INFORMATICA, 2016, Vol. 27, No. 2, 229–256

DOI: http://dx.doi.org/10.15388/Informatica.2016.83

Stochastic Global Optimization: A Review on the Occasion of 25 Years of *Informatica*

Antanas ŽILINSKAS[1]*, Anatoly ZHIGLJAVSKY[2,3]
[1] *Institute of Mathematics and Informatics, Vilnius University*
Akademijos 4, LT-08663, Vilnius, Lithuania
[2] *School of Mathematics, Cardiff University, Cardiff CF21 1AG, UK*
[3] *Lobachevsky Nizhny Novgorod State University, 23 Prospekt Gagarina*
603950 Nizhny Novgorod, Russia
e-mail: antanas.zilinskas@mii.vu.lt, zhigljavskyaa@cardiff.ac.uk

Received: December 2015; accepted: April 2016

Abstract. This is a survey of the main achievements in the methodology and theory of stochastic global optimization. It comprises two complimentary directions: global random search and the methodology based on the use of stochastic models about the objective function. The main attention is paid to theoretically substantiated methods and mathematical results proven in the last 25 years.

Key words: global optimization, statistical models, extreme value statistics, random search, kriging.

1. Introduction

Global optimization (GO) is a broad and active field of research including mathematical analysis of problems, development of algorithms and software, and applications of the corresponding software to various real world problems. In the present paper we review those global optimization methods (for continuous problems) which are based on stochastic methods and stochastic models of the objective function. Let us note that the founding editor of *Informatica* Jonas Mockus was one of the originators of these research directions.

The roots of mathematical treatment of optimization problems can be traced even in Antiquity. Some basic concepts of optimization theory were proposed by the classics of calculus: Lagrange, Cauchy and others. However, the global optimization methods are inseparable from their presentation and analysis as computer algorithms. One of the first mathematically substantiated method of random search for the global optimum was proposed by Jonas Mockus (1963) in the proceedings of the symposium on multiextremal problems organized by himself. The original idea to construct a global optimization algorithm using a statistical model of objective functions was proposed in Kushner (1962), and

*Corresponding author.

later Lithuanian researchers started the active development of the underlying theory and corresponding algorithms; see Mockus (1972), Žilinskas and Mockus (1972). The other group of researchers who also actively started the investigation of this approach (using the title information methods) worked in Nizhnij Novgorod (Neimark and Strongin, 1966; Strongin, 1969). The results of early development are well presented in several monographs (Mockus, 1967, 1988; Strongin, 1978; Törn and Žilinskas, 1989; Zhigljavsky, 1985; Žilinskas, 1986), and will not be reviewed here.

Our review covers the period from the nineties of the last century. Then the development of global optimization accelerated by the establishment of the *Journal of Global Optimization* in 1991. At the same time, in 1990, the journal Informatica was established where many important results of the considered approaches to global optimization have been published.

Formally, the problem of global minimization can be stated as follows:

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

where $f(\cdot)$ is the objective function and $\mathbf{A}$ is a feasible region. Let x_* be a global minimizer of $f(\cdot)$; that is, x_* is a point in $\mathbf{A}$ such that $f(x_*) = f_*$ where $f_* = \min_{x \in \mathbf{A}} f(x)$. Global optimization problems are usually formulated so that the structure of the feasible region $\mathbf{A}$ is relatively simple; this can be done on the expense of increased complexity of the objective function (for example, using penalty functions). The objective function $f(x)$ will be assumed continuous; further assumptions about the objective function and feasible region will be made where appropriate.

A global minimization algorithm is a rule for constructing a sequence of points $x_1, x_2, \ldots$ in $\mathbf{A}$ such that the sequence of record values

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

approaches the minimum f_* as k increases (for convenience, we will assume $y_{o,o} = +\infty$). In addition to approximating the minimal value f_*, one often needs to approximate at least one of the minimizers x_*.

Deterministic and stochastic global optimization methods constitute two separate classes of methods (despite the fact that many ideas, like the branch and bound idea, are shared by many methods in these two classes). For the state of the art in the theory and methodology of deterministic global optimization we refer to Floudas (2000), Horst *et al.* (2000).

Stochastic global optimization methods, which the present paper is devoted to, are methods for solving the global optimization problem incorporating probabilistic (stochastic) elements, either in the problem data (the objective function, the constraints, etc.), or in the algorithm itself, or in both.

If the objective function is given as a 'black box' computer code, the optimization problem is especially difficult and stochastic approaches can often deal with problems of this kind much easier and more efficiently than the deterministic algorithms.

We distinguish the following three classes of stochastic optimization algorithms.

- *Global random search* (*abbreviated below by GRS*). GRS algorithms involve random decisions in the process of choosing the observation points. Theory and methodology of GRS is reviewed in Section 2.
- *Stochastic assumptions about the objective function*. Much research have been done in stochastic global optimization where stochastic assumptions about the objective function are used in a manner similar to how the Lipschitz condition is used in deterministic algorithms. A typical example of a stochastic assumption of this kind is the postulation that $f(\cdot)$ is a realization of a certain stochastic process. This part of stochastic optimization is dealt with in Section 3 of this article.
- *Heuristics or meta-heuristics*. Many stochastic optimization algorithms where randomness is involved have been proposed heuristically. Some of these algorithms are based on analogies with natural processes; the well-known examples are simulated annealing and genetic algorithms. Heuristic global optimization algorithms are very popular in applications, especially in discrete optimization problems. Unfortunately, there is a large gap between practical efficiency of heuristic optimization algorithms and their theoretical rigour. There is a lot of literature on heuristic algorithms but this literature will not be reviewed in this paper.

2. Global Random Search

2.1. *Main Concepts*

2.1.1. *General*

A generic GRS algorithm assumes that a sequence of random points $x_1, x_2, \ldots, x_n$ is generated, where for each $j \geqslant 1$ the point x_j has some probability distribution P_j (we write this $x_j \sim P_j$). For each $j \geqslant 2$, the distribution P_j may depend on the previous points $x_1, \ldots, x_{j-1}$ and on the results of the objective function evaluations at these points (the function evaluations may not be noise-free). The number of points n (the stopping rule) can be either deterministic or random and may depend on the results of function evaluation at the points $x_1, \ldots, x_n$, see Section 2.3.4. In order for an algorithm to be classified as a GRS algorithm, at least one of the distributions P_j should be non-degenerate (so that at least one of x_j is a random point in $\mathbf{A}$).

Attractive features of GRS: (a) the structure of GRS algorithms is usually simple; (b) these algorithms are often rather insensitive to the irregularity of the objective function behaviour, to the shape of the feasible region, to the presence of noise in the objective function evaluations, and even to the growth of dimensionality; (c) it is very easy to construct GRS algorithms guaranteeing theoretical convergence.

Drawbacks of GRS: (a) practical efficiency of GRS algorithms often depends on a number of parameters, but the problem of the choice of these parameters frequently has little relevance to the theoretical results concerning the convergence of the algorithms; (b) for many GRS algorithms an analysis on good parameter values is lacking or just impossible; (c) the convergence rate can be painfully slow, see Section 2.2.

Improving the convergence rate (or efficiency of the algorithms) is a problem that much research in the theory of global random search is devoted to.

Historical remarks. The first theoretical study of properties of the simplest GRS algorithms was performed in Brooks (1958, 1959). In the sixties, seventies and the beginning of eighties of the last century, the methodology of GRS have been extensively developed by Leonard Andreevich Rastrigin and his school in Riga; see, for example, Rastrigin (1964, 1968). Many methodological developments of L.A. Rastrigin have been reinvented in the West by scholars working on meta-heuristics in GRS years after the death of L.A. Rastrigin.

The first proper theoretical study of convergence of a generic GRS algorithm was made in Solis and Wets (1981); see also Section 2.2. A very comprehensive theoretical investigation of different probabilistic and statistical schemes related to the GRS was made in Zhigljavsky (1985, 1991), which are Russian and English (much extended) versions of the same book. The results published in Zhigljavsky (1991) are still largely unknown to specialists on global optimization and especially to those working on meta-heuristics.

2.1.2. *Main Principles of GRS*

A very large number of specific global random search algorithms exist, but only a few main principles form their basis. These principles can be summarized as follows:

P1: random sampling of points at which $f(\cdot)$ is evaluated,
P2: random covering of the space,
P3: combination with local optimization techniques,
P4: use of different heuristics including cluster-analysis techniques to avoid clumping of points around a particular local minima,
P5: more frequent selection of new trial points in the vicinity of 'good' previous points,
P6: use of statistical inference, and
P7: decrease of randomness in the selection rules for the trial points.

Principle P1 classifies an optimization algorithm as a GRS algorithm. P2 makes sure that the search is global while P3 looks after local improvements in the process of search. Good local solutions improve the record values (2) which work as thresholds for the new points and are used in many algorithms for defining the prospectiveness of subsets of **A** for further search.

Right balance between globality and locality of search 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 descent) and on efficient use of all available information (prior information and information obtained during the process of search) about $f(\cdot)$ and **A**. Processing of information about $f(\cdot)$ and **A**, obtained during the process of search, can be achieved by the methodologies associated with Principles P4, P5 and P6. The standard reference for P4 is Kan and Timmer (1987); P6 is considered in Section 2.3. Principle P7 is discussed in Section 2.1.4. It is argued in that section that, in a certain sense, any decrease of randomness in the choice of points x_j leads to better (that is, more efficient) optimization algorithms.

2.1.3. *Several Important Classes of GRS Algorithms*

Consider a general GRS algorithm with $x_j \sim P_j$, $j = 1, 2, \ldots$. Construction of a particular GRS algorithm involves construction of a distribution P_j (based on all available information at time j) plus a stopping rule. For a general evolutionary GRS algorithm, in the so-called population-based algorithms the distributions P_j are updated regularly; for population-based algorithms, this updating is made only after a certain number of points with previous distribution have been generated. Let us also mention four popular classes of GRS methods where the updating of the distributions P_j is very simple.

PRS (*Pure Random Search*). Random points $x_1, x_2, \ldots$ are independent and have the same distribution: $x_j \sim P$ (so that $P_j = P$ for all j).

MGS (*Markovian Global Search*). The distribution P_j depends on x_{j-1} and the objective function value at this point but does not depend on the values of $f(\cdot)$ computed earlier.

PAS (*Pure Adaptive Search*). P_j is uniform on the set $\mathbf{A}_j = \{x \in \mathbf{A} : f(x) \leqq y_{0,j-1}\}$, where $y_{0,j-1}$ is the record value at time $j-1$, see (2).

RMS (*Random Multi-Start*). Local descents are performed from a number of random points in $\mathbf{A}$.

PBS (*Population-Based Search*). Similar to MGS but groups (populations) of points are probabilistically transformed into subsequent groups rather than points to points in MGS.

Simplicity of PRS allows detailed investigation of this algorithm, see Sections 2.2 and 2.3. Another very attractive feature of PRS is its worst-case optimality (when P is the uniform distribution on $\mathbf{A}$). Indeed, any sophistication in the way of construction of the distributions P_j leads to a waste of efforts in the worst-case scenario; this is a probabilistic version of the celebrated result of A. Sukharev, see Sukharev (1971, 1972) and Chapter 4 in Sukharev (2012).

MGS algorithms, including the celebrated simulated annealing, are more clever than the primitive PRS. At the same time, MGS algorithms are simple enough (they are simply Markov Chains) to allow a thorough theoretical investigation. There are very many papers on simulated annealing and other MGS but the practical efficiency of these algorithms is rather poor: indeed, MGS algorithms are too myopic and they waste almost all information about the objective function which is collected in the process of search. Note that some very advanced theoretical results concerning the so-called monotonous MGS (when the last observation point is always the point of current record) are contained in basically unknown papers of Alexey Tikhomirov from Novgorod State University, see, for example Tikhomirov (2006, 2007), Tikhomirov *et al.* (2007); see also Section 3.4 in Zhigljavsky and Žilinskas (2008), where some results of A. Tikhomirov are reviewed. MGS algorithms can be naturally generalized so that the distributions P_j depend not only on x_{j-1} and $f(x_{j-1})$ but also on the current record $y_{0,j-1}$ and the point where this record has been computed. Practically, these algorithms can be much more efficient than the pure MGS algorithm but the theoretical study of these algorithms is generally very complicated in view of dependence of P_j on all previous points. An obvious exception is PAS considered next.

The idea of PAS is extremely simple and different versions of PAS have been known long before publication of Patel *et al.* (1989), where PAS has received the name and high (but short-lived!) popularity, see for example Zabinsky and Smith (1992), Baritompa *et al.* (1995) and the book (Zabinsky, 2003) almost exclusively devoted to PAS. This high popularity of PAS is due to the fact that under some natural assumptions about $f(\cdot)$ and $\mathbf{A}$, PAS has exponential rate of convergence for any dimension of $\mathbf{A}$. This popularity was short-lived because people have quickly realized that generating random points in the sets $\mathbf{A}_j = \{x \in \mathbf{A} : f(x) \leqq y_{0,j-1}\}$ is extremely difficult. In particular, if to obtain a uniform random point in $\mathbf{A}_j$ we generate independent random points in $\mathbf{A}$ and wait until the first point arrives to $\mathbf{A}_j$, then, as will be discussed in Section 2.3, the expected waiting time is infinite, even for $j = 2$. There is, however, no other clear way of getting random points in $\mathbf{A}_j$.

RMS is an extremely popular algorithm in practical optimization. It is very clear and very easy to program. Efficiency of RMS depends on the complexity of computing derivatives of $f(\cdot)$ and other properties of $f(\cdot)$ such as the number of its local minimizers and the volumes of different regions of attraction of local minimizers. It is advisable to use the clustering heuristic of Kan and Timmer (1987) in order to avoid clumping of points around local minimizers. Concerning theoretical studies of RMS, we are only aware of one good paper (Zieliński, 1981), see also Section 4.5 in Zhigljavsky (1991) and Section 2.6.2 in Zhigljavsky and Žilinskas (2008). Under some natural assumptions concerning the volumes of the regions of attraction of local minimizers, R. Zieliński has derived a way of making statistical inferences about the number of local minimizers after some number of minimizers have been found (most of them, a few times).

In PBS methods, populations (that is, bunches of points) evolve rather than individual points. There is a lot of literature on PBS but the majority of publications devoted to PBS deal with metaheuristics rather than theory and generic methodology. Several probabilistic models where populations are associated with probability distributions are proposed and investigated in Chapter 5 of Zhigljavsky (1991). Some additional insights into the theoretical understanding of the asymptotic behaviour of PBS methods is given in Section 3.5 of Zhigljavsky and Žilinskas (2008). It is shown there that similar to the MGS methods, where the so-called Gibbs distributions are of fundamental importance (they are the limiting measures for the points x_j), eigen-measures of certain non-linear integral operators play a similar role in many PBS. Space limitation refrains us from talking more here about this fascinating direction of research which is not well known to specialists in GO.

2.1.4. *Choice of Points: Random or Non-Random?*

There are many attractive features of good GRS methods but how important is randomness of the points x_j? In another words, what do we gain by choosing these 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 you find in other areas of applied mathematics like Monte Carlo methods for estimation of integrals: we gain simplicity of the methods and a possibility to make statistical inferences but if we care more about efficiency (that is, the rate of convergence) then we have to sacrifice (in a clever way) as much randomness as we possibly can.

First of all, we need to perform local descents (which are parts of many GRS algorithms) using standard deterministic routines like the conjugate gradient method (local random search algorithms would never be able to compete with such methods). In the global stage of GRS methods, where we explore the whole **A** or some prospective subsets of **A**, purely deterministic or quasi-random sequences of points would do this exploration much more efficiently than random sequences. In particular, if in place of random points in PRS we use any of the quasi-random sequences (either low-discrepancy or even better low-dispersion) well described in a wonderful book (Niederreiter, 2010), then we will (i) dramatically improve the rate of convergence of PRS investigated in Section 2.2.2, (ii) avoid very long waiting times with infinite expectation for getting new records (that would cure somehow GRS methods like PAS), and (iii) gain the reproducibility of results. If we use some semi-random sequences like the stratified sample in place of i.i.d. sample in the PRS, then we will still be able to use some of the statistical procedures outlined below in Section 2.3. More precisely, consider a version of PRS where the sample $\{x_1, \dots, x_n\}$ is stratified rather than independent. Assume that the distribution $P = P_U$ is uniform on **A** and the set **A** 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 $f(\cdot)$, exactly the same estimator (7) can again be used. The accuracy of this estimator is better than the accuracy of the same estimator for the case of an independent sample, see Zhigljavsky (1991, Section 3.2).

2.2. *Convergence and Rate of Convergence of GRS Algorithms*

2.2.1. *Convergence*

In the nineteen seventies and eighties (when there were only very few results known on stochastic methods of global optimization), a number of papers were published establishing sufficient conditions for convergence of GRS algorithms; see, for example, Solis and Wets (1981) and Pintér (1984) and Auger and Hansen (2010) and Section 9.4 in Tempo *et al.* (2012) for a much more modern discussion. The main idea in most of these, and in many other results on convergence of GRS algorithms, is the classical, in probability theory, ‘zero-one law’. The following theorem stated and proved in Zhigljavsky (1991, Section 3.2) (in a more general form), illustrates this technique in a rather general setup.

Theorem 1. *Let* **A** *be a compact set and* $f(\cdot)$ *be a continuous function on* **A** *satisfying the Lipschitz condition. Assume that*

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

for any $\varepsilon > 0$, *where* $q_j(\varepsilon) = \inf P_j(B(x, \varepsilon))$, *with* $B(x, \varepsilon) = \{z \in \mathbf{A}\colon \|z - x\| \leqq \varepsilon\}$; *the infimum in the expression for* $q_j(\varepsilon)$ *is taken over all* $x \in \mathbf{A}$, *all possible previous evaluation points and the results of the objective function evaluations at them. Then, for any* $\delta > 0$, *the sequence of points x_j with distributions P_j falls infinitely often into the set $W(\delta) = \{x \in \mathbf{A}\colon f(x) - f_* \leqslant \delta\}$, with probability one.*

Theorem 1 holds in the 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 the theorem ensure that the sequence $\{x_j\}$ converges to the set $\mathbf{A}_* = \{\arg\min f\}$ of global minimizers with probability 1; similarly, the sequence of records y_{oj} converges to $f_* = \min f$ with probability 1.

For PRS with $P = P_U$ uniform on $\mathbf{A}$, $q_j(\varepsilon) = \text{const} > 0$ and hence (3) trivially holds. In practice, a very popular rule for selecting the probability measures P_j is

$$P_{j+1} = \alpha_{j+1} P_U + (1 - \alpha_{j+1}) Q_j, \tag{4}$$

where $0 \leqq \alpha_{j+1} \leqslant 1$, P_U is the uniform distribution on $\mathbf{A}$ and Q_j is an arbitrary probability measure on $\mathbf{A}$ which may depend on the results of the evaluation of the objective function at the points $x_1, \dots, x_j$. As an example, sampling from Q_j may correspond to performing several iterations of a local descent from the current record point $x_{o,j}$; that is, a point with $f(x_{o,j}) = y_{o,j}$. Sampling from the distribution (4) corresponds to taking a uniformly distributed random point in $\mathbf{A}$ with probability α_{j+1} and sampling from Q_j with probability $1 - \alpha_{j+1}$.

If the probability measures P_j are chosen according to (4), then the condition $\sum_{j=1}^{\infty} \alpha_j = \infty$ implies (3) and hence convergence of the corresponding GRS algorithm.

Unless some regularity conditions about f like the Lipschitz condition are imposed and used for guaranteeing that $\mathbf{A}$ is covered by the balls with centres at the observations points x_j, the statements like Theorem 1 are the only tools which guarantee convergence of GRS algorithms. An implication of that is that the PRS with $P = P_U$ is the fastest (in the worst-case scenario) GRS algorithm. Its rate of convergence we consider next.

2.2.2. *Rate of Convergence of PRS*

Consider the PRS defined in Section 2.1.3. Let $\varepsilon, \delta > 0$ be fixed, $x_* = \arg\min f$ be a global minimizer of $f(\cdot)$ and $W(\delta) = \{x \in \mathbf{A}\colon f(x) - f_* \leqq \delta\}$. If we want to study the rate of convergence towards x_* we set $B = B(x_*, \varepsilon)$. Otherwise, if we study convergence with respect to the function values (that is, convergence of $y_{o,n} - f_*$), then we set $B = W(\delta)$.

Assume that our objective is hitting the set B with at least one point x_j $(j = 1, \dots, n)$, where n is the total number of points generated by PRS. We will call the event 'a point x_j hits the set B' success. In this notation, PRS generates a sequence of independent Bernoulli trials with success probability $\mathbb{P}\{x_j \in B\} = P(B)$; note that under natural assumptions about P we have $P(B) > 0$ for any $\varepsilon > 0$. In view of the independence of x_j, $\mathbb{P}\{x_1 \notin B, \dots, x_n \notin B\} = (1 - P(B))^n$ and therefore $\mathbb{P}\{x_j \in B \text{ for at least one } j,\ 1 \leqslant j \leqslant n\} = 1 - (1 - P(B))^n$. Since $P(B) > 0$, this probability tends to one as $n \to \infty$.

Let us assume that we wish to reach the set B with probability at least $1 - \gamma$ for some $0 < \gamma < 1$. This gives the following inequality for n, the number of points required in the PRS: $1 - (1 - P(B))^n \geqslant 1 - \gamma$. Solving it we obtain $n \geqslant n(\gamma) = \ln\gamma / \ln(1 - P(B))$. If $P(B)$ is small (which is always the case in practice), then $\ln(1 - P(B)) \cong -P(B)$, and we can replace the previous inequality with $n \geqslant (-\ln\gamma)/P(B)$; that is, one needs to make at least $\lceil -\ln\gamma/P(B)\rceil$ points in PRS to reach the set B with probability $1-\gamma$.

Consider the quantity $n_\gamma = (-\ln\gamma)/P(B)$. Multiplier $(-\ln\gamma)$ depends on γ but it cannot be too large in practical computations. For example, $-\ln\gamma \simeq 2.99573$ for $\gamma = 0.05$. Unlike $(-\ln\gamma)$, the multiplier $1/P(B)$ in the formula for n_γ can be astronomically large.

To give an example, assume $\mathbf{A} = [0,1]^d$, $P = P_U$ is uniform on $\mathbf{A}$ and $B = B(x_*, \varepsilon)$. Then $\mathrm{vol}(B) \leqq \varepsilon^d V_d$, where $V_d = \pi^{\frac{d}{2}}/\Gamma(\frac{d}{2}+1)$ is the volume of the unit ball (of radius 1) in $\mathbb{R}^d$. In view of the multiplier ε^d in the upper bound for the volume $\mathrm{vol}(B)$, this volume can be extremely small even when d is not very large (say, $d = 10$) and ε is not very small (say, $\varepsilon = 0.1$). This is much larger than the total number of atoms in the universe (which is estimated to be smaller than 10^{81}).

2.2.3. *Rate of Convergence of a General GRS Method*

As it was discussed in Section 2.2.1, the main way to guarantee convergence of a general GRS algorithm is to choose the probabilities P_j in the form (4) where α_j satisfy (3). Let us modify the arguments provided above for the case of PRS. As a replacement of the starting equality $\mathbb{P}\{x_j \in B\} = P(B)$, for all $j \geqslant 1$ we now have $\mathbb{P}\{x_j \in B\} \geqslant \alpha_j P_U(B)$, with equality if we consider the worst-case scenario. Modifying the other arguments above correspondingly we can define $n(\gamma)$ as the smallest integer such that the following inequality is satisfied: $\sum_{j=1}^{n(\gamma)} \alpha_j \geqslant -\ln\gamma/P_U(B)$. Assume $\alpha_j = 1/j$, which is a common recommendation. Using for simplicity the approximation $\sum_{j=1}^{n} \alpha_j \simeq \ln n$, we approximately obtain $n(\gamma) \simeq \exp\{-\ln\gamma/P_U(B)\}$. If $\mathbf{A} = [0,1]^d$, then this gives $n(\gamma) \simeq \exp\{-\ln\gamma/P_U(B)\}$. Assuming further $B = B(x_*, \varepsilon, \rho_2)$ we obtain $n(\gamma) \simeq \exp\{\mathrm{const}\cdot\varepsilon^{-d}\}$, where $\mathrm{const} = (-\ln\gamma)/V_d$ (note also that if x_* lies closer to the boundary of $\mathbf{A}$ than ε in any direction, then the constant above and hence $n(\gamma)$ are even larger). For example, for $\gamma = 0.1$, $d = 10$ and $\varepsilon = 0.1$, $n(\gamma)$ is a number larger than $10^{1000000000}$. Even for a small dimension $d = 3$, $\gamma = 0.1$ and $\varepsilon = 0.1$, the value of $n(\gamma)$ is huge: $n(\gamma) \simeq 10^{238}$.

The main conclusion of this discussion is: even for moderate dimensions, general GRS algorithms do not guarantee convergence worst-case scenario in practical computations. Convergence could only be seriously discussed if the Lipschitz-type conditions are assumed and used in the process of search.

2.3. *Statistical Inference in GRS*

2.3.1. *Statistical Inference in PRS*

Assume $\mathbf{A}$ is a compact in $\mathbb{R}^d$ and $x_1, \ldots, x_n$ are i.i.d. with $x_j \sim P$, where n is a large number and P is a probability measure on $\mathbf{A}$ with some density $p(x)$, which is a piecewise continuous function on $\mathbf{A}$ and $p(x) > 0$ for all $x \in \mathbf{A}$. The following two types of statistical inference based on prior information about $f(\cdot)$ and the information contained in the values $\{y_j = f(x_j),\ j = 1\ldots, n\}$ related to the sample $\{x_1, \ldots, x_n\}$ could be made.

Type 1. Either a parametric or non-parametric estimator of $f(\cdot)$ can be constructed. In PRS, such an estimator can only be used for defining a stopping rule (as the rule for choosing points x_j's is fixed). Implicitly, non-parametric estimators of $f(\cdot)$ are constructed in most evolutionary GRS algorithms (these estimators are used for the construction of the rules for choosing the next points x_j's).

Type 2. Statistical inference about $f_* = \min f$ based on several smallest values extracted from the sample $\{y_j = f(x_j),\ j = 1\dots, n\}$.

We will not consider inferences of Type 1 here as it would take us too far into the direction of metaheuristic, see a comprehensive discussion in Zhigljavsky (1991). Below we only consider inferences of Type 2. In this exposition, we closely follow the material of Chapter 7 in Zhigljavsky (1991), Sections 2.3–2.6 in Zhigljavsky and Žilinskas (2008) and Zhigljavsky (1993).

Since x_j are i.i.d. with distribution P, the elements of the sample $Y = \{y_j = f(x_j),\ j = 1\dots, n\}$ are i.i.d. with c.d.f.

$$F(t) = \mathbb{P}\bigl\{x \in \mathbf{A}\colon f(x) \leqslant t\bigr\} = \int_{f(x)\leqslant t} P(dx) = P\bigl(W(t - f_*)\bigr), \tag{5}$$

where $t \geqq f_*$ and $W(\delta) = \{x \in \mathbf{A}\colon f(x) \leqq f_* + \delta\}$, $\delta \geqq 0$. This c.d.f. is concentrated on the interval $[f_* = \min f, f_* = \max f]$ and our main interest is the unknown value f_*. The analytic form of $F(t)$ is either unknown or incomprehensible (unless f is very simple) and we need to use asymptotic considerations. Luckily, the asymptotic distribution of the order statistics is unambiguous and the conditions on $F(t)$ (and hence on f) when this asymptotic law works are very mild and can always be assumed true. Specifically, for a very wide class of functions f and distributions P, the c.d.f. F can be shown to have the following representation for $t \simeq f_*$:

$$F(t) = c(t - f_*)^\alpha + \mathrm{o}\bigl((t - f_*)^\alpha\bigr), \quad t \downarrow f_*, \tag{6}$$

here c and α are some positive constants; more generally, $c = c(t)$ can be a slowly varying function for $t \simeq f_*$ and the results cited below are also valid for this slightly more general case. The value of c is irrelevant but the value of α, which is called 'tail index', is very important. We shall assume that the value of α is known. As discussed below in Section 2.3.2, this is usually indeed the case.

Denote by η a random variable which has c.d.f. (5) and by $y_{1,N} \leqq \dots \leqq y_{N,N}$ the order statistics corresponding to the sample Y. The parameter $f_* = \min f$ is at the same time the lower endpoint of the random variable η, i.e. $f_* = \operatorname{ess\,inf} \eta$.

There are many good books on the theory of extreme order statistics and in this survey we are not reviewing it. We refer the reader to the excellent book (Nevzorov, 2001), which provides an introduction not only to the theory of extreme order statistics but also to the related theory of record moments, which we are going to use in Section 2.3.4. A review of both theories, fully sufficient for all our purposes, is contained in Section 2.3 of Zhigljavsky and Žilinskas (2008). The most important result in the theory of extreme order statistics states that if (6) holds then 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, α, which is called 'the tail index'.

Several good estimates of f_* are known for given α, see Zhigljavsky and Žilinskas (2008, Section 2.4). We highly recommend one of them, the optimal linear estimator based on the use of k order statistics. This estimator has the form

$$\widehat{f}_{N,k} = c \sum_{i=1}^{k} \left[u_i / \Gamma(i + 2/\alpha)\right] y_{i,N}, \tag{7}$$

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

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

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

Under the assumption (6), for given k and α and for $N \to \infty$, $\widehat{f}_{N,k}$ is a consistent and asymptotically unbiased estimator of f_* and $\mathbb{E}(\widehat{M}_{N,k} - f_*)^2$, its asymptotic mean squared error, has maximum possible rate of decrease in the class of all consistent estimates including the maximum likelihood estimator of M, see Chapter 7 in Zhigljavsky (1991) for a comprehensive treatment of the theory.

Under the same assumptions, the following confidence interval for f_* has asymptotic (as $N \to \infty$) confidence level $1 - \delta$:

$$\left[y_{1,N} - (y_{k,N} - y_{1,N})/c_{k,\delta}, y_{1,N}\right], \quad \text{where } c_{k,\delta} = \left[1 - (1-\delta)^{1/k}\right]^{-1/\alpha} - 1. \tag{8}$$

Procedures of testing hypotheses about f_* are based on constructing confidence intervals for f_*. Indeed, if we want to test the hypothesis $H : f_* \leqq c$, then we construct a c.i. like (8) and if c belongs to this c.i., then the hypothesis H gets accepted. Those interested in more material related to construction of confidence intervals for f_* and testing hypotheses about f_* are advised to consult (Weissman, 1981, 1982) and especially Chapter 7 in Zhigljavsky (1991). A comparison of accuracy of statistical procedures outlined above and similar to them was performed in Hamilton *et al.* (2007).

2.3.2. *Tail Index*

As it was mentioned above, the assumption (6) can always be assumed true. The main issue is whether the value of the tail index α can be gathered or has to be estimated. The second option is not good as the estimation of α is notoriously difficult; see De Haan and Peng (1998) for a survey of a comparison of different estimators of α. The sample size N should be astronomically large if we want to have an accurate estimator of f_* obtained from (7) or any other estimator of f_* after replacing α with any (even best possible) estimator, see for example Section 2.5 in Zhigljavsky and Žilinskas (2008). In PRS, however, we can usually have enough knowledge about f 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 is locally quadratic around x_*, then the condition (6) holds with $\alpha = d/2$. This result and many generalizations of it have been established independently in De Haan (1981) and Zhigljavsky (1981). For a detailed exposition of these results and the related theory see Zhigljavsky (1985, 1991).

An important implication of the result $\alpha = d/2$ is the so-called 'curse of dimensionality'. Indeed, if α increases, the quality of all estimators, and in particular (7), fastly deteriorates; for fixed $y_{1,N}$ and $y_{k,N}$, the length of the confidence interval (8) also grows fast, see Chapter 7 in Zhigljavsky (1991).

2.3.3. *Branch and Probability Bound (BPB) Methods*

Branch and bound optimization methods are widely known. These methods consist of several iterations, each including the following stages: (i) branching the optimization set into a tree of subsets (more generally, decomposing the original problem into subproblems), (ii) making decisions about the prospectiveness of the subsets for further search, and (iii) selecting the subsets that are recognized as prospective for further branching.

To make a decision at stage (ii) prior information about $f(\cdot)$ and values of $f(\cdot)$ at some points in $\mathbf{A}$ are used, deterministic lower bounds concerning the minimal values of $f(\cdot)$ on the subsets of $\mathbf{A}$ are constructed, and those subsets $Z \subset \mathbf{A}$ are rejected (considered as non-prospective for further search) for which the lower bound for $f_{Z*} = \inf_{x \in Z} f(x)$ exceeds an upper bound $\hat{f}_*$ for $f_* = \min f$; the record value (2) of $f(\cdot)$ in $\mathbf{A}$ is a natural upper bound $\hat{f}_*$ for f_*. A standard recommendation for improving this upper bound is to use a local descent algorithm, starting at the new record point, each time we obtain such a point.

Let us briefly consider a version of the branch and bound technique, which was introduced in the first issue of *Informatica*, see Zhigljavsky (1990). We call these methods 'branch and probability bound methods' or shortly BPB methods.

At each iteration of a BPB method, an independent sample from the uniform distribution in the current search region is generated and the statistical procedure mentioned above for testing the hypothesis $H_Z : f_{Z*} \leqslant \hat{f}_*$ is applied to make a decision concerning the prospectiveness of sets Z at stage (ii). Rejection of the hypothesis H_Z corresponds to the decision that the global minimum f_* cannot be reached in Z. If such a rejection is erroneous, then it may result in losing the global minimizer. An attractive feature of the BPB methods is that the asymptotic level for the probability of false rejection can be controlled and kept on a prescribed low level.

The stages (i) and (iii) above can be implemented in exactly the same fashion as in the classical branch and bound methods. When the structure of $\mathbf{A}$ is not too complicated, the following technique has been proven to be convenient and efficient. Let A_j be a search region at iteration j, $j \geqslant 1$ (so that $A_1 = A$). At iteration j, in the search region $\mathbf{A}_j$ we first isolate a subregion Z_{j1} with centre at the point corresponding to the record value of $f(\cdot)$. The point corresponding to the record value of $f(\cdot)$ over $\mathbf{A}_j \backslash Z_{j1}$ is the centre of a subregion Z_{j2}. Similar subregions Z_{ji} $(i = 1, \ldots, I)$ are isolated until either $\mathbf{A}_j$ is covered or the hypothesis that the global minimum can occur in the residual set $\mathbf{A}_j / \bigcup_{i=1}^{I} Z_{ji}$

is rejected. The search region $\mathbf{A}_{j+1}$ in the next $(j+1)$-th iteration is naturally either $Z^{(j+1)} = \bigcup_{i=1}^{I} Z_{ji}$, a hyperrectangle covering $Z^{(j+1)}$, or a union of disjoint hyperrectangles covering $Z^{(j+1)}$. Note that at subsequent iterations all previously used points can still be used as they are uniformly distributed at the reduced regions.

The BPB methods are both practically efficient for small d (say, $d \leqq 5$) and theoretically justified in the sense that under natural assumptions about $f(\cdot)$, they asymptotically converge with a given probability, which can be chosen close to one. However, as d (and therefore α) increases, the efficiency of the statistical procedures deteriorates. Therefore, for large d the BPB methods are both hard to implement and their efficiency is poor.

BPB methods have recently been extended for solving multi-criteria optimization problems, see Žilinskas and Zhigljavsky (2016). The main idea in this extension is the use of the estimators (7) and the procedures and the confidence intervals (8) simultaneously for all objective functions in the augmented weighted Tchebycheff optimization problems. This has allowed us to estimate the Pareto front even when the original objective functions are multi-extremal. Construction schemes of related BPB methods are then similar to the standard single-criterion case.

2.3.4. *Other GRS Algorithms Using Statistical Procedures of Section 2.3.1*

BPB methods are not the only possible random search methods that benefit from the statistical procedures described in Section 2.3.1. Any population-based GRS method can be complemented with these statistical procedures. They could be useful for (a) making a stopping rule (see below), (b) helping to make decisions when to stop creating the current population and start making a new one, and (c) complementing rules for deciding on prospectiveness of different subsets of $\mathbf{A}$. Possibilities are enormous, in particular for those who like metaheuristical GRS algorithms as an example of such algorithm see a recent paper of Kulczycki and Lukasik (2014).

The use of statistical inferences outlined in Section 2.3.1 for creating GRS algorithms that can be useful for solving multicriteria optimization problems with non-convex objectives has already been mentioned, see two paragraphs above. Let us also note a potential usefulness of these statistical inferences for making stopping rules in population-based GRS algorithms and deciding whether it is worthwhile to carry on generating random points with given distribution. The idea developed in Zhigljavsky and Hamilton (2010) is as follows. Consider the confidence interval (8) and assume that it is our main characteristic of accuracy of the current sample. Assume we have made n observations so far. Then the length of the confidence interval (8) is proportional to $y_{k,n} - y_{1,n}$. We want to make a decision: stop with the current strategy of generating i.i.d. points x_j or carry on doing this hoping that our c.i. will (significantly) decrease. It is pointless waiting for an update of the new record as the waiting time is infinite. However, each time $y_{k,n}$ is updated, the length of the confidence interval (8) changes. The expected waiting time to the next update is not that long, it is $n/(k-1)$. The distribution of the change in the length of the c.i. should also be taken into account when we decide whether to wait for the next update. Zhigljavsky and Hamilton (2010) contain also many details concerning the corresponding stopping rules. For other techniques of devising stopping criteria in GRS algorithms see Dorea (1990), Hart (1998) and Yamakawa and Ohsaki (2013).

3. Statistical Models Based Global Optimization

3.1. *Basic Ideas*

Methods based on statistical models of objective functions are oriented to the problems which are generally described as 'expensive' 'black box' problems. Such a class of global optimization problems is difficult to tackle theoretically as well as to develop algorithms suitable to a broad field of applications. The theory of rational decisions under uncertainty well suits to substantiate methods for such problems because of the correspondence of the basic concepts: 'black box'/uncertainty, 'expensive'/rational, and optimization/decision. Statistical model of uncertainty means here an appropriate statistical model of aimed objective functions, selected taking into account not only the representativeness of available information but also the conformity with the efficiency of implementation of the corresponding algorithms. The algorithms are defined as sequences of decisions under uncertainty, and the ideas of rational decision making theory are used to find an appropriate algorithm.

A model of functions under uncertainty considered in the probability theory is a stochastic function; the terms 'stochastic process' and 'random field' also are used to specify stochastic functions of one and several variables correspondingly. Let $\xi(x)$, $x \in \mathbf{A}$ be a stochastic function selected for a statistical model of the aimed objective functions. The main algorithms based on the statistical models of objective functions are the P-algorithm (Žilinskas, 1985), and one-step Bayesian algorithm (Mockus, 1972; Žilinskas, 1975). We refer to the cited above monographs and papers of Žilinskas (1990, 1992) for a review and the theoretical substantiation of these algorithms.

To define the considered algorithms we need the following notation. At the $k+1$ step of search, the objective function values y_j are computed at the points x_j, $j = 1, \dots, k$, and the current minimum computed value is denoted $y_{o,k}$. The P-algorithm computes the next objective function value at the point

$$x_{k+1} = \arg\max_{x \in \mathbf{A}} \mathbb{P}\big(\xi(x) \leqq y_{o,k} - \varepsilon \big| x_j, y_j,\ j = 1, \dots, k\big), \tag{9}$$

meaning that it is aimed at maximal probability of the improvement of the current estimate of global minimum; here $\mathbb{P}(\cdot|\cdot)$ denotes the conditional probability. For the Gaussian stochastic function, the algorithm (9) is defined by a simpler formula

$$x_{k+1} = \arg\max_{x \in \mathbf{A}} \frac{y_{o,k} - \varepsilon - m(x|x_j, y_j,\ j = 1, \dots, k)}{\sigma(x|x_j, y_j,\ j = 1, \dots, k)}, \tag{10}$$

where $m(x|x_j, y_j,\ j = 1, \dots, k)$ and $\sigma(x|x_j, y_j,\ j = 1, \dots, k)$ denote the conditional mean and conditional standard deviation of the stochastic function at the point x.

The one-step Bayesian algorithm computes the next objective function where the expected improvement is maximum

$$x_{k+1} = \arg\max_{x \in \mathbf{A}} \Delta Y_{k+1}(x),$$

$$\Delta Y_{k+1}(x) = \mathbb{E}\big(\max\{y_{o,k} - \xi(x),\ 0\}\big|x_j, y_j,\ j = 1, \dots, k\big), \tag{11}$$

where $\Delta Y_{k+1}(x)$ means expected improvement at $k+1$ step in case of computing function value at the point x, $\mathbb{E}(\cdot|\cdot)$ denotes the conditional expectation. This algorithm in later papers was also called 'kriging' and 'EGO' (efficient global optimization) (Jones *et al.*, 1998). The reasons for renaming remained without an explanation. We note only that the term 'kriging' originally was used to call the prediction method by the name of its author D.G. Krige; to our best knowledge, Krige has not considered statistical models based global optimization methods. On the other hand, the substantiation for adding the pretentious attribute 'efficient' is expected, at least the theoretical analysis of conditions of the real efficiency and of the limitations of the algorithm; unfortunately, such results were not presented by the inventors of the name EGO.

In the implementations of the one-step Bayesian algorithm with a Gaussian stochastic functions for a statistical model, the computations normally are performed according to the following formula

$$x_{k+1} = \arg\max_{x \in \mathbf{A}} \sigma(x|x_j, y_j,\ j = 1, \dots, k)\left(v_k(x)\Phi(v_k(x)) + \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v_k^2(x)}{2} \right)\right),$$

$$v_k(x) = \frac{y_{o,k} - m(x|x_j, y_j,\ j = 1, \dots, k)}{\sigma(x|x_j, y_j,\ j = 1, \dots, k)},$$

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\big\{ -t^2/2 \big\} dt, \tag{12}$$

where the notation $m(x|x_j, y_j,\ j = 1, \dots, k)$ and $\sigma^2(x|x_j, y_j,\ j = 1, \dots, k)$ has the same meaning as in (10).

Various modifications of (11), usually with some numerical examples, can be found in numerous publications with EGO and 'kriging' in the titles and lists of key words.

3.2. *Selection of a Statistical Model*

In the majority of publications on the subject in question, statistical models were selected from the stochastic functions well researched by the probability theoreticians. Besides stochastic functions extensively studied in the probability theory, a specific stochastic model oriented to the information global optimization algorithm was proposed in Strongin (1978); we do not consider this model here since it is well presented and thoroughly discussed in Grishagin *et al.* (1997), Strongin and Sergeyev (2000).

The selection of an appropriate statistical model is guided by the theoretical properties of potential candidates known from the probabilistic literature, and by the computational complexity of algorithms for computing parameters of the conditional distributions. The latter criterion determines the priority of a Gaussian stochastic function, and of a Markovian processes in the single variable case.

The selection of statistical methods for constructing GO algorithms in the considered time period predominantly was the same as in the time before 1990. For the single variable models frequently selected was the Wiener process, specifically the Brownian bridge.

The Wiener model is especially appropriate for the theoretical analysis of properties of GO algorithms constructed using this model (Calvin, 2001, 2007, 2011; Calvin and Žilinskas, 2005; Zhigljavsky and Žilinskas, 2008). Let us note, that this model was appropriate for investigation of the average complexity of algorithms for other classes of numerical problems, e.g. for approximation in Ritter (1990). Although the generalization of Wiener model to multidimensional problems is widely used in investigation of the complexity of multidimensional approximation and integration (see e.g. Traub and Weschulz, 1998), no attempts have been made to extend the Wiener model based single variable GO results to the multidimensional case.

The homogeneous isotropic Gaussian random fields are a natural generalization of stationary stochastic processes to the multidimensional case. These statistical models are used for the construction of GO algorithms from the very beginning of the development of the considered approach. To our best knowledge, the non-Gaussian random fields were not used for the construction of GO algorithms. However, the algorithms based on the Gaussian fields, the characteristics of which are not invariant with respect to translations and/or rotations, are described in literature. For example, a Gaussian random field with constant mean and variance but with coordinate dependent correlation function

$$\rho(x_i, x_j) = \exp\left(- \sum_{k=1}^{d} \beta_k |x_{i,k} - x_{j,k}|^{\alpha_k} \right)$$

was used to construct a GO algorithm in Jones *et al.* (1998), where $x_{i,k}$ denotes the kth component of vector x_i, and $\alpha_k > 0$, $\beta_k > 0$. Although such a generalization seems attractive the problem of estimation of many parameters from a modest number of observations can be problematic. The maximum likelihood method is supposed in Jones *et al.* (1998), Jones (2001) as well as in the papers of a number of other authors. Despite numerical values in that estimation problem can be computed by the maximization of the likelihood function, the real worth of such estimates can be doubtful; see e.g. the analysis in Pepelyshev (2011) where the drawbacks of the estimates, obtained using values of a typical objective function, are highlighted in the case of even a single parameter of the correlation function in question.

The new idea of the construction of GO algorithms on the basis of generalized statistical models, introduced in Žilinskas (1982), is to adjust the model to a simplicial partition of the feasible region. The adjustment mimics the Markovian property of stochastic processes enabling implementation of the algorithm in branch and bound framework similarly as in the case of Lipschitz optimization (Paulavičius and Žilinskas, 2014). Let at the $k+1$ step of search the feasible region A be partitioned into m simplices, and s^i_j, $i = 1, \dots, m$, $j = 1, \dots, d+1$ be the simplices of the i simplex. In Žilinskas and Žilinskas (2002, 2010), Calvin and Žilinskas (2014) algorithms are considered which are based on generalized statistical models defined for every simplex independently, using information related to its vertices, and the maximum improvement probability approximately computed for every simplex. The simplex with maximum probability is subdivided thus making the subdivision more refined. Let us note, that a random field model can be applied in a similar way restricting the information, used in computing of conditional characteristics, by the information related to the vertices of the relevant simplex. Similarly, an algorithm can be constructed using the partitioning of the feasible region into sub-rectangles (Gimbutienė and Žilinskas, 2015). The computational burden of those algorithms is reduced with respect to the original one (9) mainly by decreasing the complexity of the computation of the improvement probability, and by the possibility to store the collected during the search information in a convenient data structure.

Finally, there exists some equivalence between the P-algorithm and the GO algorithm based on the model of radial basis functions (RBF) (Gutmann, 2001; Žilinskas, 2010). Therefore, the favourable properties of RBF interpolants can be taken into account defining the parameters of a Gaussian random field selected as a statistical model of objective functions, e.g. an approved RBF can be reasonable to select as a covariance function.

3.3. *Modifications*

The interest in modifications of the considered above original methods is caused by the difficulties to cope with their complexity as well as by the desire to extend the field of their applicability.

One of the important counterparts of the algorithms based on the statistical models of objective functions is the method of estimation of model parameters, and the choice of non appropriate parameters can crucially degrade the performance of the corresponding algorithm. The authors of Kleijnen *et al.* (2012) claim that the version of algorithm (11) performs very well with the parameters estimated by the bootstrapping (DenHertog and Kleijnen, 2006).

Some statistical models, e.g. Gaussian stochastic functions, are prone to the including noisy observations. Already in the first paper on the construction of GO algorithms based on statistical models of objective functions (see Kushner, 1962) the possibility was indicated to construct an algorithm for minimization in the presence of noise. Although theoretical definition of an algorithm for the optimization in such a case is almost coincident with that for the case without noise, implementation problems of algorithms for noisy optimization can be essentially more complicated. For example, the computational advantages of Markovian processes are loosen in case random errors in the objective function values are taken into account. We refer to the paper (Calvin and Žilinskas, 2005) and the references therein for the description of a special implementation of a one dimensional algorithm based on the Wiener process model for the optimization in the presence of noise. This implementation is sufficiently efficient for applications, although it is of notably higher complexity than its analog for the non-noisy optimization. However, the implementation in Calvin and Žilinskas (2005) is not generalizable to the multidimensional case. Some experimentation results with multidimensional noisy version of (11) and its modification are reported in Huang *et al.* (2006b). Although, for the used in the experimentation test functions, the results seem promising, the computational burden is too high for a large number of evaluations of function values needed for higher dimensions and noise level. The computational burden at a current step can be reduced using a simplicial (possibly also hypercubic) statistical model enabling the implementation of an algorithm in the branch and bound framework with the partition of the feasible region; in other words, it seems promising to generalize the idea of Žilinskas and Žilinskas (2002, 2010) to the optimization in the presence of noise. An implementation of this idea, although with a slightly different interpretation of the P-algorithm, is described in Rullière *et al.* (2013) where the testing of the developed algorithm was restricted with two dimensional examples. Finally, it should be noted that, in the algorithms based on the statistical models, the objective function values, evaluated with different level of noise, can be treated uniformly but with respect to the noise level. An algorithm demonstrating such possibilities is proposed in Huang *et al.* (2006a) where the collected during the search information is supposed of different fidelity.

The statistical models of objective functions considered in the present paper take into account and forecast the values of objective functions. The derivatives, if available, can supplement important information about the objective function, and thus improve the performance of corresponding algorithms. Some results about the one-dimensional P-algorithm operating with derivatives can be found in the book (Zhigljavsky and Žilinskas, 2008). An attempt to develop a multidimensional statistical model with derivatives was made in Makauskas (1991), however further results concerning a corresponding algorithm are not known.

In principle, statistical models can be applied for the construction of algorithms for constrained optimization where a stochastic function $\eta_i(x)$ is chosen for a model of the left hand side function in the inequality constraint $g_i(x) \leqslant 0, i = 1, \dots, K$. Then the algorithm (13) would be modified to the following one

$$x_{k+1} = \arg\max_{x \in \mathbf{A}} \mathbb{P}\big(\xi(x) \leqq y_{o,k} - \varepsilon,\ \eta_i(x) \leqq 0,\ i = 1, \dots, K \mid x_j, y_j,\ j = 1, \dots, k\big).$$

Other versions could be also defined using e.g. the average improvement (11); see Sasena (2002), Picheny (2014). However, the most difficult problem here is in choosing appropriate stochastic functions for the models of $g_i(x)$.

The research in multi-objective optimization is presently very active. The extension of statistical model based GO algorithms to multi-objective problems was natural since problems with expensive black-box objectives are no less acute than the single-objective ones. The multi-objective P-algorithm as a direct generalization of single-objective one is proposed in Žilinskas (2014). The multi-objective P-algorithm is based on the assumptions reviewed above which imply the acceptance of a vector valued random field $\Xi(x) \in \mathbb{R}^m$, $x \in \mathbf{A}$ as a statistical model of m objective functions. The algorithm is defined similarly to (9)

$$x_{k+1} = \arg\max_{x \in \mathbf{A}} \mathbb{P}\big(\Xi(x) \leqq y_{o,k} \mid x_j, y_j,\ j = 1, \dots, k\big), \tag{13}$$

where y_j, $j = 1, \dots, k$, denote the m dimensional vectors of the values of objective functions computed at previous search steps, and $y_{o,k}$ is a reference vector intended to improve at $k + 1$ step. The selection of $y_{o,k}$ enables a user to express his preferences with respect to objectives as well as to control distribution of solutions along the Pareto front. Various modifications of the criterion (13) are possible; see e.g. Picheny (2015).

The single-objective algorithms based on statistical models were used for multi-objective optimization in combination with scalarization methods. The EGO (which corresponds to (11)) was proposed in Knowles (2006) to hybridize with the Chebyshev's scalarization; this method was extended for the noisy objectives in Knowles *et al.* (2009). Similarly, the combination of the hypervolume scalarization with a statistical model based algorithm is proposed in Emerich *et al.* (2016). A version of hybridization of methods based on statistical models of objective functions with evolutionary methods is proposed in Emerich *et al.* (2006).

Because of high computational complexity of Bayesian methods, they are mainly oriented to the expensive low dimensional objective functions. Therefore, the Bayesian methods are appropriate for the application at upper level of the two level algorithms where the upper level algorithm is used to optimize the parameters of a lower level algorithm enhancing their performance; for the details of the concrete algorithms and their applications we refer to Mockus *et al*. (1994, 1997).

3.4. *Investigation of Efficiency*

The efficiency of algorithms in computer science is understood as reciprocal time complexity where the latter means the computing time (or number of computer operations) needed to complete the computations. Such assessment is well applicable to assess the efficiency of algorithms for combinatorial optimization as well as for algorithms for optimization problems solvable in finite time, e.g. problems of linear programming. The other approach of the efficiency assessment used in optimization theory is typical for numerical analysis: it is the assessment of the asymptotic convergence rate. Both approaches are theoretically appropriate to assess the performance of GO methods where large number of computations of function values can be made. However, such assessment methods are not fully relevant to assess the efficiency of algorithms supposed for expensive objective functions where the number of computations of function values is scarce.

The GO methods considered here are supposed for expensive objective functions, and they are theoretically substantiated by the original definition. Nevertheless, their asymptotic properties are of interest, at least for the comparison with the other algorithms of similar destination. The convergence of the methods in question were considered in two frameworks, in the probabilistic sense assuming that an objective function is a randomly generated sample function of the underlying stochastic function, and in deterministic sense assuming that an objective function is an arbitrary function from an appropriate class of functions.

Let us recall that the considered algorithms, although based on statistical models, are deterministic. The algorithm, applied to a concrete objective function, generates a deterministic sequence of sites $x_k \in \mathbf{A}$, $k \to \infty$ for computing values of the objective function. To guarantee the convergence of the candidate solution to the point of global minimum for an arbitrary continuous function, the sequence x_k should be every dense sequence in $\mathbf{A}$

(Törn and Žilinskas, 1989). Such a convergence of the P-algorithm and of one-step optimal Bayesian algorithms were proved in the originating papers; for the details we refer to Mockus (1988), Törn and Žilinskas (1989). To assess the rate of convergence, more strict assumptions should be made. For example, in Calvin and Žilinskas (2000) the single variable optimization problem is considered, and it is shown there that the P-algorithm (9) based on a smooth statistical model with the threshold sequence $\varepsilon_k = k^{-1+\gamma}$, $\gamma > 0$ converges to global minimum of an arbitrary twice continuously differentiable function with convergence rate $O(k^{-3+\gamma})$. The term 'smooth statistical model' is used here for a Gaussian stationary stochastic process with correlation function satisfying the following conditions:

$$\rho(t) = 1 - \frac{1}{2}\lambda_2 t^2 + \frac{1}{4}\lambda_4 t^4 + o\big(t^4\big),$$

as $t \to 0$, $\lambda_2 > 0$, $\lambda_4 > 0$, and

$$\left|\frac{d^4\rho(t)}{dt^4} - \lambda_4\right| = O\big(|t|\big), \qquad -\frac{d^2\rho(t)}{dt^2} = \lambda_2 + O\big(\big|\log^{-\alpha}|t|\big|\big),$$

for some $\alpha > 1$ as $t \to 0$, and also $\rho(t)\log(t) \to 0$ as $t \to \infty$. These assumptions guarantee that the underlying statistical model is compatible with the assumption made, in the statement on convergence, about the considered objective function. In other words, the made assumptions guarantee that the sample functions of the underlying stochastic process are twice continuously differentiable with probability 1. For the comparative analysis of the convergence rates of the univariate P-algorithms, based on different statistical models, we refer to Zhigljavsky and Žilinskas (2008).

The investigation of the convergence rate of multivariate methods based on statistical models is more difficult than of univariate ones. To the best knowledge of the authors, the only publication in question is Calvin and Žilinskas (2014) where a two variable P-algorithm is considered based on a simplical statistical model. The feasible region is iteratively partitioned into simplices by means of Delaunay triangulation, and the improvement probability is approximated using asymptotic expressions of the conditional mean and variance of the underlying statistical model (Žilinskas and Gimbutienė, 2015). Let an objective function $f(x)$ be twice continuously differentiable, $x_* = \arg\min_{x\in\mathbf{A}} f(x)$ be a unique global minimizer, and $\delta_k = y_{ok} - f(x_*)$ denotes, the error of estimation of the global minimum. As shown in Calvin and Žilinskas (2014), the following inequality is satisfied

$$\liminf_{k\to\infty} \sqrt{k}\log\left(\frac{1}{\delta_k}\right) \geqslant \frac{(\lambda_1\lambda_2)^{1/4}}{2\sqrt{6q\pi}}, \tag{14}$$

where λ_1, λ_2 are eigenvalues of the Hessian of $f(x)$ at the point x_*, and q is a measure of the triangulation quality. The result in less detail can be expressed in the following form:

$$\delta_k \leqq \exp(-c\sqrt{k}),$$

where $c > 0$ aggregates all constants of the expression in right hand side of (14).

The investigation of algorithms' performance in probabilistic sense is based on the assumption that the objective functions are randomly selected sample functions of the underlying stochastic function. The probability of an attribute of search by a deterministic algorithm means the probability to select randomly the function for which this attribute occurs during the search. The difficulties in the analysis of the probabilistic convergence are implied mainly by the complexity of the computation of conditional distribution of the relevant functionals with respect to the currently known values of the objective function. Therefore, it is not surprising that the analysis of probabilistic convergence was started using for the statistical model of objective functions the Wiener process which is Gaussian, Markovian, and with independent increments.

The nonadaptive (passive) algorithms, although believable of low efficiency, are simpler than adaptive ones, and therefore were investigated foremost. It was shown in Ritter (1990) that the uniform grid is order optimal method, and its asymptotic error is given by the following formula:

$$\delta_k = \Theta\big(k^{-1/2}\big),$$

where δ_k denotes the average error for the uniform grid of k points. Let us note, that the average error is defined with respect to the underlying statistical model. The disadvantage of the uniform grids is in their non-compositivity, i.e. the uniform grid of k point is not extensible to the $k + 1$ one. It is important to mention that for a fixed k a uniform grid is not necessary optimal (Zhigljavsky and Žilinskas, 2008); this fact highlights difference between average case and worst case optimality where uniform grids are generally optimal (Sukharev, 1971; Žilinskas, 2013).

For the detailed review of the efficiency in probabilistic sense we refer to the papers (Calvin, 2016a, 2016b), and for a discussion on the average complexity of Bayesian algorithms we refer to Mockus (1995).

3.5. *New Ideas and Open Problems*

The GO methods, based on statistical models of objective functions, are theoretically substantiated by the definition expressing their optimality. The high convergence rate, although proved for few special cases, complements to the theoretical soundness of the P-algorithm. However, some theoretical drawbacks, not appearing in publications, are worth to discuss.

Let us consider the one-step Bayesian algorithm (11) applied to the objective functions corresponding to the true statistical model. For example, let the algorithm be defined using the Gaussian stationary stochastic process with exponential correlations function, and sample functions of this process are used for objective functions. It is easy to check that after a modest number (say 20) of steps, the maximal average improvement normally does not exceed 0.01% of the standard deviation of the process since the average improvement (12) is computed via tails of the Gaussian density and cumulative distribution functions. Therefore, the rationality of maximization of such a totally small average improvement seems doubtful. If the rationality of the application of a method even in the situation of true mathematical model is doubtful, the theoretical investigation of this problem should be recognized as an urgent one. Doubts in the rationality of search casts also the rather frequent selection of a point for current computing an objective function value in a vicinity of a point of previous computation. Such a selection at early stage of search causes the instability of the whole process because the subsequently used covariance matrices become ill-conditioned.

The P-algorithm, when applied to a Gaussian statistical model, suffers a similar problem: in a situation described above, the maximal improvement probability is very low after small number of steps. However, the P-algorithm is not as sensitive to the probability distribution as (11). Let us compare the formulas (10) and (12). The formula (10) remains not changed for any distribution function in (9) of the form $\Pi(\frac{y-m}{\sigma})$ where m and σ denote mean value and standard deviation of the random variable in question. Therefore, the site of next computation of the objective function is the same for any distributions of the mentioned class, thus with considerably higher probability for the distributions with heavy tails. Recently an idea was proposed to derive an algorithm from the assumption of the search invariance

Let us consider the choice of a point for the current computation of the objective function value. Such a choice in the 'black box' situation is a decision under uncertainty, and the rational decision theory can be applied to make the choice rationally. The theory suggests to make the decision by maximization of the average utility. To compute the latter a statistical model of uncertainty is needed as well as an utility function. The axioms in Žilinskas (1982) substantiate the acceptance of a random variable as a model of uncertainty for the unknown value of the objective function. Accordingly a family of random variables ξ_x is acceptable as a statistical model of the objective function. In the previous investigation, summarized in Törn and Žilinskas (1989), Zhigljavsky and Žilinskas (2008), we proceeded with the specification of a distribution of ξ_x and of the utility function. Let at the moment limit ourselves by the assumption that there exist appropriate distribution and utility functions. We intend to construct an algorithm bypassing the necessity to specify the utility function and the distribution of ξ_x.

Any characterization of a random variable normally includes a location parameter (e.g. mean) and a spread parameter (e.g. standard deviation); we use minimal description of ξ_x by these two parameters which are denoted by $m(x)$ and $s(x)$. The dependence of both parameters on the information available at the current optimization step (x_j, y_j, $j = 1, \ldots, k$) will be included into the notation where needed. Let us assume that the average utility $u_{k+1}(x)$ of computation of the current objective function value at the point x depends on x via $m(x)$ and $s(x)$. A value desired to reach $y_{o,k}$, $y_{o,k} < \min_{1 \leq j \leq k} y_j$, is also assumed as a parameter which defines $u_{k+1}(x)$

$$u_{n+1}(x) = U\big(m(x), s(x), y_{o,k}\big), \tag{15}$$

and the point of current computation is defined as the maximizer of $u_{n+1}(x)$. The following assumptions on $U(\cdot)$ express rationality of invariance of the average utility with respect to the scales of the objective function values:

$$U\big(m(x) + c, s(x), y_{o,k} + c\big) = U\big(m(x), s(x), y_{o,k}\big),$$

$$U\big(m(x)\cdot C, s(x)\cdot C, y_{o,k}\cdot C\big) = U\big(m(x), s(x), y_{o,k}\big), \quad C > 0. \tag{16}$$

Since the minimization problem is considered, the current computed objective function value will be more valuable if smaller; therefore we postulate that

$$m < \mu \text{ implies } U(m, s, z) > U(\mu, s, z). \tag{17}$$

The postulated properties are inherent for several well known optimization algorithms; see e.g. Žilinskas (2012).

It can be shown that the average utility function satisfying the postulated properties has the following structure

$$U\big(m(x), s(x), y_{o,k}\big) = \Pi\left(\frac{y_{o,k} - m(x)}{s(x)}\right), \tag{18}$$

where $\pi(\cdot)$ is an increasing function. Therefore, the site for computing the current value of the objective function is given by the formula

$$x_{k+1} = \arg\max_{x\in\mathbf{A}} \frac{y_{o,k} - m(x|x_j, y_j, \ j = 1, \dots, k)}{\sigma(x|x_j, y_j, \ j = 1, \dots, k)}, \tag{19}$$

which is coincident with (10). Thus, the P-algorithm is rational for a broad range of statistical models and utility functions, and the criterion for choosing the current point is not necessary interpretable as the improvement probability.

The applications of numerical algorithms are tightly related to the available computing power. The fast increase of computing power influences also the development and applications of the algorithms based on statistical models of objective functions. Besides the enhancement of the classical computers, there occur also principally new possibilities, e.g. those related to the infinity computer (Sergeyev, 2010). The compatibility of the algorithm (9) and (11) with the arithmetic of infinity was shown in Žilinskas (2012), however the full scale exploitation of the potential of this perspective technology is still a challenging future project.

4. Conclusions

The theoretical foundation of the stochastic global optimization was laid in the seventies of the last century. At the beginning, many different ideas were proposed and tested. Later, the attention of the theory-oriented researchers focussed on the following directions: convergence of and statistical inference in random global search, and substantiation of the use of statistical models and properties of the model-based algorithms. The present review covers these topics as well as some recent generalizations to multi-objective optimization.

Acknowledgements. The work of the first author was supported by the Research Council of Lithuania under Grant No. MIP-051/2014. The work of the second author was supported by the Russian Science Foundation, project No. 15-11-30022 "Global optimization, supercomputing computations, and applications".

References

Auger, A., Hansen, N. (2010). Theory of evolution strategies: a new perspective. In: Auger, A., Doerr, B. (Eds.), *Theory of Randomized Search Heuristics: Foundations and Recent Developments*. World Scientific Publishing, New York, pp. 289–325.

Baritompa, W.P., Baoping, Z., Mladineo, R., Wood, G., Zabinsky, Z. (1995). Towards pure adaptive search. *Journal of Global Optimization*, 7(1), 93–110.

Brooks, S.H. (1958). A discussion of random methods for seeking maxima. *Operations Research*, 6(2), 244–251.

Brooks, S.H. (1959). A comparison of maximum-seeking methods. *Operations Research*, 7(4), 430–457.

Calvin, J. (2001). A one-dimensional optimization algorithm and its convergence rate under the Wiener measure. *Journal of Complexity*, 17, 306–344.

Calvin, J. (2007). A lower bound on complexity of optimization on the Wiener space. *Theoretical Computer Science*, 383, 132–139.

Calvin, J. (2011). An adaptive univariate global optimization algorithm and its convergence rate under the Wiener measure. *Informatica*, 22(4), 471–488.

Calvin, J. (2016a). On asymptotic tractability of global optimization. In: Pardalos, P., Zhigljavsky, A., Žilinskas, J. (Eds.), *Advances in Stochastic and Global Optimization*. Springer, pp. 3–12.

Calvin, J. (2016b). Probability models in global optimization. *Informatica*, 27(2).

Calvin, J., Žilinskas, A. (2000). A one-dimensional P-algorithm with convergence rate $O(n^{-3+\delta})$ for smooth functions. *Journal of Optimization Theory and Applications*, 106, 297–307.

Calvin, J., Žilinskas, A. (2005). A one-dimensional global optimization for observations with noise. *Computers and Mathemaics with Applications*, 50, 157–169.

Calvin, J., Žilinskas, A. (2014). On a global optimization algorithm for bivariate smooth functions. *Journal of Optimization Theory and Applications*, 163(2), 528–547.

De Haan, L. (1981). Estimation of the minimum of a function using order statistics. *Journal of the American Statistical Association*, 76(374), 467–469.

De Haan, L., Peng, L. (1998). Comparison of tail index estimators. *Statistica Neerlandica*, 52(1), 60–70.

DenHertog, D., Kleijnen, S. (2006). The correct kriging variance estimated by bootstrapping. *Journal of Operations Research Society*, 57, 400–409.

Dorea, C. (1990). Stopping rules for a random optimization method. *SIAM Journal on Control and Optimization*, 28(4), 841–850.

Emerich, M., Giannkoglou, K., Naujoks, B. (2006). Single-and multiobjective evolutionary optimization assisted by Gaussian random field metamodels. *IEEE Transactions on Evolutionary Computing*, 10, 421–439.

Emerich, M., Yang, K., Deutz, A., Wang, H., Fonseca, M. (2016). Multicriteria generalization of Bayesian global optimization. In: Pardalos, P., Zhigljavsky, A., Žilinskas, J. (Eds.), *Advances in Stochastic and Global Optimization*. Springer, pp. 223–236.

Floudas, C. (2000). *Deterministic Global Optimization: Theory, Methods and Applications*. Springer, Dordrecht.

Gimbutienė, G., Žilinskas, A. (2015). A two-phase global optimization algorithm for black-box functions. *Baltic Journal of Modern Computing*, 3(3), 214–224.

Grishagin, V., Sergeyev, Y., Strongin, R. (1997). Parallel characteristical global optimization algorithms. *Journal of Global Optimization*, (10), 185–206.

Gutmann, H.M. (2001). A radial basis function method for global optimization. *Journal of Global Optimization*, 19, 201–227.

Hamilton, E., Savani, V., Zhigljavsky, A. (2007). Estimating the minimal value of a function in global random search: Comparison of estimation procedures. In: *Models and Algorithms for Global Optimization*. Springer, pp. 193–214.

Hart, W.E. (1998). Sequential stopping rules for random optimization methods with applications to multistart local search. *SIAM Journal on Optimization*, 9(1), 270–290.

Horst, R., Pardalos, P., Thoa, N. (2000). *Introduction to Global Optimization*, 2nd. edition. Springer, Dordrecht.

Huang, D., Allen, T., Notz, W., Miller, R. (2006a). Sequential kriging optimization using multiple-fidelity evaluations. *Structural and Multidisciplinary Optimization*, 32, 369–382.

Huang, D., Allen, T., Notz, W., Zeng, N. (2006b). Global optimization of stochastic black-box systems via sequential kriging meta-models. *Journal of Global Optimization*, 34, 441–466.

Jones, D. (2001). A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, 21, 345–383.

Jones, D.R., Schonlau, M., Welch, W. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13, 455–492.

Kan, A.R., Timmer, G. (1987). Stochastic global optimization methods, part I. Clustering methods. *Mathematical Programming*, 39(1), 27–56.

Kleijnen, J., van Beers, W., van Nieuwenhuyse, I. (2012). Expected improvement in efficient global optimization through bootstrapped kriging. *Journal of Global Optimization*, 54, 59–73.

Knowles, J. (2006). ParEGO: a hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation*, 10(1), 50–66.

Knowles, J., Corne, D., Reynolds, A. (2009). Noisy multiobjective optimization on a budget of 250 evaluations. In: Ehrgott, M., et al. (Eds.), *Lecture Notes in Computer Science*, Vol. 5467. Springer, Berlin, pp. 36–50.

Kulczycki, P., Lukasik, S. (2014). An algorithm for reducing the dimension and size of a sample for data exploration procedures. *International Journal of Applied Mathematics and Computer Science*, 24(1), 133–149.

Kushner, H. (1962). A versatile stochastic model of a function of unknown and time-varying form. *Journal of Mathematical Analysis and Applications*, 5, 150–167.

Makauskas, A. (1991). On a possibility to use gradients in statistical models of global optimization of objective functions. *Informatica*, 2(2), 248–254.

Mockus, J. (1963). On a method for the most advantageous allocation of observations when solving multiextremal optimal design problems of complex systems by the Monte-Carlo method. In: *Symposium on Multiextremal Problems*, Kaunas, pp. 4–5 (in Russian).

Mockus, J. (1967). *Multiextremal Problems in Design*. Nauka, Moscow (in Russian).

Mockus, J. (1972). On Bayes methods for seeking an extremum. *Avtomatika i Vychislitelnaja Technika*, 3, 53–62 (in Russian).

Mockus, J. (1988). *Bayesian Approach to Global Optimization*. Kluwer Academic Publishers, Dordrecht.

Mockus, J. (1995). Average complexity and the bayesian heuristic approach to discrete optimization. *Informatica*, 6(2), 193–224.

Mockus, A., Mockus, J., Mockus, L. (1994). Bayesian approach adapting stochastic and heuristic methods of global and discrete optimization. *Informatica*, 5(1–2), 123–166.

Mockus, J., Eddy, W., Mockus, A., Mockus, L., Reklaitis, G. (1997). *Bayesian Heuristic Approach to Discrete and Global Optimization*. Kluwer Academic Publishers, Dordrecht.

Neimark, J., Strongin, R. (1966). An information approach to the problem of search of an extremum of functions. *Engineering Cybernetics* 1, 17–26 (in Russian).

Nevzorov, V.B. (2001). *Records: Mathematical Theory*. American Mathematical Society, Washington.

Niederreiter, H. (2010). *Quasi-Monte Carlo Methods*. Wiley Online Library.

Patel, N.R., Smith, R.L., Zabinsky, Z.B. (1989). Pure adaptive search in Monte Carlo optimization. *Mathematical Programming*, 43(1–3), 317–328.

Paulavičius, R., Žilinskas, J. (2014). *Simplicial Global Optimization*. Springer.

Pepelyshev, A. (2011). Fixed-domain asymtotics of the maximum likelihood estiomator and the Gaussian process approach for deterministic models. *Statistical Methodology*, 8(4), 356–362.

Picheny, V. (2014). A stepwise uncertainty reduction approach to constrained global optimization. In: *AISTATS, Conference Proceedings*, pp. 787–795.

Picheny, V. (2015). Multiobjective optimization using Gaussian process emulators via stepwise uncertainty reduction. *Statistics and Computing*, 25, 1265–1280.

Pintér, J.N. (1984). Convergence properties of stochastic optimization procedures. *Optimization*, 15(3), 405–427.

Rastrigin, L. (1964). Convergence of random search method in extremal control of many-parameter system. *Automation and Remote Control*, 24(11), 1337–1342.

Rastrigin, L. (1968). *Statistical Methods of Search*. Nauka, Moscow (in Russian).

Ritter, K. (1990). Approximation and optimization on the Wiener space. *Journal of Complexity*, 6, 337–364.

Rullière, D., Faleh, A., Planchet, F., Youssef, W. (2013). Exploring or reducing noise? A global optimization algorithm in the presence of noise. *Structural and Multidisciplinary Optimization*, 47, 921–936.

Sasena, M. (2002). *Flexibility and efficiency enhancements for constrained global design optimization with kriging approximations*, Dissertation. Michigan University, Ann Arbor.

Sergeyev, Y. (2010). New computational methodology using infinite and infinitesimal numbers. In: Bandini, S., et al. (Eds.), *Lecture Notes in Computer Science*, Vol. 6350. Springer, Berlin, pp. 646–649.

Solis, F.J., Wets, R.J.B. (1981). Minimization by random search techniques. *Mathematics of Operations Research*, 6(1), 19–30.

Strongin, R. (1969). Information method of global minimization in the presence of noise. *Engineering Cybernetics*, 6, 118–126 (in Russian).

Strongin, R.G. (1978). *Numerical Methods of Multiextremal Minimization*. Nauka, Moscow (in Russian).

Strongin, R.G., Sergeyev, Ya.D. (2000). *Global Optimization with Non-Convex Constraints: Sequential and Parallel Algorithms*. Kluwer Academic Publishers.

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

Sukharev, A.G. (1972). Best sequential search strategies for finding an extremum. *USSR Computational Mathematics and Mathematical Physics*, 12(1), 39–59.

Sukharev, A.G. (2012). Minimax models in the theory of numerical methods. In: *Theory and Decision Library*, Vol. 21. Springer Science & Business Media.

Tempo, R., Calafiore, G., Dabbene, F. (2012). *Randomized algorithms for analysis and control of uncertain systems: with applications*. Springer Science & Business Media.

Tikhomirov, A., Stojunina, T., Nekrutkin, V. (2007). Monotonous random search on a torus: Integral upper bounds for the complexity. *Journal of Statistical Planning and Inference*, 137(12), 4031–4047.

Tikhomirov, A.S. (2006). On the Markov homogeneous optimization method. *Computational Mathematics and Mathematical Physics*, 46(3), 361–375.

Tikhomirov, A.S. (2007). On the convergence rate of the Markov homogeneous monotone optimization method. *Computational Mathematics and Mathematical Physics*, 47(5), 780–790.

Törn, A., Žilinskas, A. (1989). Global optimization. *Lecture Notes in Computer Science*, 350, 1–252.

Traub, J., Weschulz, A. (1998). *Complexity and Information*. Cambridge University Press, Cambridge.

Weissman, I. (1981). Confidence intervals for the threshold parameter. *Communications in Statistics-Theory and Methods*, 10(6), 549–557.

Weissman, I. (1982). Confidence intervals for the threshold parameter, II. Unknown shape parameter. *Communications in Statistics-Theory and Methods*, 11(21), 2451–2474.

Yamakawa, M., Ohsaki, M. (2013). Worst-case design of structures using stopping rules in k-adaptive random sampling approach. In: *Proceedings of the 10th World Congress on Structural and Multidisciplinary Optimization*, Orlando, Florida, USA, May 20–24.

Zabinsky, Z.B. (2003). *Stochastic Adaptive Search for Global Optimization*. Kluwer, Boston.

Zabinsky, Z.B., Smith, R.L. (1992). Pure adaptive search in global optimization. *Mathematical Programming*, 53(1–3), 323–338.

Zhigljavsky, A. (1981). *Monte Carlo Methods in Global Optimization*. PhD thesis. Leningrad University, Leningrad.

Zhigljavsky, A. (1985). *Mathematical Theory of Global Random Search*. Leningrad University Press, Leningrad (in Russian).

Zhigljavsky, A. (1990). Branch and probability bound methods for global optimization. *Informatica*, 1(1), 125–140.

Zhigljavsky, A. (1991). *Theory of Global Random Search*. Kluwer, Dordrecht.

Zhigljavsky, A.A. (1993). Semiparametric statistical inference in global random search. *Acta Applicandae Mathematica*, 33(1), 69–88.

Zhigljavsky, A., Hamilton, E. (2010). Stopping rules in k-adaptive global random search algorithms. *Journal of Global Optimization*, 48(1), 87–97.

Zhigljavsky, A., Žilinskas, A. (2008). *Stochastic Global Optimization*. Springer.

Zieliński, R. (1981). A statistical estimate of the structure of multi-extremal problems. *Mathematical Programming*, 21(1), 348–356.

Žilinskas, A. (1975). One-step Bayesian method for the search of the optimum of one-variable functions. *Cybernetics*, 1, 139–144 (in Russian).

Žilinskas, A. (1982). Axiomatic approach to statistical models and their use in multimodal optimization theory. *Mathematical Programming*, 22, 104–116.

Žilinskas, A. (1985). Axiomatic characterization of a global optimization algorithm and investigation of its search strategies. *Operations Research Letters,*, 4, 35–39.

Žilinskas, A. (1986). *Global Optimization: Axiomatic of Statistical Models, Algorithms, Applications*. Mokslas, Vilnius (in Russian).

Žilinskas, A. (1990). Statistical models of multimodal functions and construction of algorithms for global optimization. *Informatica*, 1(1), 141–155.

Žilinskas, A. (1992). A review of statistical models for global optimization. *Journal of Global Optimization*, 2, 144–153.

Žilinskas, A. (2010). On similarities between two models of global optimization: statistical models and radial basis functions. *Journal of Global Optimization*, 48, 173–182.

Žilinskas, A. (2012). On strong homogeneity of two global optimization algorithms based on statistical models of multimodal objective functions. *Applied Mathematics and Computation*, 218(16), 8131–8136.

Žilinskas, A. (2013). On the worst-case optimal multi-objective global optimization. *Optimization Letters*, 7, 1921–1928.

Žilinskas, A. (2014). A statistical model-based algorithm for black-box multi-objective optimisation. *International Journal of System Science*, 45(1), 82–92.

Žilinskas, A., Gimbutienė, G. (2015). On asymptotic property of a simplicial statistical model of global optimization. In: Migdalas, A., Karakitsiou, A. (Eds.), *Optimization, Control, and Applications in the Information Age*. Springer, pp. 383–392.

Žilinskas, A., Mockus, J. (1972). On a Bayesian method for seeking the minimum. *Avtomatika i Vychislitelnaja Technika*(4), 42–44 (in Russian).

Žilinskas, A., Žilinskas, J. (2002). Global optimization based on a statistical model and simplicial partitioning. *Computers & Mathematics with Applications*, 44(7), 957–967.

Žilinskas, A., Žilinskas, J. (2010). P-algorithm based on a simplicial statistical model of multimodal functions. *TOP*, 18, 396–412.

Žilinskas, A., Zhigljavsky, A. (2016). Branch and probability bound methods in multi-objective optimization. *Optimization Letters*, 10, 341–353.

A. Žilinskas is a member of Lithuanian Academy of Sciences and holds the positions of a principal researcher and professor of Vilnius University. His scientific interests are statistical theory of global optimization, optimization based modelling and design, and analysis of multidimensional data by means of visualization. His research is oriented to develop statistical models of global optimization, implement and investigate the corresponding algorithms, and apply them to practical problems. He is a member of editorial boards of *Journal of Global Optimization*, *Informatica*, *Baltic Journal of Modern Computing*, *International Journal of Grid and High Performance Computing, Statistics, Optimization and Information Computing*, and *The Open Cybernetics and Systemics Journal*. He is a member of IFIP working group Optimization Based Computer Aided Modeling and Design and of American Mathematical Society.

A. Zhigljavsky was born in 1953. Graduated from the Faculty of Mathematics, St. Petersburg State University, in 1976. PhD on applied probability in 1981. Professor of statistics at the St. Petersburg State University during 1989–1997. Since 1997: Professor, Chair in Statistics at Cardiff University. Author or co-author of 9 monographs on the topics of stochastic global optimization (4), time series analysis (3), optimal experimental design (1) and dynamical systems (1); editor/co-editor of 8 books or special issue of renowned journals on various topics, author of about 150 research papers in refereed journals, organizer of several major conferences on time series analysis, experimental design and global optimization. Member of the editorial board of two journals: *Journal of Global Optimization* and *Statistics and Its Interface*. Director of the Centre for Optimisation and Its Applications at Cardiff University.

Stochastinė globali optimizacija: apžvalga *Informatica* dvidešimtmečio proga

Antanas ŽILINSKAS, Anatolij ZHIGLJAVSKY

Apžvelgti stochastinės globaliosios optimizacijos teoriniai ir metodologiniai pasiekimai. Apžvalgą sudaro dvi dalys: atsitiktinė globalioji paieška ir statistinių tikslo funkcijų modelių taikymas. Pagrindinis dėmesys kreipiamas į paskutiniųjų 25 metų pasiekimus teoriškai pagrįstų metodų ir matematinių rezultatų srityse.