JOGO manuscript No.
(will be inserted by the editor)

Performance of global random search algorithms for large dimensions

Andrey Pepelyshev · Anatoly Zhigljavsky · Antanas Žilinskas

Received: date / Accepted: date

Abstract We investigate the rate of convergence of general global random search (GRS) algorithms. We show that if the dimension of the feasible domain is large then it is impossible to give any guarantee that the global minimizer is found by a general GRS algorithm with reasonable accuracy. We then study precision of statistical estimates of the global minimum in the case of large dimensions. We show that these estimates also suffer the curse of dimensionality. Finally, we demonstrate that the use of quasi-random points in place of the random ones does not give any visible advantage in large dimensions.

Keywords Global optimization · statistical models · extreme value statistics · random search

1 Introduction

Let us consider the problem of global minimization which we formulate as

$$f(x) \to \min_{x \in \mathbf{X}} . \tag{1}$$

Here $f(\cdot)$ is the objective function and $\mathbf{X} \subset \mathbb{R}^d$ is a feasible domain. The set $\mathbf{X}$ is assumed to be closed, bounded and having non-empty interior and the

A. Pepelyshev
School of Mathematics, Cardiff University, Cardiff, CF24 4AG, UK and St.Petersburg State University, University Embankment, 7/9, St.Petersburg, Russia, E-mail: pepelyshevan@cardiff.ac.uk
A. Zhigljavsky
School of Mathematics, Cardiff University, Cardiff, CF24 4AG, UK and Lobachevsky Nizhny Novgorod State University, 23 Prospekt Gagarina, 603950, Nizhny Novgorod, Russia, E-mail: zhigljavskyaa@cardiff.ac.uk
A. Žilinskas
Institute of Mathematics and Informatics, Vilnius University, Akademijos 4, Vilnius, LT-08663, Lithuania, E-mail: antanas.zilinskas@mii.vu.lt

objective function $f(\cdot)$ is assumed to satisfy some smoothness conditions which will be discussed below.

Let $f_* = \min_{x\in\mathbf{X}} f(x)$ be the minimal value of $f(\cdot)$ and x_* be a global minimizer; that is, x_* is any point in $\mathbf{X}$ such that $f(x_*) = f_*$. Global optimization problems are often stated so that the domain $\mathbf{X}$ has a relatively simple shape by using, for example, penalty functions.

A convergent global minimization algorithm constructs a sequence of points $x_1, x_2, \ldots$ in $\mathbf{X}$ such that the sequence of record values

$$y_{1,n} = \min_{i=1,\ldots,n} f(x_i) \tag{2}$$

tends to f_* as n increases. In addition to finding the minimal value f_*, at least one of the minimizers x_* has usually to be determined. In practice, the optimization iterative algorithms are always accompanied with stopping rules.

If the objective function is given as a 'black box' computer code and Lipschitz-type information about this function is unavailable then good stochastic approaches often perform better than deterministic algorithms, see [13, 15, 16, 18]. Moreover, stochastic algorithms are typically simpler than their deterministic counterparts.

In the following sections, we consider performance of various methods of global random search (GRS) in the case when the dimension d is large. As a rough guide, we consider dimensions 1, 2 and 3 as small, dimensions 10, 20 as moderate and dimensions 50 and above as large.

A generic GRS algorithm constructs a sequence of random points $x_1, x_2, \ldots$ such that the point x_j has some probability distribution P_j, $j = 1, 2, \ldots$; we write this as $x_j \sim P_j$. For each $j \geqslant 2$, the distribution P_j may depend on the previous points $x_1, \ldots, x_{j-1}$ and on $f(x_1), \ldots, f(x_{j-1})$. The stopping rule for GRS can be either deterministic or random and may depend on the evaluations $f(x_1), f(x_2), \ldots$. The distribution P_j should maintain the right balance between globality and locality of search. This balance is one of the main ingredients of algorithm's efficiency. Achieving the right balance depends on the complexity of computing derivatives of $f(\cdot)$ for performing fast local descents and on the efficient use of all available information, which is a prior information about $f(\cdot)$ and $\mathbf{X}$ and the information contained in the evaluations $f(x_1), f(x_2), \ldots$. Construction of a particular GRS algorithm involves setting of a distribution P_j (based on all available information before time j) along with a stopping rule.

In the present paper, we will mostly concentrate on the so-called pure random search (PRS) algorithm, where the points $x_1, x_2, \ldots$ are independent and have the same distribution $P = P_j$ for all j. Simplicity of PRS allows detailed investigation of this algorithm, see Sections 2 and 3.

The present paper is organized as follows. Convergence of GRS is reviewed in Section 2. Statistical inference procedures in GRS are discussed in Section 3 and the use of the low-dispersion sequences for global search is considered in Section 4.

2 Convergence and the rate of convergence

In this section we illustrate the following two points:

(i) it is very easy to construct a GRS algorithm which has the theoretical property of convergence;
(ii) on the other hand, even for moderate dimensions it is impossible to guarantee that the global minimum is found in the worst-case scenario with reasonable accuracy.

2.1 Convergence

Consider a general GRS algorithm defined by a sequence of probability distributions P_j, $j = 1, 2, \ldots$. We say that this algorithm converges if for any $\delta > 0$, the sequence of points x_j arrives at the set $W(\delta) = \{x \in \mathbf{X}\colon f(x) - f_* \leqslant \delta\}$ with probability one. If the objective function is evaluated without error then this obviously implies convergence (as $n \to \infty$) of record values $y_{1,n}$ to f_* with probability 1.

Conditions on the distributions P_j, $j = 1, 2, \ldots$ ensuring convergence of the GRS algorithms are well understood; see, for example, [1,8,9]. The results on convergence of GRS algorithms are usually formulated in the form of the 'zero-one law', which is classical in probability theory. The following theorem provides an illustration of such results in a very general setup and is proved in [13, Sect. 3.2] in a more general form.

Theorem 1 *Consider a GRS algorithm with $x_j \sim P_j$ for the minimization problem* (1)*, where $\mathbf{X}$ is a compact set and $f(\cdot)$ is a function on $\mathbf{X}$ satisfying the Lipschitz condition. Let $B(x, \varepsilon) = \{z \in \mathbf{X}\colon \|z - x\| \leqslant \varepsilon\}$ be a ball centered at x. Define $q_j(\varepsilon) = \inf P_j(B(x, \varepsilon))$, where the infimum is taken over all $x \in \mathbf{X}$, all possible points $x_1, \ldots, x_{j-1}$ all evaluations of the objective function at these points. Assume that*

$$\sum_{j=1}^{\infty} q_j(\varepsilon) = \infty \tag{3}$$

for any $\varepsilon > 0$. Then for any $\delta > 0$, the sequence of points x_j falls infinitely often into the set $W(\delta)$, with probability one.

Note that Theorem 1 holds in the very general case where evaluations of the objective function $f(\cdot)$ are noisy and the noise is not necessarily random. If the function evaluations are noise-free, then the conditions of Theorem 1 ensure that the corresponding algorithm converges; that is, that the sequence of records $y_{1,j}$ converges to f_* with probability 1 and the corresponding subsequence of points $\{x_{i_j}\}$ (where the new records are attained) of the sequence $\{x_j\}$ converges (with probability 1) to the set $\mathbf{X}_* = \{x \in \mathbf{X}\colon f(x) = f_*\}$ of global minimizers.

If the objective function is evaluated with random error then the algorithm of generation of points x_j should be accompanied with an algorithm of estimation of the objective function estimation, see [13, Sect. 4.1.3]. Then the rate of convergence of the corresponding algorithm will also depend on the smoothness of the objective function and the chosen approximation routine.

If we use PRS with the uniform distribution $P = P_U$ on $\mathbf{X}$, we obtain $q_j(\varepsilon) = \text{const} > 0$ and therefore the condition (3) trivially holds. In practice, a usual choice of the distribution P_j is

$$P_j = \alpha_j P_U + (1 - \alpha_j) Q_j \,, \tag{4}$$

where $0 \leqslant \alpha_j \leqslant 1$ and Q_j is a specific probability measure on $\mathbf{X}$ which usually depends on evaluations of the objective function at the points $x_1, \ldots, x_{j-1}$. Sampling from the distribution (4) corresponds to taking a uniformly distributed random point in $\mathbf{X}$ with probability α_j and sampling from Q_j with probability $1 - \alpha_j$.

Note that $\sum_{j=1}^{\infty} \alpha_j = \infty$ yields the fulfilment of (3) for distributions P_j in the form (4) and therefore the GRS algorithm with such P_j is theoretically converging. On the other hand, if $\sum_{j=1}^{\infty} \alpha_j < \infty$ then there is a non-zero probability that the neighbourhood of the global minimizer will never be reached.

Unless smoothness conditions about $f(\cdot)$ like the Lipschitz condition are imposed, the statements like Theorem 1 are the only tools which are ensuring convergence of the GRS algorithms. Note that one of the implications of these arguments is that the PRS with $P = P_U$ is the GRS algorithm enjoying the fastest convergence in the worst-case scenario.

2.2 Rate of convergence of PRS

Consider a PRS algorithm with $x_j \sim P$. Let $\varepsilon, \delta > 0$ be fixed and $W(\delta) = \{x \in \mathbf{X} \colon f(x) - f_* \leqslant \delta\}$. Define a set B as $B = B(x_*, \varepsilon)$ if we are studying convergence towards x_*, and as $B = W(\delta)$ if the purpose of study is the convergence with respect to the function values to f_*. Assume that P is such that $P(B) > 0$; for example, $P = P_U$ is the uniform probability measure on $\mathbf{X}$.

Define the Bernoulli trials where the success in the j-th trial means that $x_j \in B$. PRS generates a sequence of independent Bernoulli trials with the same success probability $\Pr\{x_j \in B\} = P(B)$. In view of the independence of x_j, we have $\Pr\{x_1 \notin B, \ldots, x_n \notin B\} = (1 - P(B))^n$ and therefore the probability

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

tends to one as $n \to \infty$. We also assume that $P(B)$ is small.

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,

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

Solving it we obtain

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

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

We can see that the numerator in the expression of n_γ depends on γ but it is not large; for example, $-\ln \gamma \simeq 2.996$ for $\gamma = 0.05$. But the denominator, which is approximately $P(B)$, can be extremely small and hence n_γ could be astronomically large in most real-life optimization problems.

Consider an example with the set $\mathbf{X} = [0,1]^d$, the uniform distribution $P = P_U$ on $\mathbf{X}$ and the set $B = B(x_*, \varepsilon)$. Then $P(B) = \mathrm{vol}(B) \leqslant \varepsilon^d V_d$, where $V_d = \pi^{\frac{d}{2}} / \Gamma\left(\frac{d}{2} + 1\right)$ is the volume of the unit ball in $\mathbb{R}^d$, where $\Gamma(t)$ is the gamma function. In view of the upper bound of the form "const $\cdot\, \varepsilon^d$", the probability $P(B)$ can be extremely small even when ε is not very small (say, $\varepsilon = 0.1$) and $d \geqslant 10$. The number n_γ in this case is shown in Figure 1.

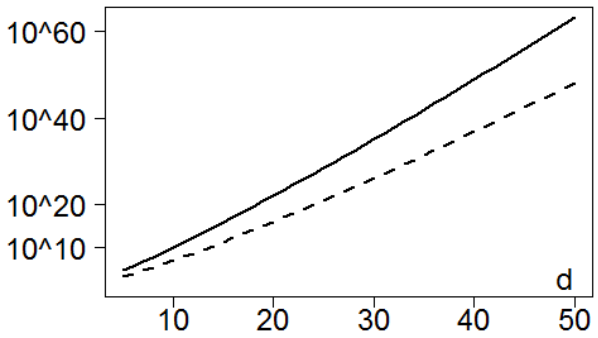

Fig. 1 The number n_γ of points which are required for PRS to reach the set $B = B(x_*, \varepsilon)$ with probability at least $1 - \gamma = 0.95$ for $\varepsilon = 0.1$ (solid) and $\varepsilon = 0.2$ (dashed) as the dimension d varies in $[5, 50]$.

2.3 Rate of convergence of a general GRS method

The easiest way to ensure convergence of a general GRS algorithm is to choose the probabilities P_j in the form (4), where the coefficients α_j satisfy the condition (3), see Section 2.1.

Let us generalize the arguments given in Section 2.2 for the case of PRS to the case of GRS. Instead of the equality $\Pr\{x_j \in B\} = P(B)$ for all $j \geqslant 1$, we now have the inequality $\Pr\{x_j \in B\} \geqslant \alpha_j P_U(B)$, where the equality holds in the worst-case scenario. Further we define $n(\gamma)$ as the smallest integer such that the inequality $\sum_{j=1}^{n(\gamma)} \alpha_j \geqslant -\ln \gamma / P_U(B)$ 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 $\mathbf{X} = [0,1]^d$ 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 $\mathbf{X}$ 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, namely, $n(\gamma) \simeq 10^{238}$. Figure 2 shows the behaviour of $n(\gamma)$ as the dimension grows.

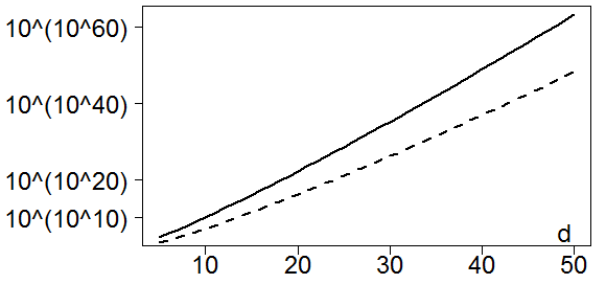

Fig. 2 The number $n(\gamma)$ of points which are required for GRS to reach the set $B = B(x_*, \varepsilon)$ with probability at least $1-\gamma = 0.95$ for $\varepsilon = 0.1$ (solid) and $\varepsilon = 0.2$ (dashed) as the dimension d varies in $[5, 50]$.

3 Statistical inference about f_*

Let $\mathbf{X}$ be a compact in $\mathbb{R}^d$ and $x_1, \ldots, x_n$ be points constructed by PRS with $x_j \sim P$, where n is a large number and P is a probability measure on $\mathbf{X}$ with some density $p(x)$, which is a piece-wise continuous function on $\mathbf{X}$ and $p(x) > 0$ for all $x \in \mathbf{X}$. Using prior information about $f(\cdot)$ and considering the values $\{f(x_j)\}_{j=1,\ldots,n}$ as a sample, we can make statistical inference of the two types:

1. Building either a parametric or non-parametric estimator of $f(\cdot)$, e.g., a kriging estimator or an estimator based on the Lipschitz condition.
2. Construction of an estimator and a confidence interval for f_*.

Inferences of Type 1 will not be considered in the present paper because it requires much metaheuristic for explanation, see a comprehensive discussion in [13]. Below we only consider inferences of Type 2 following an approach in [13, Ch. 7], [16, Sect. 2.3–2.6] and [17]. Statistical inference about f_*, the minimal value of the objective function $f(\cdot)$, can serve for the following purposes:

(i) devising specific GRS algorithms like the branch and probability bounds methods, see [12,19] and [13, Sect. 4.3],
(ii) constructing stopping rules, see [14], and

(iii) increasing efficiency of population-based GRS methods, see discussion in [16, Sect. 2.6.1].

Another direction where the use of statistical inferences in GRS algorithms can be very helpful is solving multi-objective optimization problems with non-convex objectives, see [19].

3.1 Statistical inference in PRS: the main assumption

Since the points x_j in PRS are independent identically distributed (i.i.d.) with distribution P, the elements of the sample $Y = \{y_1, \ldots, y_n\}$ with $y_j = f(x_j)$ are i.i.d. with cumulative distribution function (c.d.f.)

$$F(t) = \Pr\{x \in \mathbf{X} : \, f(x) \leqslant t\} = \int_{f(x) \leqslant t} P(dx) = P(W(t - f_*)) \,, \qquad (5)$$

where $t \geqslant f_*$ and $W(\delta) = \{x \in \mathbf{X} \colon f(x) \leqslant f_* + \delta\}$, $\delta \geqslant 0$. Note that the c.d.f. $F(t)$ is concentrated on the interval $[f_*, f^*]$, where $f^* = \max_{x \in \mathbf{X}} f(x)$, and our main interest is the unknown value f_*, which is the lower bound of this interval. Since the analytic form of $F(t)$ is either unknown or incomprehensible (unless f is very simple), for making statistical inferences about f_* we need to use asymptotic considerations based the record values of the sample Y. It is well known, see e.g. [6], that the asymptotic distribution of the order statistics is unambiguous and the conditions on $F(t)$ and $f(\cdot)$, when this asymptotic law works, are very mild and typically hold in real-life problems. Specifically, for a very wide class of functions $f(\cdot)$ and distributions P, the c.d.f. $F(t)$ can be represented as

$$F(t) = c_0 (t - f_*)^\alpha + \mathrm{o}((t - f_*)^\alpha), \; t \downarrow f_*, \qquad (6)$$

where c_0 and α are some positive constants. Moreover, for more general $f(\cdot)$ and P, the coefficient $c_0 = c_0(t)$ can be a slowly varying function for $t \simeq f_*$ and the results given below are also valid for this more general case. In our constructions the value of c_0 is not important but the value of α is absolutely essential. The coefficient α is called 'tail index' and its value is often known, as discussed in Section 3.2.

Denote by η a random variable which has c.d.f. (5) and by $y_{1,n} \leqslant \ldots \leqslant y_{n,n}$ the order statistics corresponding to the sample Y. Note that f_* is the lower endpoint of the random variable η.

One of the fundamental results in the theory of extreme order statistics states (see e.g. [6] and [16, Sect. 2.3]) that if (6) holds then the c.d.f. $F(t)$ belongs to the domain of attraction of the Weibull distribution with density $\psi_\alpha(t) = \alpha \, t^{\alpha - 1} \exp\{-t^\alpha\} \,, \;\; t > 0 \,.$ This distribution has only one parameter, the tail index α.

3.2 Tail index

As stated in Section 3.1, the representation (6) holds in most real-life problems. The main issue is whether the value of the tail index α can be specified or has to be estimated. The second option, that is estimation of α, is notoriously difficult; see [4] for a survey about comparison of different estimators of α. The number n of points must be astronomically large (even for small dimension d) if we want to accurately estimate f_* after replacing α with any (even best possible) estimator. Practically, n should be extremely large to see any convergence of estimators, see discussion in [16, Sect. 2.5.1]. Asymptotically, as $n \to \infty$, if α is estimated then the asymptotic mean squared error (MSE) of the maximum likelihood estimator (MLE) of f_* is at least $(\alpha - 1)^2$ times larger than the MSE of the MLE of f_* in the case when α is known. As for large d estimation of f_* is extremely hard even when α is known and α is large when d is large (see below) we can conclude that estimation of f_*, when d is large and α unknown, is virtually impossible.

In PRS, however, we can usually have enough knowledge about $f(\cdot)$ to get the exact value of the tail index α. In particular, the following result holds: if the global minimizer x_* of $f(\cdot)$ is unique and $f(\cdot)$ is locally quadratic around x_* then the representation (6) holds with $\alpha = d/2$. Moreover, if the global minimizer x_* of $f(\cdot)$ is attained at the boundary of $f(\cdot)$ and the gradient of $f(\cdot)$ is has all non-zero components at x_*, then the representation (6) holds with $\alpha = d$. This result, as well as some of its generalizations, has been established in [3] and [10] independently; see also [11,13] for a detailed exposition of the related theory.

The fact that α has the same order as d when d is large implies the phenomena called 'the curse of dimensionality'. We theoretically study this in the following sections but in this section we illustrate this curse of dimensionality on a simple numerical example.

Consider the minimization problem with the objective function $f(x) = e_1^T x$, where $e_1 = (1, 0, \dots, 0)^T$, and the set $\mathbf{X}$ is the unit ball: $\mathbf{X} = \{x \in \mathbb{R}^d : ||x|| \le 1\}$. It is easy to see that the minimal value is $f_* = -1$ and the global minimizer $z_* = (-1, 0, \dots, 0)^T$ is located at the boundary of $\mathbf{X}$. Consider the PRS algorithm with points x_j generated from the uniform distribution P_U on $\mathbf{X}$. In Figure 3 we depict projections of points $x_1, \dots, x_n$ on a fixed two-dimensional plane for $n = 10^3$ and $n = 10^5$ and the dimension $d = 20$. We can see that even if the number of simulated points is large, there is a thick ring inside the unit circle with no projections of points although the points are uniformly distributed in the unit hyperball.

Define $r_j = ||x_j||$. It is well known that $\Pr(r_j < t) = t^d$. Thus, the distribution of r_j satisfies the representation (6) with $\alpha = d$. We are interested in the record values for the sample with $y_j = e_1^T x_j$, $j = 1, \dots, n$.

Let us give some numerical values. In a simulation with $n = 10^3$ and $d = 20$, we have received $y_{1,n} = -0.64352$, $y_{2,n} = -0.61074$, $y_{3,n} = -0.60479$ and $y_{4,n} = -0.60208$. In a simulation with $n = 10^5$ and $d = 20$, we have obtained $y_{1,n} = -0.74366$, $y_{2,n} = -0.73894$, $y_{3,n} = -0.73228$ and $y_{4,n} =$

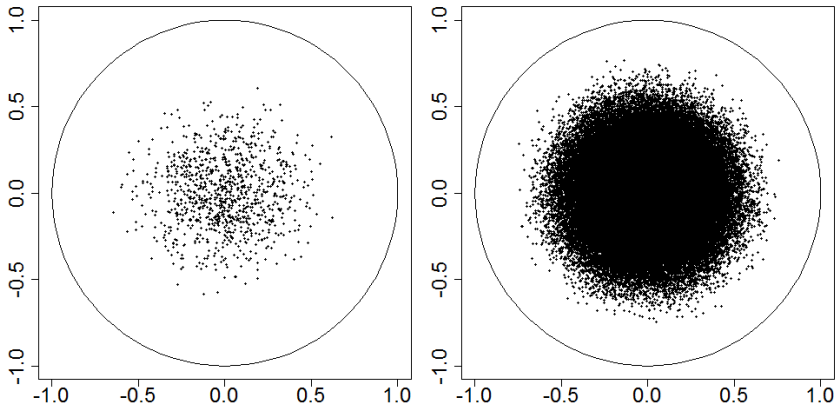

Fig. 3 Projections of points $x_1, \ldots, x_n$ with uniform distribution in the unit hyperball on a plane for $n = 10^3$ (left) and $n = 10^5$ (right) and the dimension $d = 20$.

-0.72603. In Figure 4 we depict the differences $y_{k,n} - f_*$ for $k = 1, 4, 10$ and $n = 10^3, \ldots, 10^{13}$, where the horizontal axis has logarithmic scale. We can clearly see that the difference $y_{k,n} - y_{1,n}$ is much smaller than the difference $y_{1,n} - f_*$; that shows that the problem of estimating the minimal value of f_* is very difficult.

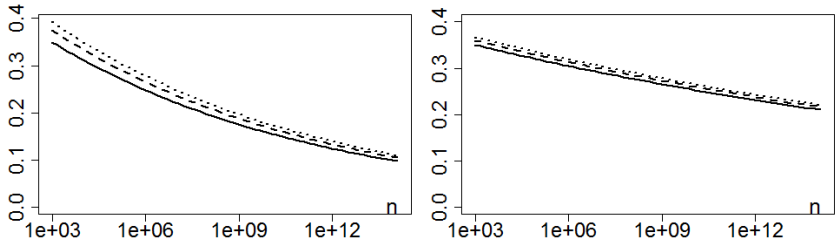

Fig. 4 Differences $y_{1,n} - f_*$ (solid), $y_{4,n} - f_*$ (dashed) and $y_{10,n} - f_*$ (dotted), where $y_{k,n}$, $k = 1, 4, 10$, are records of evaluations of the function $f(x) = e_1^T x$ at points $x_1, \ldots, x_n$ with uniform distribution in the unit hyperball in the dimension $d = 20$ (left) and $d = 50$ (right).

In Figure 5 we show that the difference $y_{1,n} - f_*$ increases as the dimension d grows, for fixed n. Thus, the minimization problem becomes harder in larger dimensions. Also, Figure 5 shows that difference $y_{10,n} - y_{1,n}$ is much smaller than the difference $y_{1,n} - f_*$.

3.3 Estimation of the minimal value of f

In this section, we review the asymptotic properties of two estimators of the minimal value f_*, the MLE and the best linear estimator. We also discuss

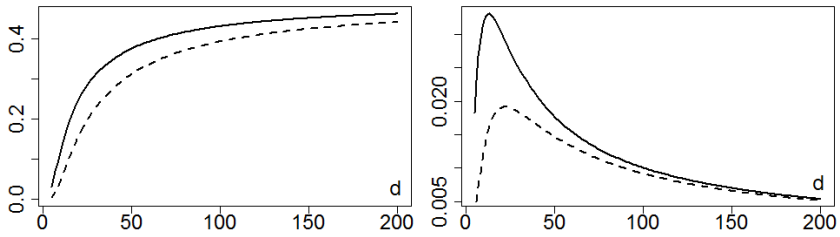

Fig. 5 The difference $y_{1,n} - f_*$ (left) and $y_{10,n} - y_{1,n}$ (right) for $n = 10^6$ (solid) and $n = 10^{10}$ (dashed), where $y_{j,n}$ is the j-th record of evaluations of the function $f(x) = e_1^T x$ at points $x_1, \dots, x_n$ with uniform distribution in the unit hyperball in the dimension d; d varies in $[5, 200]$.

properties of these estimators in the case of large dimension d (and hence large α).

If the representation (6) holds, $\alpha \geq 2$ is fixed, $k \to \infty$, $k/n \to 0$ as $n \to \infty$, then the MLE $\hat{f}_{mle}$ of f_* is asymptotically normal and asymptotically efficient in the class of asymptotically normal estimators, and the MSE has the asymptotic form

$$\mathrm{E}(\hat{f}_{mle} - f_*)^2 \approx \begin{cases} (1 - \frac{2}{\alpha})(\kappa_n - f_*)^2 \, k^{-1+2/\alpha}, & \alpha > 2, \\ (\kappa_n - f_*)^2 / \ln k, & \alpha = 2\,, \end{cases} \tag{7}$$

where κ_n is the $(1/n)$-quantile of the c.d.f. $F(\cdot)$. In view of (6) we have

$$\kappa_n - f_* = (c_0 n)^{-1/\alpha} \, (1 + o(1)) \quad \text{as} \quad n \to \infty. \tag{8}$$

Linear estimators of f_* are simpler than the MLE. Nevertheless, the best linear estimators have the same asymptotic properties. To define a linear estimator, we introduce the vectors $a = (a_1, \dots, a_k)^T \in \mathbb{R}^k$, $\mathbf{1} = (1, 1, \dots, 1)^T \in \mathbb{R}^k$, $b = (b_1, \dots, b_k)^T \in \mathbb{R}^k$, where $b_i = \Gamma(i + 1/\alpha) \, / \, \Gamma(i)$, and the matrix $\Lambda = \|\lambda_{ij}\|_{i,j=1}^k$, where $\lambda_{ji} = \lambda_{ij} = u_i v_j$, $i \geqslant j$, and

$$u_i = \frac{\Gamma(i+2/\alpha)}{\Gamma(i+1/\alpha)}, \quad v_j = \frac{\Gamma(j+1/\alpha)}{\Gamma(j)}.$$

The matrix Λ in such form can be inverted analytically, see [5, Lemma A.1].

A general linear estimator of f_* can be written as

$$\hat{f}_{n,k}(a) = \sum_{i=1}^{k} a_i y_{i,n}\,, \tag{9}$$

where $a = (a_1, \dots, a_k)^T$ is the vector of coefficients. Then using explicit expressions for moments of order statistics, for any linear estimator $\hat{f}_{n,k}(a)$ of the form (9) we obtain

$$\mathrm{E}\hat{f}_{n,k}(a) = \sum_{i=1}^{k} a_i \mathrm{E} y_{i,n} = f_* \sum_{i=1}^{k} a_i + (\kappa_n - f_*) a^T b + \mathrm{o}(\kappa_n - f_*), \; n \to \infty. \tag{10}$$

Since $\kappa_n - f_* \to 0$ as $n \to \infty$ and the variances of all $y_{i,n}$ are finite, the estimator $\hat{f}_{n,k}(a)$ is a consistent estimator of f_* iff $a^T\mathbf{1} = \sum_{i=1}^k a_i = 1$. Using explicit expressions for the moments of order statistics and the expression (8), we obtain the following expression for the MSE of the estimator $\hat{f}_{n,k}(a)$:

$$\mathrm{E}(\hat{f}_{n,k}(a) - f_*)^2 = (c_0 n)^{-2/\alpha}\, a^T \Lambda a\, (1 + o(1))\,, \quad n \to \infty\,. \tag{11}$$

The asymptotic mean squared error (11) is a natural optimality criterion for choosing the vector of coefficients a, whose minimum is attained at

$$a^* = \arg \min_{a: a^T\mathbf{1}=1} a^T \Lambda a = \frac{\Lambda^{-1}\mathbf{1}}{\mathbf{1}^T\Lambda^{-1}\mathbf{1}}\,, \tag{12}$$

and

$$\min_{a: a^T\mathbf{1}=1} a^T \Lambda a = (a^*)^T \Lambda a^* = 1/\mathbf{1}^T\Lambda^{-1}\mathbf{1}\,.$$

The estimator $\hat{f}_{n,k}(a^*)$ is called the optimal linear estimator; it has been proposed in [2], where the form (12) was obtained.

As shown in [13, Th. 7.3.2] and could also be derived from a general Lemma A.1 from [5], the components of the vector $a^* = (a_1^*, \ldots, a_k^*)^T$ can be evaluated explicitly as follows: $a_i^* = v_i/\mathbf{1}^T\Lambda^{-1}\mathbf{1}$ for $i = 1, \ldots, k$, where

$$\begin{aligned} v_1 &= (\alpha+1)\,/\,\Gamma(1+2/\alpha), \\ v_i &= (\alpha-1)\,\Gamma(i)/\Gamma(i+2/\alpha), \qquad i = 2, \ldots, k-1, \\ v_k &= -(\alpha k - \alpha + 1)\Gamma(k)\,/\,\Gamma(k+2/\alpha). \end{aligned}$$

and

$$\mathbf{1}^T\Lambda^{-1}\mathbf{1} = \begin{cases} \frac{1}{\alpha-2}\left(\frac{\alpha\Gamma(k+1)}{\Gamma(k+2/\alpha)} - \frac{2}{\Gamma(1+2/\alpha)}\right), & \alpha \neq 2, \\ \sum_{i=1}^k 1/i, & \alpha = 2\,. \end{cases} \tag{13}$$

Note that the expression (13) is valid for all $\alpha > 0$ and $k = 1, 2, \ldots$. Using the Taylor series

$$\Gamma(k + 2/\alpha) = \Gamma(k) + \frac{2}{\alpha}\Gamma'(k) + O(1/\alpha^2)$$

for large values of α, we obtain

$$\min_{a: a^T\mathbf{1}=1} a^T \Lambda a = \frac{1}{\mathbf{1}^T\Lambda^{-1}\mathbf{1}} \simeq \frac{1}{k} + \frac{2(\psi(k) - 1 + 1/k)}{\alpha k}\,, \tag{14}$$

for large α, where $\psi(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$ is the psi-function. In view of (11), the coefficient $(a^*)^T\Lambda a^*$ in the expression for the MSE of the optimal linear estimator $\hat{f}_{n,k}(a^*)$ is nearly constant for large dimensions and has little effect on the rate of convergence of the MSE of $\hat{f}_{n,k}(a^*)$. The quality of approximation (14) is illustrated in Figures 6 and 7.

The asymptotic properties (when both n and k are large) of the optimal linear estimators coincide with the properties of the MLE and hold under the same regularity conditions, as proved in [13, Sect. 7.3.3]. In particular,

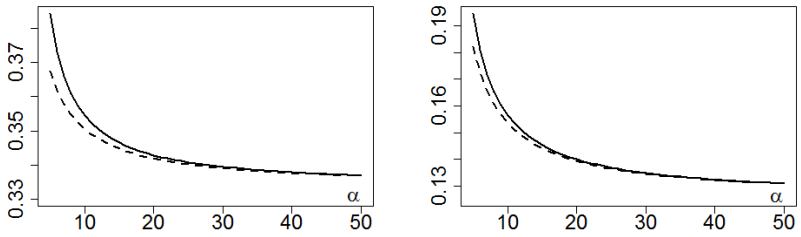

Fig. 6 The exact expression of $1/\mathbf{1}^T \Lambda^{-1}\mathbf{1}$ (solid) and the approximation (14) (dashed) for $k = 3$ (left) and $k = 8$ (right); α varies in $[5, 50]$.

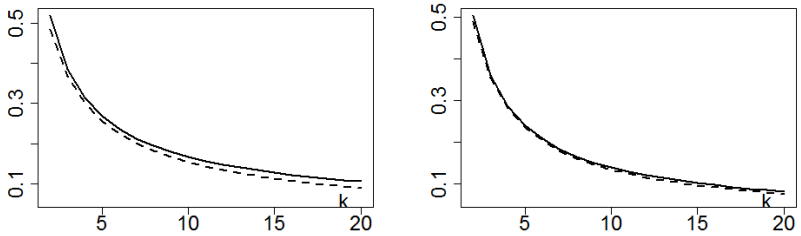

Fig. 7 The exact expression of $1/\mathbf{1}^T \Lambda^{-1}\mathbf{1}$ (solid) and the approximation (14) (dashed) for $\alpha = 5$ (left) and $\alpha = 8$ (right); as k varies in $[2, 20]$.

the optimal linear estimator $\hat{f}_{n,k}(a^*)$ is asymptotically normal (as $n \to \infty$, $k \to \infty$, $k/n \to 0$) and the mean square error $\mathrm{E}(\hat{f}_{n,k}(a^*) - f_*)^2$ asymptotically behaves like (7).

Note that the efficiency of the estimators $\hat{f}_{mle}$ and $\hat{f}_{n,k}(a^*)$ can be low if an incorrect value of α is used for computing this estimator, see [16, Sect. 2.5.2].

In practice of global optimization, the standard estimator of f_* is the current record $y_{1,n}$ computed by (2). This estimator can be written as $y_{1,n} = \hat{f}_{n,k}(e_1)$ where $e_1 = (1, 0, 0, \ldots, 0)^T$. By (11), the MSE of $y_{1,n}$ is

$$\mathrm{E}(\hat{f}_{n,k}(e_1) - f_*)^2 = \Gamma(1 + 2/\alpha)(c_0 n)^{-2/\alpha} \, (1 + o(1)) \,, \quad n \to \infty \,.$$

Asymptotic efficiency of $y_{1,n}$ is therefore

$$\mathrm{eff}(y_{1,n}) = \left[\Gamma(1 + 2/\alpha) \cdot \mathbf{1}^T \Lambda^{-1} \mathbf{1}\right]^{-1} . \tag{15}$$

In view of (14), this efficiency tends to $1/k$ if k is fixed and $\alpha \to \infty$. The asymptotic behaviour of the efficiency (15) is illustrated in Figure 8.

4 Comparison of random and quasi-random sequences

GRS algorithms compared with deterministic optimisation procedures have a very attractive feature: in GRS algorithms we can use statistical procedures for increasing efficiency of the algorithms and devising stopping rules. But do

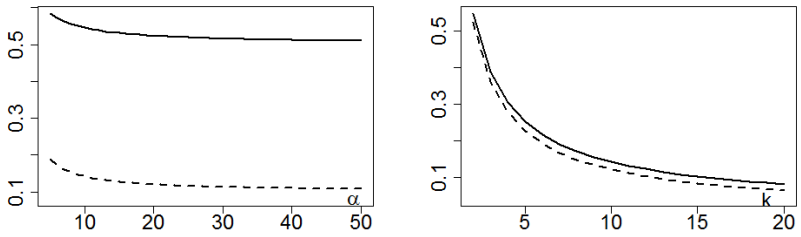

Fig. 8 Asymptotic efficiency (15) of $y_{1,n}$. Left: $k = 2$ (solid) and $k = 10$ (dashed); as α varies in $[5, 50]$. Right: $\alpha = 10$ (solid) and $\alpha = 20$ (dashed); as k varies in $[2, 20]$.

we gain much by choosing the points at random and can we improve the efficiency of GRS algorithms if we sacrifice some randomness? The answer to this question is similar to what we know from other areas of applied mathematics like estimation of integrals using Monte-Carlo methods and cubature formulas: randomness provides simplicity of methods and possibility to make statistical inferences but for small dimensions we can significantly improve efficiency by reducing randomness and making the best possible deterministic decisions; however, this is not so for large dimensions.

One possible recommendation for combining random search and deterministic procedures can be formulated as follows. First of all, if a global optimization method requires local descents then for doing this we must use standard deterministic routines like the conjugate gradient method (since local random search algorithms would never be able to compete with such methods). In the stage, where a GRS algorithm explores the whole $\mathbf{X}$ or some prospective subsets of $\mathbf{X}$, purely deterministic or quasi-random sequences of points may do this exploration more efficiently than random sequences, especially in small dimensions. If PRS will use any of the quasi-random sequence instead of random points, then this will improve the rate of convergence of PRS in low dimensions only, avoid very long waiting times with infinite expectation for getting new records (in the purely random version of PRS) and gain the reproducibility of results.

If we use appropriate semi-random sequences like a stratified sample in place of an i.i.d. sample in the PRS, then we still be able to use some of the statistical procedures. More precisely, consider a version of the PRS where the sample $\{x_1, \dots, x_n\}$ is stratified rather than independent. Assume that the distribution $P = P_U$ is uniform on $\mathbf{X}$ and the set $\mathbf{X}$ is split into m subsets of equal volume. Assume also that in each subset we generate l independent uniformly distributed points. The sample size is then $n = ml$. In particular, under the assumption $l > k$ and exactly the same assumptions about $f(\cdot)$, the estimators (9) can again be used. The accuracy of these estimator is better than the accuracy of the same estimators for the i.i.d. sample, see [13, Sect. 3.2].

We claim, however, that if the dimension d is large then the use of quasi-random points instead of purely random does not bring any advantage. Let us try to illustrate this using simulation experiments.

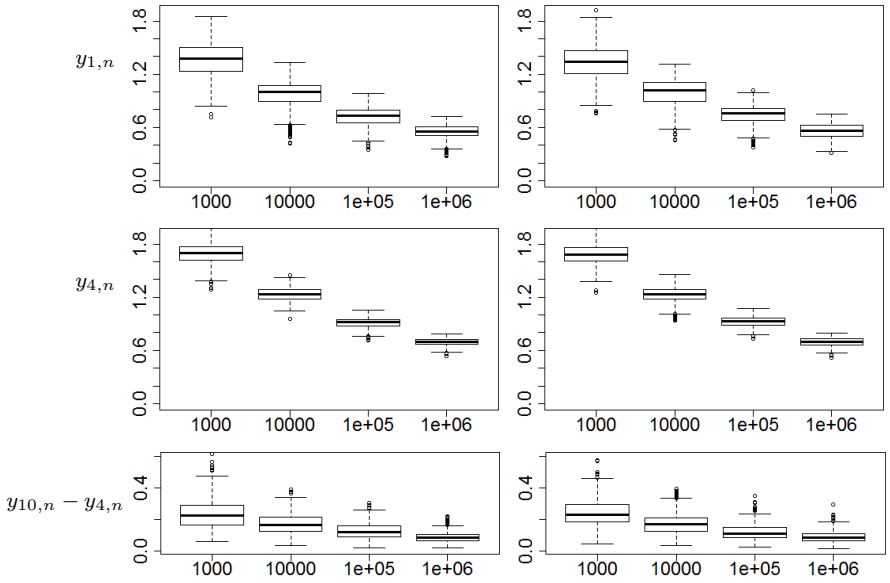

Fig. 9 Boxplot of records $y_{1,n}$ (top) and $y_{4,n}$ (middle) and the difference $y_{10,n} - y_{4,n}$ (bottom) for 500 runs of the PRS algorithm with points generated from the Sobol low-dispersion sequence (left) and the uniform distribution (right), $d = 20$.

Assume $\mathbf{X} = [0,1]^d$ and consider an algorithm of global search, in which points are create a low-dispersion sequence relative to the L_∞-metric in the multi-dimensional case. As shown in [7, Th. 6.8 and 6.9], for every dimension d and any n-point sequence $X_n = \{x_1, \ldots, x_n\}$, the dispersion (with respect to L_∞-metric ρ_∞)

$$d'(X_n) = \max_{x \in \mathbf{X}} \min_{i=1,\ldots,n} \rho_\infty(x, x_i)$$

satisfies the inequality $d'(X_n) \geq 0.5 n^{-1/d}$ and there exists a sequence X_n^* such that

$$\lim_{n \to \infty} n^{1/d} d'(X_n^*) = 1/(2 \ln 2).$$

This means that the rate of covering of the set $\mathbf{X}$ by points from the best low-dispersion sequence has the order $O(n^{-1/d})$, which coincides with the rate achieved by PRS with uniform distribution P_U.

Using simulation we now compare the performance of the PRS algorithm with $P = P_U$ and quasi-random points generated from the Sobol low-dispersion sequence. We consider the minimization problem with the objective function $f(x) = \sum_{s=1}^{d} (x_s - |\cos(s)|)^2$ and the set $\mathbf{X} = [0,1]^d$ in the dimension $d = 20$. It is easy to see that the global minimum $f_* = 0$ is attained at the internal point $x_* = (|\cos(1)|, \ldots, |\cos(d)|)$. For each run of the PRS algorithm, we obtain n points and compute the records $y_{1,n}$ and $y_{j,n}$ $(j > 1)$, for $n = 10^3, 10^4, 10^5, 10^6$. We repeat this procedure 500 times and depict the obtained records as box-plots in Figure 9.

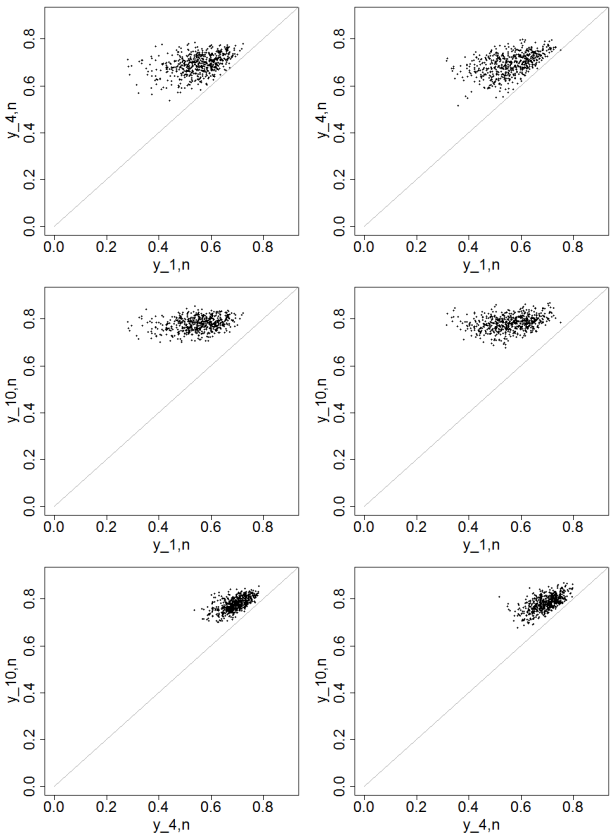

Fig. 10 Scatterplots of records $y_{1,n}$, $y_{4,n}$ and $y_{10,n}$ with $n = 10^6$ for 500 runs of the PRS algorithm with points generated from the Sobol low-dispersion sequence (left) and the uniform distribution (right), $d = 20$.

We can see that the performance of the PRS algorithm with points generated from the Sobol low-dispersion sequence and the uniform distribution is very similar. We also note that the variability of $y_{1,n}$ is larger than variability of $y_{4,n}$ and variability of the difference $y_{10,n} - y_{4,n}$ is small. Figure 9 shows additionally that the convergence of the record $y_{1,n}$ to the minimal value $f_* = 0$ is very slow, with the rate $O(1/n^{2/d})$, as n increases.

In Figures 10 and 11 we show the joint empirical distribution of the records $y_{k,n}$ with $n = 10^6$ and Figure 12 shows the averaged values of these records. We can see that the records $y_{1,n}$ and $y_{10,n}$ are almost independent and records $y_{4,n}$ and $y_{10,n}$ are higly correlated. We also note that the record $y_{4,n}$ is close to $y_{1,n}$ in few simulation trials and the record $y_{2,n}$ is close to $y_{1,n}$ in many simulation trials. These figures also show that the global minimum $f_* = 0$ is very far from the cloud of points corresponding to the joint empirical distribution of records.

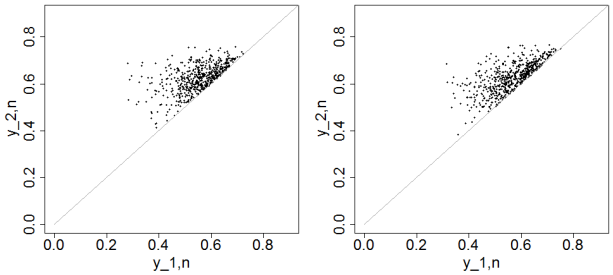

Fig. 11 Scatterplots of records $y_{1,n}$, $y_{2,n}$ with $n = 10^6$ for 500 runs of the PRS algorithm with points generated from the Sobol low-dispersion sequence (left) and the uniform distribution (right), $d = 20$.

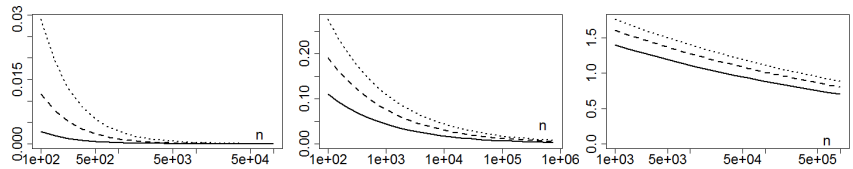

Fig. 12 Averaged records $y_{1,n}$ (solid), $y_{4,n}$ (dashed) and $y_{10,n}$ (dotted), where averaging is taken over 500 runs of the PRS algorithm with points generated from either the Sobol low-dispersion sequence or the uniform distribution, in the dimension $d = 2$ (left), $d = 5$ (middle) and $d = 20$ (right).

5 Conclusions

We have investigated the rate of convergence of a general global random search algorithm. We have shown that if the dimension of the feasible domain is large then it is virtually impossible to guarantee that the global minimizer is reached by a general global random search algorithm. We have studied precision of statistical estimates of the global minimum in the case of large dimensions. We have shown that these estimates suffer the curse of dimensionality. Finally, using extensive computer study we have demonstrated that the use of quasi-random points in place of the random ones does not give any visible advantage in large dimensions.

Acknowledgements

The work of the first author was partially supported by the SPbSU project No. 6.38.435.2015 and the RFFI project No. 17-01-00161. The work of the second author was supported by the Russian Science Foundation, project No. 15-11-30022 'Global optimization, supercomputing computations, and applications'. The work of the third author was supported by the Research Council of Lithuania under Grant No. MIP-051/2014.

References

1. Auger, A., Hansen, N.: Theory of evolution strategies: a new perspective. Theory of Randomized Search Heuristics: Foundations and Recent Developments pp. 289–325 (2010)
2. Cooke, P.: Optimal linear estimation of bounds of random variables. Biometrika **67**(1), 257–258 (1980)
3. De Haan, L.: Estimation of the minimum of a function using order statistics. Journal of the American Statistical Association **76**(374), 467–469 (1981)
4. De Haan, L., Peng, L.: Comparison of tail index estimators. Statistica Neerlandica **52**(1), 60–70 (1998)
5. Dette, H., Pepelyshev, A., Zhigljavsky, A.: Optimal designs in regression with correlated errors. The Annals of Statistics **44**(1), 113–152 (2016)
6. Nevzorov, V.B.: Records: mathematical theory. American Mathematical Soc. (2001)
7. Niederreiter, H.: Random number generation and quasi-monte carlo methods, cbms-nsf reg. In: Conf. Series Appl. Math, vol. 63 (1992)
8. Pintér, J.n.: Convergence properties of stochastic optimization procedures. Optimization **15**(3), 405–427 (1984)
9. Solis, F.J., Wets, R.J.B.: Minimization by random search techniques. Mathematics of operations research **6**(1), 19–30 (1981)
10. Zhigljavsky, A.: Monte-Carlo methods in global optimization, PhD thesis. Leningrad University (1981)
11. Zhigljavsky, A.: Mathematical Theory of Global Random Search. Leningrad University Press (1985). In Russian
12. Zhigljavsky, A.: Branch and probability bound methods for global optimization. Informatica **1**(1), 125 – 140 (1990)
13. Zhigljavsky, A.: Theory of global random search. Kluwer Academic Publishers (1991)
14. Zhigljavsky, A., Hamilton, E.: Stopping rules in k-adaptive global random search algorithms. Journal of Global Optimization **48**(1), 87–97 (2010)
15. Zhigljavsky, A., Žilinskas, A.: Methods of seeking a global extremum. Nauka, Moscow (1991)
16. Zhigljavsky, A., Žilinskas, A.: Stochastic Global Optimization. Springer (2008)
17. Zhigljavsky, A.A.: Semiparametric statistical inference in global random search. Acta Applicandae Mathematica **33**(1), 69–88 (1993)
18. Žilinskas, A.: A statistical model-based algorithm for black-box multi-objective optimisation. International Journal of System Science **45**(1), 82–92 (2014)
19. Žilinskas, A., Zhigljavsky, A.: Branch and probability bound methods in multi-objective optimization. Optimization Letters **10**(2), 341–353 (2016)