Statistical estimation in global random search algorithms in case of large dimensions

Andrey Pepelyshev[1], Vladimir Kornikov[2], and Anatoly Zhigljavsky[1,3]

[1]School of Mathematics, Cardiff University, Cardiff, CF24 4AG, UK
[2]Faculty of Applied Mathematics, St.Petersburg State University
[3]Lobachevsky Nizhny Novgorod State University, 23 Prospekt Gagarina, 603950, Nizhny Novgorod, Russia
pepelyshevan@cardiff.ac.uk, vkornikov@mail.ru, zhigljavskyaa@cardiff.ac.uk

Abstract. We study asymptotic properties of optimal statistical estimators in global random search algorithms when the dimension of the feasible domain is large. The results obtained can be helpful in deciding what sample size is required for achieving a given accuracy of estimation.

Keywords: Global optimization, extreme value, random search, estimation of end-point

1 Introduction

We consider the problem of global minimization $f(x) \to \min_{x \in \mathbf{X}}$, where $f(\cdot)$ is the objective function and $\mathbf{X} \subset \mathbb{R}^d$ is a feasible domain. The set $\mathbf{X}$ is a compact set with non-empty interior and the 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 $f(x_*) = f_*$.

If the objective function is given as a 'black box' computer code and there is no information about this function available of Lipschitz type, then good stochastic approaches often perform better than deterministic algorithms, especially in large dimensions; see for example [4,3]. Moreover, stochastic algorithms are usually simpler than deterministic algorithms.

A general Global Random Search (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 $f(x_1), \ldots, f(x_{j-1})$.

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 enables detailed examination of this algorithm.

2 Statistical inference about f_* in pure random search

Consider a PRS with $x_j \sim P$. Statistical inference about f_* can serve for the following purposes: (i) devising specific GRS algorithms like the branch and

probability bounds methods, see [2,6] and [4, Sect. 4.3], (ii) constructing stopping rules, see [5], and (iii) increasing efficiency of population-based GRS methods, see discussion in [3, Sect. 2.6.1]. Moreover, the use of statistical inferences in GRS algorithms can be very helpful in solving multi-objective optimization problems with non-convex objectives, see [6].

Since the points x_j in PRS are independent identically distributed (i.i.d.) with distribution P, the elements of the sample $Y = \{y_1, \dots, y_n\}$ with $y_j = f(x_j)$ are i.i.d. with cumulative distribution function (c.d.f.) $F(t) = \Pr\{x \in \mathbf{X} : f(x) \leqslant t\} = \int_{f(x)\leqslant t} P(dx) = P(W(t - f_*))$, where $t \geqslant f_*$ and $W(\delta) = \{x \in \mathbf{X}\colon f(x) \leqslant f_* + \delta\}$, $\delta \geqslant 0$. Since the analytic form of $F(t)$ is either unknown or intractable (unless f is very simple), for making statistical inferences about f_* we need to use the asymptotic approach based the record values of the sample Y. It is known that (i) the asymptotic distribution of the order statistics is unambiguous, (ii) the conditions on $F(t)$ and $f(\cdot)$ when this asymptotic law works are very mild and typically hold in real-life problems, (iii) for a broad class of functions $f(\cdot)$ and distributions P, the c.d.f. $F(t)$ has the representation

$$F(t) = c_0(t - f_*)^\alpha + \mathrm{o}((t - f_*)^\alpha),\ t \downarrow f_*, \tag{1}$$

where c_0 and α are some positive constants. The value of c_0 is not important but the value of α is essential. The coefficient α is called 'tail index' and its value is usually known, as discussed below.

Let η be a random variable which has c.d.f. $F(t)$ and $y_{1,n} \leqslant \ldots \leqslant y_{n,n}$ be the order statistics for the sample Y. By construction, f_* is the lower endpoint of the random variable η.

One of the most important result in the theory of extreme order statistics states (see e.g. [3, Sect. 2.3]) that if (1) 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.

In PRS we can usually have enough knowledge about $f(\cdot)$ to get the exact value of the tail index α. Particularly, the following statement holds: if the global minimizer x_* of $f(\cdot)$ is unique and $f(\cdot)$ is locally quadratic around x_* then the representation (1) holds with $\alpha = d/2$. However, if the global minimizer x_* of $f(\cdot)$ is unique and $f(\cdot)$ is not locally quadratic around x_* then the representation (1) may hold with $\alpha = d$. See [4] for a comprehensive description of the related theory.

The result that α has the same order as d when d is large implies the phenomena called 'the curse of dimensionality'. Let us first illustrate this curse of dimensionality on a simple numerical example.

3 Numerical examples

We investigate 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: $\mathbf{X} = \{x \in$

$\mathbb{R}^d : ||x|| \leq 1\}$. The minimal value is $f_* = -1$ and the global minimizer $z_* = (-1, 0, \ldots, 0)^T$ is located at the boundary of $\mathbf{X}$. Consider the PRS algorithm with points x_j generated from the uniform distribution P_U on $\mathbf{X}$.

Let us give some numerical values. In a simulation with $n = 10^3$ and $d = 20$, we have received $y_{1,n} = -0.6435$, $y_{2,n} = -0.6107$, $y_{3,n} = -0.6048$ and $y_{4,n} = -0.6021$. In a simulation with $n = 10^5$ and $d = 20$, we have obtained $y_{1,n} = -0.7437$, $y_{2,n} = -0.7389$, $y_{3,n} = -0.7323$ and $y_{4,n} = -0.726$. In Figure 1 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 see that the difference $y_{k,n} - y_{1,n}$ is much smaller than the difference $y_{1,n} - f_*$; that demonstrates that the problem of estimating the minimal value of f_* is very hard.

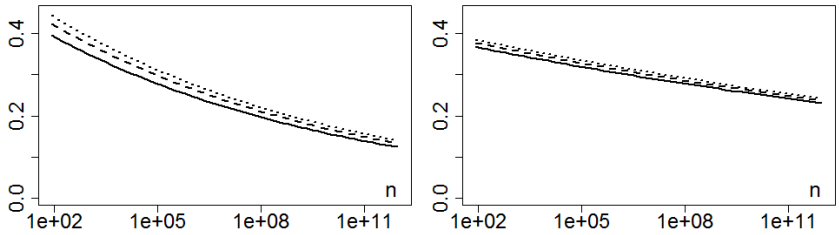

Fig. 1. Differences $y_{1,n} - f_*$ (solid), $y_{4,n} - f_*$ (dashed) and $y_{10,n} - f_*$ (dotted), where $y_{k,n}$, $k = 1, 4, 10$, are records of evaluations of the function $f(x) = e_1^T x$ at points $x_1, \ldots, x_n$ with uniform distribution in the unit hyperball in the dimension $d = 20$ (left)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

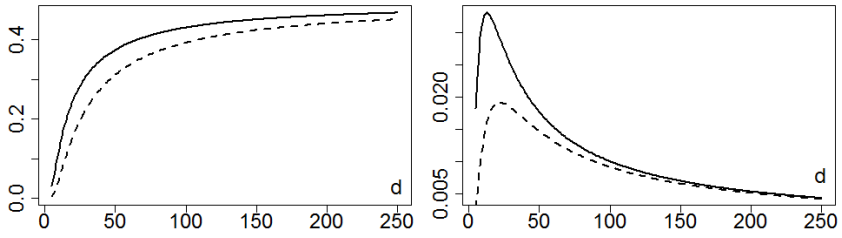

Fig. 2. The difference $y_{1,n} - f_*$ (left) and $y_{10,n} - y_{1,n}$ (right) for $n = 10^6$ (solid) and $n = 10^{10}$ (dashed), where $y_{j,n}$ is the j-th record of evaluations of the function $f(x) = e_1^T x$ at points $x_1, \ldots, x_n$ with uniform distribution in the unit hyperball in the dimension d; d varies in $[5, 250]$.

In Figure 2 we observe that the difference $y_{1,n} - f_*$ increases as the dimension d grows, for fixed n. Thus, the minimization problem becomes more difficult in

larger dimensions. Also, Figure 2 shows that difference $y_{10,n} - y_{1,n}$ is much smaller than the difference $y_{1,n} - f_*$.

Consider now the optimal linear estimator based on the use of k order statistics; this estimator, as shown in [1,4], has the form

$$\hat{f}_{n,k} = \frac{1}{C_{k,\alpha}} \sum_{i=1}^{k} \frac{u_i}{\Gamma(i+2/\alpha)} y_{i,n}\,, \tag{2}$$

where $\Gamma(\cdot)$ is the Gamma-function,

$$u_i = \begin{cases} \alpha+1\,, & i=1\,,\\ (\alpha-1)\Gamma(i), & i=2,\dots,k-1\,,\\ (\alpha-\alpha k-1)\Gamma(k)\,, & i=k\,, \end{cases}$$

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

If the representation (1) holds, then for given k and α and as $n\to\infty$, the estimator $\hat{f}_{n,k}$ is a consistent and asymptotically unbiased estimator of f_* and its asymptotic mean squared error $E(\hat{f}_{n,k} - f_*)^2$ has maximum possible rate of convergence in the class of all consistent estimators including the maximum likelihood estimator of f_*, as shown in [4, Ch 7]. This mean squared error has the following asymptotic form:

$$E(\hat{f}_{n,k} - f_*)^2 = C_{k,\alpha}(c_0 n)^{-2/\alpha}\,(1+o(1))\,,\quad n\to\infty\,. \tag{3}$$

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

$$C_{k,\alpha} \simeq \frac{1}{k} + \frac{2(\psi(k) - 1 + 1/k)}{\alpha k}\,, \tag{4}$$

for large α, where $\psi(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$ is the psi-function. Quality of this approximation is illustrated on Figures 3 and 4.

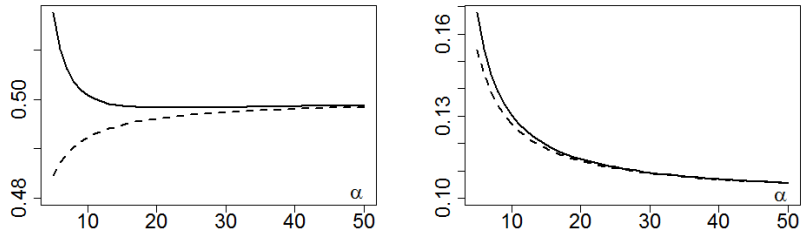

Fig. 3. The exact expression of $C_{k,\alpha}$ (solid) and the approximation (4) (dashed) for $k=2$ (left) and $k=10$ (right); α varies in $[5,50]$.

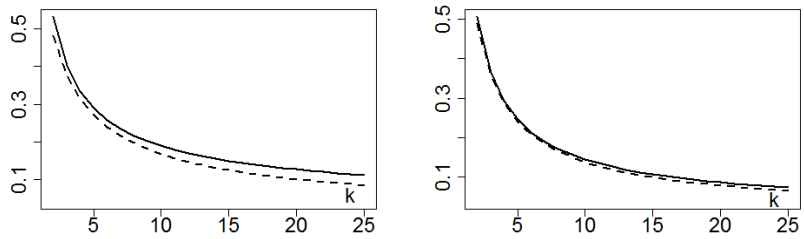

Fig. 4. The exact expression of $C_{k,\alpha}$ (solid) and the approximation (4) (dashed) for $\alpha = 4$ (left) and $\alpha = 7$ (right); as k varies in $[2, 25]$.

In practice of global optimization, the standard estimator of f_* is the current record $y_{1,n} = \min_{i=1,\dots,n} f(x_i)$. Its asymptotic mean squared error 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 $\mathrm{eff}(y_{1,n}) = C_{k,\alpha}/\Gamma(1 + 2/\alpha)$. This efficiency is illustrated on Fig. 5.

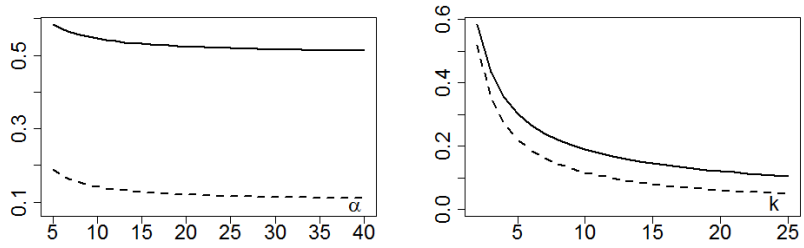

Fig. 5. Asymptotic efficiency $\mathrm{eff}(y_{1,n})$ of $y_{1,n}$. Left: $k = 2$ (solid) and $k = 10$ (dashed); as α varies in $[5, 40]$. Right: $\alpha = 5$ (solid) and $\alpha = 25$ (dashed); as k varies in $[2, 20]$.

GRS algorithms have a very attractive feature in comparison with deterministic optimisation procedures. Specifically, in GRS algorithms we can use statistical procedures for increasing efficiency of the algorithms and devising stopping rules. But do we lose much by choosing the points at random? We claim 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 some simulation experiments.

Using simulation studies we now investigate the performance of the PRS algorithm with $P = P_U$ and quasi-random points generated from the Sobol low-dispersion sequence. We examine 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 $d = 15$. In this problem, 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 generate n points and compute the records $y_{1,n}$ and $y_{2,n}$, for $n = 10^3, 10^4, 10^5, 10^6$. We repeat this procedure 500 times and show the obtained records as boxplots in Figure 6.

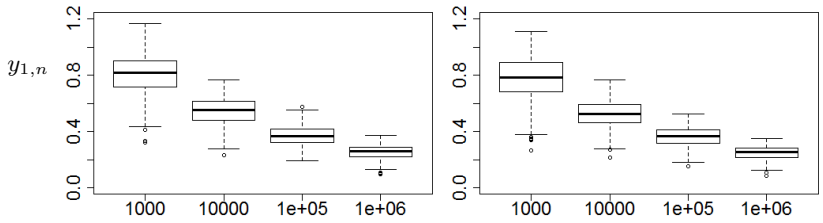

Fig. 6. Boxplot of records $y_{1,n}$ for 500 runs of the PRS algorithm with points generated from the Sobol low-dispersion sequence (left) and the uniform distribution (right), $d = 15$.

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 the difference $y_{10,n} - y_{4,n}$ has a small variability.

Acknowledgements. The work of the third author was supported by the Russian Science Foundation, project No. 15-11-30022 'Global optimization, supercomputing computations, and applications'.

References

1. Zhigljavsky, A.: Mathematical Theory of Global Random Search. Leningrad University Press (1985), in Russian
2. Zhigljavsky, A.: Branch and probability bound methods for global optimization. Informatica 1(1), 125 – 140 (1990)
3. Zhigljavsky, A., Žilinskas, A.: Stochastic Global Optimization. Springer (2008)
4. Zhigljavsky, A.: Theory of global random search. Kluwer Academic Publishers (1991)
5. Zhigljavsky, A., Hamilton, E.: Stopping rules in k-adaptive global random search algorithms. Journal of Global Optimization 48(1), 87–97 (2010)
6. Zilinskas, A., Zhigljavsky, A.: Branch and probability bound methods in multi-objective optimization. Optimization Letters pp. 1–13 (2014)