal{X}$. We restrict ourself to hypercubes $\mathcal{X} = [\ell_1, u_1] \times \cdots \times [\ell_d, u_d]$ in this work.

Definition 1 is actually also identical to that of the minimax distance design criterion as defined by Johnson *et al.* (1990). Based on this definition, Niederreiter (1992, p. 149) proves an error bound on the estimate $\hat{f}^* = f(\hat{\mathbf{x}}^*)$ of the global minimum $f(\mathbf{x}^*)$ of a function $f : \mathcal{X} \to \mathbb{R}$. In this estimate, $\hat{\mathbf{x}}^* = \arg\min\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{P}\}$ denotes the point in the finite approximation set $\mathcal{P} \subset \mathcal{X}$ for which the best objective value could be observed.

Theorem 1. *(See Niederreiter, 1992, p. 149.) If $(\mathcal{X}, \rho)$ is a bounded metric space, then, for any point set $\mathcal{P}$ of n points in $\mathcal{X}$ with covering radius $d_n = d_n(\mathcal{P}, \mathcal{X})$, we have*

$$\hat{f}^* - f\left(\mathbf{x}^*\right) \leqslant \omega(f, d_n)\,,$$

where

$$\omega(f,t) = \sup_{\substack{\mathbf{x}_i,\mathbf{x}_j\in\mathcal{X}\\ \rho(\mathbf{x}_i,\mathbf{x}_j)\leqslant t}} \left\{\left|f(\mathbf{x}_i) - f(\mathbf{x}_j)\right|\right\}$$

is, for $t \geqslant 0$*, the* modulus of continuity *of* f.

This means that the prediction error for the global optimum is bounded by a function only depending on f and the covering radius of the set $\mathcal{P}$. Unfortunately, it is much more difficult to create samples with low covering radius than with low discrepancy (Santner *et al.*, 2003, p. 149). The difficulty in calculating and thus optimizing $d_n(\mathcal{P},\mathcal{X})$ is due to the involvement of the uncountable $\mathcal{X}$. Although no explicit formula is known for arbitrary $\mathcal{X}$, Pronzato and Müller (2012) give an algorithm for calculating $d_n(\mathcal{P},[0,1]^d)$ regarding Euclidean distance with run time $O((dn)^{\lfloor d/2\rfloor})$, based on Delaunay tessellation. (Another resort would be using a Monte Carlo approximation, because if $\mathcal{X}$ is finite with $|\mathcal{X}| = m$, calculation of the indicator becomes straightforward with run time $O(mdn)$.)

Instead, we are going to take another approach, based on an observation of Johnson *et al.* (1990).

2. Distance to the Boundary

A popular approach to generate point sets with a good distribution is to maximize the minimal distance between points in the set (thus known as maximin approach). Unfortunately, maximin approaches are plagued by a drift towards the boundary (Johnson *et al.*, 1990), which means that the point density at the boundary is higher than in the interior of the space $\mathcal{X}$ (Illian *et al.*, 2008, p. 145). Johnson *et al.* (1990) point out that in contrast to maximin designs, minimax designs do tend to avoid the boundaries. Therefore, it would also be interesting to find out more about the boundary behaviour of other sampling methods. And it would be even more desirable to have sampling methods with controllable behaviour. As a first step in this direction, let us define a measure to quantify the proximity of a point set to the boundary.

Proposition 1. *The distance between a point* $\mathbf{x}\in\mathcal{X}$ *and the nearest neighbour on the boundary* $\mathcal{B} = \{\mathbf{x}\in\mathcal{X} \mid \exists i\in\{1,\ldots,d\} : x_i = u_i \vee x_i = \ell_i\}$ *is under every* L_p *norm*

$$d_{\mathrm{nn}}(\mathbf{x},\mathcal{B}) = \min_{1\leqslant i\leqslant n} \left\{\min\{x_i - \ell_i, u_i - x_i\}\right\}.$$

Proof. Because the boundaries of $\mathcal{X}$ are paraxial, there exists a $\mathbf{y}\in\mathcal{B}$ with $\rho(\mathbf{x},\mathbf{y}) = d_{\mathrm{nn}}(\mathbf{x},\mathcal{B})$, which only differs from $\mathbf{x}$ in one variable i. As in one-dimensional space all L_p distances are identical to the absolute difference, the distance must be the smaller one of $x_i - \ell_i$ and $u_i - x_i$. □

Proposition 2. *The expected distance between a random uniform point X in $[0,1]^d$ and the boundary $\mathcal{B}$ is*

$$\delta_d := \mathsf{E}\big[d_{\mathrm{nn}}(X,\mathcal{B})\big] = \frac{1}{2}\cdot\frac{1}{1+d}.$$

Proof. The expected distance to the lower boundaries is identical to the first order statistic $U_{1:d}$ (the minimum) of a random sample $U_1,\ldots,U_d$ from $U(0,1)$. $U_{1:d}$ has Beta$(1,d)$ distribution (Arnold *et al.*, 1992, pp. 13–14) with density

$$f_{U_{1:d}}(x) = d\cdot(1-x)^{d-1}\cdot\mathbb{1}_{[0,1]}(x)\,dx \tag{1}$$

and expectation $\mathsf{E}[U_{1:d}] = 1/(1+d)$. To account for the upper boundaries, too, it is sufficient to consider $Y_i \sim U(0,\frac{1}{2})$ instead, because $0 \leqslant Y_i = \min\{U_i, 1-U_i\} \leqslant \frac{1}{2}$. Therefore, the sought quantity is $\mathsf{E}[Y_{1:d}] = \mathsf{E}[\frac{1}{2}U_{1:d}] = \frac{1}{2}\cdot\mathsf{E}[U_{1:d}]$. □

As we can see, the expected distance to the boundary decreases with increasing dimension. This is just another manifestation of the curse of dimensionality and shows us that in high dimensions, almost all of the space is in the boundary regions (Illian *et al.*, 2008, p. 183). We can now use the sample mean

$$\bar{d}_{\mathcal{B}} = \frac{1}{n}\sum_{i=1}^{n} d_{\mathrm{nn}}(\mathbf{x}_i,\mathcal{B}) \tag{2}$$

to estimate how much emphasis a point set puts on the boundary in comparison with the uniform distribution. The interesting thing about this criterion is that we are using a Monte Carlo estimate to assess the quality of our point set. Although it alone is not sufficient for getting the whole picture, it is attractive because it is a necessary condition for uniformity and can be computed in linear time. Finally, we formalize the observation of Johnson *et al.* (1990), regarding the boundary behaviour, in the following conjecture.

Conjecture 1. *Point sets with maximal separation distance (maximin designs) possess $\bar{d}_{\mathcal{B}} < \delta_d$ and point sets with minimal covering radius (minimax designs) exhibit $\bar{d}_{\mathcal{B}} > \delta_d$.*

This conjecture may be only a rule of thumb, but it is certainly true for the conventional grid and the Sukharev grid, which are the optimal solutions under the L_∞ norm regarding separation distance and covering radius, respectively, Sukharev (1971), LaValle (2006, pp. 202–203). The distance to the boundary is shown in Fig. 1, where we can also see that $\bar{d}_{\mathcal{B}}$, the Monte Carlo estimate for δ_d, becomes more precise with increasing n, but this estimate is often less accurate than what we would expect in 95% of the cases with random uniform points. For other norms, the optimal point sets can in the general case only be approximated.

Niederreiter (1992, p. 152) shows that every low-discrepancy point set also is a low-covering-radius point set (but not vice versa). This raises hope that combining discrepancy

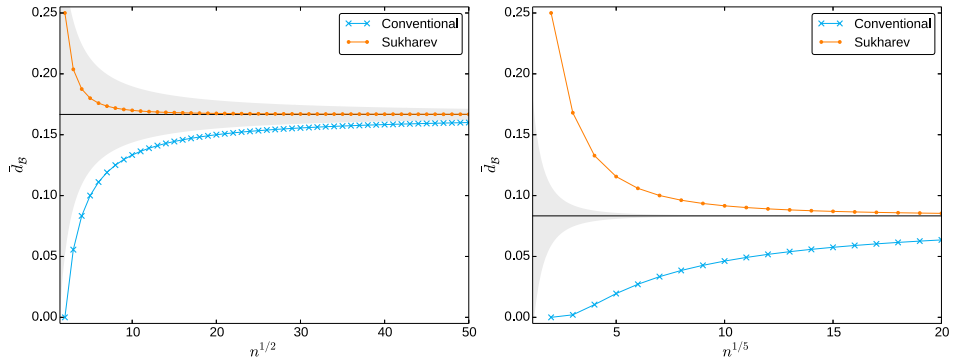


Fig. 1. Distance to the boundary for conventional and Sukharev grids in two (left) and five dimensions (right). The horizontal black line indicates δ_d. The gray area represents a 95% confidence interval for δ_d under the conditions of Proposition 3.

(or even a maximin criterion) with a side constraint $\bar{d}_{\mathcal{B}} - \delta_d \geqslant \varepsilon$ yields a good criterion to obtain point sets with a small covering radius. If, on the other hand, an exactly uniform point set is sought, we can compare the deviation $|\bar{d}_{\mathcal{B}} - \delta_d|$ to the expected value for random uniform point sets. This expected value can be calculated exactly. Note that $\bar{d}_{\mathcal{B}}$ can be rewritten to

$$\bar{d}_{\mathcal{B}} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} U_{1:d}^{(i)}$$

where $U_{1:d}^{(i)}$ is the ith of n independent and identically distributed r.v.s. representing the minimum of d independent standard uniform r.v.s. Therefore holds

$$\begin{aligned} \mathsf{E}[|\bar{d}_{\mathcal{B}} - \delta_d|] &= \mathsf{E}\left[\left|\frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} U_{1:d}^{(i)} - \delta_D\right|\right] = \frac{1}{2n} \mathsf{E}\left[\left|\underbrace{\sum_{i=1}^{n} U_{1:d}^{(i)}}_{=:S_n} - \underbrace{\frac{n}{d+1}}_{=:\mu_n}\right|\right] \\ &= \frac{1}{2n} \mathsf{E}[|S_n - \mu_n|] \end{aligned} \tag{3}$$

where the p.d.f. of S_n is a convolution of the densities of n independent and identically distributed Beta random variables each with p.d.f. as given in (1). According to Nadarajah *et al.* (2015, p. 103), no one has been able to derive the exact distribution of the sum of more than two independent Beta random variables. Of course, this statement does not exclude that it might be possible for special cases.

Recall that the p.d.f. of S_1 is just the p.d.f. of $U_{1:d}$ and therefore a polynomial of degree $(d-1) \in \mathbb{N}$ with rational coefficients. As a consequence, the p.d.f. of S_n can be determined recursively via

$$f_{S_n}(s) = \int_0^n f_{S_{n-1}}(t) \cdot f_{S_1}(s-t)\, dt$$

for $n \geqslant 2$ yielding a polynomial of degree $n\,d - 1$ with rational coefficients. Since

$$\begin{aligned}\mathsf{E}\big[|S_n - \mu_n|\big] &= \int_0^n |s - \mu_n|\, f_{S_n}(s)\, ds \\ &= \int_0^{\mu_n} (\mu_n - s)\, f_{S_n}(s)\, ds + \int_{\mu_n}^n (s - \mu_n)\, f_{S_n}(s)\, ds\end{aligned}$$

and $\mu_n \in \mathbb{Q}$, the value of the expression can be calculated exactly in form of a rational number. Insertion in (3) leads to the exact value of $\mathsf{E}[|\bar{d}_{\mathcal{B}} - \delta_d|]$ for given d and n. Admittedly, this is a tedious procedure but it has to be done only once to build a table up to some limits for d and n. If these limits are large enough, approximations should be sufficient.

Proposition 3. *For a set of random uniform points $\mathcal{P}$ in $\mathcal{X} = [0,1]^d$, $|\mathcal{P}| = n$, with $\bar{d}_{\mathcal{B}}$ computed as in (2), the mean absolute deviation around the mean δ_d is for large n*

$$\mathsf{E}\big[|\bar{d}_{\mathcal{B}} - \delta_d|\big] \approx \frac{1}{\sqrt{2\,\pi}} \cdot \frac{1}{\sqrt{n}} \cdot \frac{1}{d+1} \cdot \sqrt{1 - \frac{2}{d+2}}. \tag{4}$$

Proof. As shown in the proof of Proposition 2, we have $d_{\mathrm{nn}}(X, \mathcal{B}) \stackrel{d}{=} \frac{1}{2}\, U_{1:d}$ and hence (Johnson *et al.*, 1995, p. 217)

$$\mathsf{V}\big[d_{\mathrm{nn}}(X, \mathcal{B})\big] = \mathsf{V}\left[\frac{1}{2}\, U_{1:d}\right] = \frac{1}{4}\, \mathsf{V}[U_{1:d}] = \frac{1}{4}\, \frac{d}{(d+2)\,(d+1)^2}\,. \tag{5}$$

As n is typically large, we can now apply the central limit theorem, which says that the mean of n independent identically distributed random variables with mean μ and variance σ^2 is asymptotically normally distributed with mean μ and variance σ^2/n (Billingsley, 1995, p. 357). Therefore, the distribution of $\bar{d}_{\mathcal{B}}$ is approximately $N(\delta_d, \sigma^2/n)$ for large n where σ^2 is the variance given in (5). The mean absolute deviation around the mean then approximately follows the half-normal distribution, whose expected value in this case is $\sqrt{2/\pi} \cdot \sigma/\sqrt{n}$. Insertion of (5) leads to the result. □

3. Distance Between Centre of Mass and Centroid of the Hypercube

The distance of the sample's centre of mass (i.e. the sample mean)

$$\bar{\mathbf{c}}_{\mathcal{P}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

from the centroid $\mathbf{c} = (\mathbf{l} + \mathbf{u})/2$ of the hypercube $\mathcal{X}$ may be used as an additional quality measure. It is another Monte Carlo estimate that can be computed in linear time and we are able to analytically derive the expected value for random uniform sets as a reference.

Suppose that $\mathcal{X} = [0, 1]^d$ with centroid $\mathbf{c} = (1/2, \ldots, 1/2)^T$. If the sample of size n is drawn from a uniform distribution on $\mathcal{X}$, then each component of the sample mean is built by the sum of n independent uniform random variables with support $[0, 1]$ divided by n. Consequently, all components of the sample mean are independent and identically distributed with a well known distribution.

Let $S_n = U_1 + U_2 + \cdots + U_n$ be the sum of n independent uniform random variables (r.v.s) with support $[0, 1]$. Then S_n has Irwin-Hall distribution (Johnson *et al.*, 1995, p. 296) and $B_n = \frac{1}{n} S_n$ has Bates[1] distribution (Johnson *et al.*, 1995, p. 297) with probability density function (p.d.f.) $f_{B_n}(x) = n\, f_{S_n}(n\, x)$ given by

$$f_{B_n}(x) = \frac{n^n}{2\Gamma(n)} \sum_{k=0}^{n} (-1)^k \binom{n}{k} \left(x - \frac{k}{n}\right)^{n-1} \left(1 - 2 \cdot 1_{[0,k/n)}(x)\right) \cdot 1_{[0,1]}(x)$$

with $\mathsf{E}[B_n] = \frac{1}{2}$, $\mathsf{V}[B_n] = \frac{1}{12n}$ and excess kurtosis $\mathsf{K}[B_n] = -\frac{6}{5n}$.

The next result ensures that the distance between sample mean and expected mean in L_1 norm can be determined from the mean deviation of a single Bates random variable.

Proposition 4. *Let $\mathbf{c} \in \mathbb{R}^d$ with $c_i = c_0$ for $i = 1, \ldots, d$. If $U^{(i)} \sim U[0, 1]^d$ with $i = 1, \ldots, n$ are independent and identically distributed (i.i.d.) random vectors, then*

$$\mathsf{E}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} U^{(i)} - \mathbf{c} \right\|_1\right] = d \cdot \mathsf{E}\big[|B_n - c_0|\big] \tag{6}$$

$$\mathsf{V}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} U^{(i)} - \mathbf{c} \right\|_1\right] = d \cdot \mathsf{V}\big[|B_n - c_0|\big] \tag{7}$$

where B_n is Bates-distributed.

Proof. Since

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^{n} U^{(i)} - \mathbf{c} \right\|_1 &= \left\| \left(\frac{1}{n} \sum_{i=1}^{n} U_1^{(i)} - c_0, \ldots, \frac{1}{n} \sum_{i=1}^{n} U_d^{(i)} - c_0 \right)' \right\|_1 \\ &= \left\| \left(B_n^{(1)} - c_0, \ldots, B_n^{(d)} - c_0 \right)' \right\|_1 \\ &= \sum_{j=1}^{d} \left| B_n^{(j)} - c_0 \right| \end{aligned} \tag{8}$$

and the Bates r.v.s $B_n^{(j)}$ are i.i.d. for $j = 1, \ldots, d$, the result follows by applying standard rules for expectation and variance in (8). □

[1] Bates herself (Bates, 1955, p. 713) credits the derivation of this distribution to Pierre Simon de Laplace in 1774.

The mean absolute deviation $\mathsf{E}[|B_n - \mu|]$ with $\mu = \mathsf{E}[B_n]$ can be calculated exactly according to the following result, provided that the integral in (9) can be determined exactly.

Proposition 5. *If B is a continuous r.v. with support $[0,1]$ and $\mathsf{E}[B] = \mu = \frac{1}{2}$ where $f_B(\mu + x) = f_B(\mu - x)$ for all $x \in \mathbb{R}$, then*

$$\mathsf{E}\big[|B-\mu|\big] = \mu - 2\int_0^\mu x\, f_B(x)\,dx \quad \textit{and} \tag{9}$$

$$\mathsf{V}\big[|B-\mu|\big] = \mathsf{V}[B] - \mathsf{E}\big[|B-\mu|\big]^2. \tag{10}$$

Proof. Since

$$\begin{aligned}
\mathsf{E}\big[|B-\mu|\big] &= \int_0^1 |x-\mu|\, f_B(x)\,dx \\
&= \int_0^\mu |x-\mu|\, f_B(x)\,dx + \int_\mu^1 |x-\mu|\, f_B(x)\,dx \\
&= \int_0^\mu (\mu-x)\, f_B(x)\,dx + \int_\mu^1 (x-\mu)\, f_B(x)\,dx \\
&= \underbrace{\mu\int_0^\mu f_B(x)\,dx - \mu\int_\mu^1 f_B(x)\,dx}_{=0} \\
&\quad + \int_\mu^1 x\, f_B(x)\,dx - \int_0^\mu x\, f_B(x)\,dx \\
&= \underbrace{\int_0^1 x\, f_B(x)\,dx}_{=\mu} - 2\int_0^\mu x\, f_B(x)\,dx,
\end{aligned}$$

the result for the expectation (9) has been shown. As for the variance, note that

$$\mathsf{V}\big[|B-\mu|\big] = \mathsf{E}\big[|B-\mu|^2\big] - \mathsf{E}\big[|B-\mu|\big]^2 = \underbrace{\mathsf{E}\big[(B-\mu)^2\big]}_{=\mathsf{V}[B]} - \mathsf{E}\big[|B-\mu|\big]^2$$

which proves (10). □

If B has Bates distribution, i.e. $B = B_n$, the p.d.f. f_{B_n} in (9) is a polynomial with rational coefficients so that the integral can be evaluated *exactly* as a rational number. Table 1 contains the values for $n = 1, \ldots, 8$.

Exact values for larger n can be obtained and tabulated with the assistance of symbolic mathematical software. But since tables are inevitably of finite size an appropriate *quick* approximation of the mean deviation for large n is desirable.

Table 1
Exact mean deviation of Bates r.v.s
for $n = 1, \ldots, 8$.

n	$\mathsf{E}[\lvert B_n - \mu \rvert]$
1	$\frac{1}{4} = 0.2500$
2	$\frac{1}{6} \approx 0.1666$
3	$\frac{13}{96} \approx 0.1354$
4	$\frac{7}{60} \approx 0.1166$
5	$\frac{1199}{11520} \approx 0.1040$
6	$\frac{239}{2520} \approx 0.0948$
7	$\frac{113149}{1290240} \approx 0.0876$
8	$\frac{1487}{18144} \approx 0.0819$

Proposition 6. *The expected L_1 distance between $\bar{\mathbf{c}}_\mathcal{P}$ and $\mathbf{c}_\mathcal{X} = (0.5, \ldots, 0.5)^\top$ for a set of random uniform points $\mathcal{P}$ in $\mathcal{X} = [0,1]^d$, $|\mathcal{P}| = n$, is for large n*

$$\mathsf{E}\big[\|\bar{\mathbf{c}}_\mathcal{P} - \mathbf{c}_\mathcal{X}\|_1\big] \approx d \cdot \sqrt{\frac{2}{\pi}} \cdot \frac{1}{\sqrt{12n}} =: \hat{m}_1. \tag{11}$$

Proof. First of all, we note that it suffices to regard the one-dimensional distances, because

$$\mathsf{E}\big[\|\bar{\mathbf{c}}_\mathcal{P} - \mathbf{c}_\mathcal{X}\|_1\big] = \mathsf{E}\left[\sum_{i=1}^{d} |\bar{c}_{\mathcal{P},i} - 0.5|\right] = \sum_{i=1}^{d} \mathsf{E}\big[|\bar{c}_{\mathcal{P},i} - 0.5|\big].$$

The standard deviation of a random uniform variable on $[0,1]$ is $1/\sqrt{12}$ (Johnson *et al.*, 1995, p. 279). Again using the central limit theorem, we obtain that $\forall i \in \{1, \ldots, d\}$, $\bar{c}_{\mathcal{P},i}$ converges in distribution to $N(0.5, \sigma^2_{\bar{c}_{\mathcal{P},i}})$, where $\sigma_{\bar{c}_{\mathcal{P},i}} = 1/\sqrt{12n}$ is both the standard error of the mean and the standard deviation of the estimate $\bar{c}_{\mathcal{P},i}$. Finally, $|\bar{c}_{\mathcal{P},i} - 0.5|$ again approximately follows a half-normal distribution with expected value $\sqrt{2/\pi} \cdot \sigma_{\bar{c}_{\mathcal{P},i}}$, which we only have to multiply by d to obtain the expected L_1 distance in d dimensions. □

The data depicted in Fig. 2 was generated exactly by symbolic mathematical software up to $n = 300$ before subtracting the approximating value (11) with 50 significant digits. The deviation is less than 10^{-5} for $n > 110$ and about $2.2 \cdot 10^{-6}$ for $n = 300$. Thus, the tabulated values can be used for small values of n whereas the approximation is taken for larger values.

In the light of the results of Aggarwal *et al.* (2001), this reference value may be useful in high dimensions. On the other hand, it seems inappropriate to assess a $\bar{\mathbf{c}}_\mathcal{P}$ with radially symmetric distribution using the L_1 distance, which is not radially symmetric. Luckily, we can also derive a similar result for the L_2 distance:

Proposition 7. *Let B_n be Bates-distributed and $\mathbf{c} \in \mathbb{R}^d$ with $c_i = \mathsf{E}[B_n]$ for $i = 1, \ldots, d$. If $U^{(i)} \sim U[0,1]^d$ with $i = 1, \ldots, n$ are independent and identically distributed (i.i.d.)*

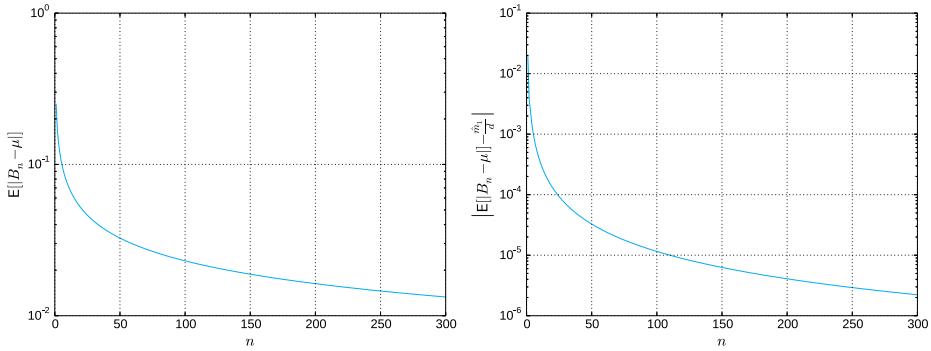


Fig. 2. The development of $\mathsf{E}[|B_n-\mu|]$ for values of $n=1,\ldots,300$ (left) and the corresponding approximation error of the normality assumption (right).

random vectors, then

$$\mathsf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}U^{(i)}-\mathbf{c}\right\|_2^2\right] = d\cdot\mathsf{V}[B_n]=\frac{d}{12\,n} \tag{12}$$

$$\mathsf{V}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}U^{(i)}-\mathbf{c}\right\|_2^2\right] = d\cdot\mathsf{V}[B_n]^2(2+\mathsf{K}[B_n])=\frac{d\,(5\,n-3)}{360\,n^3}. \tag{13}$$

Proof. Since analogous to the previous proof

$$\left\|\frac{1}{n}\sum_{i=1}^{n}U^{(i)}-\mathbf{c}\right\|_2^2 = \left\|\left(B_n^{(1)}-c_0,\ldots,B_n^{(d)}-c_0\right)'\right\|_2^2=\sum_{j=1}^{d}\left(B_n^{(j)}-c_0\right)^2 \tag{14}$$

and the Bates r.v.s $B_n^{(j)}$ are i.i.d. for $j=1,\ldots,d$, we obtain $d\cdot\mathsf{E}[(B_n-c_0)^2]=d\,\mathsf{V}[B_n]$ for the expectation of (14) and $d\cdot\mathsf{V}[(B_n-c_0)^2]=d\cdot(\mathsf{E}[(B_n-c_0)^4]-\mathsf{E}[(B_n-c_0)^2]^2)=\mathsf{V}[B_n]^2\,(\mathsf{K}[B_n]+3)-\mathsf{V}[B_n]^2$ for the variance of (14) where $c_0=\mathsf{E}[B_n]$. □

Again, note that there is no point in directly optimizing $\|\bar{\mathbf{c}}_{\mathcal{P}}-\mathbf{c}_{\mathcal{X}}\|$ or $|\bar{d}_{\mathcal{B}}-\delta_d|$, because they are not sufficient conditions for uniformity. However, it should be useful to apply these measures to given point sets to detect potentially undesired deviations from uniformity.

4. Application

Quasirandom sequences are often analyzed experimentally by plotting discrepancy against the number of points (Morokoff and Caflisch, 1994). We are going to do the same thing here and compare the resulting figures with figures for our linear-time estimators. As discrepancy, the measure called T_n by Morokoff and Caflisch (1994) is used. Discrepancy

is calculated for up to 2^{18} points, while the linear-time measures are applied until 2^{20}. The following quasirandom sequences are used: besides the original Halton sequence, we employ the generalized (or scrambled) Halton sequence (Faure and Lemieux, 2009), which arises from inserting certain permutations into the definition of the original Halton sequence. Concretely, we use the implementation of De Rainville *et al.* (2012) with optimized two-dimensional projections.[2] As a third representative, we use a state-of-the-art Sobol' sequence[3] by Joe and Kuo (2008). It, too, offers optimized two-dimensional projections, but not treating all projections equally.

Figure 3 compares the three sequences regarding discrepancy and the two new measures $|\bar{d}_{\mathcal{B}} - \delta_d|$ and $\|\bar{\mathbf{c}}_{\mathcal{P}} - \mathbf{c}_{\mathcal{X}}\|_2$ in five and twenty dimensions. (The comparison has also been carried out for $d = 2, 3, 10, 40$.) The good news is that a reasonable estimation of $\mathbf{c}_{\mathcal{X}}$ seems to be possible with all sequences with as few as 2^{12} to 2^{14} points (see Fig. 3(b)). This holds even in forty dimensions. However, the Sobol' sequence seems to have a burn-in period where the other measures considerably deviate from uniformity (see Fig. 3(a)). In forty dimensions, an estimated number of 2^{23} points is necessary to even reach the discrepancy of random uniform points (not shown here), although in experiments of Jäckel (2002, pp. 91–96), Sobol' sequences with "pattern-breaking" initializations could always provide at least the same discrepancy as pseudorandom numbers. Results by Carter (2011) indicate that this contradiction to Jäckel's results stems from the different discrepancy formulation and not from the different Sobol' sequence.

Sobol' sequences typically exhibit local optima around powers of two if the number of points is sufficiently high in relation to the dimension (see Fig. 3(a)). These optima can also be detected in Fig. 3(b) for $\|\bar{\mathbf{c}}_{\mathcal{P}} - \mathbf{c}_{\mathcal{X}}\|$.

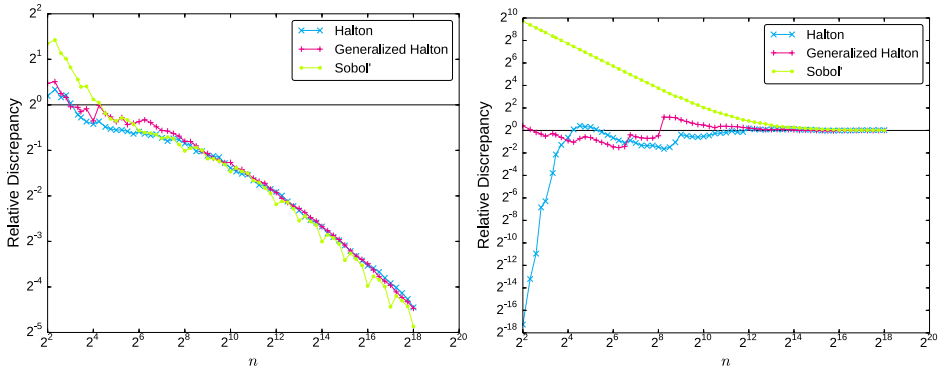
The Halton sequence initially obtains a suspiciously low discrepancy in high dimensions (see Fig. 3(a)). This may indicate that T_n gives misleading results if $n \ll 2^d$, because simultaneously the value of $\|\bar{\mathbf{c}}_{\mathcal{P}} - \mathbf{c}_{\mathcal{X}}\|$ of the Halton sequence is bad. In any case this problem was predicted by Matoušek (1998) for a related discrepancy formulation.

Finally, the results regarding $|\bar{d}_{\mathcal{B}} - \delta_d|$ are surprising (see Fig. 3(c)). While the absolute values of $\bar{d}_{\mathcal{B}}$ exhibit a nice progression towards δ_d (not shown), the impression changes when we put the deviation from δ_d into relation with the expected deviation for random uniform point sets. In five dimensions, the results seem noisy, but acceptable. For $d = 20$, however, we see a wavy pattern with deep spikes (which is even more pronounced in forty dimensions), and also the convergence rate seems no better than the random uniform expectation. Especially the $\bar{d}_{\mathcal{B}}$ of the Sobol' sequence is too large for low n, which causes the bad performance in Fig. 3(c).

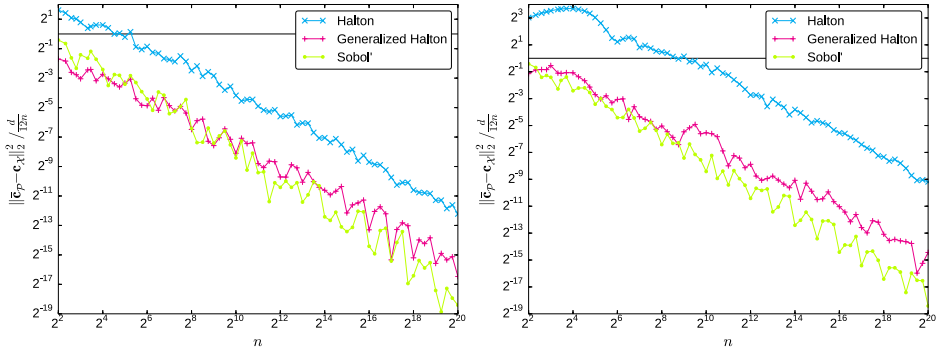
We conclude that our linear-time measures can give additional insights into the behaviour of quasirandom sequences, especially for large point sets where discrepancy cannot be computed conveniently any more. In total, the generalized Halton sequence seems to have the most stable performance, considering all three measures.

[2] Available in the ghalton Python library at `https://github.com/fmder/ghalton`, Version 0.6.

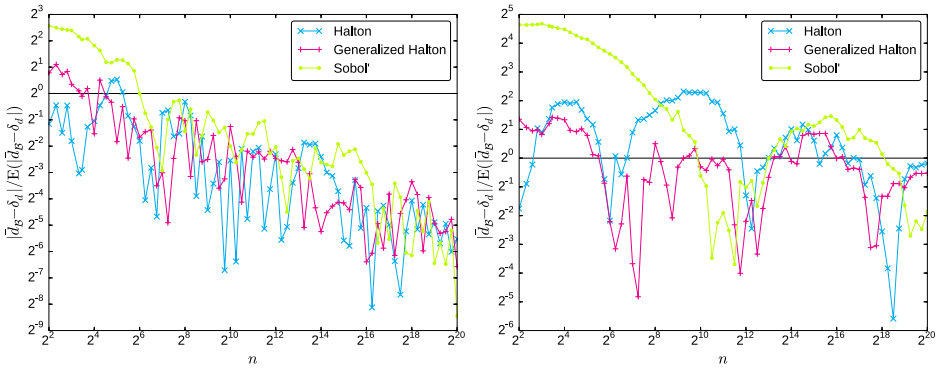
[3] Generated with the direction numbers "new-joe-kuo-6.21201" and software from `http://web.maths.unsw.edu.au/~fkuo/sobol/`, Version from 16 September 2010.



(a) Relative discrepancy.



(b) Relative squared Euclidean distance between $\mathcal{P}$'s centre of mass and the centroid of the hypercube.



(c) Relative deviation from the expected distance to the boundary.

Fig. 3. Quasirandom points evaluated with three summary characteristics for $d = 5$ (left) and $d = 20$ (right). The horizontal line indicates the reference value for each measure.

5. Conclusion and Outlook

In this work we proposed to gain information about the uniformity of point sets indirectly, by calculating measures for which we know reference values of the uniform distribution.

In the future, it may also be promising to incorporate these measures into statistical tests. For example, we might apply a nonparametric test for the hypothesis that some of the characteristic values calculated from the sample are equal to the value under uniformity assumption. If the test rejects the hypothesis, then we have strong evidence that the sample is nonuniform. Moreover, we could test if the values $2 \cdot d_{\mathrm{nn}}(\mathbf{x}_1, \mathcal{B}), \ldots, 2 \cdot d_{\mathrm{nn}}(\mathbf{x}_n, \mathcal{B})$, corresponding to a sample $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, are taken from a $\mathrm{Beta}(1, d)$ distribution. This approach provides the ability to detect the nonuniformity of a d-dimensional sample in linear time and statistically sound manner.

The general idea of our approach might also be transferable to more complicated sets, e.g. hyperballs. However, it seems difficult to obtain the required reference values for the uniform distribution in most cases. Especially non-convex objects do not seem to be tractable. In this case, one would have to revert to direct calculation of uniformity, as, e.g. shown by Hardin and Saff (2004).

Another important summary characteristic in spatial statistics is the nearest-neighbour distance (Illian *et al.*, 2008, pp. 126–127). Knowing the expected value for random uniform points in arbitrary hypercubes would be very useful, although the Monte Carlo estimator cannot be computed in linear time. Unfortunately, even special cases exhibit considerable intricacy (Žilinskas, 2003).

References

Aggarwal, C,C., Hinneburg, A., Keim, D.A. (2001). On the surprising behavior of distance metrics in high dimensional space. Van den Bussche, J., Vianu, V. (Eds.), *Database Theory – ICDT 2001. Lecture Notes in Computer Science*, Vol. 1973. Springer, pp. 420–434.

Arnold, B.C., Balakrishnan, N., Nagaraja, H.N. (1992). *A First Course in Order Statistics*. Wiley, New York.

Bates, G.E. (1955). Joint distributions of time intervals for the occurrence of successive accidents in a generalized Polya urn scheme. *Annals of Mathematical Statistics*, 26(4), 705–720.

Billingsley, P. (1995). *Probability and Measure*, third ed. Wiley, New York.

Carter, M. (2011). A toolbox for quasirandom simulation. *The Mathematica Journal*, 13. doi:10.3888/tmj.13-21.

Damelin, S.B., Hickernell, F.J., Ragozin, D.L., Zeng, X. (2010). On energy, discrepancy and group invariant measures on measurable subsets of Euclidean space. *Journal of Fourier Analysis and Applications*, 13(6), 813–839.

De Rainville, F.-M., Gagné, C., Teytaud, O., Laurendeau, D. (2012). Evolutionary optimization of low-discrepancy sequences. *ACM Transactions on Modeling and Computer Simulation*, 22(2), 9:1–9:25.

Erhan, E. (1990). The discrete p-dispersion problem. *European Journal of Operational Research*, 46(1), 48–60.

Faure, H., Lemieux, C. (2009). Generalized Halton sequences in 2008: a comparative study. *ACM Transactions on Modeling and Computer Simulation*, 19(4), 15:1–15:31.

Hardin, D.P., Saff, E.B. (2004). Discretizing manifolds via minimum energy points. *Notices of the American Mathematical Society*, 51(10), 1186–1194.

Illian, J., Penttinen, A., Stoyan, H., Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. Wiley.

Joe, S., Kuo, F.Y. (2008). Constructing Sobol sequences with better two-dimensional projections. *SIAM Journal on Scientific Computing*, 30(5), 2635–2654,

Johnson, M.E., Moore, L.M., Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26(2), 131–148.

Johnson, N.L., Kotz, S., Balakrishnan, N. (1995). *Continuous Univariate Distributions*, Vol. 2, 2nd ed. Wiley, New York.

Jäckel, P. (2002). *Monte Carlo Methods in Finance*. Wiley, New York.

LaValle, S.M. (2006). *Planning Algorithms*. Cambridge University Press, Cambridge.

Lunacek, M., Whitley, D. (2006). The dispersion metric and the CMA evolution strategy. In: *Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation, GECCO '06*. ACM, pp. 477–484.

L'Ecuyer, P. (2008). Comparison of Point sets and sequences for quasi-Monte Carlo and for random number generation. In Golomb, S.W., Parker, M.G., Pott, A., Winterhof, A. (Eds.), *Sequences and Their Applications – SETA 2008*, *Lecture Notes in Computer Science*, Vol. 5203. Springer, pp. 1–17.

Matoušek, J. (1998). On the L2-discrepancy for anchored boxes. *Journal of Complexity*, 14(4), 527–556.

Meinl, T., Ostermann, C., Berthold, M.R. (2011). Maximum-score diversity selection for early drug discovery. *Journal of Chemical Information and Modeling*, 51(2), 237–247,

Morokoff, W.J., Caflisch, R.E.. (1994). Quasi-random sequences and their discrepancies. *SIAM Journal on Scientific Computing*, 15(6), 1251–1279.

Nadarajah, S., Jiang, X., Chu, J. (2015). A saddlepoint approximation to the distribution of the sum of independent non-identically beta random variables. *Statistica Neerlandica*, 69(2), 102–114.

Niederreiter, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods. CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics.

Pronzato, L., Müller, W.G. (2012). Design of computer experiments: space filling and beyond. *Statistics and Computing*, 22(3), 681–701.

Santner, T.J., Williams, B.J., Notz, W.I. (2003). *The Design and Analysis of Computer Experiments*, Springer.

Sukharev, A.G. (1971). Optimal strategies of the search for an extremum. *USSR Computational Mathematics and Mathematical Physics*, 11(4), 119–137.

Žilinskas, A. (2003). On the distribution of the distance between two points in a cube. *Random Operators and Stochastic Equations*, 11(1), 21–24.

G. Rudolph was awarded a *Diplom-Informatiker* degree in 1991 and a doctoral degree (*Dr. rer. nat.*) in computer science in 1996, both at the University of Dortmund. He is a professor of TU Dortmund University at the Department of Computer Science. His research interests include the development and theoretical analysis of bio-inspired methods applied to difficult optimization problems encountered in engineering sciences, logistics, and economics.

S. Wessing was awarded a *Diplom-Informatiker* degree in 2009 and a doctoral degree (*Dr. rer. nat.*) in computer science in 2015, both at the University of Dortmund. Currently he is a research associate at the Department of Computer Science, University of Dortmund. He is mainly interested in global optimization and decision making, with a focus on improving experimental research.

Taškų hiperkube tolygumo įvertinimas tiesiniu laiku

Günter RUDOLPH, Simon WESSING

Nagrinėjamas uždavinys, kaip įvertinti taškų pasiskirstymo hiperkube nuokrypį nuo tolygaus pasiskirstymo. Geras tolygumas yra svarbus Monte-Karlo metodų, skirtų integralams skaičiuoti, tikslumui, taip pat globalios optimizacijos algoritmų blogiausio atvejo paklaidoms įvertinti. Didelio matavimo skaičiaus erdvėje tenka generuoti dideles taškų aibes ir tenka naudoti greitus kvazi atsitiktinius metodus. Deja, tolygumo įvertinimo algoritmų sudėtingumas kvadratinis. Siūlome kelias tolygumo charakteristikas, kurios gali būti apskaičiuotos tiesinio sudėtingumo algoritmais. Nors šios charakteristikos nėra tiesioginis tolygumo matas, jos gerai indikuoja nuokrypius nuo tolygumo. Straipsnyje pateikiamos kritinės šių charakteristikų reikšmės apskaičiuotos, kur įmanoma, tiksliai arba apytiksliai.