A Review of Statistical Models for Global Optimization

A. ŽILINSKAS
Dept. Optimal Decision Theory, Institute of Mathematics and Informatics, 232600, Vilnius, K. Poželos 54, Lithuania

(Received: 14 August 1990; accepted: 24 May 1991)

Abstract. A review of statistical models for global optimization is presented. Rationality of the search for a global minimum is formulated axiomatically and the features of the corresponding algorithm are derived from the axioms. Furthermore the results of some applications of the proposed algorithm are presented and the perspectives of the approach are discussed.

Key words. Global optimization, stochastjc models, optimal design, rational choice.

1. Introduction

Stochastic functions are used as models of complicated functions with elements of uncertainty in hydrodynamics, theory of automatic control, radar theory, etc. Some algorithms of global optimization are also based on stochastic functions (see, e.g., [3, 6, 7, 4, 10, 11, 8, 1]). However, the use of such models in global optimization needed some theoretical justification. The proof of the stability of frequencies, as it is supposed in classical statistics, seems to be unrealistic for the characteristics of the class of real objective functions. Therefore, a justification of such a stochastic approach required the development of a general theory of statistical models for global optimization. Stochastic processes with Markovian property have been proven to be constructive models for the development of one-dimensional algorithms in the papers cited above. However, the use of stochastic functions as models for the multi-dimensional case is restricted due to numerical problems in inverting the correlation matrices, whose dimension is equal to the number of trial points. That is why a generalization of classical stochastic models was necessary to simplify the computation of their characteristics. In this paper a review concentrating on the main problems of the axiomatic development of such a theory is presented.

A model helps to interpret the results of the previous optimization steps and to plan the next ones. However, the definition of a rational algorithm remains not trivial. The algorithms, which are optimal with respect to the obviously rational criteria, are too complicated for the computer realization. If an optimal algorithm is simplified or approximated by a computer algorithm, the approximation errors remain unclear. An example is the substitution of the optimal algorithm by the one-step algorithm. The latter, although obviously simpler than the original one,

Journal of Global Optimization **2**: 145–153, 1992.

is not well justified: e.g., a one-step Bayesian algorithm may have a too local behaviour in the initial phase of the optimization [9]. Asymptotic features of a global algorithm, e.g., the asymptotic rate of convergence, are not fully adequate to the real efficiency of the algorithm, since the global search normally is stopped far before the asymptotic features could effect. The final refinement of the global and main local minima is performed by some well known local techniques defining therefore the rate of convergence. Because of the difficulties mentioned above, it is reasonable to construct the algorithm axiomatically, formalizing simple and intuitively obvious requirements to the algorithm at the current minimization step.

The use of classical stochastic models, e.g., stochastic functions, is constructive only in the one-dimensional case. To apply the approach to multidimensional problems the one-dimensional algorithms are combined with dimension reduction techniques [8]. In the Bayesian approach the optimal algorithms are defined in a classical way, but the computer realizations are based on considerable simplifications (one-step optimality, turning the complicated formulas for the calculation of the characteristics of the stochastic function into simple ones) [5].

The axiomatic theory of rational choice allows to construct statistical models and optimization algorithms within the framework of a unified approach of 'average rationality'. Some crucial computational difficulties are removed here, but not all of them, of course. Due to the fact, that the methodology of this approach is rather different from the customarily accepted one, it may be interesting for experts in global optimization to grasp it in the condensed form of a review.

2. The Statistical Model

Let the unique objective information on the function $f(x)$, $x \in A \subset R^n$, be the values of $f(\cdot)$ at the points $x_i \in A: y_i = f(x_i)$, $i = 1, \ldots, k$. In addition we have some subjective information (e.g., from the experience of solving similar problems in the past) concerning multimodality and complexity of $f(x)$. The weakest, but still reasonable assumption on available information is the comparability of likelihood of the intervals of the possible values $f(x)$, $x \neq x_i$, $i = 1, \ldots, k$, often called *comparative probability* (CP) (see [2]). Let the binary relation CP be given and denoted by $\succeq_x$, where $(a, a') \succeq_x (b, b')$ means, that the event $f(x) \in (a, a')$is at least as likely as the event $f(x) \in (b, b')$. The index x may be omitted if it is apparent from the context. The impossible event $\varnothing$ is introduced formally and considered in a way similar to the other events. The event $((a, a') \succeq_x (b, b')) \wedge ((b, b') \succeq_x (a, a'))$ is denoted as $(a, a') \sim_x (b, b')$. The shorter expression $(a, a') \succ_x (b, b')$ is used for $((a, a') \succeq_x (b, b')) \wedge \neg((a, a') \sim_x (b, b'))$. Let the point $x \neq x_i$, $i = 1, \ldots, k$ be fixed. The information on $f(\cdot)$ normally does not contradict the following assumptions on the rationality of CP:

A1. For arbitrary intervals (a, a'), (b, b') there holds either $(a, a') \succeq (b, b')$ or $(b, b') \succeq (a, a')$.

A2. If $(a, a') \succeq (b, b')$ and $(b, b') \succeq (c, c')$ then $(a, a') \succeq (c, c')$.

A3. The statement $(a, a') > 0$ is true if and only if $\mu[a, a'] > 0$, where $\mu(\cdot)$ denotes a Lebesgue measure; $(a, a') \sim [a, a') \sim (a, a'] \sim [a, a']$.

A4. Let the following relations hold: $B = [a, a'] \cap [b, b'] \neq \emptyset$, $C = [a, a'] \cap [c, c'] \neq \emptyset$, $\mu(B \cup C) = 0$. The relation $[b, b'] \succeq [c, c']$ is true if and only if $[a, a'] \cup [b, b'] \succeq [a, a'] \cup [c, c']$.

A5. If $(a, a') > (b, b') > \emptyset$ holds, then a_1, a_2, $a < a_i < a'$, $i = 1, 2$ exist such that $(a, a_1) \sim (a_2, a') \sim (b, b')$.

Since in the axiom A1 only simple sets (intervals) are involved in the comparison, A1 is weaker than assumed usually. The transitivity axiom A2 is discussed by many authors (see [2]) and it is one of the fundamental assumptions regarding the rationality of CP. The intuitive acceptability of the axioms A1 and A2 for solving complicated optimization problems is shown by the results of psychological experiments (summarized in [9]). The axiom A4 expresses the additivity of CP and is a normal rationality assumption for CP. The axioms A3 and A5 are specific for this approach. The axiom A3 expresses the complexity of the function and states that the exact prediction of $f(x)$ is impossible, as well as the choice of an interval (a, a') such that $\mu(a, a') > 0$ and the event $f(x) \in (a, a')$ is equivalent to $\emptyset$. The continuity of CP with respect to intervals seems quite natural, the axiom A5 expresses this continuity in the most obvious way. The CP, defined by A1–A5 for intervals, may be extended to the algebra of finite unions of intervals in a rather natural way, implying the existence of a unique probability density $p_x(\cdot)$ compatible with CP. Let X_i, $i = 1, 2$ denote the finite unions of disjointed intervals. The density $p(\cdot)$ is called *compatible* with $\succeq$ if $X_1 \succeq X_2 \Leftrightarrow \int_{X_1} p(t)\,dt \geqslant \int_{X_2} p(t)\,dt$. This result implies the interpretation of an unknown value $f(x)$ as a random variable Y_x with probability density $p_x(\cdot)$ and finally the acceptability of a family Y_x, $x \in A$, $x \neq x_i$, $i = 1, \ldots k$ for the statistical model of $f(x)$. The axioms discussed here imply the existence and uniqueness of $p_x(\cdot)$, however, the constructive form of $p_x(\cdot)$ (i.e., of the probability density and its dependency on x) is necessary to construct the optimization algorithms. The results of a psychological experiment show, that the CP for researchers and designers solving technical optimization problems in their daily work may be approximated with acceptable accuracy by means of a Gaussian probability density [9].

A stochastic function may be considered as a family of random variables, therefore the stochastic functions are a specific case of the models defined above. The generalization of axioms on CP for the case of multidimensional intervals (of the values of $f(\cdot)$ at finite sets of points in A) formalise similar assumptions and imply the existence and uniqueness of a stochastic function compatible with CP. However, the reformulation of the axioms is more complicated and not so obvious intuitively.

The main practical conclusion from the axiomatic theory is the possibility to construct well defined statistical models of multimodal functions which are simpler from the computational point of view than the stochastic Gaussian functions.

3. Characteristics of the Statistical Model

The assumptions regarding the information about $f(\cdot)$ are sufficiently natural and imply that the family of Gaussian random variables Y_x, $x \in A$, $x \neq x_i$, $i = 1, \ldots, k$ forms an acceptable model of $f(x)$. For a further characterization of this statistical model it is necessary to define the expected value of $f(x)$, which is denoted by $m_k(x, (x_i, y_i), i = 1, \ldots, k)$. Informally $m_k(x, \cdot)$ may be termed as the average value or the most likely value or the representative value of the function at the point x. If Y_x corresponds to a random function, then the conditional mean of it corresponds to this wording. As shown in [12] such a definition of $m_k(\cdot)$ is of interest also when extrapolating under uncertainty independently of the underlying statistical model. The rationality of the extrapolation can be understood as the invariance of the expected value of $f(x)$ with respect to some transformations of the available information:

- invariance with respect to the scale of measuring y_i,
- invariance with respect to the choice of the zero point of measuring y_i,
- invariance with respect to the numeration of (x_i, y_i), and
- a restriction of the complexity of an extrapolation is formulated as the admissibility of data aggregation.

The strict formulation of the axioms may be found in [12]. The unique extrapolator compatible with the axioms is

$$m_k(x, (x_i, y_i), i = 1, \ldots, k) = \sum_{i=1}^{k} y_i w_i(x, x_j, j = 1, \ldots, k), \tag{1}$$

where the weights $w_i(x, \cdot)$ have some natural properties.

The second characteristic of the model $s_k(x, (x_i, y_i), i = 1, \ldots, k)$, the variance of Y_x, may be characterized by similar axioms, implying the following expression

$$s_k(x, (x_i, y_i), i = 1, \ldots, k) = \gamma_k \sum_{i=1}^{k} \|x - x_i\| w_i(x, x_j, j = 1, \ldots, k),$$

where γ_k may depend on (x_i, y_i), $i = 1, \ldots, k$.

The investigation of expression (1) using the weights given below has shown that such an extrapolator is rather precise and that it can be implemented efficiently. The weights are:

$$w_i^k(x, x_j, j = 1, \ldots, k) = 0, \quad i \notin I(x),$$

$$w_i^k(x, x_j, j = 1, \ldots, k) = \mathrm{d}(x, x_i) / \sum_{j \in I(x)} \mathrm{d}(x, x_j), \quad i \in I(x),$$

where $I(x)$ is the set of indices of the r nearest neighbours of x,

$$d(x, x_i) = \exp(-c\|x - x_i\|^2)/\|x - x_i\|, \quad c > 0,$$

$\|\cdot\|$ is the Euclidean norm in R^n, $r = 5$ and the value $c = 3.3$ is appropriate if R^n is scaled by normalizing the components of x by the mean-square-root deviations of the corresponding components of the vectors x_i, $i = 1, \ldots, k$.

The expression of the conditional mean of a Gaussian random field is a special case of (1) where the weights are defined by the inversion of the correlation matrix. It is interesting to specify this case axiomatically. Two specific axioms proposed in [12] imply the expression (1) coinciding with the expression of the conditional mean of a Gaussian random field. The latter results show the relations between the proposed statistical models and classical ones and express the features, which imply the difficulties of numerical realization of the extrapolation.

4. The Optimization Algorithm

Assume that the function $f(x)$, $x \in A \subset R^n$, is to be minimized. Let k evaluations of $f(\cdot)$ be given by $y_i = f(x_i)$, $i = 1, \ldots, k$. The preceding discussion implies that the family of Gaussian random variables Y_x, $x \in A$ with the probability density $p_x(\cdot)$ depending on x_i, y_i, $i = 1, \ldots, k$ is an acceptable statistical model of $f(\cdot)$. The choice of the next point $x_{k+1} \in A$ for evaluation of $f(\cdot)$ may be interpreted as a choice of a particular probability density $p_{x_{k+1}}(\cdot)$. If the preference of choosing between the two densities p_{x_1} and p_{x_2} satisfies some rationality requirements, it may be possible to construct a utility function $u(\cdot)$ compatible with the preference of choice between them, i.e.,

$$p_{x_1} \succeq p_{x_2} \Leftrightarrow \int_{-\infty}^{+\infty} u(t)p_{x_1}(t)\,dt \geqslant \int_{-\infty}^{+\infty} u(t)p_{x_2}(t)\,dt$$

Since the probability densities are Gaussian, i.e.,

$$p_x(t) = n(t|m_k(x, .), s_k(x, .))$$

these preferences are equivalent to preferences between the vectors (m, s) where m denotes the mean value and s^2 the variance of Y_x. The construction of a utility function $u(\cdot)$ obviously implies the construction of a utility function $U(m, s)$ for vectors (m, s), i.e.,

$$U(m, s) = \int_{-\infty}^{+\infty} u(t)n(t|m, s)\,dt$$

The axiomatic definition of the preference relation and the corresponding interpretation are given in [9]. Here only the ideas of the axioms are presented:

- a current observation may be rational at the point x with a large mean value m only in the case of a sufficiently large uncertainty measure s,
- it is not rational to choose a point for current observation with value of $f(\cdot)$ which is with guarantee larger than the best value found during the previous iterations,
- the preference relation is continuous with respect to m,
- the utility function is continuous from the left.

The unique utility function compatible with these assumptions is $u(t) = I(z_{0k} - t)$ where $z_{0k} < \min_{(1 \leq i \leq k)} y_i$ and $I(\cdot)$ is a unit-step function. Therefore, the current observation of the minimization algorithm corresponding to all the assumptions is defined by the relation $x_{k+1} = \operatorname{argmax}_{x \in A} P(Y_x < z_{0k})$. In the one-dimensional case the maximum point of the probability $P(\cdot)$ may be expressed by a simple formula. In the multidimensional case the problem is not so easy and usually is attached by a combination of Monte-Carlo and local techniques. A choice of the statistical model and of some parameters of the original algorithm are rather arbitrary. Therefore, a high accuracy in solving the auxiliary maximization problem is not reasonable. A global optimization algorithm is used to obtain the points in a region of attraction of the global minimum, and the refinement of the solution is performed by the local algorithm. Therefore a variation of the coordinates of a global trial point is negligible.

The efficiency of the algorithm crucially depends on the transition from the global search to the local one. In the algorithm considered here the transition is performed if the local inadequacy of the statistical model and the obtained data is detected. In the one-dimensional algorithm the transition condition is rested as a statistical hypothesis. In the multi-dimensional case it is based on heuristic and empiric rules.

The convergence of the axiomatically defined algorithm is considered for very weak assumptions: only the continuity of the objective function is supposed. Therefore, the convergence may be guaranteed if the trial points are dense everywhere in A. It is not always easy to prove this fact for sophisticated algorithms because they place the trial point in the 'promising' subregions of A more often than in 'non-promising' ones aiming at efficient search. However, it seems reasonable to perform observations (although seldom) in the 'not-promising' subregions to be sure not to miss a sharp deep hole (global minimum for the 'worst case' objective function).

5. Applications to Optimal Design

The test results may be summarized as follows: the constructed algorithms are very efficient in respect with the number of objective function evaluations necessary to find the global minimum. It is interesting to note that even for one-dimensional functions with analytical estimates of the Lipschitz constant (or

the bound on the second derivative) such an algorithm is more efficient than that based on the Lipschitzian model. However, the computer realization of these algorithms in the multidimensional case is impossible without time consuming auxiliary computations. Therefore, the field of rational applications of the algorithms is the optimization of expensive (time consuming) multimodal functions whose dimensional does not exceed 10 [9].

Such problems are quite often in optimal design. An example is the optimal design of magnetic deflection system (MDS) for a colour TV. An important criterion of MDS quality is given by the aberration of the electron beam, i.e., the dispersion of electrons while deflecting them by the MDS. The aberration depends on the configuration of the magnetic field. The latter may be defined by a choice of the currents in the sections of MDS. Therefore, the minimization of the aberration with respect to the currents in the sections of the MDS is one of the important parts in the optimal MDS design. The algorithm for the calculation of the objective function $f(\cdot)$ (aberration) includes a numerical integration of the system of differential equations describing the motion of an electron in the magnetic field of the MDS. The computing time of one value of $f(\cdot)$ in the real problems often exceeds 20 sec on a BESM-6 computer. An analytical investigation of the features of $f(\cdot)$ including the regions of attraction of local minima is impossible, because only the computer algorithm for computing the values of $f(\cdot)$ is available. The application of gradient type methods to solve the problem is difficult. First, the time needed for evaluating only one gradient vector is very large, in general n-times 20 sec in case of n decision variables. Second, the errors of the computation of the values of $f(\cdot)$ may be too large for acceptable estimate of the gradient by means of numerical differentiation. The experiment showed that variable metric techniques, which are very efficient in case of analytically given test functions, cannot reach an acceptable solution in reasonable time (1–2 hours).

The application of well known Nelder–Meed algorithm which is simpler and more robust than the gradient type methods, also did not give an acceptable result either. Therefore, to solve the problem, global optimization algorithms should be used. The comparative analysis showed, that an algorithm based on the axiomatic approach is rather efficient [9].

The second example of an efficient application of the algorithm constructed here is the optimal synthesis of pigmental compositions (colours). The set of pigments (whose spectral characteristics are known) should be used to produce a colour similar to a given standard colour. There are several criteria of similarity, e.g. spectral distance, colour distance, etc. Investigations of real problems with 9 pigments showed that the solutions obtained by a local algorithm essentially depend on the chosen initial point. The process of local descent takes a considerable computing time. The application of our algorithm yields acceptable solutions of different versions of the problem within 5–6 minutes [9].

Several versions of the algorithm, based on statistical models, are coded in

FORTRAN. Some of them are included in libraries developed at the institute of Mathematics and Informatics, Lithuanian Academy of Sciences.

6. Perspectives

The axiomatic approach to the construction of statistical models and optimization algorithms originated as an attempt to achieve rationality of the global search *in average*. It has grown from the Bayesian approach presented in [4], but it is different from the latter in the methodology of construction. In the axiomatic approach simple (from the computational point of view) expressions of $m_k(\cdot)$, $s_k(\cdot)$ are defined as the characteristics of an extrapolator under uncertainty. The minimization algorithm is defined axiomatically for the statistical model. In the Bayesian approach the optimal algorithm is defined for a stochastic function. Further simplifications, necessary for numerical realization, are described in [5]. The algorithms based on both approaches are similar concerning their efficiency as well as their complexity of realization. Both are oriented towards minimization of expensive multimodal functions.

One of the main problems in the axiomatic approach is the reduction of auxiliary computations necessary to implement the algorithm. The amount of auxiliary computations is growing very fast while the information stored is growing, i.e., the number of iteration is increasing. It may be attractive to use qualitatively different information than the values of $f(\cdot)$, e.g., the gradients of $f(\cdot)$ which are very important in local optimization. Known statistical models even implicitly do not contain any information on gradients. Therefore it is supposed to extend the known system of axioms for $m_k(\cdot)$, $s_k(\cdot)$ by postulating the features of differentiability in the framework of the statistical model.

The other direction of development concerns the construction of statistical models and optimization algorithms in the presence of noise. The experience of local optimization with noise shows that this case is much more complicated than the optimization without the noise. One may even consider noisy multimodal problems as unsolvable at all. However, the approach based on statistical models opens some new possibilities. Since the one-dimensional algorithm in the presence of noise has been proved to be quite efficient, one may expect similar efficiency for the multidimensional algorithm as well [9].

Experts in applied mathematics recently started to extend various methods for parallel computers. Some general problems of parallel computing in global optimization are discussed in [9]. The parallelization of general algorithms based on statistical models is difficult. However, some simple but efficient algorithms may be used in parallel schemata, e.g., the one-dimensional algorithm may be applied for the multidimensional case by using random search directions, where different one-dimensional searches are performed on different processors exchanging some information. The parallelization principle used here is task parallelism. The code was developed in parallel FORTRAN 77. When a processor

becomes idle, a task specified by two points determining the line in A and the best value known is sent by the master to the idle processor (worker) who then sends a new best point back to the master. It is easy to implement such parallel algorithms on a processor farm. The expected speedup and efficiency were obtained, however, only after a long investigation. Partly this is due to the novelty of the subject, and to the fact that programming parallel algorithms is more difficult than programming sequential ones. A fundamental difficulty is given by the limited parallelism achievable in the algorithmic parallelization of most sequential algorithms. Due to the fact that global optimization is suitable for Monte Carlo and geometric parallelization, which do not have the drawbacks of algorithmic parallelization, methods based on these concepts and on task parallelism seem worth future investigation. Seemingly, the way to combine mathematical and heuristic ideas is most promising in this field.

References

1. Archetti, F. and Betro, B. (1979), A Probabilistic Algorithm for Global Optimization, *Calcolo* **16**, 335–343.
2. Fine, T. (1973), *Theories of Probability*, AP, NY.
3. Kushner, H. (1964), A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise, *Trans. ASME, ser. D* **86**(1), 97–105.
4. Mockus, J. (1972), On Bayesian Methods of Search for Extremum, *Automatics and Computers* **3**, 53–62.
5. Mockus, J. (1989), *Bayesian Approach to Global Optimization*, Kluwer, Dordrecht.
6. Neimark, J. and Strongin, R. (1969), Informational Approach to the Problem of Search for Extremum of a Function, *Engineering Cybernetics* **4**, 17–26.
7. Strongin, R. (1978), *Numerical Methods of Multimodal Optimization*, Nauka, Moskow.
8. Šaltenis, V. (1971), On a Method for Multimodal Optimization, *Automatics and Computers* **3**, 33–38.
9. Törn, A. and Žilinskas, A. (1989), *Global Optimization*, Springer, Berlin.
10. Žilinskas, A. (1976), A Method of One-Dimensional Multiextremal Minimization, *Engineering Cybernetics*, 71–74.
11. Žilinskas, A. (1978), Optimization of One-Dimensional Multimodal Functions, Algorithm AS 133, *Applied Statistics* **23**, 367–375.
12. Žilinskas, A. (1979), Axiomatic Approach to Extrapolation Problem under Uncertainty, *Automatics and Remote Control* **12**, 66–70.