A SURVEY ON THE GLOBAL OPTIMIZATION PROBLEM: GENERAL THEORY AND COMPUTATIONAL APPROACHES

F. ARCHETTI and F. SCHOEN
Department of Mathematics, University of Milan, Via L. Cicognara 7
I-20129 Milan, Italy

and

C.N.R., Istituto per le Applicazioni della Matematica e dell'Informatica
Via L. Cicognara 7, I-20129 Milano, Italy

Abstract

Several different approaches have been suggested for the numerical solution of the global optimization problem: space covering methods, trajectory methods, random sampling, random search and methods based on a stochastic model of the objective function are considered in this paper and their relative computational effectiveness is discussed. A closer analysis is performed of random sampling methods along with cluster analysis of sampled data and of Bayesian nonparametric stopping rules.

Keywords and phrases

Global optimization, statistical optimization, Bayesian statistics, random search.

1. Introduction

Let K be a compact set in the N-dimensional Euclidean space R^N. We define the global optimization problem with the following task:

find the couple $x^\star, f^\star$ such that:

$$f^\star = f(x^\star) \leqslant f(x) \quad \forall x \in K \tag{1.1}$$

where $f: K \longrightarrow R,\ f \in \mathcal{C}(K)$.

Many approaches, often of a very different nature, have been suggested to solve this problem: even if, by now, a class of methods is widely agreed upon as the most effective one, a general computational paradigm is still lacking.

First, we note that the very definition of the problem as given in (1.1) is not well posed in the classical sense. Indeed, even within infinitely differentiable functions with a unique global minimum point, it is possible to choose functions which are as close to one another as we want, yet their global minima are far apart. As a simple example we can take $f_\delta(x) = \cos(x) + \delta x,\ x \in (-2\pi, 2\pi)$. The optimum is located near $-\pi$ or $+\pi$ when δ is a small positive or negative constant, respectively. So, while two such functions $f_{\delta_1}(x)$ and $f_{\delta_2}(x)$ characterized by $\delta_1 < 0 < \delta_2$ have a supremum norm difference bounded by $2\pi|\delta_2 - \delta_1|$, the distance between the solutions is approximately 2π. On the other hand, the value $f^\star$ of the global minimum depends continuously on the data of the problem, in the sense that, as it is easy to show, for any continuous functions f and g we have:

$$|f^\star - g^\star| \leqslant \operatorname*{Sup}_{k \in K} |f(x) - g(x)| = \|f - g\|_\infty ,$$

so that the basic requirement of well-posedness of the global optimization problem is fulfilled if we restrict ourselves to the search for the global optimum value $f^\star$.

From now on, when speaking of the 'global optimization problem', we will refer to the following task:

find $f^\star$ such that:

$$f^\star \leqslant f(x) \quad \forall x \in K \tag{1.2}$$

where $f: K \longrightarrow R, f \in \mathcal{C}(K)$.

A basic difficulty of problem (1.2) is the impossibility of bounding the error in the approximate solution: in fact, if we let any algorithm choose n points in the domain K, it is always possible to build a smooth function, interpolating f in those points, whose global minimum value can be kept as far as desirable from $f^\star$.

If we restrict the function class to which f belongs, it is possible to build deterministic algorithms (space covering techniques), which have the property of finite convergence to an estimate whose distance from the optimum can be deterministically controlled. The price to be paid for this is, apart from the necessity of restricting the function class, the exponential increase of the computational effort with the dimension of problem (1.2), and the necessity of giving precise bounds to the variation of the objective function: thus their numerical usefulness is restricted to very small sized (say 1 or 2 dimensions) problems. Nevertheless, we shall discuss them in some detail in sect. 2 because the analysis of their performance can give some insight into the very nature of global optimization problems and its hidden intractability.

In sect. 3 we shall consider other deterministic methods, namely trajectory techniques and the tunneling approach, which are widely investigated but require a very complex implementation, and still do not offer a reliable numerical performance.

Probabilistic methods can be seen as a tool for overcoming the basic difficulty of problem (1.2) by allowing some kind of uncertainty to the final result. Thus a 'solution' found by any of these methods can be thought of as having a certain 'probability' of being the true one. This uncertainty does not seem to be too heavy a price in order to have information about the solution of problems which would otherwise be intractable. A class of probabilistic methods, namely those based on random sampling, which we shall be discussing in sect. 4, are now considered as the most reliable tool for solving global optimization problems. In sect. 5 we shall briefly be concerned with random search techniques which have received early attention in the literature [1,39], and can be regarded as an effective tool, at least in some instances, of global optimization. Section 6 will be devoted to the analysis of a particular class of probabilistic methods, in which the objective function is modeled as a sample path of a stochastic process.

We must also remark that the distinction between deterministic and probabilistic methods is quite crude as random elements are often introduced, as we shall see later on, into deterministic schemes to improve their performance.

We conclude this brief introduction by observing that because of the high computational cost of global optimization problems, some research is being devoted to the design and analysis of parallel algorithms, i.e. algorithms which can be implemented on multiprocessor systems and array processors [32,11,37].

2. Space covering techniques

Space covering techniques originate from the desire to provide deterministic bounds on the approximation error. This can be done by restricting the function class to which f belongs to a subclass where some bound on the variation of the objective function or its derivatives is known *a priori*. The easiest way to do so is to impose a Lipschitz condition on f:

$$f \in L_{\ell,\rho}(K) = \{ g : |g(x) - g(y)| \leqslant \ell \rho(x,y) \quad \forall x, y \in K \}, \tag{2.1}$$

where ℓ is a known positive constant and ρ is a continuous distance on K.

DEFINITION 1

An n-step deterministic strategy S_n is an n-dimensional vector of functions $\langle y_1, y_2, \ldots, y_n \rangle$ such that

$$y_1(f;K) = x_1 \in K,$$

$$y_{i+1}(x_1,f_1,x_2,f_2,\ldots,x_i,f_i) = x_{i+1} \in K f_i = f(x_i) \qquad i = 1,\ldots,n-1 .$$

We shall use the term 'passive' or '*a priori*' to denote those strategies which are constant in $L_{\ell,\rho}(K)$. In the following, we shall denote by $\mathcal{S}_n$ the class of strategies, and by $\mathcal{P}_n$ the subclass of passive ones.

In order to measure the effectiveness of a strategy, the *a posteriori* and *a priori* errors are introduced as follows:

DEFINITION 2

The accuracy of an n-step strategy S_n is:

$$\mathcal{A}(S_n,f) = \min_{i=1,\ldots,n} f(x_i) - f^\star .$$

DEFINITION 3

The accuracy guaranteed by an n-step strategy S_n over the class $L_{\ell,\rho}(K)$ is:

$$A(S_n) = \operatorname*{Sup}_{f \in L_{\ell,\rho}(K)} \mathcal{A}(S_n,f) .$$

In this class, any strategy guarantees a finite accuracy. In fact, let $(x_1, x_2, \ldots, x_n)$ be the points chosen by a strategy S_n and let

$$d = \operatorname*{Sup}_{x \in K} \ \min_{i = 1,n} \ \rho(x, x_i) .$$

Then it is easy to show that S_n guarantees the accuracy ld over the class $L_{\ell,\rho}(K)$. The following optimality criteria can be given for a deterministic strategy.

DEFINITION 4

A strategy $S_n^\star$ is A-optimal in $L_{\ell,\rho}(K)$ if:

$$A(S_n^\star) = \operatorname*{Inf}_{S_n \in \mathcal{S}_n} \quad \operatorname*{Sup}_{f \in L_{\ell,\rho}(K)} \quad \mathcal{A}(S_n, f) .$$

DEFINITION 5

Let $\epsilon > 0$. A strategy $S_{n^\star}$ such that $A(S_{n^\star}) \leqslant \epsilon$ is n-optimal in $L_{\ell,\rho}(K)$ if

$$n^\star = \min\left\{n : \exists S_n \in \mathcal{S}_n : A(S_n) \leqslant \epsilon\right\} .$$

By the above optimality criteria, strategies are evaluated on a 'worst case' basis, i.e. their effectiveness is measured on that function over which they display the poorest performance. Thus, not surprisingly, the following equivalence results can be shown to hold:

THEOREM 1 [47]

$$\operatorname*{Inf}_{P_n \in \mathcal{P}_n} \quad A(P_n) = \operatorname*{Inf}_{S_n \in \mathcal{S}_n} \quad A(S_n) .$$

THEOREM 2 [7]

Let $\epsilon > 0$.

$$\begin{aligned} &\min\left\{n : \exists\, S_n \in \mathcal{S}_n, \quad A(S_n) \leqslant \epsilon\right\} \\ &= \min\left\{n : \exists\, P_n \in \mathcal{P}_n, \quad A(P_n) \leqslant \epsilon\right\} . \end{aligned}$$

After these results, it is clear that an *a priori* analysis is not sufficient to give an effective criterion to select 'good' strategies, that is, strategies that, apart from pathological worst cases, display a better behaviour than passive ones.

What can be done, still in the framework of worst-case analysis, is to characterize those strategies which take advantage in an optimal way of the function evaluations already performed [48].

Let

$$L_{\ell,\rho;x_1,x_2,\dots,x_m;f_1,f_2,\dots,f_m}(K)$$

$$= \left\{\phi \in L_{\ell,\rho}(K):\phi(x_i) = f_i, \quad i = 1,\dots,m\right\},$$

and let 'choice' functions $C_m : K^m \longrightarrow K$ be given.

DEFINITION 6

A strategy $\Sigma_n^\star$ is sequentially A-optimal in $L_{\ell,\rho}(K)$ if:

$$y_{i+1} = C_{n-i} \circ P_{n\,|\,i}^\star ,$$

where $P_{n\,|\,i}^\star$ is an $(n-i)$-step A-optimal strategy in $L_{\ell,\rho;\,x_1,x_2,\dots,x_i;\,f_1,\dots,f_i}(K)$ and $\circ$ represents the composition operator.

DEFINITION 7

A strategy $\Sigma_{n\star}$ is sequentially n-optimal in $L_{\ell,\rho}(K)$ if:

$$y_{i+1} = C_{n^\star\,|\,i} \circ P_{n^\star\,|\,i} ,$$

where $P_{n^\star\,|\,i}$ is an n-optimal strategy in $L_{\ell,\rho;\,x_1,x_2,\dots,x_i;\,f_1,f_2,\dots,f_i}(K)$ and $n^\star|\,i$ is the number of steps it requires.

These strategies (which are easily seen to be optimal in the sense of Definitions 4 and 5 are designed in such a way as to adapt their behaviour to an 'updated' worst case and they can be considered as a good way of exploiting sequentiality while retaining *a priori* optimality. Unfortunately, the computational cost of actually building the optimal (passive) strategies required at each step of these algorithms makes them unfeasible for practical problems, even of very low dimension.

In [45] a method for one-dimensional optimization is described in which each new observation is placed where the uncertainty about the value of the objective function is maximum. Allowing $K = [0, 1]$, we have:

$$x_1 = 1/2 \ ,$$

$$x_{i+1} = \arg \max_{x \in K} \left\{ \min_{j = 1,i} f_j + \ell \, |x - x_j| \right\} \quad i = 1, 2, \ldots \ .$$

This strategy could be generalized in a straightforward way to higher dimensional space, but only in the one-dimensional case, the optimization problem whose solution is required in order to find each new point, allows for a simple solution.

Shubert's strategy is not optimal in the sense of Definition 5: it is easy to see that when the objective function is constant, an optimal (passive) strategy requires a number of steps which is roughly one half of those necessary for Shubert's algorithm to stop.

Another well-known deterministic strategy [25] is based on the idea of growing around each observation point a hypersphere whose radius is such that we can guarantee that the minimum of the objective function inside these spheres is bounded from below by $f_i^\star - \epsilon$, where $f_i^\star$ is the minimum observed value. Then each new observation is placed outside these hyperspheres, whose radius is easily found to be $R_i = (\epsilon + f(x_i) - f_i^\star)/\ell$. The algorithm stops when a covering of K has been performed with such spheres. If $N > 1$, in order to keep down the computational overhead and the memory requirements, the algorithm utilizes a covering made of N-dimensional hypercubes inscribed in the above-mentioned spheres, and each new point is chosen by changing one component of the last observation point by the quantity $(\epsilon/\ell + R_i)/\sqrt{N}$. Only in the one-dimensional case is Evthushenko's strategy optimal (more, it is sequentially n-optimal). It should be noticed, however, that the definition of the choice functions C_m in Definitions 6 and 7 is a crucial factor in determining the performance of sequentially optimal strategies in that a bad choice of C_m can completely annhilate the advantage of sequentially optimal schemes. Computational experience shows that, apart from pathological cases, Shubert's algorithm out-performs the one by Evtushenko. In [43] a scheme where the one-step optimality of Shubert's scheme is embedded in a sequentially n-optimal algorithm is presented.

3. Trajectory techniques and the tunneling approach

Trajectory techniques (see [17,18,29,51,53]) are based on the idea of finding many, hopefully all, local minima by means of the numerical integration of a differential equation. In [17,18], the solutions of the set of nonlinear equations:

$$\operatorname{grad} f(x) = 0 \tag{3.1}$$

are sought following the trajectories of the system of differential equations:

$$\mathrm{d}(\mathrm{grad}\, f)/\mathrm{d}t + \mathrm{grad}\, f = 0\,. \tag{3.2}$$

The relationship between this system and (3.1) can be easily recognized by observing that in the analytic solution

$$\mathrm{grad}\, f(x(t)) = \mathrm{grad}\, f(x(0)) \exp(-t) \tag{3.3}$$

the vector $x(t)$ converges, as $t \to \infty$, to a solution of (3.1). By making equation (3.2) explicit with respect to t, we obtain:

$$\mathrm{d}x/\mathrm{d}t = -H^{-1}(f(x))\, \mathrm{grad}\,(f(x))\,, \tag{3.4}$$

where $H(f(x))$ is the Hessian matrix of f in x. System (3.4) is not defined on the hypersurface $\det(H) = 0$. Since $H^{-1} = \mathrm{adj}(H)/\det(H)$, where $\mathrm{adj}(H)$ denotes the adjoint matrix of H, we can replace (3.4) by

$$\mathrm{d}x/\mathrm{d}t = -\,\mathrm{adj}(H)\, \mathrm{grad}\,(f(x))\;, \tag{3.5}$$

which amounts to a scale change in the parameter t and to a reversal of trajectory direction when the region $\det(H) = 0$ is crossed. Of course, once a local minimum has been found, a change of sign in (3.5) is necessary in order to escape from its region of attraction [26,27].

The methods based on this idea, although theoretically appealing, have serious drawbacks both from a theoretical and from a practical point of view. In [52] a counterexample is exhibited, where a region of non-convergence of Branin's method is shown to exist for a function whose contours are topologically equivalent to spheres. Secondly, this, as well as the other proposed trajectory approaches, fail in giving a precise answer to the most peculiar problem of global optimization, that is when to stop the computation; in other words, unless an *a priori* knowledge of the number of local minima of f is available, the algorithm cannot be stopped with the certainty of having located the global minimum.

In [58] an improvement to the basic scheme of trajectory methods has been proposed: random elements are introduced by associating to problem (1.2) the stochastic differential equation:

$$\begin{cases} \mathrm{d}x = -\,\mathrm{grad}\,(f(x)) + \epsilon(t)\mathrm{d}w \\ x(0) = x_0 \end{cases},$$

where w is an N-dimensional Wiener process. It is possible to prove, at least for some classes of functions that, provided that $\epsilon(t)$ approaches zero slowly enough as t tends to infinity, the probability for a trajectory of the above dynamic stochastic system leaving x_0 at $t = 0$ to reach a point x at time t is such that:

$$\lim_{t \to \infty} \; p(x,t;x_0,0) = \sum_i c_i \delta(x - x_i^\star),$$

where $x_i^\star$ are the global minimizers of f, $0 \leqslant c_i \leqslant 1$ and $\Sigma_i c_i = 1$. From a computational point of view, it seems more convenient to follow simultaneously several (say $M > 1$) trajectories, keeping $\epsilon(t)$ constant. After a number of steps of numerical integration, the M trajectories are compared. The 'worst' one is discarded and $\epsilon(t)$ is decreased. The numerical integration is continued after splitting one of the remaining $M - 1$ trajectories into two: the splitting is obtained very easily since, because of the stochastic term, each initial value problem is solved by an infinite number of trajectories.

More recently, Griewank [28] designed a trajectory method considering a model of the objective function given by the sum of a smooth unimodal function and a bounded perturbation. The algorithm consists of following the trajectories of the second order differential equation

$$x''(t) = -A(I - x'(t)x'^T(t))\nabla f(x(t))/(f(x(t)) - c) ,$$

where $A > 0$ and $c \geqslant f^\star$. As long as $f(x(t)) \gg c$, the trajectory is little affected by the perturbation.

Finally we mention a method which came to be termed the 'tunneling approach' [54,55]. The basic idea is that, starting from a local minimum point $\bar{x}$, a point x^0 is sought such that $f(x^0) \leqslant f(\bar{x})$. Starting from this point, a local optimization routine, a local optimum with function value not greater than $f(\bar{x})$ will be found. If $\{x_i^\star\}_{i=1}^{\ell}$ are those local minima already found whose function value is $f(\bar{x})$, this starting point can be found by solving the equation

$$T(x,\Lambda) = (f(x) - f(\bar{x}))/\left(\prod_{i=1}^{\ell} [(x - x_i^\star)^T(x - x_i^\star)]^{\lambda_i}\right) = 0 \qquad (3.6)$$

provided that the parameters $\Lambda = (\lambda_1, \ldots, \lambda_\ell)$, inserted to ensure that the solution of (3.6) will be different from the $x_i^\star$'s, are properly set. Even if computational results are reportedly good, a basic weakness is inherent to the stopping rule of this algorithm: deciding that no root of $T(x,\Lambda)$ exists in K can be as hard a problem as the global optimization problem (1.2) itself.

4. Methods based on random sampling

Probabilistic methods have been proposed since the earliest studies in global optimization and they have been gaining increasing attention in the last decade. Some of them, namely those based on a clever combination of random sampling and cluster analysis, can now be regarded as effective tools for the numerical solution of global optimization problems and display the following positive features:

(a) the required number of function evaluations grows rather slowly with the dimension of the problem;

(b) they are sequential in nature, i.e. at each stage an approximation to the global optimum is given and the decision is taken whether to stop and accept the approximation, or to reject it and perform further sampling;

(c) they require little *a priori* information about the objective function.

Methods based on random sampling can be structured after the following pattern:

(i) Draw q points $\{x_j \in K\}$, $j = 1, \ldots, q$ from a uniform distribution in K and compute $f(x_j)$.

(ii) Select the 'most promising points' and start from them a local optimization routine obtaining a least value f_c.

(iii) Test whether f_c is the global minimum of f in K.

The above cycle is repeated until the test is satisfied. In the simplest such algorithm, the point yielding the least sampled value is used in (ii) as the starting point, and the test in phase (iii) is satisfied if no improvement over f_c is observed in one (or more) further executions of phase (i). This simple test relies upon the fact that, as the sample size increases, the probability of not improving over f_c decreases unless f_c is the global minimum. In this method no more than one local optimization is performed in the region of attraction of a minimum: this good feature, unfortunately, is obtained at the cost of wasting most of the information contained in the sample. Effective algorithms require a more balanced compromise between the conflicting goals of making good use of the available information and of reducing the risk of converging to a local minimum already found: this can now be properly accomplished using cluster analysis. The third step is clearly the critical part of these algorithms: it is very difficult to evaluate the probability of the result being exact or, more generally, the probability that some meaningful index of the error does not exceed a prefixed level. Only very recently, some important steps have been taken in order to frame the third step into correct statistical terms.

We now proceed to a closer analysis of steps (i)–(iii).

(i) GENERAL PROPERTIES OF UNIFORM SAMPLING

We have already remarked, while discussing the general structure of probabilistic methods, that the uniform sampling of f in K is meant to provide a sample for the inferential processes to be carried out on in the following stages of the algorithm, rather than to provide directly, as in the crude Monte Carlo, an approximation to $f^\star$. Still, a 'per se' analysis of crude Monte Carlo can be performed, albeit only to some extent and in mainly negative terms. Let $K_0 \subseteq K$ be such that $m(K_0)/m(K) = \alpha$, where m denotes the Lebesgue measure on K: the probability $P(K_0\,; q)$ that at least one point in the sample $\{x_j\}$, $j = 1, \ldots, q$ will belong to K_0 is given by

$$P(K_0\,; q) = 1 - (1 - \alpha)^q \ .$$

Now, let K_0 be a neighborhood of $x^\star$: one could derive, given a value α and a probability level $\bar{P}$, a sample size $\bar{q} = \log(1 - \bar{P})/\log(1 - \alpha)$ and assume $f^\star_{\bar{q}} = \min\limits_{j=1,\bar{q}} f(x_j)$ as an approximation to $f^\star$.

A Monte Carlo procedure based upon the sequential application of this formula has been proposed in [2]; nothing, however, can be said, for finite values of q and without specific assumptions about f, about the probability that the error $f^\star_{\bar{q}} - f^\star$ exceeds a prefixed value, nor can we assure that the value $f^\star_{\bar{q}}$ has been achieved in K_0.

If $f \in L_{\ell,\rho}(K)$, the effectiveness of random sampling can be evaluated against that given by grid search, also for finite values of q. Uniform random sampling is argued to be more effective than grid search for $N > 6$ in [3], where the statistical distribution of the largest gap of an N-dimensional uniform sample and therefore the expected value of the error $f^\star_{\bar{q}} - f^\star$ are computed. A partially conflicting claim – possibly due to more general assumptions about the location of $x^\star$ – is in [6], where the performance of uniform random sampling after the criterion of guaranteed expected accuracy is computed and shown to be poorer than that of grid search, which is an optimal passive deterministic strategy.

(ii) CLUSTER ANALYSIS

By cluster analysis we mean a set of statistical techniques aimed at dividing a set of data into subsets (clusters) of 'similar' objects (see [30,24] for a general reference). Its relevance to the numerical solution of global optimization problems, first stressed in [49,50], has been subsequently substantiated by a number of highly successful implementations, of which we quote [15]. In the following, we shall briefly outline a possible way of performing the basic steps in the application of a clustering procedure to global optimization problems.

The original sample $\{x_j\}$, $j = 1, \ldots, q$ is modified, discarding those points whose function value is larger than $(1 - \gamma)f_\ell + \gamma f_\mu$, where $0 \leqslant \gamma \leqslant 1$ and f_ℓ and f_μ

are, respectively, the smallest and largest observed value of the objective function.

Let $\{y_j\}$, $j = 1, \ldots, p$ be the new sample. Let Y be a random variable uniformly distributed in $\bar{K} = \{x \in K: f(x) < (1 - \gamma)f_\ell + \gamma f_\mu\}$,$\{y_j\}$ is a sample of Y: clusters can now be associated to subsets of K where the distribution of Y is uniform, i.e. to the connected components of the support of the density function of Y.

By means of clustering techniques it is then possible to allocate the points $\{y_j\}$ to different clusters in such a way that the points in the region of attraction of the same local minimum are assigned to the same cluster.

Now we briefly mention a clustering technique as it has recently been implemented in [19] which blends the positive features of the two main approaches known in the literature, namely density and single-linkage clustering.

Clusters are grown around a center $\bar{y}$, called seed point. The control of the growth of the cluster is based upon the following result.

Let $\{x_j\}$, $j = 1, \ldots, q$ be uniformly distributed in K; let also $x \in K$ be a given point and ρ_j be the distance between x_j and x. Then it can be shown that

$$2\lambda(q)\rho_{(k)}^{N} \sim \chi^2_{2k}$$

where $\lambda(q) = \pi^{N/2}q/(m(K)\cdot\Gamma(\frac{1}{2}N + 1))$ is the expected number of points in $\{x_j\}$, $j = 1, \ldots, q$ falling in the unit N-dimensional hypersphere, and $\rho_{(k)}$ is the kth order statistics from $\{\rho_j\}$, $j = 1, \ldots, q$.

Therefore, given a probability level $1 - \epsilon$, we can compute $r_{(k)}$ such that $P\left\{\rho_{(k)} \leqslant r_{(k)}\right\} \geqslant 1 - \epsilon$. A cluster is built including those points of the sample $\{y_j\}$ belonging to $D_k \backslash D_{k-1}$ $(D_0 = \phi)$, where D_k is the hypersphere centered in $\bar{y}$ with radius $r_{(k)}$. When no more points can be added in this way, i.e. the set $(D_{\ell+1} \backslash D_\ell) \cap \{y_j\} = \phi$, the same cluster is enlarged by restarting the procedure from each point in $D_\ell \backslash D_{\ell-1}$.

This basic scheme can be improved by the use of local searches along the clustering procedure. A termination criterion of the whole clustering procedure can be naturally embedded in a scheme of sequential sampling of f:

let $X^\star_{n-1} = (x^\star_1, x^\star_2, \ldots, x^\star_{n-1})$ be the set of local minima already found; when clusters have been grown around $X^\star_{n-1}$, a local search is applied to the unclustered point with the least function value, yielding a local optimum $\bar{x}$. If $\bar{x} \in X^\star_{n-1}$, a cluster is grown around the starting point of the last local search detecting $\bar{x}$; if $\bar{x} \notin X^\star_{n-1}$, we set $x^\star_n = \bar{x}$, perform a new sample $\{y_j\}$ and grow clusters around $X^\star_n$. When all points have been clustered, or the local search has been applied to all unclustered points, the clustering procedure terminates.

(iii) STOPPING RULES

A basic fact about global optimization algorithms which can now be properly understood after the discussion of the clustering approach, is that the global part of a successful algorithm, actually the probabilistic part of it, is connected not with the numerical approximation of the global optimum, which is far more effectively performed by local searches, but rather with the control of these local searches and the decision whether a local minimum can be accepted as the global one. This fact, clearly behind the computational success of the clustering approach, amounts to a recognition of the multimodal structure of the objective function. This fact is clearly recognized in the latest papers about global optimization, whose main concern is the development of proper stopping rules. The main alternative approaches so far suggested will be dealt with in the remainder of this section. An important step towards a precise mathematical formulation of the stopping problem has been taken in [56].

Let $X_k^* = (x_1^*, x_2^*, \dots, x_k^*)$ be the set of local minima of f in K and $L : K$—X_k^* be an operator defined by $L(x) \in X_i^*$, which can be naturally interpreted as the application of a local search starting in x and converging to that local minimum x_i^* in whose region of attraction x lies.

Let $A_i^* = \{x \in K : L(x) = x_i^*\}$, $i = 1, \dots, k$ and $\theta_i = m(A_i^*)/m(K)$. The normalization condition $\Sigma_{i=1}^k \, m(A_i^*) = m(K)$ holds and the value θ_i will be called the 'share' of the ith local minimum.

By solution of a global optimization problem we mean, in this context, the determination of the values k and $\{\theta_i, x_i^*\}$, $i = 1, \dots, k$. A Monte Carlo method for finding this solution can be set up computing from the original sample $\{x_j\}$, $j = 1, \dots, q$ the values $L(x_j)$: let Q_i be the cardinality of the set $L^{-1}(x_i^*) \cap \{x_j\}_{j=1}^q$.

Then the random variable $(Q_1, Q_2, \dots, Q_k)$ has the multinomial distribution:

$$\text{Prob}\left\{Q_1 = q_1, Q_2 = q_2, \dots, Q_k = q_k\right\}$$

$$= \begin{pmatrix} q \\ \\ q_1, q_2, \dots, q_k \end{pmatrix} \theta_1^{q_1} \, \theta_2^{q_2}, \dots, \theta_k^{q_k} \, .$$

Hence, a set of local extrema obtained by performing a number of local searches from uniformly distributed starting points can be interpreted as a sample drawn from this multinomial distribution. If the value k is known in advance, then the random vector $(Q_1/q, \; Q_2/q, \dots, Q_k/q)$ is the standard minimum variance unbiased estimator of $(\theta_1, \theta_2, \ldots, \theta_k)$. If an upper bound U of k is known, we can still apply the above estimator and compute the probability for each $k \leqslant U$ of being the true number of local minima.

If nothing can be said *a priori* about k, a Bayesian framework is required, after which an *a priori* distribution is given on the set $\Theta = \{\theta_1, \theta_2, \ldots, \theta_k; k = 1, \ldots, \infty$ $0 \leqslant \theta_i \leqslant 1, \Sigma_{i=1}^{k} \theta_i = 1\}$ and after a suitable loss function has been defined, the optimal decision can be made about the value k and the share of the local minima.

Therefore, as in [16], the *a posteriori* probability can be computed that another local search will lead to the identification of a new local minimum. This information can be used to determine optimal Bayesian stopping rules to balance the costs of premature termination and that of further sampling.

Let us now reformulate the global optimization problem as that of finding the essential infimum of f in the compact set K:

$$f^\star = \max\{t : m(x \in K : f(x) \leqslant t)/m(K) = 0\}.$$

The above definition makes sense if f is Lebesgue measurable; if $f \in \mathcal{C}(K)$, the essential infimum coincides with the global optimum of definition (1.2). Thus, apart from pathological cases of no computational interest, the two definitions can be regarded as equivalent.

The interest of this new definition is that it is more sensible from a probabilistic viewpoint – functions which differ on a set of measure zero are indistinguishable by a random sampling process; thus, it will be used to introduce the main approaches to the inferential process about the objective function and, more specifically, about its global optimum.

Let:

$$\psi(t) = m\{x \in K : f(x) \leqslant t\}\{/m(K)\},$$

which is defined if f is Lebesgue measurable. It is possible to prove that

(i) ψ is non-decreasing.

(ii) ψ is a.e. differentiable.

(iii) ψ is continuous in R, provided that no set $H \subseteq K$ exists such that $m(H) > 0$ and $f(x) =$ constant on H.

The essential infimum $f^\star$ of f can thus be characterized by the condition

$$f^\star = \max\{t : \psi(t) = 0\}.$$

We define the ϵ-approximation $f^\star_\epsilon$ to $f^\star$ as:

$$f^\star_\epsilon = \max\{t : \psi(t) \leqslant \epsilon\} ,$$

and accept f_c as an approximation to $f^\star$ within an 'accuracy' ϵ if:

$$f_c \leqslant f^\star_\epsilon, \quad \text{i.e.} \quad \psi(f_c) \leqslant \epsilon , \tag{4.1}$$

since ψ is a non-decreasing function of t.

Condition (4.1) bears no implication about the value of the error $f_c - f^\star$: it only implies that the probability of finding, by further uniform samples, a function value lower than f is smaller than ϵ and thus makes for a sensible stopping criterion for an algorithm based on random sampling.

Should $\psi(t)$ be known, then the validation of f_c could be solved; but this is not the case, apart from some trivial cases. The first idea which has been explored in this framework was of deriving an analytical approximation to ψ, based on random sampling.

Let Z be a random n-dimensional vector uniformly distributed in K. If $P(t) = \text{Prob}\{f(Z) \leqslant t\}$ is the probability of hitting the set $E(t) = \{x \in K : f(x) \leqslant t\}$, then the uniformity of the distribution of Z implies $\psi \equiv P$. Thus, ψ is also the distribution function of the random variable $f(Z)$.

Given a value $\bar{t}$ and a sample $S_q = \{f(z_j)\}, j = 1, \ldots, q$, where z_j are samples from Z, if p is the number of points hitting $E(t)$, then p/q is an unbiased estimator of $\psi(t)$. One can construct a least squares spline approximation $\tilde{\psi}$ to the data thus obtained, for different values of t, update the approximation $\tilde{\psi}$ as new samples are generated, and show that this regression function converges uniformly to ψ. Once $\tilde{\psi}$ can be regarded as a satisfactory approximation it can be used to control whether $\tilde{\psi}(f_c) \leqslant \epsilon$.

Even if this approach is sensible and performs reasonably well on the usual test problems, still it has two main drawbacks:

(1) no statistical measure of the confidence one can place in the result can be derived;

(2) the choice of the approximation model is quite arbitrary.

A major step towards obtaining a simpler, more general and statistically meaningful procedure has recently been taken by Betrò [12,13].

As ψ is the probability distribution function of $f(Z)$, then the ϵ-approximation $f^\star_\epsilon$ is its quantile of order ϵ. Thus, testing whether $f_c \leqslant f^\star_\epsilon$ is a problem of testing about a quantile of an unknown distribution function. The approach is based on the idea of modelling the distribution of the sampled values by a suitable family of random distribution functions. These distribution functions are the sample paths of a suitable stochastic process and allow the making of inferences about the unknown distribution on the basis of the sampled values. The family considered in [13,14] is that of neutral to the right random probabilities, which enjoys the properties of being 'wide enough', closed under conditioning on the observed values and computationally manageable. The test in (4.1) is then formulated as a problem of optimal decision: let d_0 denote the decision of accepting the hypotheses $f_c \leqslant f_\epsilon^\star$ and d_1 the decision of rejecting it; let the losses connected with the wrong decisions be measured by two positive constants w_0 and w_1. Then the optimal decision, i.e. the one which minimizes the expected loss, is:

$$d_0 \text{ when } P(f_c \leqslant f_\epsilon^\star) \geqslant w_0/(w_0 + w_1) ,$$

$$d_1 \text{ otherwise.}$$

This approach has been implemented in connection with a clustering technique and has proven extremely successful [14].

Another approach employs tools from order statistics [20.5,21.15]. Let $\{x_j\}$, $j = 1, \ldots, q$ be uniformly distributed points in K; $f_c = \min_{j=1,q} f_j$ is a random variable whose distribution is given by:

$$G_q(t) = 1 - [1 - \psi(t)]^q .$$

For large values of q, $G_q(t)$ takes the following asymptotic form

$$G_q(t) = \begin{cases} 0 & t < f^\star \\ & , \\ \exp(1 - (t - f^\star)^\alpha / aq) & t \geqslant f^\star \end{cases} \tag{4.2}$$

where the influence of $\psi(t)$ is expressed by 3 parameters. It is possible to identify a wide class of functions which this result applies to and to compute for those functions the value α. The other unknown parameters can be estimated by a sequential sampling process. Once G_q is fully identified, it can in principle be used to provide confidence statements, at different levels of probability, for the estimate to the global optimum approximation as yet obtained. A main difficulty connected with the use of order statistics is that an exceedingly large sample is required to test that the distribution of the extremum can be modelled by the asymptotic expression (4.2).

5. Random search methods

The basic structure of these methods can be simply outlined by the following iterative formula:

$$\eta_k = x_k + \xi_k$$

$$x_{k+1} = \begin{cases} \eta_k & \text{if} \quad f(\eta_k) < f(x_k) \\ x_k & \text{otherwise}\,. \end{cases}$$

Here ξ_k is drawn from some probability distribution μ_k given in K. Different random search techniques depend on the choice of μ_k: it is natural to allow μ_k to be dependent on $x_1, x_2, \ldots, x_k$ and $f(x_1), f(x_2), \ldots, f(x_k)$ obtaining an adaptive random search [44]. These features can make random search techniques of some computational interest in some instances of large scale nonlinear programming [4] where we trade the typically slow convergence of random search techniques for their reduced requirements of computer storage, in stochastic optimization [22,42], where we capitalize on their robustness, and in some instances of global optimization problems, where the random nature of the search allows us to jump over local minima, converging, albeit very slowly, to the global one [38]. General convergence conditions, as well as computational results of specific choices of μ_k, are reported in [46].

Some instances of random search techniques can be viewed as stochastic gradient techniques (see [23] for a general survey on the use of stochastic gradient methods in systems optimization), and sensibly analyzed within the framework of the properties of convolution operators, which can be shown to smooth out the local minima of f for suitable choices of the kernel of the convolution and its parameters [41].

6. Methods based on a stochastic model of the objective function

The basic idea of the methods considered in this section is to model the objective function f as a sample path of a stochastic process $\phi(x, \omega)$ – the stochastic model of f – where ω is an unknown index which belongs to some sample space Ω. A probability distribution on Ω is given by means of the finite dimensional joint distributions

$$P_{x_1, x_2, \ldots, x_n}(Y_1, Y_2, \ldots, Y_n)$$

$$= P\{\omega : \phi(x_j, \omega) \leqslant y_j, \quad j = 1, \ldots, n\} \; \forall n > 0; \; x_1, x_2, \ldots, x_n \in K\,.$$

This amounts to giving a probability distribution on the instances of global optimization problems. Thus in this scheme, which was first suggested in [31,33], we can, at least in principle, move from worst case analysis to average case analysis, computing the *a posteriori* expected performance of algorithms (e.g. of space covering techniques), and we can design algorithms on the basis of average, rather than guaranteed accuracy. A consequence of modelling the objective function as a sample path of a stochastic process is that probabilistic bounds for the error $f^* - f_\epsilon^*$ can, at least in principle, be computed.

Algorithms in this framework can be developed according to two criteria. Let us assume that we can evaluate the objective function in n points; after the first criterion the points are chosen in such a way as to minimize the expected value of the objective function f; after the second scheme, evaluation points are chosen in such a way as to minimize the probability that the final error exceeds a prefixed quantity ϵ. In this framework algorithms, or strategies (borrowing the notation of space-covering techniques), can be considered vectors $d_1, d_2, \dots, d_n$, such that d_j, $j = 1, \dots, n$ is a measurable function of $d_1, d_2, \dots, d_{j-1}$ and maps $(K \times R)^j$ into K. An n-step strategy d, when applied to a function f, produces an n-tuple

$$x_1^d, x_2^d = d_2(x_1, f(x_1)), \dots, x_n^d = d_n(x_1, f(x_1), \dots, x_{n-1}, f(x_{n-1})) .$$

We know that $E(f(x) \mid f(x_1), f(x_2), \dots, f(x_n))$ is the least squares approximation to $f(x)$ given the observations $f(x_1), f(x_2), \dots, f(x_n)$. Then we can proceed, following a dynamic programming scheme, by first computing

$$u_n = \min_{x_n \in K} E(f(x_n) \mid f(x_1), \dots, f(x_{n-1})) ,$$

$$d_n = \arg \min_{x_n \in K} E(f(x_n) \mid f(x_1), \dots, f(x_{n-1})) .$$

Further steps backward lead to the determination of

$$u_{n-1} = \min_{x_{n-1} \in K} E(u_n(x_1, f(x_1), \dots, x_{n-1}, f(x_{n-1})) \mid f(x_1), \dots, f(x_{n-2})) ,$$

$$d_{n-1} = \arg \min_{x_{n-1} \in K} E(u_n(x_1, f(x_1), \ldots, x_{n-1}, f(x_{n-1})) \mid$$

$$f(x_1), \ldots, f(x_{n-2})) \, ,$$

until the first evaluation point $x_1^d = d_1$ is found. The points $x_2^d, \ldots, x_n^d$ are then found by substitution in:

$$x_j^d = d_j(x_1^d, f(x_1^d), \ldots, x_{j-1}^d, f(x_{j-1}^d)) \qquad j = 2, \ldots, n \, .$$

It is easy to show that this strategy d is such that

$$E(f(x_n^d)) = \operatorname*{Inf}_{\bar{d}} E(f(x_n^{\bar{d}})) \, ,$$

where $\bar{d}$ ranges over all n-step strategies.

One can show, under wide conditions [10], that $E(f(x_n^d) - f^\star) \to 0$ for $n \to \infty$ and that $f(x_n^d) \to f^\star$ in probability as n grows to infinity. In particular, $P\{f(x_n^d) - f^\star > \epsilon\} \geqslant \operatorname*{Inf}_{\bar{d}} P\{f(x_n^{\bar{d}}) - f^\star > \epsilon\}$.

If a strategy d exists for which the Inf is attained, then d is termed P-optimal (optimal in probability): P-optimal strategies can be computed by a recursive scheme analogous to that of E-optimal ones.

It should be clear that these optimality schemes do not lend themselves easily to the design of actual optimization algorithms, which should rather use the common one-step approximations, after which each point is chosen as it should be if it were the last one. After this framework, the scheme for the choice of the next point, corresponding to E-optimal schemes, can be defined as follows:

$$x_{i+1} = \arg \min_{x \in K} E(f(x) \mid f(x_1), f(x_2), \ldots, f(x_i))$$

and analogously for the P-optimal scheme.

Let us first consider the case $N = 1$, where one can take full advantage of the characteristic properties of these schemes. Let $K = [x_0, \bar{x}]$; as a stochastic model for f, the Wiener process is assumed for which

$$f(x_0) = \mu \ ,$$

$$f(x) - f(y) \sim N(0, \sigma^2 \, |x - y|) \quad \forall x, y \in K \, .$$

This choice is sensible for the following reasons:

(1) it is a good model of the global behaviour of f and its goodness of fit can be easily checked by testing the randomness and the normality of a sample of size $\bar{N}$ of values of the objective function. These statistical tests lead to the rejection of the Wiener model when the sample size $\bar{N}$ exceeds some value $\hat{N}$: the rejection of the Wiener process can be easily explained: when the number of observations is large, the local feature of f becomes noticeable and the sample paths of the Wiener process, due to its a.e. non-differentiability, are a very crude model of the local behaviour of a smooth objective function;

(2) the following simple formulae hold: let

$$Z_n = \left\{ f(x_j), \ j = 1, \ldots, n \right\}, \quad x \in \Delta_i = [x_i, x_{i+1}], \quad i = 0, \ldots, n-1 \ ,$$

$$\mu(x) = E(f(x) \,|\, Z_n) = f_i(x_{i+1} - x)/(x_{i+1} - x_i) + f_{i+1}(x - x_i)/(x_{i+1} - x_i) \, ,$$

$$\sigma^2(x) = \operatorname{var}(f(x) \,|\, Z_n) = \sigma^2(x - x_i)(x_{i+1} - x)/(x_{i+1} - x_i) \ ,$$

while, for $i = n$ and $x \in \Delta_n = [x_n, \bar{x}]$ we have:

$$\mu(x) = E(f(x) \,|\, Z_n) = f(x_n) \ ,$$

$$\sigma^2(x) = \operatorname{var}(f(x) \,|\, Z_n) = \sigma^2(x - x_n) \ .$$

The following result can be shown to hold: let $f(x)$, $x \in [a, b]$ be a Wiener process such that $f(a) = f_a$ and consider the distribution

$$F(z) = P\left\{ \min_{a \leqslant x \leqslant b} f(t) \leqslant z \,|\, f(b) = f_b \right\} .$$

Then:

$$f(z) = \begin{cases} 1 & \text{if} \quad z \geqslant \min\{f_a ; f_b\} \\ \exp(-2(f_a - z)(f_b - z)/(\sigma^2(b - a)) & z < \min\{f_a ; f_b\} \, . \end{cases}$$

An effective one-dimensional algorithm can be designed which exploits naturally the above formula [9].

The same approach could be extended to multivariable functions ($N > 1$) considering gaussian random fields as stochastic models. The application of two such models is reported in [34,40]: their numerical usefulness is severely limited by the significant computational overhead connected with the updating of the inverse of the covariance matrix.

A radical alternative for avoiding the large amount of computation required to update the model is to use simplified models, based on heuristic considerations rather than derived from a general random function [8,35,36].

Let $f(x_1), f(x_2), \ldots, f(x_n)$ be the values of f in the points $x_1, x_2, \ldots, x_n$; after a simple heuristic model we can assume $\forall x \in K$, $x \neq x_j$, $j = 1, \ldots, n$, that $f(x)$ is a normal random variable with expected value

$$\mu_n(x) = \sum_{i=1}^{n} (f(x_i)/\|x - x_i\|) / \sum_{i=1}^{n} (1/\|x - x_i\|) \qquad (6.1)$$

and variance

$$\sigma_n^2(x) = \sigma^2 \min_{i=1,n} \left\{ \|x - x_i\| \right\}, \qquad (6.2)$$

where $\|\cdot\|$ is the euclidean norm and σ^2 is a parameter of the model. Even if (6.1) and (6.2) are not derived from a random function, but assumed only on heuristic grounds, we shall speak, allowing for some impropriety, of conditional probability of the r.v. $f(x)$, meaning the probability, computed under the normality assumption to $f(x)$, and related through (6.1) and (6.2) to the values of $f(x)$ already observed. A theoretical analysis of simplified models has been developed in [57] in the framework of the theory of rational choice.

The models considered in this last section have, at least in the authors' opinion, two major drawbacks with negative computational implications: first, the complexity of these models grows with the dimension of the problem, while the statistical techniques outlined in sect. 4 are rather insensitive to the dimension of the problem. Moreover, the design of the algorithms of this section is tuned to a reduction (either in expected value, or in probability) in the error in the approximation to f. This seems to be much more effectively accomplished by means of local searches, and this fact is clearly recognized in the algorithms based on random sampling, where the probabilistic part of the algorithm performs the 'decisions' and local searches perform the approximation.

References

[1] R.L. Anderson, Recent advances in finding best operating conditions, J. Amer. Stat. Assoc. 48(1953)80.

[2] R.S. Anderssen, Global optimization, in: Anderssen, Jennings and Ryan, Optimization (University of Queensland Press, 1972) p. 26.

[3] R.S. Anderssen and P. Bloomfield, Properties of the random search in global optimization, J.O.T.A. 16, no. 5/6 (1975)91.

[4] F. Archetti, Evaluation of random gradient techniques for unconstrained optimization, Calcolo, Vol. XII, f.1 (1975)83.

[5] F. Archetti, B. Betró and S. Steffé, A theoretical framework for global optimization via random sampling, Quaderni del Dipartimento di Ricerca Operativa e Scienze Statistiche A – 25 (Universitá di Pisa, 1975).

[6] F. Archetti and B. Betró, On the effectiveness of uniform random sampling in global optimization problems, Quaderni del Dipartimento di Ricerca Operativa e Scienze Statistiche A – 32 (Universitá di Pisa, 1977).

[7] F. Archetti and B. Betró, A priori analysis of determinsitic strategies for global optimization problems, in: Towards Global Optimization 2, ed. L.C.W. Dixon and G.P. Szego (North-Holland, Amsterdam, 1978) p. 31.

[8] F. Archetti, A stopping criterion for global optimization algorithms, Quaderni del Dipartimento di Ricerca Operativa e Scienze Statistiche A – 61 (Universitá di Pisa, 1979).

[9] F. Archetti and B. Betró, A probabilistic algorithm for global optimization, Calcolo Vol. XVI, III (1979)335.

[10] F. Archetti and B. Betró, Stochastic models and optimization, Bollettino della Unione Matematica Italiana 5, 17 – A (1980) p. 295.

[11] F. Archetti and F. Schoen, Asynchronous parallel search in global optimization problems, in: Proc. X IFIP Conf. on System Modeling and Optimization, Lecture Notes on Control and Information Sciences, Vol. 38 (Springer-Verlag, 1982) p. 500.

[12] B. Betró, A Bayesian nonparametric approach to global optimization, Methods of operations research, ed. P.S. Stähly (Athenäum Verlag, 1983) p. 45, 47.

[13] B. Betró, Bayesian testing of nonparametric hypotheses and its application to global optimization problems, J.O.T.A. 42(1984)31.

[14] B. Betró and R. Rotondi, A Bayesian algorithm for global optimization, Oper. Res. 1(1984)111.

[15] C.G.E. Boender, A.H.G. Rinnooy Kan, L. Stougie and G.T. Timmer, A stochastic method for global optimization, Math. Progr. 22(1982)125.

[16] C.G.E. Boender and A.H.G. Rinnooy Kan, Optimal stopping rules for random sampling global optimization procedures, contributed talk at I.I.S.O. (1982).

[17] F.H. Branin, jr. and S.K. Hoo, A method for finding multiple extrema of a function of N variables, in: Numerical Methods of Nonlinear Optimization (Academic Press, 1982).

[18] F.H. Branin, jr., Widely convergent method for finding multiple solutions of simultaneous nonlinear equations, IBM J. Res. Develop. (September, 1972) p. 504.

[19] P. Chiappa, G. Remotti and R. Rotondi, A clustering technique based on *k*th nearest neighbour distribution (1983), private communication.

[20] D. Clough, An asymptotic extreme value sampling theory for estimation of global maximum, Can. Oper. Res. Soc. J. (1969)102.

[21] L. De Haan, Estimation of the minimum of a function using order statistics, Report 7902/S (Econometric Institute, Erasmus University, Rotterdam, 1979).

[22] Yu. M. Ermoliev, On the stochastic quasi-gradient method and stochastic quasi-Feyer sequences, Kibernetika 3(1969)18.

[23] Yu. M. Ermoliev, Stochastic quasi-gradient methods and their application in systems optimization, Working Paper WP – 81 – 2, I.I.A.S.A. (1981).

[24] B. Everitt, Cluster Analysis (Heinemann, 1974).

[25] Y.G. Evtushenko, Numerical methods for finding global extrema (Case of a non-uniform mesh), Zh. Vychisl. Mat. Fiz. 11,6(1971)1390.

[26] J. Gomulka, Remarks on Branin's method for solving nonlinear equations, in: Towards Global Optimization, ed. L.C.W. Dixon and G.P. Szegö (North-Holland, Amsterdam, 1975) p. 96.

[27] J. Gomulka, Two implementations of Branin's method: numerical experience, in: Towards Global Optimization 2, ed. L.C.W. Dixon and G.P. Szegö (North-Holland, Amsterdam, 1978) p. 151.

[28] A.O. Griewank, Generalized descent for global optimization, J.O.T.A. 34, n.1 (1981)11.

[29] J.W. Hardy, An implemented extension of Branin's method, in: Towards Global Optimization, ed. L.C.W. Dixon and G.P. Szegö (North-Holland, Amsterdam, 1975) p. 117.

[30] J. Hartigan, Clustering Algorithms (Wiley, 1975).

[31] M.J. Kushner, A new method for locating the maximum point of an arbitrary multipeak curve in presence of noise, J. Basic Engineering (1964) 97.

[32] J.J. McKeown, Aspects of parallel computation in numerical optimization, in: Numerical Techniques for Stochastic Systems, ed. F. Archetti and M. Cugiani (North-Holland, Amsterdam, 1980) p.297.

[33] J. Mockus, On a method for allocation of observations for the solution of extremal problems, USSR Comp. Mat. and Mat. Fiz. 2(1964)103.

[34] J. Mockus, V. Tiesis and A. Žilinskas, The application of Bayesian method for seeking the extremum, in: Towards Global Optimization 2, ed. L.C.W. Dixon and G.P. Szegö (North-Holland, Amsterdam, 1978) p. 117.

[35] J. Mockus, The simple Bayesian algorithm for multidimensional Bayesian optimization, in: Numerical Techniques for Stochastic Systems, ed. F. Archetti and M. Cugiani (North-Holland, Amsterdam, 1980) p. 369.

[36] J. Mockus, The Bayesian approach to global optimization, in: Proc. 10th IFIP Conf. on System Modeling and Optimization (Springer, 1981) p. 473.

[37] K.D. Patel, Parallel computations and numerical optimization. Oper. Res. 1(1984)135.

[38] J. Pinter, Sztochastikus modszerek optimalizalasi feladatok megoldasara, Alkalmazott Matematikai Lapok 7(1981)217.

[39] L.A. Rastrigin, The convergence of the random search method in the extremal control of a many parameter system, Automat. Remote Control 24(1963)216.

[40] R. Rotondi, Valutazione numerica di un algoritmo probabilistico di ottimizzazione globale, Rendiconti dell'Istituto Lombardo di Scienze e Lettere (1983) to appear.

[41] R.Y. Rubinstein, Simulation and the Monte Carlo method (Wiley, 1981).

[42] R.Y. Rubinstein and G. Samorodnitsky, Efficiency of the random search method, Math. and Comp. in Sim. 24(1982)257.

[43] F. Schoen, On a sequential search strategy in global optimization problems, Calcolo III (1982)321.

[44] M.A. Schumer and K. Steiglitz, Adaptive step size random search, IEEE Transactions AC, Vol. AC – 13(1968)351.

[45] B.O. Shubert, A sequential method seeking the global maximum of a function, SIAM J. Numer. Anal. 9:3(1972)379.

[46] F.J. Solis and R.B. Wets, Minimization by random search techniques, Math. Oper. Res. no. 1 (1981)19.

[47] A.G. Sukharev, Optimal strategies for the search of an extremum, Zh. Vychisl. Mat. Fiz. 11, no. 4 (1971)910.

[48] A.G. Sukharev, Best sequential strategies for finding an extremum, Zh. Vychisl. Mat. Fiz. 18, no. 1 (1972)35.

[49] A. Torn, Cluster analysis using seed points and density-determined hyperspheres with an application to global optimization, in: Proc. 3rd Int. Joint Conf. on Pattern Recognition (1976) p. 394.

[50] A. Torn, Probabilistic global optimization, a cluster analysis approach, in: Proc. 2nd European Congress on Operations Research (North-Holland, Amsterdam, 1976) p. 521.

[51] G. Treccani, A new strategy for global minimization, in: Towards Global Optimization, ed. L.C.W. Dixon and G.P. Szegö (North-Holland, Amsterdam, 1975) p. 143.

[52] G. Treccani, On the convergence of Branin's method: a counter example, in: Towards Global Optimization, ed. L.C.W. Dixon and G.P. Szegö (North-Holland, Amsterdam, 1975) p. 107.

[53] G. Treccani, A global descent optimization strategy, in: Towards Global Optimization 2, ed. L.C.W. Dixon and G.P. Szegö (North-Holland, Amsterdam, 1978) p. 165.

[54] A. Velasco Levy and S. Gomez, The tunneling algorithm for the global optimization of constrained functions, IIMAS – UNAM Tech. Rep. no. 231 (1980).

[55] A. Velsco Levy and A. Montalvo, A modification to the tunneling algorithm for finding the global minima of an arbitrary one-dimensional scalar function, IIMAS – UNAM Tech. Rep. no. 240 (1980).

[56] R. Zielinski, A statistical estimate of the structure of multi-extremal problems, Math. Progr. 21(1981)348.

[57] A. Zilinskas, Axiomatic approach to statistical models and their use in multimodal optimization theory, Math. Progr. 22(1982)

[58] F. Zirilli, The use of stochastic differential equations in global optimization (1982), private communication.