Improving exploration strategies in large dimensions and rate of convergence of global random search algorithms

Jack Noonan[1] · Anatoly Zhigljavsky[1]

Abstract

We consider global optimization problems, where the feasible region $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$ with $d \geq 10$. For these problems, we demonstrate that the actual convergence of global random search algorithms is much slower than that given by the classical estimates, based on the asymptotic properties of random points, and that the usually recommended space exploration schemes are inefficient in the non-asymptotic regime. Moreover, we show that uniform sampling on entire $\mathcal{X}$ is much less efficient than uniform sampling on a suitable subset of $\mathcal{X}$, and that the effect of replacement of random points by low-discrepancy sequences can be felt in small dimensions only.

1 Introduction

Consider the general problem of continuous global minimization $f(x) \rightarrow \min_{x \in \mathcal{X}}$ with objective function $f(\cdot)$ and feasible region $\mathcal{X}$, which is assumed to be a compact subset of $\mathbb{R}^d$ with $\mathrm{vol}(\mathcal{X}) > 0$. In order to avoid unnecessary technical difficulties, we assume that $\mathcal{X}$ is convex. In all numerical examples, we use $\mathcal{X} = [0, 1]^d$.

Any global optimization algorithm combines two key strategies: exploration and exploitation. Performing exploration is equivalent to what we call "space-filling"; that is, choosing points which are well-spread in $\mathcal{X}$. Exploitation strategies use local information about f (and perhaps derivatives of f) and differ greatly for different types of global optimization algorithms. In this paper, we are only concerned with the exploration stage. Although many of our findings can be generalized to other space-filling schemes (where space-filling is not random and the space-filling strategy changes in the course of receiving more information about the objective function), in this paper we concentrate on simple exploration schemes like pure random search, where space-filling is performed by covering $\mathcal{X}$ with balls of given radius centered at the chosen points. Moreover, we assume that the points chosen at the

✉ Anatoly Zhigljavsky
ZhigljavskyAA@cardiff.ac.uk

[1] Cardiff University, Cardiff, UK

exploration stage are independent. That is, we associate the exploration stage with a global random search (GRS) algorithm producing a sequence of random points $x_1, x_2, \ldots, x_n$, where each point $x_j \in \mathcal{X}$ has some probability distribution P_j (we write this $x_j \sim P_j$) and $x_1, x_2, \ldots, x_n$ are independent. The value n is determined by a stopping rule. We assume that $1 \leq n_{\min} \leq n \leq n_{\max} < \infty$, where $n_{\min}$ and $n_{\max}$ are two given numbers. The number $n_{\max}$ determines the maximum number of function evaluations at the exploration stage and the fact that $n_{\max} < \infty$ determines what we call "the non-asymptotic regime". In the numerical study of Sect. 4, we also use Sobol's sequence, the most widely used low-discrepancy sequence.

We distinguish between 'small', 'medium' and 'high' dimensional problems depending on the following relations between d and $n_{\max}$:

(S) small dimensions: $n_{\min} \geq 2^d$, $n_{\max} \gg 2^d$ (hence, $\log_2 n_{\max} \gg d$);
(M) medium dimensions: $n_{\max}$ is comparable to 2^d: $c_1 d \leq \log_2 n_{\max} \leq c_2 d$ with suitable constants c_1 and c_2: $0 \ll c_1 \leq 1 \leq c_2 \ll \infty$;
(H) high dimensions: $n_{\max} \ll 2^d$.

Of course, there are in-between situations and the classification above depends on the cost of function evaluation. In case of non-expensive observations and $10^3 \leq n_{\max} \leq 10^6$, typical values of d in the three cases are: $(S): d \leq 10$; $(M): 10 \leq d \leq 20$; $(H): d > 20$. Values of $d \approx 10$ are border-line cases between (S) and (M) whereas $d \approx 20$ are border-line cases between (M) and (H).

In this study, we leave out the situation (S) of small dimensions and concentrate on situations (M) and (H). The reasons why we are not interested in the situation (S) of small dimensions are: (a) there are too many exploration schemes available in literature in the case of small dimensions, and (b) we are interested in the situations when the asymptotic regime is out of reach, and these are the situations (M) and (H).

In all considerations below we assume that the aim of the exploration stage is to reach a neighbourhood of an unknown point $x_* \in \mathcal{X}$ with high probability $\geq 1 - \gamma$ (with some $\gamma > 0$). We assume that x_* is uniformly distributed in $\mathcal{X}$ and by a neighbourhood of x_* we mean the ball $B = \mathcal{B}(x_*, \varepsilon)$ with suitable $\varepsilon > 0$. In other words, we will be interested in the problem of construction of partial coverings defined as follows.

Let $x_1, \ldots, x_n$ be some points in $\mathbb{R}^d$. Denote $X_n = \{x_1, \ldots, x_n\}$ and

$$B(X_n, r) = \bigcup_{i=1}^{n} B(x_i, r), \qquad (1.1)$$

where $r > 0$ is the radius of the balls $\mathcal{B}(x_i, r)$ and $B(x_i, r) = \mathcal{X} \cap \mathcal{B}(x_i, r)$. We will call $B(X_n, r)$ partial (or weak) covering of $\mathcal{X}$ of level $1 - \gamma$ if $\mathrm{vol}(B(X_n, r))/\mathrm{vol}(\mathcal{X}) \geq 1 - \gamma$.

If $\gamma = 0$ then $B(X_n, r)$ would make a full (strong) covering of $\mathcal{X}$. As demonstrated in [7–9], for any n and any given $\gamma > 0$, one can construct partial coverings of $\mathcal{X}$ with significantly smaller radii r than for the case $\gamma = 0$ (assuming that d is not too small). This is the main reason why we are not interested in strong coverings. The second reason is that numerically checking whether the set (1.1) makes a full covering (for a generic X_n) is extremely hard in situations (M) and (H) whereas simple Monte-Carlo gives very accurate estimates of γ for partial coverings, even for very high dimensions. For a short discussion concerning full covering and its role in optimization, see Sect. 2.1.

The main technique of construction of partial coverings will be generation of independent random points $x_1, \ldots, x_n$ in $\mathcal{X}$ with $x_j \sim P$, where P is a distribution concentrated either on the whole $\mathcal{X}$ or a subset of $\mathcal{X}$. It follows from Proposition 3.2.3 in [1] that using points outside $\mathcal{X}$ for construction of coverings is not beneficial when $\mathcal{X}$ is convex and hence we will always assume that $x_j \in \mathcal{X}$ for all j.

The following are the main messages of the paper.

1. Classical results on convergence rates of GRS algorithms are based on the asymptotic properties of random points uniformly distributed in $\mathcal{X}$; see Sect. 2. In the non-asymptotic regime, however, these results give estimates on the convergence rates which are far too optimistic. We show in Sect. 3 that for medium and high dimensions, the actual convergence rate of GRS algorithms is much slower.
2. The usually recommended sampling schemes (these schemes are based on the asymptotic properties of random points) are inefficient in the non-asymptotic regime. In particular, as shown in Sect. 4, uniform sampling on entire $\mathcal{X}$ is much less efficient than uniform sampling on a suitable subset of $\mathcal{X}$ (we will refer to this phenomena as the 'δ-effect').
3. In situations (M) and (H), the effect of replacement of random points by low-discrepancy sequences is negligible; see Sect. 4.2.

We also make certain practical recommendations concerning the best exploration schemes in the situations (M) and (H) in the case $\mathcal{X} = [0, 1]^d$. Our main recommendations will concern the situation (M) of medium dimensions, which we consider as the hardest for analysis. The situation (H) is simpler than (M) in the sense that the optimization problems in case (H) are so complicated that very simple space-filling schemes outlined in Sect. 6 provide relatively effective sampling schemes.

The structure of the paper is as follows. In Sect. 2, which contains no new results, we discuss the importance of covering and review classical results on convergence and rate of convergence of general GRS algorithms. The purpose Sect. 3 is to demonstrate that for medium and high dimensions the asymptotic regime is unachievable, and hence the actual convergence rate of GRS algorithms is much slower than the classical estimates of the rate of convergence indicate. In Sect. 4 we compare several exploration strategies and show that standard recommendations (such as: "use a low-discrepancy sequence") are inaccurate (for medium and high dimensions). In Sect. 5, we develop accurate approximations for the volume of intersection of a cube and a ball (with arbitrary centre and any radius). The approximations of Sect. 5 are used throughout numerical studies of Sects. 3 and 4. In Sect. 6 we summarize our findings and give recommendations on how to perform exploration of $\mathcal{X}$ in medium and large dimensions.

2 Importance of covering and classical results on convergence and rate of convergence of GRS algorithms

2.1 Covering radius

Consider $X_n = \{x_1, \ldots, x_n\}$, a set of n points in $\mathcal{X}$. The covering radius of $\mathcal{X}$ for X_n is $\mathrm{CR}(X_n) = \max_{x \in \mathcal{X}} \rho(x, X_n)$, where

$$\rho(x, X_n) = \min_{x_j \in X_n} \rho(x, x_j) \tag{2.1}$$

is the distance between a point $x \in \mathcal{X}$ and the point set X_n. Covering radius is also the smallest $r \geq 0$ such that the union of the balls with centers at $x_j \in X_n$ and radius r fully covers $\mathcal{X}$; that is, $\mathrm{CR}(X_n) = \min_{r>0}$ such that $\mathcal{X} \subseteq \mathcal{B}(X_n, r)$, where $\mathcal{B}(X_n, r) = \bigcup_{j=1}^n \mathcal{B}(x_j, r)$ and $\mathcal{B}(x, r) = \{z \in \mathbb{R}^d : \rho(x, z) \leq r\}$ is the ball of radius r and centre $x \in \mathbb{R}^d$. Optimal n-point covering is the point set X_n^* such that $\mathrm{CR}(X_n^*) = \min_{X_n} \mathrm{CR}(X_n)$. Most of the general considerations in the paper are valid for a general distance ρ, but all numerical studies

are conducted for the Euclidean distance only. We will thus assume that the distance ρ is Euclidean.

Other common names for the covering radius are: fill distance (in approximation theory; see [18, 23]), dispersion (in Quasi Monte Carlo; see [6, Ch. 6]), minimax-distance criterion (in computer experiments; see [16, 17]) and coverage threshold (in probability theory; see [11]).

Point sets with small covering radius are very desirable in theory and practice of global optimization and many branches of numerical mathematics. In particular, the celebrated results of A.G.Sukharev imply that any n-point optimal covering design X_n^* provides the following: (a) min-max n-point global optimization method in the set of all adaptive n-point optimization strategies, see [20] and [21, Ch.4,Th.2.1], (b) worst-case n-point multi-objective global optimization method in the set of all adaptive n-point algorithms, see [29], and (c) the n-point min-max optimal quadrature, see [21, Ch.3,Th.1.1]. In all three cases, the class of (objective) functions is the class of Liptshitz functions, and the optimality of the design is independent of the value of the Liptshitz constant. Sukharev's results on n-point min-max optimal quadrature formulas have been generalized in [10] for functional classes different from the class of Liptshitz functions; see also formula (2.3) in [2].

2.2 Convergence of a general GRS algorithm

Consider the general problem of continuous global minimization $f(x) \to \min_{x \in \mathcal{X}}$. Assume that $f_* = \inf_{x \in \mathcal{X}} f(x) > -\infty$ and $f(\cdot)$ is continuous at all points $x \in W(\delta)$ for some $\delta > 0$, where $W(\delta) = \{x \in \mathcal{X} : f(x) - f_* \leq \delta\}$. That is, we assume that $f(\cdot)$ is continuous in the neighbourhood of the set $\mathcal{X}_* = \{x_* \in \mathcal{X} : f(x_*) = f_*\}$ of global minimizers of $f(\cdot)$, which is non-empty but may contain more than one point x_*. To avoid technical difficulties, we assume that there are only a finite number of global minimizers of $f(\cdot)$; that is, the set $\mathcal{X}_*$ is finite.

Consider a general GRS algorithm producing a sequence of random points $x_1, x_2, \ldots$, where each point $x_j \in \mathcal{X}$ has some probability distribution P_j (we write this $x_j \sim P_j$), where for $j > 1$ the distributions P_j may depend on the previous points $x_1, \ldots, x_{j-1}$ and on the results of the objective function evaluations at these points (the function evaluations may not be noise-free). We say that this algorithm converges if for any $\delta > 0$, the sequence of points x_j arrives at the set $W(\delta) = \{x \in \mathcal{X} : f(x) - f_* \leq \delta\}$ with probability one. If the objective function is evaluated without error then this obviously implies convergence (as $n \to \infty$) of record values $f_{o,j} = \min_{i=1\ldots j} f(x_i)$ to f_* with probability 1.

In view of continuity of $f(\cdot)$ in the neighbourhood of $\mathcal{X}_*$, the event of arrival of sequence of points x_j at the set $W(\delta)$ with given $\delta > 0$, is equivalent to the arrival of this sequence at the set $B_*(\varepsilon) = \cup_{x_* \in \mathcal{X}_*} B(x_*, \varepsilon)$ for some $\varepsilon > 0$ depending on δ.

Conditions on the distributions P_j ($j = 1, 2, \ldots$) ensuring convergence of the GRS algorithms are well understood; see, for example, [15, 19, 28] and [25, Sect. 3.2]. Such results are consequences of the classical in probability theory 'zero–one law' or Borel-Cantelli lemmas (see e.g. [4, Section 7.3]) and provide sufficient conditions on convergence. We follow [27, Theorem 2.1] to provide the most general sufficient conditions for convergence of GRS algorithms.

Theorem 1 *Consider a GRS algorithm with $x_j \sim P_j$ and let $B \subset \mathcal{X}$ be a Borel subset of $\mathcal{X}$. Assume that*

$$\sum_{j=1}^{\infty} q_j(B) = \infty, \tag{2.2}$$

where $q_j(B) = \inf P_j(B)$ and the infimum is taken over all locations of previous points x_i ($i = 1, \ldots, j - 1$) and corresponding results of evaluations of $f(\cdot)$. Then the sequence of points $\{x_1, x_2, \ldots\}$ falls infinitely often into the set B, with probability 1.

Note that Theorem 1 does not make any assumptions about observations of $f(\cdot)$ and hence is valid for the very general case where evaluations of the objective function $f(\cdot)$ are noisy and the noise is not necessarily random.

Consider the following three particular cases.

(a) If in (2.2) we use $B = B_*(\varepsilon)$ or $B = W(\varepsilon)$ with some $\varepsilon > 0$, then Theorem 1 gives a sufficient condition that the corresponding GRS algorithm converges; that is, there exists a subsequence $\{x_{i_j}\}$ of the sequence $\{x_j\}$ which converges (with probability 1) to the set $\mathcal{X}_*$ in the sense that the distance between x_{i_j} and $\mathcal{X}_*$ tends to 0 as $j \to \infty$. For this subsequence $\{x_{i_j}\}$, we have $f(x_{i_j}) \to f_*$ as $j \to \infty$.

If the evaluations of $f(\cdot)$ are noise-free, then we can use the sequence of record points (that is, the points where the records $f_{o,j} = \min_{\ell < j} f(x_\ell)$ are attained) as $\{x_{i_j}\}$; in this case, $f(x_{i_j}) = f_{o,j}$ is the sequence of records converging to f_* with probability 1. By the dominated convergence theorem (see e.g. [4, Section 7.2]), convergence of the sequence of records $f_{o,j}$ to f_* with probability 1 implies other important types of convergence of $f_{o,j}$ to f_* — in mean and mean square: $Ef_{o,j} \to f_*$ and $E(f_{o,j} - f_*)^2 \to 0$ as $j \to \infty$.

(b) If (2.2) holds for $B = B(x, \varepsilon)$ with any $x \in \mathcal{X}$ and any $\varepsilon > 0$, then Theorem 1 gives a sufficient condition that the sequence of points $\{x_1, x_2, \ldots\}$ is dense with probability 1. As this is a stronger sufficient condition than in (a), all conclusions of (a) are valid.

(c) If we use pure random search (PRS) with $P = P_U$, the uniform distribution on $\mathcal{X}$ (that is, $P_j = P_U$ for all j and the points $x_1, x_2, \ldots$ are independent), then the assumption that $\mathcal{X}$ is convex implies $\Pr(B(x, \varepsilon)) \geq \text{const}_\varepsilon > 0$ for all $x \in \mathcal{X}$ any $\varepsilon > 0$ and therefore the condition (2.2) trivially holds for any $B = B(x, \varepsilon)$, as in (b) above. In practice, the usual choice of the distribution P_j is

$$P_j = \alpha_j P_U + (1 - \alpha_j) Q_j, \tag{2.3}$$

where $0 \leq \alpha_j \leq 1$ and Q_j is a specific probability measure on $\mathcal{X}$ which may depend on previous evaluations of the objective function. Sampling from the distribution (2.3) corresponds to taking a uniformly distributed random point in $\mathcal{X}$ with probability α_j and sampling from Q_j with probability $1 - \alpha_j$. In case of distributions (2.3), the condition $\sum_{j=1}^{\infty} \alpha_j = \infty$ yields the fulfilment of (2.2) for all $B = B(x, \varepsilon)$ and therefore the GRS algorithm with such P_j is theoretically converging.

2.3 Rate of convergence

Consider first a PRS algorithm, where x_j are i.i.d. with distribution P. Let $\varepsilon, \delta > 0$ be fixed and B be the target set we want to hit by points $x_1, x_2, \ldots$. For example, we set $B = W(\delta) = \{x \in \mathcal{X} : f(x) - f_* \leq \delta\}$ in the case when the accuracy is expressed in terms of closeness with respect to the function value, $B = B(x_*, \varepsilon)$ if we are studying

convergence towards a particular global minimizer x_*, and $B = B_*(\varepsilon)$ if the aim is to approach a neighbourhood of $\mathcal{X}_*$.

Assume that P is such that $P(B) > 0$. In particular, if $P = P_U$ is the uniform probability measure on $\mathcal{X}$, then, as $\mathcal{X}$ has Lipschitz boundary, we have $P(B) = \mathrm{vol}(B)/\mathrm{vol}(\mathcal{X}) > 0$. Note that in all interesting instances the value $P(B)$ is positive but small, and this will be assumed below.

Define the Bernoulli trials where the success in the trial j means $x_j \in B$. PRS generates a sequence of independent Bernoulli trials with the same success probability $\mathrm{Pr}\{x_j \in B\} = P(B)$. In view of independence of $x_1, x_2, \ldots$, we have

$$\mathrm{Pr}\{x_1 \notin B, \ldots, x_n \notin B\} = (1 - P(B))^n$$

and therefore the probability

$$\mathrm{Pr}\{x_j \in B \text{ for at least one } j, \ 1 \leq j \leq n\} = 1 - (1 - P(B))^n$$

tends to one as $n \to \infty$.

Let n_γ be the number of points which are required for PRS to reach the set B with probability at least $1 - \gamma$, where $\gamma \in (0, 1)$; that is,

$$n_\gamma = \min\{n : \ 1 - (1 - P(B))^n \geq 1 - \gamma\}.$$

Solving the equation $1 - (1 - P(B))^n \geq 1 - \gamma$ with respect to n, we obtain

$$n_\gamma = \left\lceil \ln \gamma / \ln (1 - P(B)) \right\rceil \cong (-\ln \gamma)/P(B) \qquad (2.4)$$

as $P(B)$ is small and $\ln (1 - P(B)) \cong -P(B)$ for small $P(B)$.

The numerator $-\ln \gamma$ in the expression (2.4) for n_γ depends on γ but it is not large; for example, $-\ln \gamma \simeq 4.605$ for $\gamma = 0.01$. However, the denominator $P(B)$ (depending on ε, d and the shape of $\mathcal{X}$) can be very small.

Assuming that $B = B(x_*, \varepsilon)$, where the norm is standard Euclidean, and B is fully inside $\mathcal{X}$, we have

$$\mathrm{vol}(B(x_*, \varepsilon)) = \mathrm{vol}(\mathcal{B}(x_*, \varepsilon)) = V_d \, \varepsilon^d, \qquad (2.5)$$

where $V_d = \pi^{d/2}/[\Gamma(d/2 + 1)]$ is the volume of the unit Euclidean ball $\mathcal{B}(0, 1)$ and $\Gamma(\cdot)$ is the gamma-function. The resulting version of the expression (2.4) for n_γ in the case $B = B(x_*, \varepsilon)$ and $\mathrm{vol}(\mathcal{X}) = 1$ becomes

$$n_\gamma^{\mathrm{as}} = -\ln \gamma / \left(\varepsilon^d V_d \right). \qquad (2.6)$$

As $\varepsilon \to 0$, the ball $B = B(x_*, \varepsilon)$ lies fully inside $\mathcal{X}$ for P_U-almost all x_*. Indeed, asymptotically, as $n \to \infty$, the covering radius computed for uniformly distributed random points x_j, tends to 0 and hence the equality (2.5) is valid asymptotically for almost all x_*. This is the reason for superscript 'as' in (2.6). As shown below in Sect. 3, in the non-asymptotic regime in situations (M) and (H), the volume $\mathrm{vol}(B(x_*, \varepsilon))$ is necessarily smaller than given by (2.5) and therefore the true n_γ is (much) larger than n_γ^{as} in (2.6).

Consider now general GRS algorithms where the probabilities P_j are chosen in the form (2.3), where the coefficients α_j satisfy the condition (2.2). Instead of the equality $\mathrm{Pr}\{x_j \in B\} = P(B)$ for all $j \geq 1$, we now have the inequality $\mathrm{Pr}\{x_j \in B\} \geq \alpha_j P_U(B)$, where the equality holds in the worst-case scenario. We define $n(\gamma)$ as the smallest integer such that the inequality $\sum_{j=1}^{n(\gamma)} \alpha_j \geq -\ln \gamma / P_U(B)$ is satisfied. For the choice $\alpha_j = 1/j$, which

is a common recommendation, we can use the approximation $\sum_{j=1}^{n} \alpha_j \simeq \ln n$. Therefore we obtain $n(\gamma) \simeq \exp\{-\ln \gamma / P_U(B)\}$. For the case of $\mathcal{X} = [0, 1]^d$ and $B = B(x_*, \varepsilon)$, we obtain $n(\gamma) \simeq \exp\{c \cdot \varepsilon^{-d}\}$, where $c = (-\ln \gamma)/V_d$.

Note also that if the distance between x_* and the boundary of $\mathcal{X}$ is smaller than ε, then the constant c and hence $n(\gamma)$ are even larger. For example, for $\gamma = 0.1$, $d = 10$ and $\varepsilon = 0.1$, $n(\gamma)$ is larger than $10^{1000000000}$. Even for optimization problems in a small dimension $d = 3$, and for $\gamma = 0.1$ and $\varepsilon = 0.1$, the number $n(\gamma)$ of points required for the GRS algorithm to hit the set B in the worst-case scenario is huge: $n(\gamma) \simeq 10^{238}$.

3 Points uniformly distributed on $\mathcal{X}$

3.1 Asymptotic case

In this section, the point set $X_n = \{x_1, \ldots, x_n\}$ consists of the first n points of a sequence $X_\infty = \{x_1, x_2, \ldots\}$ of independent uniformly distributed random vectors in $\mathcal{X}$. Assume, without loss of generality, that $\mathrm{vol}(\mathcal{X}) = 1$.

Consider the random variable $\rho(U, X_n)$, the distance between U (the uniform random point in $\mathcal{X}$) and X_n; see (2.1) for the definition of ρ. The cdf (cumulative distribution function) of $\rho(U, X_n)$ gives the average proportions of $\mathcal{X}$ which are covered by the balls centered at X_n with radius r. That is,

$$F_d(r, X_n) := \Pr(\rho(U, X_n) \leq r) = \mathbb{E}_{X_n} \mathrm{vol}(B(X_n, r)), \tag{3.1}$$

where the set $B(X_n, r)$ is defined in (1.1). In asymptotic considerations, we need to suitably normalize the radius (which tends to zero as $n \to \infty$) in (3.1). We thus consider the following sequence of cdf's:

$$F_n(t) := \Pr(n^{1/d} V_d^{1/d} \rho(U, X_n) \leq t) = F_d\left([nV_d]^{-1/d} t, X_n\right). \tag{3.2}$$

Lemma 1

$$F_n(t) \to F(t) := 1 - \exp(-t^d) \quad as \ n \to \infty, \tag{3.3}$$

where the convergence is uniform in t and cdf's F_n are defined in (3.2).

The statement of Lemma 1 follows from Zador's arguments in his fundamental paper [24]; see the beginning of page 142. The key observation of Zador is that asymptotically, as $n \to \infty$, the covering radius computed for uniformly distributed random points x_j, tends to 0 and hence the equality (2.5) is valid asymptotically for almost all U; this is formula (19) in [24]. The statement of Lemma 1 is in fact a particular case of Theorem 9.1 in [3], if Q is chosen as the uniform distribution on $\mathcal{X}$.

In what follows, we will need the $(1 - \gamma)$-quantile $(0 < \gamma < 1)$ of the cdf F in the rhs of (3.3). This $(1 - \gamma)$-quantile is determined as $t_{1-\gamma} = [-\log(\gamma)]^{1/d}$, for which we have $F(t_{1-\gamma}) = 1 - \gamma$. The quantity $t_{1-\gamma}$ can be interpreted as the normalised asymptotic radius required for covering a subset of $\mathcal{X}$ of volume $(1 - \gamma)$ (the partial covering introduced in Sect. 1). For very small ε, to cover a subset of $\mathcal{X}$ with random centers $x_j \in X_n$ of volume which is approximately $1 - \gamma$, $n = n_\gamma$ should satisfy

$$n_\gamma = \frac{t_{1-\gamma}^d}{\varepsilon^d V_d} = \frac{-\ln(\gamma)}{\varepsilon^d V_d}, \tag{3.4}$$

which coincides with (2.6). The above result can be reformulated in terms of the asymptotic radius r as follows: for very large n the union of n balls with random centers $x_j \in X_n$ and radius

$$r_{n,1-\gamma} = n^{-1/d} V_d^{-1/d} t_{1-\gamma} = n^{-1/d} V_d^{-1/d} [-\log(\gamma)]^{1/d} \tag{3.5}$$

covers a subset of $\mathcal{X}$ of volume which is approximately $1 - \gamma$.

In the non-asymptotic (finite n) regime, the distribution function $F_d(r, X_n)$ of (3.1) can be obtained in the following way (below, for $X_n = \{x_1, \ldots, x_n\}$, the components $x_1, x_2, \ldots x_n$ are not necessarily uniform but are i.i.d.).

Conditionally on U, we have for fixed $U \in \mathcal{X}$:

$$\mathbb{P}\{U \in \mathcal{B}_d(X_n, r)\} = 1 - \prod_{j=1}^{n} \mathbb{P}\{U \notin \mathcal{B}_d(x_j, r)\}$$

$$= 1 - \prod_{j=1}^{n} \left(1 - \mathbb{P}\{U \in \mathcal{B}_d(x_j, r)\}\right)$$

$$= 1 - \left(1 - \mathbb{P}_X\{\|U - X\| \le r\}\right)^n, \tag{3.6}$$

where X has the same distribution as x_1. From (3.6), the distribution function $F_d(r, X_n)$ can be obtained by averaging over the distribution of U:

$$F_d(r, X_n) = \mathbb{E}_U \mathbb{P}\{U \in \mathcal{B}_d(X_n, r)\} . \tag{3.7}$$

For large n and small r we use an approximate equality $\mathbb{P}_X\{\|U - X\| \le r\} \simeq r^d V_d$ in (3.6). By doing so, averaging with respect to U is redundant and we arrive at the results of Sect. 2.3. If n is not so large, the quantity $\mathbb{P}_X\{\|U - X\| \le r\}$ has to be approximated by other means. This will be discussed in Sect. 5.

3.2 Bounds for $F_d(r, X_n)$

Evaluating the expectation in (3.7) is difficult but simple bounds can be obtained by applying Jensen's inequality. Here we will focus attention to the case of $\mathcal{X} = [0, 1]^d$ and $X_n = \{x_1, \ldots, x_n\}$, where $x_1, x_2, \ldots$ is a sequence of uniformly distributed random vectors on $\mathcal{X}$. From (3.7), we have

$$\mathbb{E}_U \mathbb{P}\{U \in \mathcal{B}_d(X_n, r)\} = 1 - \mathbb{E}_U\left[(1 - \mathbb{P}_X\{\|U - X\| \le r\})^n\right].$$

An immediate use of Jensen's inequality yields the bound:

$$\mathbb{E}_U \mathbb{P}\{U \in \mathcal{B}_d(X_n, r)\} \le 1 - (1 - \mathbb{P}_X\{\|\mathbf{1/2} - X\| \le r\})^n . \tag{3.8}$$

Here and below $\mathbf{a} = (a, a, \ldots, a) \in \mathbb{R}^d$ for any a. However, noticing the fact $\mathbb{P}_X\{\|U - X\| \le r\} = \mathbb{P}_Z\{\|Z - X\| \le r\}$ where Z in a uniform random vector on $[1/2, 1]^d$, we can apply Jensen's inequality to obtain:

$$\mathbb{E}_U \mathbb{P}\{U \in \mathcal{B}_d(X_n, r)\} \le 1 - (1 - \mathbb{P}_X\{\|\mathbf{3/4} - X\| \le r\})^n . \tag{3.9}$$

The forms of the bounds in (3.8) and (3.9) suggest an approximation of the following form may be useful:

$$\mathbb{E}_U \mathbb{P}\{U \in \mathcal{B}_d(X_n, r)\} \simeq 1 - \left(1 - \mathbb{P}_{U,X}\{\|U - X\| \le r\}\right)^n . \tag{3.10}$$

Here, instead of fixing U to **1/2** or **3/4**, it is a uniform random vector on $[0, 1]^d$. The probability $\mathbb{P}_{U,X} \{\|U - X\| \leq r\}$ has the interpretation of being the average intersection a ball of radius r with a random center at U has with the cube $[0, 1]^d$. For different d and r, the distribution of $\mathbb{P}_X \{\|U - X\| \leq r\}$ normalised by the volume of the ball $r^d V_d$ is shown in Figs. 9–10.

3.3 Numerical studies

In this section, we will demonstrate one of the key messages of the paper saying that in high dimensions, the asymptotic results are not attainable for reasonable values of n and consequently produce poor approximations for n not astronomically large.

In Fig. 1, we plot $F_d(r, X_n)$ as a function of d for $n = 1000$ (using blue plusses) and $n = 10000$ (using black circles). For each value of d, the radius r is chosen based on the asymptotic result given in (3.5) with $1 - \gamma = 0.9$; this is shown by the solid red line at 0.9. We see that very quickly and for n that would be deemed large, $F_d(r, X_n)$ is significantly smaller than 0.9 and quickly tends to zero in d.

The big difference between the asymptotic and finite regime is further illustrated in Figs. 2–8. In these figures, using a solid black line we depict $F_d(r, X_n)$ as a function of r for different values of d and n that are provided in the caption of each figure. In these figures, the dashed red line is the approximation obtained from the asymptotic result (3.3), that is, the approximation $F_d(r, X_n) \approx F(n^{1/d} V_d^{1/d} r)$. In Figs. 3–8, we also include two Jensen's bounds given in (3.8) (dot dashed orange) and (3.9) (dotted blue), as well as the approximation given in (3.10) (longer dashed green). From these figures, we can make the following observations.

1. Unless d is small, the asymptotic results produce poor approximations even if n is reasonably large.
2. The approximation in (3.10) is rather accurate but worsens for smaller γ.
3. For $r \leq 1/2$, the asymptotic bounds and (3.8) coincide; this follows from the equality $\mathbb{P}_X \{\|\mathbf{1/2} - X\| \leq r\} = r^d V_d$ for $r \leq 1/2$.
4. The refined Jensen's bound given in (3.9) is superior to (3.8) and especially to the asymptotic bound. This becomes particularly evident in higher dimensions; see Figs. 7 and 8.

In Figs. 3-6, the crosses on the dashed red line and solid black line mark points of interest. In Fig. 3, for $r = 0.5$ we obtain $F(n^{1/d} V_d^{1/d} r) = 0.91$ but the true value of $F_d(0.5, X_n)$ is closer to 0.41. As n increases from 1000 to 10000 as is shown in Fig. 4, for $r = 0.4$ we have $F(n^{1/d} V_d^{1/d} r) = 0.925$ and $F_d(0.4, X_n)$ is closer to 0.6. (Recall that in view of (3.3), we should have $F(n^{1/d} V_d^{1/d} r) \simeq F_d(r, X_n)$) for all r and n large enough). The respective triples $(r; F(n^{1/d} V_d^{1/d} r), F_d(r, X_n))$ for Figs. 4–6 are (0.9; 0.935, 0.08) and (0.8; 0.95, 0.13). For the case of $d = 50$ and shown in Figs. 7–8, the asymptotic properties are so far from being achieved with $n = 1000$ and $n = 10000$ that such a comparison does not even make sense.

In Figs. 9 and 10 we use $d = 10$, $d = 20$ and the values of r corresponding to the crosses in Figs. 3 and 5. In these figures, we depict the distribution of intersection a random point has with the cube normalised by the volume of the ball $r^d V_d$; that is, we plot the density of the r.v. $\kappa_U = \mathbb{P}_X \{\|U - X\| \leq r\} / (r^d V_d)$, where both U and X have uniform distribution on $[0, 1]^d$. The importance of these two figures is another illustration of inadequacy of the key assumption behind (2.6), which can be formulated as the assumption that the distribution of density of the r.v. κ_U is very close to the delta-measure concentrated at one. This assumption is indeed reasonably adequate if r can be chosen small enough. However, as Fig. 9 and especially Fig. 10 illustrate, even for relatively large values of n the required values of r
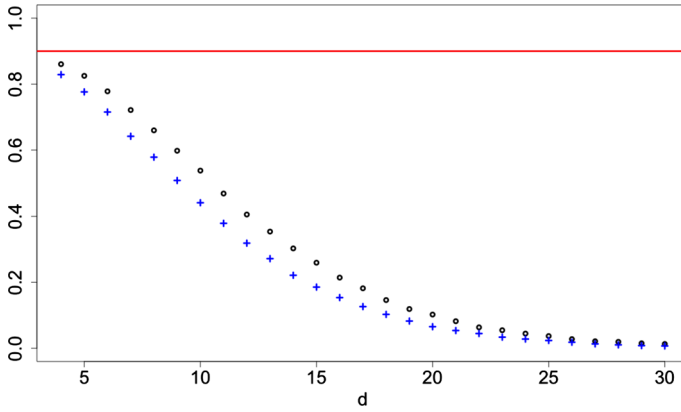
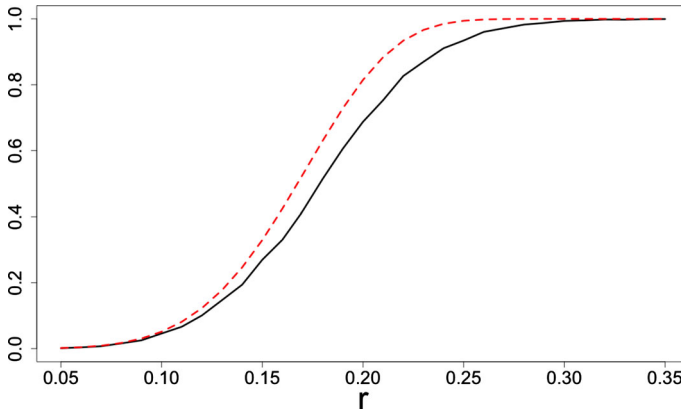**Fig. 1** Covering proportions using the asymptotic radius; $n = 1000, \ 10000$

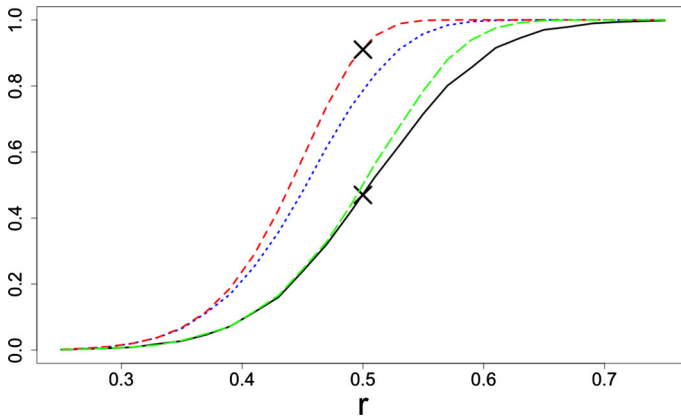**Fig. 2** $F_d(r, X_n)$ and $F(n^{1/d} V_d^{1/d} r)$ as functions of r; $d = 5$ and $n = 1000$
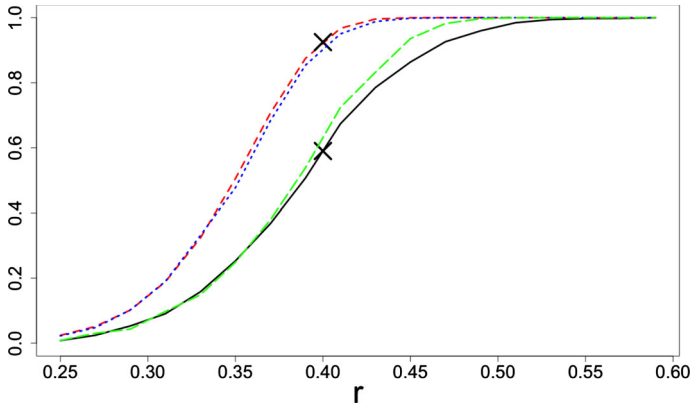

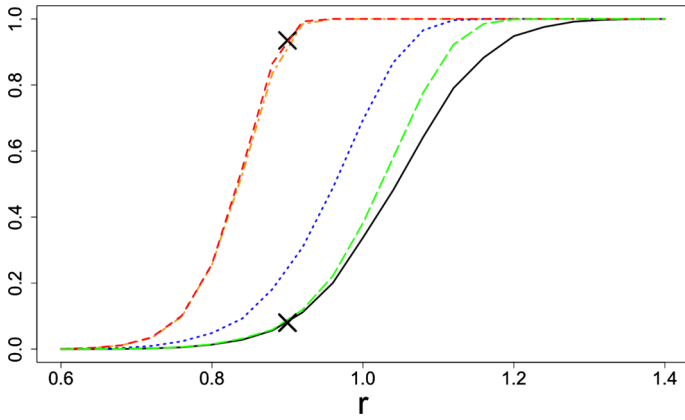**Fig. 3** $d = 10, \ n = 1000$

Fig. 4 $d = 10, \ n = 10000$
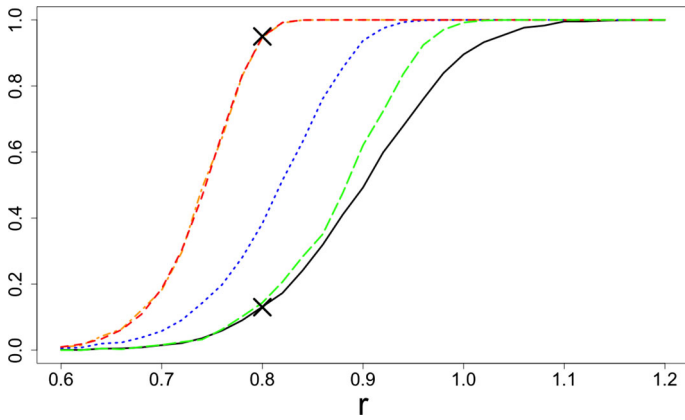

Fig. 5 $d = 20, \ n = 1000$
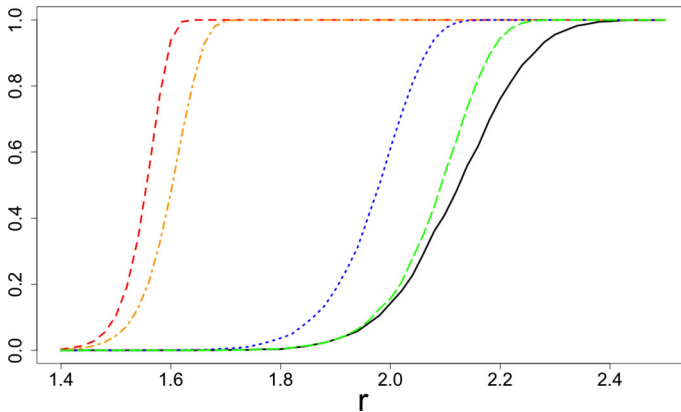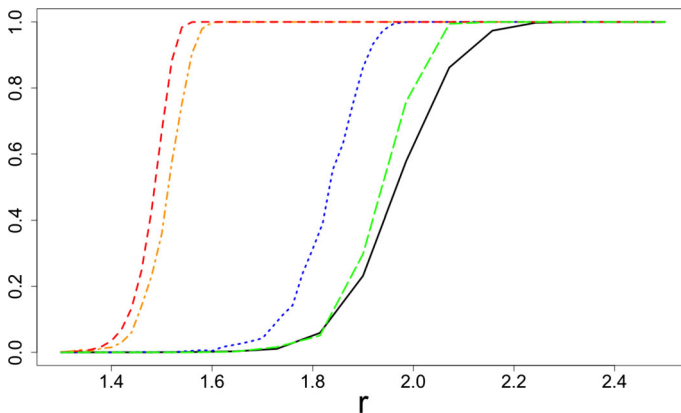

Fig. 6 $d = 20, n = 10000$

Fig. 7 $n = 1000, d = 50$

Fig. 8 $n = 10,000, d = 50$

are not small enough for this to hold even approximately. Note that in the derivation of the asymptotic values of $n_\gamma = n_\gamma^{as}$ in (2.6) we use the value 1 rather than the random variables κ_U with the densities shown in Figs. 9,10.

4 Modification of sampling schemes and non-uniform distribution of the target

In Sect. 3, we have used the principal sampling scheme where points x_j in $X_n = \{x_1, \ldots, x_n\}$ are i.i.d. uniform on $\mathcal{X} = [0, 1]^d$. In Sect. 4.1 we study a modification of this scheme where $x_j \in X_n$ are i.i.d. uniform random points in a smaller δ-cube $C_\delta = [1/2 - \delta/2, 1/2 + \delta/2]^d$ with $0 < \delta < 1$. In Sect. 4.2 we investigate the effect of replacing random points by points from a low-discrepancy sequence. The choice of a specific low-discrepancy sequence has very little impact on the results and we present the results for Sobol sequence only. In Sect. 4.3 we will investigate the effect of replacement of the uniform distribution of the target $x_* \in [0, 1]^d$ by a bowl-shaped distribution such as the product of arcsine distributions on $[0, 1]$.
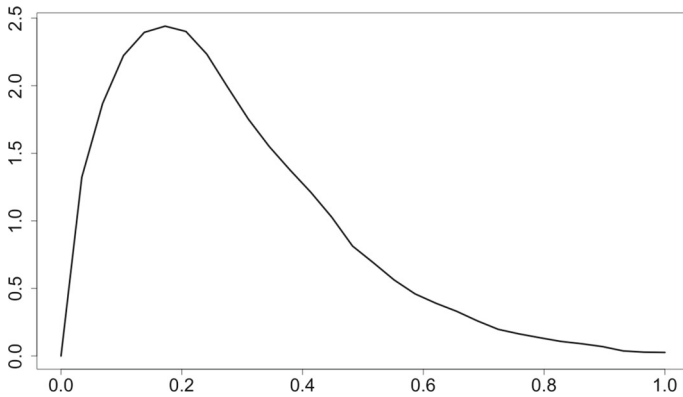
Fig. 9 Density of r.v. κ_U; $d = 10$, $r = 0.5$
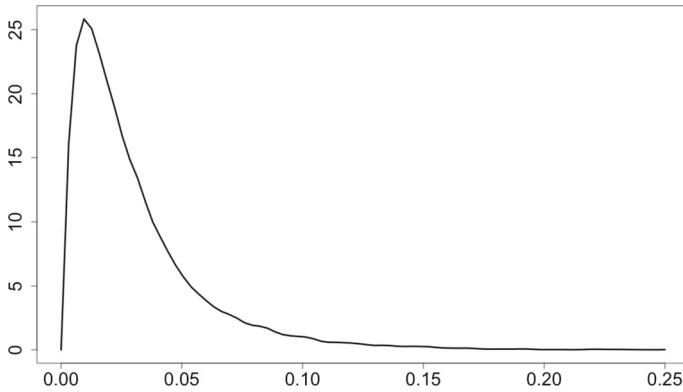
Fig. 10 Density of r.v. κ_U; $d = 20$, $r = 0.9$

4.1 Points x_j are i.i.d. uniformly distributed on C_δ

In this section we demonstrate the δ-effect, which manifests that in high dimensions sampling in a cube C_δ with suitable $0 < \delta < 1$ leads to a much more efficient covering scheme than sampling within the whole cube $[0, 1]^d$. Note that the δ-effect is not obvious being completely unknown in the literature on stochastic global optimization and perhaps in literature on global optimization in general. All existing literature recommends space-filling in the whole set X and not in its subset. Moreover, there are recommendations in literature (see, for example, [5, 22]) of choosing more points closer to the boundary of the cube rather than purely uniformly in order to improve space-filling properties of random points.

In Figs. 11–13, for different values of d and n we plot $F_d(r, X_n)$ as a function of δ. For each d and n, the value of r has been chosen such that $\max_{0 \le \delta \le 1} F_d(r, X_n) = 0.9$; these values of r (along with optimal values of δ, in brackets) can be obtained from Table 1. In these figures, the values of $F_d(r, X_n)$ for $n = 1000, 10000, 100000$ are shown with a solid black line, dashed blue line and dotted green line respectively. These figures demonstrate the 'δ-effect' formulated as the second main message in Introduction. These figures also clearly demonstrate that despite sampling uniformly in the cube $[0, 1]^d$ is asymptotically optimal, for large d it is always a poor strategy, which can be substantially improved.
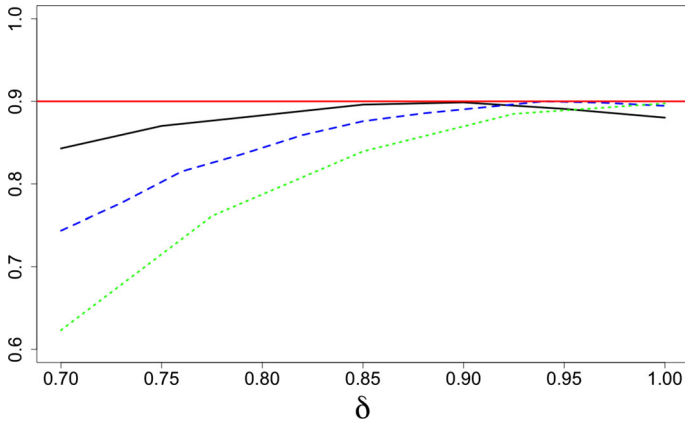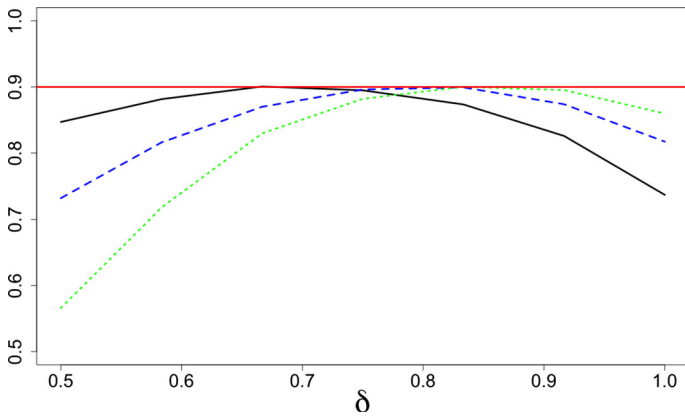
Fig. 11 $d = 10$; $n = 1000, 10, 000, 100, 000$

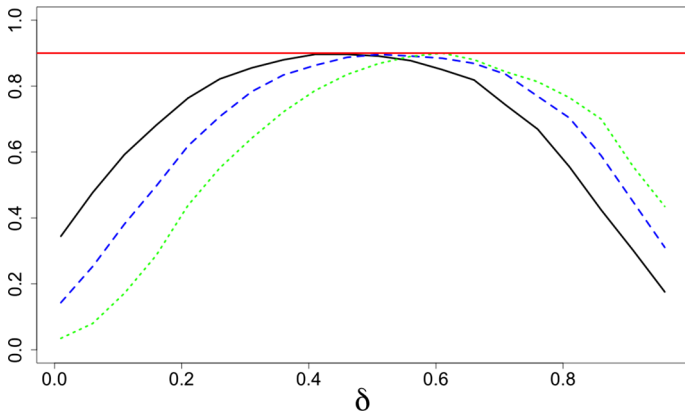**Fig. 12** $d = 20$; $n = 1000, 10, 000, 100, 000$

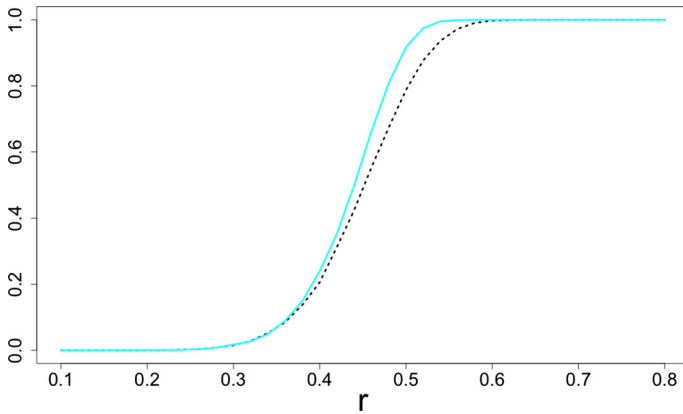**Fig. 13** $d = 50$; $n = 1000, 10, 000, 100, 000$

Fig. 14 $d = 10, n = 1000$: Jensen's bound with $\delta = 1$ and $\delta = 0.5$

Table 1 Values for $r_{n,1-\gamma}$ with $\gamma = 0.1$

	$n = 100$		$n = 1,000$		$n = 10,000$		$n = 100,000$	
	$[0,1]^d$	δ-cube	$[0,1]^d$	δ-cube	$[0,1]^d$	δ-cube	$[0,1]^d$	δ-cube
$d = 5$	0.41	0.40 (0.9)	0.24	0.24 (1.0)	0.24	0.29 (1.0)	0.09	0.089 (1.0)
$d = 10$	0.81	0.78 (0.7)	0.61	0.60 (0.9)	0.46	0.46 (1.0)	0.36	0.360 (1.0)
$d = 15$	1.13	1.04 (0.6)	0.91	0.88 (0.8)	0.76	0.74 (0.9)	0.62	0.619 (0.9)
$d = 20$	1.38	1.25 (0.5)	1.17	1.11 (0.7)	1.01	0.97 (0.8)	0.87	0.855 (0.9)
$d = 25$	1.60	1.42 (0.5)	1.39	1.30 (0.6)	1.23	1.18 (0.8)	1.09	1.060 (0.8)
$d = 50$	2.46	2.07 (0.4)	2.26	1.98 (0.5)	2.10	1.90 (0.5)	1.96	1.790 (0.6)

The discussion of Jensen's bounds given in Sect. 3.2 still apply to the case of X_n sampled uniformly within δ-cube C_δ. The only adjustment that needs to be made to the results of Sect. 3.2 is to let X be a uniform random vector in C_δ and not $[0, 1]^d$. In Fig. 14, we depict the Jensen's lower bound given in (3.9) for X uniform in $[0, 1]^d$ and for X uniform in the δ cube $[1/4, 3/4]^d$ (so that $\delta = 0.5$). We see that the lower bound for X_n sampled within the δ-cube is larger than X_n sampled from the whole cube. This further supports the conclusion that for n not astronomically large, the 'δ-effect' should always be considered.

In Table 1, for X_n chosen uniformly in the cube $[0, 1]^d$ and X_n chosen uniformly in the δ-cube, we tabulate the values of $r_{n,1-\gamma}$ with $\gamma = 0.1$ for different d and n. In the columns labeled δ-cube, the values in the brackets correspond to the approximately optimal values of δ. We can see that for small d, the δ-effect is very small (since n is relatively large in these dimensions). For larger dimensions, the δ-effect is very prominent.

In Tables 2–3, we consider an equivalent reformulation of the results of Table 1. In these tables, for a given r we specify the value of n_γ, with $\gamma = 0.1$, for X_n chosen uniformly in the cube $[0, 1]^d$ and X_n chosen uniformly in the δ-cube. We also include the approximation based on the the asymptotic arguments leading to (2.6). We see that in high dimensions, the requirement of r being small enough for (2.6) to provide sensible approximations requires n to be extremely large. Such large values of n are impractical.

Table 2 Values of n_γ: $d = 20$, $\gamma = 0.1$

r	0.9	0.95	1	1.05	1.1	1.15
n_γ with $\delta = 1$	54,000	25,000	10,800	4,600	2,700	1,300
n_γ with $\delta = \delta^*$	40,000 (0.8)	15,000 (0.8)	6,300 (0.8)	2,700 (0.7)	1,100 (0.7)	500 (0.7)
n_γ from (2.6)	734	249	89	34	13	5

Table 3 Values of n_γ: $d = 50$, $\gamma = 0.1$

r	2	2.05	2.1	2.15	2.2	2.25	2.3
n_γ with $\delta = 1$	50,000	21,000	10,000	5,000	2,200	1,200	600
n_γ with $\delta = \delta^*$	700 (0.4)	200 (0.4)	50 (0.3)	12 (0.2)	2 (0.1)	NA	NA
n_γ from (2.6)	0	0	0	0	0	0	0

4.2 Points x_j are taken from a low-discrepancy sequence

Figures 15–16 are extended versions of Fig. 1. Here we plot $F_d(r, X_n)$ as a function of d, where the radius is fixed from (3.5) with $\gamma = 0.1$ (the line $1 - \gamma$ is depicted by a red solid line). For X_n chosen uniformly in the cube $[0, 1]^d$, we depict $F_d(r, X_n)$ with blue plusses. For X_n chosen from a Sobol sequence in the whole cube $[0, 1]^d$, we use orange triangles. When the points in X_n are uniform i.i.d. within the δ-cube with optimal δ we use green crosses. Finally, when points in X_n are chosen from a Sobol sequence within the same δ-cube we use purple diamond. Figures 15 and 16 illustrate two new key messages along with the message discussed in Fig. 1. Firstly, the use of low-discrepancy sequences seem to produce slightly better results in comparison to random choice of points for small dimensions but in higher dimensions the use of low-discrepancy sequences (in our case, Sobol sequences) produces results that are almost equivalent to random sampling uniformly either in $[0, 1]^d$ or in the optimally chosen δ-cube. Secondly, in large dimensions sampling from a suitable δ-cube greatly outperforms the other schemes considered here being still far from the asymptotic results. These messages are further supported in Figs. 17 and 18. Here we plot the asymptotic approximation $F_n(r)$ from Lemma 1 (dashed red) and $F_d(r, X_n)$ as a function of r for the following choices of X_n: random in the cube $[0, 1]^d$ (blue line with plusses), chosen from a Sobol sequence in $[0, 1]^d$ (orange line with triangles), random in the δ-cube with optimal delta (green line with crosses), chosen from a Sobol sequence within the same δ-cube (purple line with diamonds). We see that in Fig. 17 for $d = 10$, the Sobol sequence is slightly advantageous to the random uniform on the whole cube and δ-cube for most interesting values of γ. Choosing X_n as uniform within the δ-cube produces better coverings than with Sobol's points in $[0, 1]^d$ for most values of γ, but slightly worse for small γ. This slight advantage of the Sobol sequence in $[0, 1]^d$ and in the δ-cube diminishes in the case $d = 20$ shown in Fig. 18.

To further study the similarities in performance between X_n chosen randomly in the δ-cube with optimal δ and X_n chosen from a Sobol sequence within the same δ-cube, in Figs. 19–20 we plot the ratio of the c.d.f.'s $F_d(r, X_{n,U})/F_d(r, X_{n,S})$ across different d. The subscript U and S respectively differentiate between X_n chosen randomly in the δ-cube with optimal δ and X_n chosen from a Sobol sequence within the same δ-cube. For each value of d, r is chosen so that $\max_{0 \leq \delta \leq 1} F_d(r, X_{n,U}) = 0.9$.
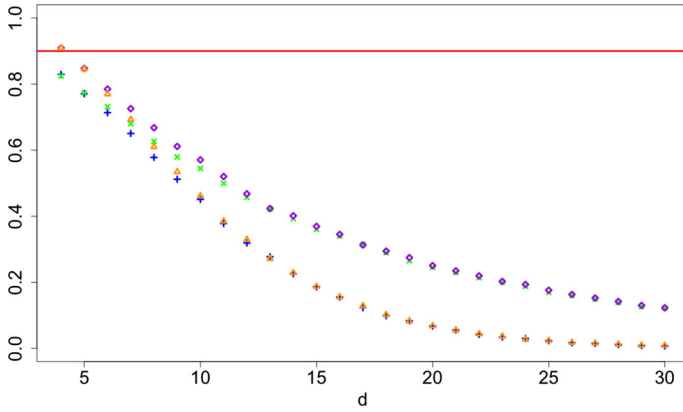
Fig. 15 Covering using the asymptotic radius with Sobol and δ-cube points: $n = 2^{10}$
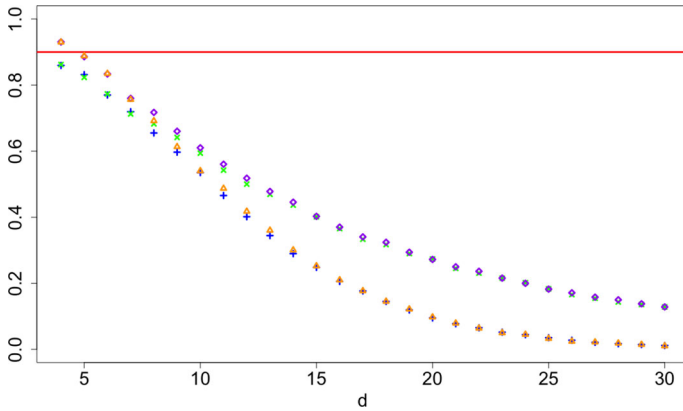

Fig. 16 Covering using the asymptotic radius with Sobol and δ-cube points: $n = 2^{13}$

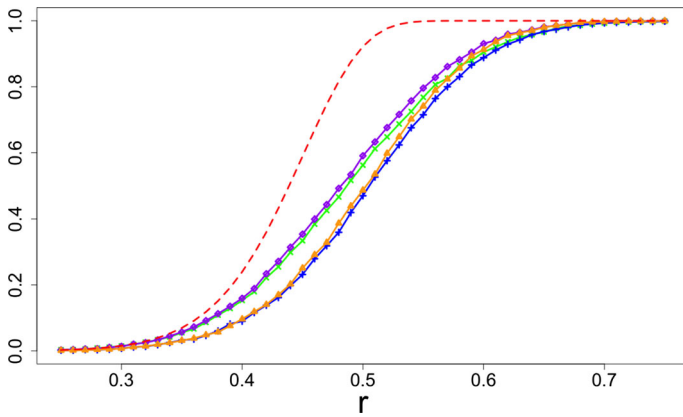**Fig. 17** Sobol and δ-cube points versus the asymptotic covering: $d = 10$, $n = 1024$
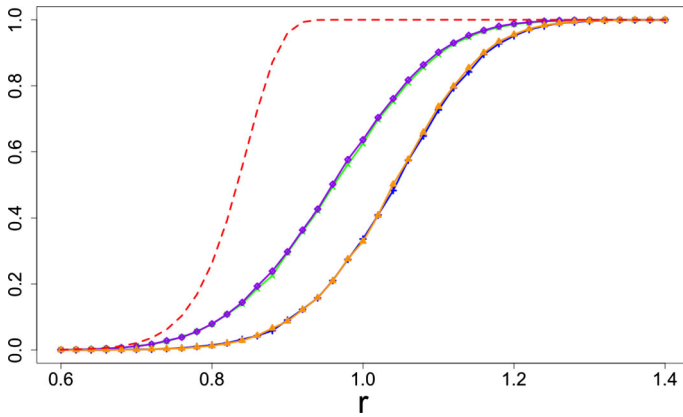
Fig. 18 Sobol and δ-cube points versus the asymptotic covering: $d = 20, n = 1024$
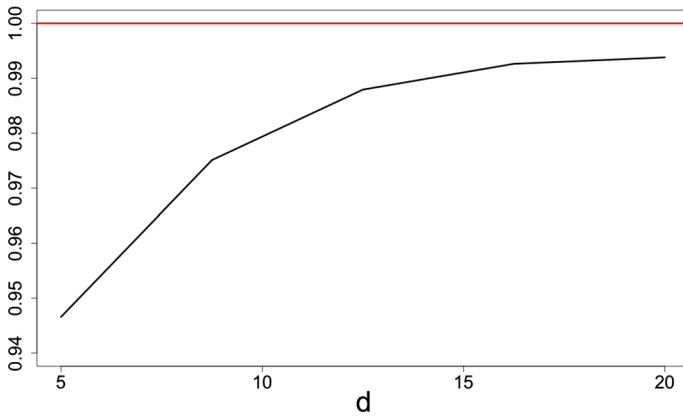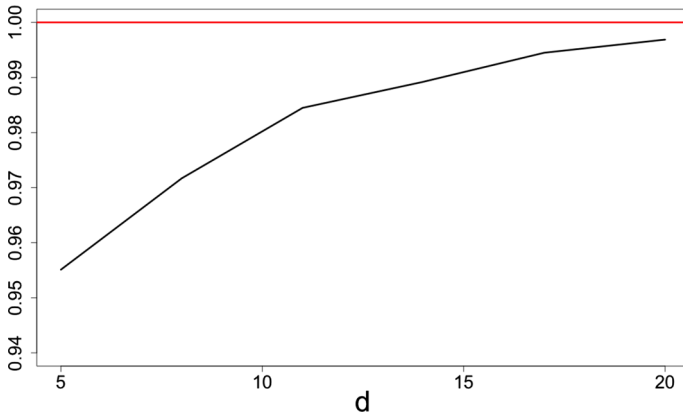

Fig. 19 Efficiency of Sobol's points, $n = 2^{10}$

Fig. 20 Efficiency of Sobol's points, $n = 2^{13}$
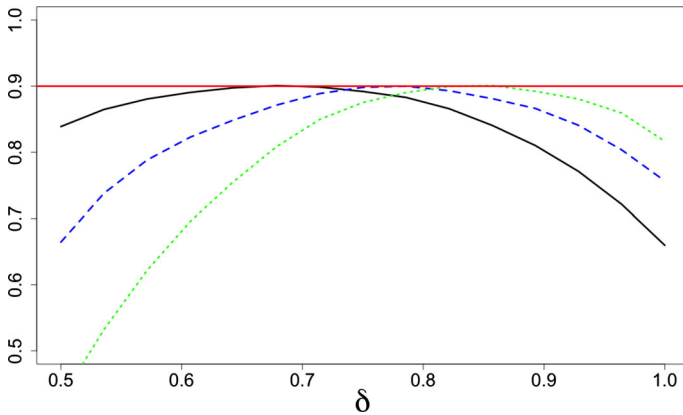
Fig. 21 $d = 20$ and $\alpha = 0.5$ with $n = 1000, 10,000, 100,000$

4.3 Non-uniform prior distribution for the target

In this section, we explore the effect a non-uniform prior distribution for $x_* \in \mathcal{X}$ has on the conclusions above formulated for the case of uniform distribution. We will assume that each component of x_* has independent components distributed according to the following symmetric beta distribution with density:

$$p_\alpha(t) = \frac{t^{\alpha-1}[1-t]^{\alpha-1}}{\text{Beta}(\alpha, \alpha)} , \text{ for some } \alpha > 0.$$

If $\alpha = 1$, the density $p_\alpha(t)$ is uniform on $[0, 1]$ while for $0 < \alpha < 1$ this density is U-shaped. In most cases below we choose the arcsine density $p_{0.5}(t)$.

We then select the distribution of random points $x_j \in X_n$ to have similar shape but constrained to the δ-cube. More precisely, we assume that x_j have independent components with the density

$$p_{\alpha,\delta}(t) = \frac{2 \cdot (2\delta)^{1-2\alpha}}{\text{Beta}(\alpha, \alpha)} \left[\delta^2 - (2t - 1)^2\right]^{\alpha-1} , \quad \frac{1-\delta}{2} < t < \frac{\delta+1}{2},$$
$$\text{for some } \alpha > 0 \text{ and } 0 \leq \delta \leq 1.$$

In the case $\alpha = 1$, the points x_j have uniform distribution on the cube C_δ.

Figs. 21–22 are similar to Figs. 11–13, but with the key difference of assuming a non-uniform prior distribution for x_*. For different values of d and n and for $\alpha = 0.5$, we plot $F_d(r, X_n)$ as a function of δ. For each d and n, the value of r has been chosen so that $\max_{0 \leq \delta \leq 1} F_d(r, X_n) = 0.9$. In these figures, the values of $F_d(r, X_n)$ for $n = 1000, 10000, 100000$ are shown with a solid black line, dashed blue line and dotted green line respectively. Figures 23–24 are similar to Figs. 21–22, but with varying values of α and fixed $n = 10000$. In these figures, we selected $\alpha = 0.25$ (dashed dark green), $\alpha = 0.5$ (dotted purple), $\alpha = 0.75$ (dot-dashed grey) and $\alpha = 1$ which gives the uniform distribution (solid black). Figures 21–22 clearly demonstrate that the 'δ-effect' is still significant in this non-uniform setting.
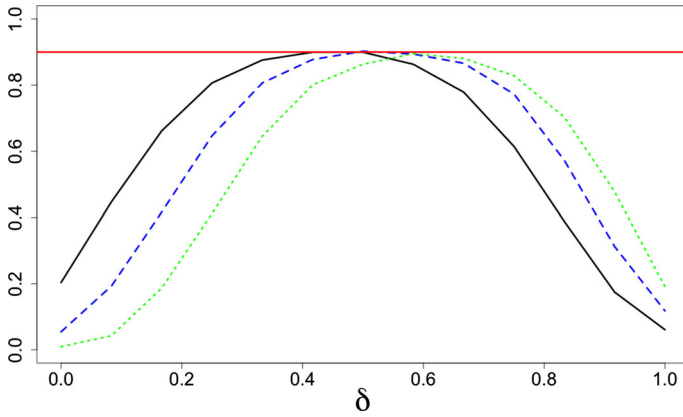
Fig. 22 $d = 50$ and $\alpha = 0.5$ with $n = 1000, 10,000, 100,000$
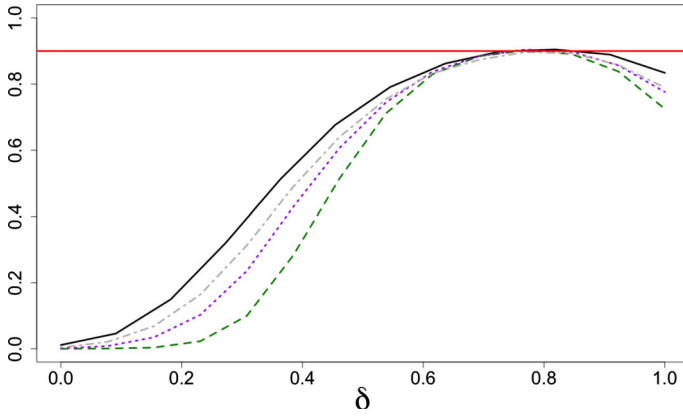

Fig. 23 $d = 20, n = 10000$ and $\alpha = 0.25, 0.5, 0.75, 1$
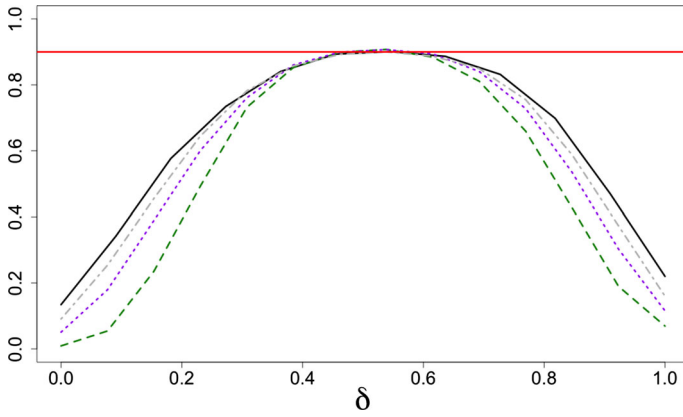

Fig. 24 $d = 50, n = 10000$ and $\alpha = 0.25, 0.5, 0.75, 1$

5 Intersection of one ball with the cube

As a result of (3.6), our main quantity of interest in this section will be the probability

$$P_{U,\delta,r} := \mathbb{P}_X \{\|U - X\| \le r\} = \mathbb{P}_X \{\|U - X\|^2 \le r^2\} = \mathbb{P}\left\{ \sum_{j=1}^{d} (u_j - x_j)^2 \le r^2 \right\} \quad (5.1)$$

in the case when X has the uniform distribution on the δ-cube $[1/2 - \delta/2, 1/2 + \delta/2]^d$ and $U = (u_1, \ldots, u_d) \in \mathbb{R}^d$ is fixed. The case of $\delta = 1$ will be directly applicable to Sect. 3.2. Because of the results of Sect. 3.2, we will bear in mind two typical choices of U will be $U = 1/2$ and $U = 3/4$ but will formulate results for general U.

For fixed $u \in \mathbb{R}^d$, consider the r.v. $\eta_{u,\delta} = (z - u)^2$, where z has density

$$p_\delta(t) = 1/\delta, \quad (1 - \delta)/2 < t < (1 + \delta)/2, \text{ for some } 0 \le \delta \le 1. \quad (5.2)$$

The first three central moments of $\eta_{u,\delta}$ are:

$$\mu_u^{(1)} = \mathbb{E}\eta_{u,\delta} = \left(u - \frac{1}{2} \right)^2 + \frac{\delta^2}{12}, \quad (5.3)$$

$$\mu_u^{(2)} = \text{var}(\eta_{u,\delta}) = \frac{\delta^2}{3} \left[\left(u - \frac{1}{2} \right)^2 + \frac{\delta^2}{60} \right], \quad (5.4)$$

$$\mu_u^{(3)} = \mathbb{E}\left[\eta_{u,\delta} - \mu_u^{(1)} \right]^3 = \frac{4\delta^4}{15} \left[\left(u - \frac{1}{2} \right)^2 + \frac{\delta^2}{252} \right]. \quad (5.5)$$

Then for given $U = (u_1, \ldots, u_d) \in \mathbb{R}^d$, consider the random variable

$$\|U - X\|^2 = \sum_{i=1}^{d} \eta_{u_i, \delta, \alpha} = \sum_{j=1}^{d} (u_j - x_j)^2,$$

where we assume that $X = (x_1, \ldots, x_d)$ is a random vector with i.i.d. components x_i with density (5.2). From (5.3), its mean is

$$\mu = \mu_{d,\delta,U} := \mathbb{E}\|U - X\|^2 = \|U - 1/2\|^2 + \frac{d\delta^2}{12}.$$

Using independence of $x_1, \ldots, x_d$ and (5.4), we obtain

$$\sigma_{d,\delta,U}^2 := \text{var}(\|U - X\|^2) = \frac{\delta^2}{3} \left[\|U - 1/2\|^2 + \frac{d\delta^2}{60} \right],$$

and from independence of $x_1, \ldots, x_d$ and (5.5) we get

$$\mu_{d,\delta,U}^{(3)} := \mathbb{E}\left[\|U - X\|^2 - \mu \right]^3 = \sum_{j=1}^{d} \mu_{u_j}^{(3)} = \frac{\delta^4}{15} \left[\|U - 1/2\|^2 + \frac{d\delta^2}{252} \right]. \quad (5.6)$$

If d is large enough then the conditions of the CLT for $\|U - X\|^2$ are approximately met and the distribution of $\|U - X\|^2$ is approximately normal with mean $\mu_{d,\delta,U}$ and variance $\sigma_{d,\delta,U}^2$. That is, we can approximate the probability $P_{U,\delta,r} = \mathbb{P}_X \{\|U - X\| \le r\}$ by

$$P_{U,\delta,r} \cong \Phi\left(\frac{r^2 - \mu_{d,\delta,U}}{\sigma_{d,\delta,U}} \right), \quad (5.7)$$

where $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution:

$$\Phi(t) = \int_{-\infty}^{t} \varphi(v)dv \ \text{ with } \ \varphi(v) = \frac{1}{\sqrt{2\pi}}e^{-v^2/2} \,.$$

The approximation (5.7) has acceptable accuracy if the probability $P_{U,\delta,r}$ is not very small; for example, it falls inside a 2σ-confidence interval generated by the standard normal distribution.

To improve on the usual CLT approximation, we use Edgeworth-type expansion in the CLT for sums of independent non-identically distributed r.v. by V.Petrov, see [13]:

$$P\left(\frac{\|U - X\|^2 - \mu_{d,\delta,U}}{\sigma_{d,\delta,U}} \le t\right) = \Phi(t) + \sum_{v=1}^{\infty} \frac{Q_{v,d}(t)}{d^{v/2}} \,, \tag{5.8}$$

where

$$Q_{v,d}(t) = -\varphi(t) \sum H_{v+2s-1}(t) \prod_{m=1}^{v} \frac{1}{k_m!} \left(\frac{\lambda_{m+2,d}}{(m+2)!}\right)^{k_m} \,,$$

$$\lambda_{v,d} = \frac{d^{(v-2)/2}}{\sigma_{d,\delta,U}^{v}} \sum_{j=1}^{d} \gamma_{v,j},$$

$\gamma_{v,j}$ is the cumulant of order v at $(u_j - x_j)^2 - \mu_{u_j}^{(1)}$, H_m is the Chebyshev-Hermite polynomial of degree m and the summation is carried out over all non-negative integer solutions of the equation

$$k_1 + 2k_2 + \cdots + vk_v = v$$
$$s = k_1 + k_2 + \cdots + k_v.$$

The partition function $p(v)$ provides the number of possible partitions of a non-negative integer v and therefore at each value of v provides the number of terms in the summation. The sequence has the generating function

$$\sum_{v=0}^{\infty} p(v)x^v = \prod_{k=1}^{\infty} \left(\frac{1}{1-x^k}\right);$$

the first few values are: 1, 1, 2, 3, 5, 7, 11, 15, 22, 30, 42, 56, 77, 101. The first few terms in the summation (including Hermite polynomials) are provided in [14, p. 139].

In the case of $U = \mathbf{1/2}$ or $U = \mathbf{3/4}$, the random variables $(u_j - x_j)^2$ will be i.i.d. For this case $\lambda_{v,d}$ does not depend on d and thus we have the slight simplification $\lambda_{v,d} = \lambda_v = \gamma_v/\sigma_{d,\delta,U}^{v}$.

In Figs. 25–30, we plot $P_{U,1,r}$ for $U = \mathbf{1/2}$ and $U = \mathbf{3/4}$ as a function of r with a solid black line. In these figures, we demonstrate the accuracy of approximation (5.7) with a dashed blue line. With a dot-dashed red line, we plot the accuracy of an approximation obtained by taking one additional term in the expansion given in (5.8); this requires use of the third central moment given in (5.6). We can see that overall, for $d = 10$ and $d = 20$, the approximations are fairly accurate. However, when considering covering by n balls it is more important to focus on the lower tail. Figures 26, 28, 29 and 30 demonstrate that taking one additional term in the Petrov's expansion (5.8) produces a significant improvement in accuracy.
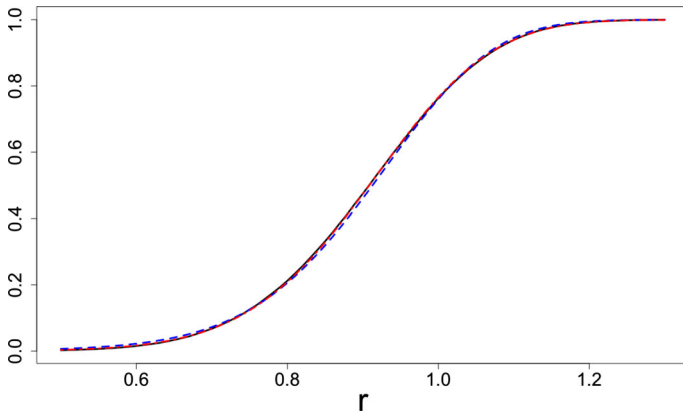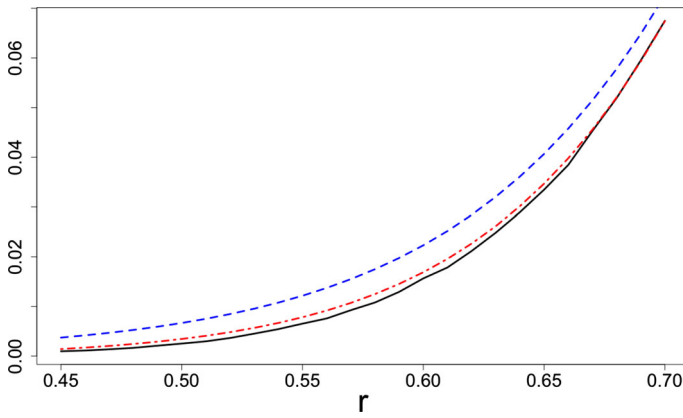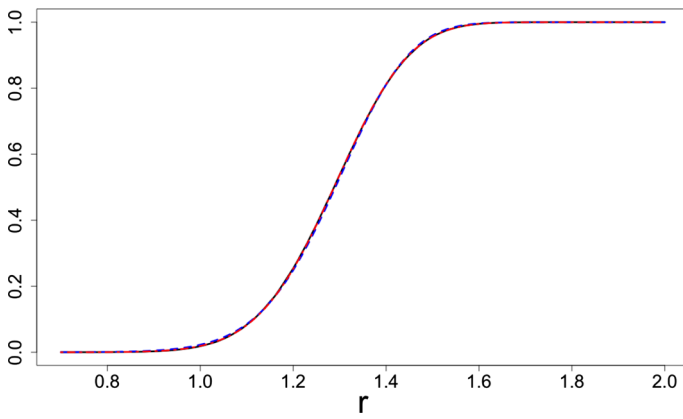
Fig. 25 $d = 10, U = \mathbf{1/2}$

Fig. 26 $d = 10, U = \mathbf{1/2}$

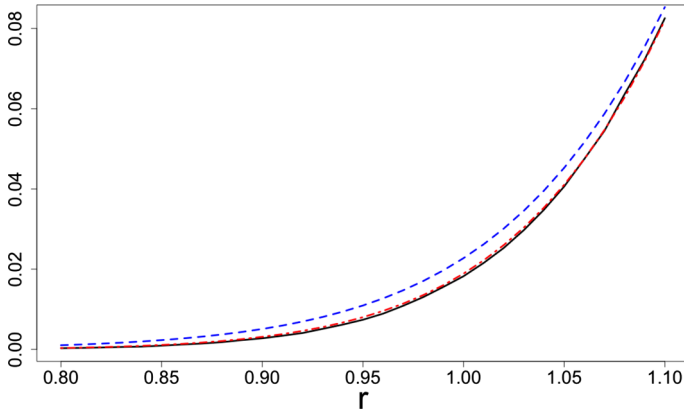**Fig. 27** $d = 20, U = \mathbf{1/2}$
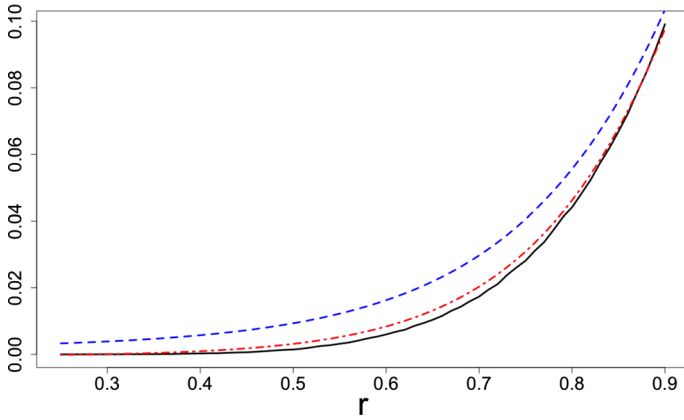
Fig. 28 $d = 20, U = 1/2$
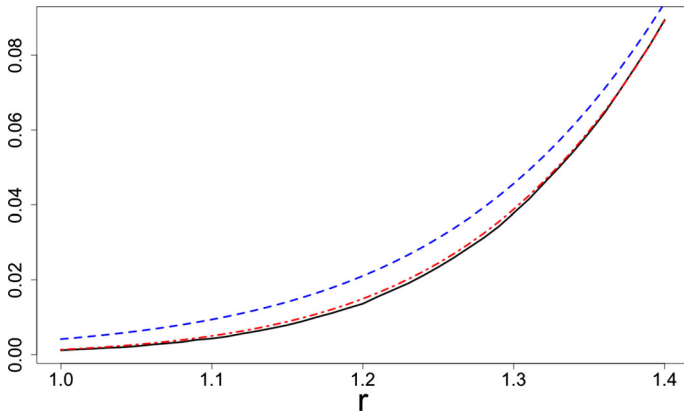

Fig. 29 $d = 10, U = 3/4$

Fig. 30 $d = 20, U = 3/4$

6 Conclusions

We have considered continuous global optimization problems, where the feasible region $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$. As a strategy for exploration, we have mostly considered sampling of i.i.d. random points either in $\mathcal{X}$ or a suitable subset of $\mathcal{X}$.

We have distinguished between between 'small', 'medium' and 'high' dimensional problems depending on the following relations between d and $n_{\max}$ (which is the maximum possible number of points available for space exploration):

(S) small dimensions: $n_{\max} \gg 2^d$ (roughly, $d < 10$);

(M) medium dimensions: $n_{\max}$ is comparable to 2^d (roughly, $10 \le d \le 20$);

(H) high dimensions: $n_{\max} \ll 2^d$ (roughly, $d > 20$).

We only considered the situations (M) and (H), where we have demonstrated the following effects: (i) the actual convergence of randomized exploration schemes is much slower than that given by the classical estimates, which are based on the asymptotic properties of random points; (ii) the usually recommended space exploration schemes are practically inefficient as the asymptotic regime is unreachable. In particular, we have shown: (ii-a) uniform sampling on entire $\mathcal{X}$ is much less efficient than uniform sampling on a suitable subset of $\mathcal{X}$, and (ii-b) the effect of replacement of random points by a low-discrepancy sequence is very small so that using low-discrepancy sequences and other deterministic constructions does not lead to significant improvements (unless the number of evaluation points $n = n_{\max}$ is fixed to some particular value like 2^d or 2^{d-1}, see [9]). We believe that the effects (i) and (ii) have not been stated in literature, at least in this generality. The effect (ii-a) has been numerically demonstrated in our previous papers [7, 8]. The effect (ii-b) enhances one of the main messages of the paper [12].

It was not the purpose of the paper to give the most effective exploration schemes. However, the results of this paper, along with studies reported in [7–9, 26], allow us to give several general recommendations on efficient organization of exploration strategies in the situations (M) and (H), at least when $\mathcal{X}$ is a cube.

In a high-dimensional cube $\mathcal{X} = [0, 1]^d$ with $d > 20$ and $1 < n_{\max} < 2^d$, we propose the following strategy of construction of nested exploration designs X_n: $x_1 = \mathbf{1/2}$ (the centre of $\mathcal{X}$) and the other points x_j are taken randomly among the vertices of a cube $[1/4, 3/4]^d$. Sampling from vertices can be done without replacement (see [8]) and, moreover, we can keep the points x_j so that the Hamming distance between them is at least $\lfloor d - \log_2(n_{\max}-1)\rfloor + 1$.

It is more difficult to be so specific in the situation (M) as there may be different relations between 2^d, $n_{\min}$ and $n_{\max}$. A good strategy would be using the product of arcsine distributions on a suitable δ-cube (see [8]); this distribution is slightly superior to the uniform on δ-cube of Sect. 4 (with different values of δ optimized for the respective distribution). An even more natural strategy would be sampling in 2^d small cubes (of side-length ε) surrounding the vertices of a δ-cube $C_\delta = [1/2 - \delta/2, 1/2 + \delta/2]^d$ (after placing $x_1 = \mathbf{1/2}$). The choice of δ and ε depends on 2^d, $n_{\min}$ and $n_{\max}$ and requires a separate study. As usual, reduction of randomness in sampling makes any of these schemes marginally more efficient.

References

1. Borodachov, S., Hardin, D., Saff, E.: Discrete Energy on Rectifiable Sets. Springer, London (2019)
2. Du, Q., Faber, V., Gunzburger, M.: Centroidal voronoi tessellations: applications and algorithms. SIAM Rev. **41**(4), 637–676 (1999)
3. Graf, S., Luschgy, H.: Foundations of Quantization for Probability Distributions. Springer, London (2007)
4. Grimmett, G., Stirzaker, D.: Probability and random processes. Oxford University Press (2020)
5. Janson, S.: Random coverings in several dimensions. Acta Math. **156**(1), 83–118 (1986)
6. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. SIAM, Philadelphia, PA (1992)
7. Noonan, J., Zhigljavsky, A.: Covering of high-dimensional cubes and quantization. SN Operat. Res. Forum **1**(3), 1–32 (2020)
8. Noonan, J., Zhigljavsky, A.: Non-lattice covering and quantization of high dimensional sets. In Black Box Optimization, Machine Learning, and No-Free Lunch Theorems, pp. 273–318. Springer: London (2021)
9. Noonan, J., Zhigljavsky, A.: Efficient quantisation and weak covering of high dimensional cubes. Discrete & Computational Geometry, pp. 1–26, (2022)
10. Pagès, G.: A space quantization method for numerical integration. J. Computat. Appl. Math. **89**(1), 1–38 (1998)
11. Penrose, M.: Random Euclidean coverage from within. Probabil. Theory Relat. Fields **185**(3–4), 747–814 (2023)
12. Pepelyshev, A., Zhigljavsky, A., Žilinskas, A.: Performance of global random search algorithms for large dimensions. J. Global Optimiz. **71**(1), 57–71 (2018)
13. Petrov, V.: Sums of independent random variables. Springer-Verlag (1975)
14. Petrov, V.: Limit theorems of probability theory: sequences of independent random variables. Oxford Science Publications (1995)
15. Pintér, J.: Convergence properties of stochastic optimization procedures. Optimization **15**(3), 405–427 (1984)
16. Pronzato, L., Zhigljavsky, A.: Bayesian quadrature, energy minimization, and space-filling design. SIAM/ASA J. Uncert. Quantif. **8**(3), 959–1011 (2020)
17. Santner, T., Williams, B., Notz, W., Williams, B.: The design and analysis of computer experiments, vol. 1. Springer (2003)
18. Schaback, R., Wendland, H.: Kernel techniques: from machine learning to meshless methods. Acta Num. **15**, 543–639 (2006)
19. Solis, F., Wets, R.: Minimization by random search techniques. Math. Operat. Res. **6**(1), 19–30 (1981)
20. Sukharev, A.: Optimal strategies of search for an extremum. USSR Computat. Math. Math. Phys. **11**(4), 910–924 (1971)
21. Sukharev, A.: Minimax models in the theory of numerical methods. Springer (1992)
22. Tsvetkov, E., Krymov, R.: Pure random search with virtual extension of feasible region. J. Optimizat. Theory Appl. **195**(2), 575–595 (2022)
23. Wendland, H.: Scattered data approximation, vol. 17. Cambridge University Press, UK (2004)
24. Zador, P.: Asymptotic quantization error of continuous signals and the quantization dimension. IEEE Trans. Inform. Theory **28**(2), 139–149 (1982)
25. Zhigljavsky, A.: Theory of global random search. Kluwer, Dordrecht (1991)
26. Zhigljavsky, A., Noonan, J.: Covering of High-Dimensional Sets, pp 1–6. Springer: Cham, (2023). https://doi.org/10.1007/978-3-030-54621-2_770-1
27. Zhigljavsky, A., Žilinskas, A.: Stochastic Global Optimization. Springer (2008)
28. Zhigljavsky, A., Žilinskas, A.: Bayesian and High-Dimensional Global Optimization. Springer (2021)
29. Žilinskas, A.: On the worst-case optimal multi-objective global optimization. Optimizat. Lett. **7**(8), 1921–1928 (2013)