INFORMATICA, 1990, Vol.1, No.1, 141–155

STATISTICAL MODELS OF MULTIMODAL FUNCTIONS AND CONSTRUCTION OF ALGORITHMS FOR GLOBAL OPTIMIZATION

Antanas ŽILINSKAS

Institute of Mathematics and Cybernetics,
Lithuanian Academy of Sciences,
232600 Vilnius, Akademijos St.4 Lithuania

Abstract. The problems and results in constructing the statistical models of multimodal functions are reviewed. The rationality of the search for global minimum is formulated axiomatically and the features of the corresponding algorithm are discussed. The results of some applications of the proposed algorithm are presented.

Key words: global optimization, stochastic models, optimal design, rational choice.

Introduction. To justify and to construct the optimization algorithm, a model of the objective function is necessary. In the local optimization theory the quadratic models have been proved to be very useful. In the multimodal case a model must be adequate to the considerably more uncertain behaviour of a function than in the local case (see e.g. Tőrn, Žilinskas, 1989). The stochastic functions are used for the models of complicated functions with the elements of uncertainty in hydrodynamics, theory of automatic control, radar theory, etc. Some algorithms of global optimization are based also on the stochastic functions in the papers by Kushner (1964), Šaltenis (1971), Mockus (1972), Neimark and Strongin (1966), Strongin (1978). However, the justification of the use of such models in global optimization was only heuristic. The proof of the stability of frequencies, as it is supposed in classical statistics, seems not realistic for the characteristics of the class of real objective functions. Therefore, a justification of statistical models in global optimization needed the development of a general theory of statistical models of multimodal objective functions. In this paper the review is presented on the main problems of axiomatic development of such a theory.

The assumptions on information on the objective functions are formulated axiomatically. The basic assumption is the possibility to compare the likelihood of the intervals of values of the objective function. It is shown, that the family of random variables is a model, corresponding to the system of rather simple axioms, formalizing available information on the class of objective functions. A further characterization of statistical model is similar to the problem of extrapolation under uncertainty. It is shown, that in the class of axiomatic models there exist such ones which are simpler from the computational point of view than Gaussian stochastic functions. Some additional axioms to specify the latter case in the class of axiomatic models are discussed.

The model helps to interpret the results of the previous optimization steps and to plan the current 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, e.g., in case of substituting the optimal algorithm by the one-step optimal algorithm. The latter, although obviously simpler than the original one, is not well justified: it seems reasonable to perform the general investigation on the features of the function at the initial optimization steps collecting information useful to organize an efficient search at the subsequent steps. Asymptotic features of the global algorithms, e.g., asymptotic rate of convergence, are not fully adequate to the real efficiency of the algorithms, since the final refinement of the global and main local minima normally is performed by the well known local techniques. Because of the difficulties mentioned above, it is reasonable to construct the algorithm axiomatically, formalizing simple and intuitively obvious requirements to the algorithm at current minimization step. A slightly different approach to the use of statistical models is considered in a book by Mockus (1989).

The results of the minimization of test and practical problems have shown, that the field of rational application of the constructed algorithms is the minimization of expensive functions, i.e., whose computation is time consuming and whose dimensionality does not exceed 10.

Construction of 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 $x_i \in A:\ y_i = f(x_i),\ i = \overline{1,k}$. Besides, we have the subjective information (e.g., the experience of solving similar problems in the past) on 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 $f(x)$, $x \neq x_i$, $i = \overline{1,k}$. Let the comparability relation (CR) be given and denoted by $\succsim_x$, where $(a, a') \succsim_x (b, b')$ does mean that the event $f(x) \in (a, a')$ is at least as like as the event $f(x) \in (b, b')$. The index x may be omitted if it is apparent from the context. The impossible event O is introduced formally and considered in a similar way with the other events. The event $(a, a') \succsim_x (b, b')$ and $(b, b') \succsim_x (a, a')$ is denoted as $(a,a') \sim_x (b,b')$. The expression $(a,a') \succ_x (b,b')$ used for shortening of $(a,a') \succsim_x (b,b')$ but $(a,a') \sim_x (b,b')$is not true. Let the point $x \neq x_i$, $i = \overline{1,k}$ be fixed. The information on $f(\cdot)$ normally does not contradict the following assumptions on rationality of CR:

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

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

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

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

A5. If there hold $(a,a') \succ (b,b') \succ O$, then exist a_1, a_2, $a < a_i < a'$, $i = 1,2$ 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 as it is customary assumed. The transitivity axiom A2 is discussed by many authors and it is one of the fundamental assumptions regarding the CR rationality. The intuitive acceptability of the axioms A1 and A2 in solving complicated optimization problems is shown by the results of psychological experiment by Žilinskas (1986). The axiom A4 expresses the additivity of CR and is a normal rationality assumption for CR. 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 $f(x) \notin (a,a')$ is equivalent to O. The continuity of CR with respect to intervals seems quite natural, the axiom A5 expresses this continuity in the most obvious way. The CR, 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 CR. The density $p(\cdot)$ is called compatible with $\succsim$ in case $X_1 \succsim X_2$ holds if and only if $\int_{X_1} p(t)dt \geqslant \int_{X_2} p(t)dt$, where $X_i, i = 1,2$ denote the finite unions of disjoined intervals. This result implies the interpretation of unknown value $f(x)$ as a random variable Y_x with probability density $p_x(\cdot)$ and finally the acceptability of a family $Y_x, \quad x \in A, \quad x \neq x_i, \quad i = \overline{1,k}$ for the statistical model of $f(x)$. The discussed axioms imply the existence and uniqueness of $p_x(\cdot)$, however, the constructive form of $p_x(\cdot)$ (i.e., of probability density and it's dependence on x) is necessary to construct the optimization algorithms. The results of a psychological experiment show, that the CR for researchers and designers solving technical optimization problems in their daily work may be expressed by means of Gaussian probability density (see Žilinskas 1986).

A stochastic function may be considered as a family of random variables, therefore, the stochastic functions are specific case of the models defined above. The axiomatic definition of this case (very important for the theory) was considered by Žilinskas and Katkauskaitė (1982). The additional axioms on CR of multidimensional intervals of the values of $f(\cdot)$ at several points have similar sense as the A1-A5 and imply the existence and uniqueness of a stochastic function compatible with CR. However, the formulation 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.

Definition of the characteristics of a statistical model.

The natural enough assumptions, regarding the information about $f(\cdot)$, imply that the family of Gaussian random variables Y_x, $x \in A$, $x \neq x_i$, $i = \overline{1,k}$ is 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 $m_k(x, (x_i, y_i), i = \overline{1,k})$. 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. Let us note that such a definition of $m_k(\cdot)$ is of interest also, when extrapolating under uncertainty independently of the underlying statistical model as shown by Žilinskas (1979). 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:

a) invariance in the scale of measuring of y_i, b) invariance in the choice of zero point of measuring of y_i, c) invariance in the numeration of (x_i, y_i), d) a restriction of complexity of an extrapolation is formulated as the admissibility of data aggregation.

The strict formulation of the axioms may be found in the paper of Žilinskas (1979). The unique extrapolator compatible with the axioms is

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

where the weights have some natural properties.

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

$$s_k(x, (x_i, y_i), i = \overline{1,k}) = \gamma_k \sum_{i=1}^{k} \|x - x_i\| w_i(x, x_j, j = \overline{1,k}), \quad (2)$$

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

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

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

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

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

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

$\|\cdot\|$ 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, $i = \overline{1,k}$.

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 correlation matrix. It is interesting to specify this case axiomatically. Two specific axioms proposed in the paper of Žilinskas (1979) imply the expression (1) coinciding with the expression of conditional mean of 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.

Construction of 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 $y_i = f(x_i)$, $i = \overline{1,k}$. The proceeding 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, $i = \overline{1,k}$ is an acceptable statistical model of

$f(\cdot)$. The choice of the next point $x_{k+1} \in A$ where to evaluate $f(\cdot)$ may be interpreted as a choice of a particular probability density $p_{x_{k+1}}(\cdot)$. If preference when choosing between the two densities p_{x_1} and p_{x_2} satisfy some rationality requirements, it may be possible to construct a utility function compatible with the preference of choice, i.e.,

$$p_{x_1} \geqslant p_{x_2} \text{ iff } \int_{-\infty}^{+\infty} u(t)p_{x_1}(t)dt > \int_{-\infty}^{+\infty} u(t)p_{x_2}(t)dt.$$

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

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

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.,

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

The axiomatic definition of the preference relation and the corresponding interpretation are given by Žilinskas (1985). Here only the ideas of the axioms are presented: a) a current observation may be rational at the point x with a large mean value m only in the case of sufficiently large uncertainty measure s, b) it is not rational to choose the point for current observation with guarantee the $f(\cdot)$ value be larger than the best value found at the previous iterations, c) the preference relation is continuous in respect with m,d) the utility function is continuous from the left.The unique utility function compatible with the assumptions is $u(t) = I(z_{0k} - t)$, where $z_{0k} < \min_{1\leqslant i\leqslant k} y_i$; where $I(\cdot)$ is a unit-step function. Therefore, the current observation of minimization algorithm corresponding to all the assumptions is defined by the relation $x_{k+1} = arg \max_{x\in A} P(Y_x < z_{0k})$. In one-dimensional case the maximum point of probability $p(\cdot)$ may be expressed by a simple formula. In multidimensional case the problem is not so easy and usually it is attacked by the combination of Monte-Carlo and local techniques. The choice of statistical model and some parameters of the original algorithm are rather arbitrary. Therefore, high accuracy in solving the auxiliary maximization problem is not reasonable. Since the 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, the variation of coordinates of global trial point is negligible.

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

The convergence of the axiomatically defined algorithms is considered by Žilinskas (1985), Žilinskas and Katkauskaitė (1987),including the noisy case. Only the continuity of the objective function is supposed. Therefore, the convergence may be guaranteed only if the trial points are dense everywhere in A. It is not always easy to prove this fact for the sophisticated algorithms,because they place the trial point in the "promising" subregions of A more often than in "not 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 lose a sharp deep hole (global minimum for the "worst case" objective function).

4.Applications to optimal design. The results of testing of the constructed algorithms are presented in a book of Žilinskas (1986), where different algorithms are compared. The results may be summarized as follows: the constructed algorithms are very efficient in respect with the number of the objective function values, necessary to find the global minimum. It is interesting to mention, that even for the one-dimensional functions with analytical estimates of Lipshitz constant (or the bound of the second derivative) such an algorithm is more efficient than that based on the Lipshtzian model. However, the computer realization of these algorithms in mutidimensional case is impossible without time consuming auxiliary computations. Therefore, the region of rational applications of the algorithms is optimization of the expensive multimodal (time consuming) functions whose dimensionality does not exceed 10.

Such problems are quite common in optimal design. An example is the optimal design of magnetic deflection system (MDS) for a coloured TV. An important criterion of MDS quality is the aberration of the electron beam, i.e., the dispersion of electrons while deflecting them by MDS. The aberration depends on configuration of a 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 sections of MDS is one of the important parts in the optimal MDS design. The algorithm of calculation of the objective function $f(\cdot)$ (aberration) includes a numerical integration of the system of differential equations, describing a motion of electron in the magnetic field of MDS. The computing time of one value of $f(\cdot)$ in the real problems often exceeds 20 sec. on BESM-6 computer. Analytical investigation of the features of $f(\cdot)$, including the regions of attraction of a 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 of evaluating only one gradient is very large, e.g., ten dimensional problem takes 200 sec. Second, the errors of numerical differentiation caused by the errors of computation of the values of $f(\cdot)$ may be too large for a gradient. The experiment shows that the techniques of variable metric type, which are very efficient for the test functions (given by analytical formulae), can not reach the acceptable solution in reasonable time (1–2 hours).

The application of simpler technique, more robust than the gradient type methods, also does not give the acceptable result. Therefore, to solve the problem, global optimization algorithms should be used. The comparative analysis in Žilinskas (1986) shows, that the algorithm based on axiomatic approach is rather efficient.

The second example of an efficient application of the constructed algorithm is the optimal synthesis of pigmental compositions (colours). The set of pigments (whose spectral characteristics are known) should be used to produce the 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 show, that the solutions obtained by the local algorithms essentially depend on the chosen initial points. The process of the local descend takes a considerable extention of computing time. The application of the constructed algorithm gives the acceptable solutions of different versions of the problem in 5–6 minutes (see Barauskas , Žilinskas, Piliavskij, Juškienė 1980).

Several versions of the algorithm, based on statistical models, are coded in FORTRAN, e.g., included in library OPTIMUM (1983). The one-dimensional algorithms (for minimization without and with noise) are published by Žilinskas (1978a, 1980).

The perspectives. The axiomatic approach to the construction of statistical models and optimization algorithms originated as an attempt to realize the rationality of global search "in average". It has grown from the Bayesian approach presented by Mockus (1989), but it is different from the latter in the methodology. From the computational point of view, simple expressions of $m_k(\cdot)$, $s_k(\cdot)$ in the axiomatic approach are defined as the characteristics of the extrapolator under uncertainty. The algorithm is defined axiomatically for statistical models.

In Bayesian approach the algorithm is justified for a stochastic function. The further simplifications, necessary for the numeric realization, are described by Mockus (1989). The algorithms based on both approaches are similar in efficiency as well as in complexity of realization. Both are oriented to the minimization of expensive multimodal functions.

One of the main problems in the axiomatic approach is the reducing of auxiliary computations necessary to realize the algorithm. To achieve the aim,it may be useful to include the gradients of the function $f(\cdot)$ in the model. Therefore,it is supposed to extend the known system of axioms for $m_k(\cdot)$, $s_k(\cdot)$,postulating the features of differentiability in the frames of a statistical model.

The other direction of the development is the construction of statistical models and algorithms for the minimization in the presence of noise. The one-dimensional case is investigated by Žilinskas (1980,1986). The initial results for the multidimensional algorithm are published by Žilinskas and Katkauskaitė (1987). Since the one-dimensional algorithm in the presence of noise has been proved to be quite efficient, one may expect the similar efficiency of the multidimensional algorithm as well.

The experts in applied mathematics recently have started to be interested in parallel computing. Some general problems of the parallel computing in global optimization are discussed by Tőrn and Žilinskas (1989). The parallelisation of general algorithms based on statistical models is difficult. However, some simple, but efficient specific algorithms may be useful in parallel schemes, e.g., the one-dimensional algorithm may be applied for a multidimensional case, using random search directions, where different one-dimensional searches are performed on different processors exchanging some information in progress. Seemingly, the way to combine mathematical and heuristic ideas is the most promising in this field.

REFERENCES

Barauskas, A., A.Žilinskas, V.Piliavskij, E.Juškienė (1980). Investigation of optimization problems in synthesis of pigmental compositions. In A.Žilinskas (Ed.), *Teorija Optimaljnych Reshenij,* Vol.6. Instit. Math. Cybern. Lithuanian Acad. Sci., Vilnius. pp. 35–39 (in Russian).

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.

Mockus, J. (1989). *Bayesian Approach to Global Optimization.* Kluwer, Dordrecht. 254pp.

Mockus, J. (1972). On Bayesian methods of search for extremum. *Automatics and Computers,* **3**, 53–62.

Neimark, J.,and R.Strongin (1966). Informational approach to the problem of search for extremum of a function. *Engineering Cybernetics,* **1**, 17–26.

OPTIMUM (1983). *The Package for Optimization OPTIMUM,* State Fund of Algorithms and Programs, No.007120. Instit. Math. Cybern. Lithuanian Acad. Sci., Vilnius (in Russian).

Strongin, R. (1978). *Numerical Methods of Multimodal Minimization.* Nauka, Moscow. 239pp. (in Russian).

Šaltenis, V. (1971). On a method for multimodal optimization. *Automatics and Computers,* **3**, 33–38.

Tőrn, A., and A.Žilinskas (1989). *Global Optimization.* Springer, Berlin. 256pp.

Žilinskas, A. (1978). On one–dimensional multimodal minimization. In *Trans. of the 8th Prague Conf. on Inform. Theory Stat. Dec. Funct. Rand. Proc.,* Vol.B. Akademia, Prague. pp. 392–402.

Žilinskas, A. (1978a). Optimization of one–dimensional multimodal functions, Algorithm AS 133. *Applied Statistics,* **23**, 367–375.

Žilinskas, A. (1980). MINIMUM–optimization of one–dimensional multimodal functions in the presence of noise, Algorithmus 44. *Aplikace Matematiky,* **25**, 234–240.

Žilinskas, A. (1985). Axiomatic characterization of a global optimization algorithm and investigation of its search strategy. *Operations Research Letters,* **4**(1), 35–39.

Žilinskas, A. (1979). Axiomatic approach to extrapolation problem under uncertainty. *Automatics and Remote Control,* **12**, 66–70.

Žilinskas, A. (1986). *Global Optimization.* Mokslas, Vilnius. 168pp. (in Russian).

Žilinskas, A., and A.Katkauskaitė (1982). On existence of stochastic function compatible with conditional relation of likelihood. *Cybernetics,* **4**, 80–83.

Žilinskas, A., and A.Katkauskaitė (1987). On statistical models of global optimization in presence of noise. In A.Žilinskas (Ed.), *Teorija Optimaljnych Reshenij,* Vol.12. pp. 43–53 (in Russian).

Received January 1990

A. Žilinskas received the Degree of Candidate of Technical Sciences from the Kaunas Polytechnic Institute, Lithuania, 1973 and the Degree of Doctor of Physical and Mathematical Sciences from the Leningrad University, 1985. He is a senior researcher at the Department of Optimal Decision Theory, Institute of Mathematics and Cybernetics, Lithuanian Acad. Sci. , and a head of the Department of Informatics and Computers at Vilnius Pedagogical Institute.