Application of Bayesian Approach to Numerical Methods of Global and Stochastic Optimization

JONAS MOCKUS
Department of Optimal Decisions Theory, Institute of Mathematics and Informatics, 2600 Vilnius, Akademijos 4, Lithuania (email: jonas.mockus@mii.lt)

(Received: 18 August 1992; accepted: 17 July 1993)

Abstract. In this paper a review of application of Bayesian approach to global and stochastic optimization of continuous multimodal functions is given. Advantages and disadvantages of Bayesian approach (average case analysis), comparing it with more usual minimax approach (worst case analysis) are discussed. New interactive version of software for global optimization is discussed. Practical multidimensional problems of global optimization are considered

Key words: Optimization, global, Bayesian, continuous, stochastic.

1. Advantages and Disadvantages of Bayesian Approach to Global Optimization

Classical approach to numerical methods is to design such sequence of points

$$x_n \in A \subset R^m, \quad n = 1, 2 \ldots$$

which converge to exact solution x^* when n is large for all problems from a given family. In some simple cases, usually connected with a convexity, the convergence rate also can be defined. In the terms of decision theory this approach can be considered as "worst case" analysis, or "minimax approach" It means that some property of method should be present always, including the worst case.

This approach seems so natural that most numerical analysts consider it as the only mathematical one. Anything else usually is classified as "heuristic," meaning the methods which may be practically quite acceptable, but without the proper mathematical justification. Mathematical justification usually is supposed to be a necessary part of serious numerical analysis. So any property of numerical methods which does not hold for all problems from a given family is often regarded at best as some empirical evidence.

The obvious advantage of the classical approach is that it helps to keep the strict standards in numerical analysis. It does not allow to flood this field with numerical methods of unknown mathematical properties.

The important disadvantage is that the minimax approach usually is to expensive. To provide something for the worst case one needs to make enormous number of iterations if the family of problems is broad enough. In a broad family of problems the worst case can be expected to be very bad.

Let us to consider for example the global optimization of the family of Lif-

Journal of Global Optimization **4**: 347–365, 1994.

schitzian functions with unknown constant. Then the best method in the minimax sense is a uniform grid on a compact feasible set, see Sukharev (1975). It means that this global optimization algorithm is of exponential complexity. The number of observations is increasing as exponent of the dimension of problem. Here "observation" means an evaluation of objective function $f(x)$ at some fixed point x.

If the Lifschitzian constant is known, then some nonuniform grid technique is preferable, see Evtushenko (1985). However, even here the complexity of algorithm apparently remains exponential, perhaps with a better constant.

In global optimization of continuous functions on a compact set we cannot apply a minimax approach at all. It is well known that maximum does not exist on a set of all continuous functions. It means that for any fixed continuous function and a fixed method of search there exists some other continuous function with a larger deviation from a global minimum. So the strong condition of uniform convergence does not apply here.

Some weaker convergence conditions usually are considered. For example a condition of convergence for any fixed continuous function. It is easy to see that to provide even this much weaker condition we need an exponential algorithm. The convergence for any continuous function can be provided only by asymptotic dense observations. Meaning that maximal distance between observations converges to zero. Otherwise we can miss a global minimum for some continuous function.

The problem becomes even more complicated if the "noise" is present. For example if we define an objective function by Monte Carlo techniques or by physical experiments.

The condition of ordinary convergence is rather a week one. It holds for any reasonable global optimization method providing asymptotic dense observations of continuous functions. It means that ordinary convergence can be regarded only as a necessary condition. Some additional conditions should be provided if we wish to justify our method sufficiently.

Following the traditions of local optimization one would like to prove not just ordinary convergence but the rate of convergence also. However, we cannot do it for a set of all continuous functions. This set is too large. The trouble is that usually we define a convergence rate using directly or indirectly a notion of supremum. This notion does not apply to a noncompact set of all continuous functions. So different ways to make the convergence conditions stronger should be investigated.

One way is to consider a density ratio instead of the rate of convergence. A density ratio we define as a ratio of density of observations in a vicinity of global minimum to an average density of observations. It defines the asymptotic efficiency of methods of global optimization reasonably well. So the density ratio in global optimization can be regarded as a replacement of rate of convergence.

So far we were talking about asymptotic. However, a useful method should be good not only asymptotic but for a finite number of observations too. It means that a good asymptotic behavior can be regarded as some desirable but not sufficient condition. To justify any claim of practical efficiency of a method of global optimization some additional "non-asymptotic" conditions should be considered. It is well known that in real life applications even the best asymptotic property can happen to be nearly useless if the number of observations is not large enough.

By "worst case" notion an "optimal" method is a method which provides minimal or at least reasonably low deviation from a global optimum for all functions from a given family. It involves a necessity to define a "worst case," what can be impossible to do if a set of functions is not compact. It is a theoretical reason against a "worst case" approach. The practical reason is that a worst case can happen to be very bad, if a family of functions is broad enough.

For example for smooth convex functions the well known variable metric methods are perhaps nearly optimal. For those methods a superlinear convergence was proven, see Powell (1971). For a family of one-dimensional unimodal functions Kiefer (1953) developed the optimal in a minimax sense method. Some optimal in a minimax sense methods were developed also for a set of Lifschitzian functions, see Evtushenko (1985), Pijavskij (1972), Shubert (1972), Sucharev (1975).

Some of those objective functions belong to a narrow sets of functions, namely smooth convex or one-dimensional unimodal. It is easy to see from the examples that here the minimax (or approximately minimax) methods can be regarded as nearly optimal from both theoretical and practical points of view. For the broader family of functions, such as Lifschitzian functions with unknown constant, the minimax approach is not so attractive. Here the provision of optimality of method for a worst case is too expensive, if a dimension of the problem is large.

For a family of continuous functions and functions with noise the worst case does not exist. So here worst case analysis is not possible at all. In global optimization of continuous functions the average case analysis seems like a reasonable way to reconcile the conditions of practical efficiency and mathematical justification.

For a theoretical justification of a numerical method it is needed to show its properties under some well defined conditions. Assuming continuity, differentiability or the existence of Lifschitzian constant of objective functions. Supposing homogeneity and independence of m-th differences of an a priori distribution, and so on. Formal testing of those conditions in real life applications is a problem about as complex as the problem of global optimization. So the correspondence of applied problems to theoretical conditions is judged mainly by intuition of experts.

It is well known that the intuition of experts depends on practical experience. It means that the theoretical analysis of methods of global optimization should be supplemented by the analysis of case studies, covering a sufficiently large family of different applied problems. It is done in the "Applications" part of the paper, considering seven real life examples from very different fields.

2. The General Ideas of Average Case Analysis in Global Optimization

The main question related to average case analysis is how to define a notion of "average." Mathematically an average is an integral. It is well known that to define an integral some measure should be fixed. The most convenient one being a probability measure P. This measure is a part of problem definition. So it should be fixed before any investigation of the problem starts. In Statistical Decision Theory, see De Groot (1970), P is called a priori distribution. An average case analysis is called Bayesian approach.

The first problem of Bayesian approach is how to define an a priori distribution P. The second one is how to update it using the results of observations $z_n = (x_i, f(x_i), i = 1, \dots, n)$. Updated distribution is called an a posteriori distribution $P(z_n)$. The third problem of Bayesian approach is how to minimize an a posteriori risk function. The risk function $R_n(x)$ is an expected deviation from global minimum at a fixed point x. The expectation is defined by an a posteriori distribution $P(z_n)$. A minimization of the risk function $R_n(x)$ defines a point of next observation x_{n+1}.

Any Bayesian method depends on a priori distribution by definition. So it is desirable to define this distribution on a basis of some clear and simple assumptions. An example: it follows from the conditions of continuity of $f(x)$, homogeneity of P and independence of m-th differences that an a priori distribution P is Gaussian with a special covariance matrix.

We update an a priori distribution by the well known formula of conditional probability. Unfortunately to update a Gaussian distribution one should inverse a covariance matrix of n-th order. Where n is number of observations. It is hardly practical if n is more than say 500. The covariance matrix represents Kolmogorov's consistency conditions. It means that the inversion can be avoided only if we replace the consistency conditions by something weaker.

Let us to replace them by the following three conditions. Continuity of risk function $R_n(x)$. Convergence of Bayesian method to a global minimum of any continuous function $f(x)$. Simplicity of expressions defining "conditional" expectation and "conditional" variance. So we define some "Bayesian" method which can be regarded as the simplest one under some assumptions. Here the term "Bayesian" has different meaning from the classical definition of Bayesian approach. The reason is that the modifyed definition of "conditional" expectation and variance do not correspond to Kolmogorov's conditions.

There are other ways to simplify the expressions of conditional expectation and conditional variance. Zilinskas, see (1986), roughly expressed them using an extrapolation theory.

We assume that a Bayesian method should converge to a global minimum of any continuous function, if an a priori distribution is chosen correctly. It means that the asymptotic of Bayesian method is at least as good as that of any classical one for a family of continuous functions. In fact it is even better. The asymptotic density of observations of Bayesian methods is considerably higher near global minimum. However, the main advantage of Bayesian methods is that they minimize an expected deviation from the global minimum for any fixed number of observations. Here asymptotic notions of complexity, such as exponential or polynomial are not directly applicable.

The problem of minimization of the risk function $R_n(X)$ even in a simplest case is multimodal. So by using the Bayesian method we are replacing the original multimodal problem by an auxiliary multimodal problem. The advantage of doing so is that the auxiliary problem is for some rough prediction of the gain expected as a result of the next observation. We need prediction just to estimate the point of next observation. The actual evaluation of objective function will be made by the next observation.

It means that there is no need for exact minimization of the risk function. We can use some simplest methods such as Monte Carlo to minimize $R_n(x)$ roughly. Even so an application of Bayesian methods can be efficient only for global optimization of "expensive" objective functions. A function $f(x)$ can be regarded as expensive if for its evaluation we need minutes of CPU time. For not expensive functions the simpler global optimization methods such as clustering, see Torn (1989), can happen to be more efficient.

Now we shall formalize all this.

3. Method of Search

We shall consider a family C_A of continuous functions $f = f(x),\ x \in A \subset R^m$. We assume a possibility to evaluate f at any fixed point $x_n, n = 1, \dots, N$, where N is a total number of observations.

A point of $n + 1$ observation is defined by decision function d_n in the following way: $x_{n+1} = d_n(z_n)$. Here we represent observed data by vector $z_n = (x_i, y_i, i = 1, \dots, n)$, $y_i = f(x_i)$. The method of search we represent by vector $d = (d_0, \dots, d_N)$. A final decision shall be denoted by $x_{N+1} = x_{N+1}(d)$. Then a deviation of a method d from a global minimum x^* can be expressed as:

$$\delta = \delta(f, d) = f(x_{N+1}(d)) - f(x^*)$$

Worst case analysis corresponds to a minimax condition:

$$\min_{d} \max_{x \in C_A} \delta(f, d) \tag{1}$$

A well known example of minimax method is a uniform grid, in a case of Lipschitzian functions with unknown Lipschitz constant, see Sukharev (1975).

Average case analysis we define by the Bayesian condition:

$$\min_{d} \int_{C_A} \delta(f, d)\, \mathrm{d}P(f) \tag{2}$$

Here P is an a priori distribution.

4. How to Define an a priori Distribution

We can see from (2) that Bayesian methods depend on an a priori distribution P. The choice of P is very wide. So first we must set some conditions which define a family of "correct" a priori distributions.

A priori distribution can be considered as correct if it provides convergence to global minimum of any continuous function when n is large. This is so, see Mockus (1989), if conditional variance converges to zero if and only if when the distance from nearest observation approaches zero. Otherwise conditional variance converges to some positive number. It means that we cannot predict exact values of objective function outside the "densely observed" area.

A large family of a priori distributions satisfy this. Some additional conditions should be introduced to narrow this family. Simple and natural are the three following conditions:

(i) Continuity of sample functions $f(x)$

(ii) Homogeneity of a priori distribution P

(iii) Independence of m-th differences

Those conditions satisfies a Gaussian a priori distribution with constant mean μ and the following covariance function:

$$\sigma_{jk} = \prod_{i=1}^{m} \left(1 - \frac{|x_j^i - x_k^i|}{2} \right) \tag{3}$$

Here $x_j^i \in [-1, 1], i = 1, \ldots, m$.

Condition (ii) means that a priori probabilities do not depend on the origin of coordinates. Condition (iii) means that the m-th differences are a sort of "white noise". Here m-th differences can be regarded as discrete approximations of derivatives of m-th order. So assumption (iii) is the weakest condition compatible with continuity of samples. It does not restrict the behavior of derivatives. As a result the sample functions are non-differentiable almost everywhere. If $m = 1$ a priori distribution (3) can be regarded as a sum of two Wiener fields running in the opposite directions plus a constant μ.

This example shows that the a priori distribution is not necessarily so "subjective" after all. It can be derived from some simple and clear assumptions.

5. How to Update an a priori Distribution

Let us denote by $p_x(y)$ an a priori probability density of objective function $f(x)$ at a fixed point x. Denote by $p_x(y|z_n)$ a conditional probability density of $f(x)$ with regard to observation vector z_n. It is called "a posteriori density".

Transformation of $p_x(y)$ to $p_x(y|z_n)$ corresponds to well known Bayesian formula. This transformation can be regarded as an updating of a priori density after observing the results represented by vector $z_n = (x_i, y_i = 1, \ldots, n), y_i = f(x_i)$.

In a Gaussian case an a posteriori density can be expressed as:

$$p_x(y|z_n) = \frac{1}{\sqrt{2\pi}\sigma_n(x)} e^{-\frac{1}{2}\frac{(y-\mu_n(x))^2}{\sigma_n^2(x)}} \tag{4}$$

Denote by $\mu_n(x)$ a conditional expectation of $f(x)$ with regard to z_n. It shows expected values of $f(x)$ after n observations.

Denote by $\sigma_n^2(x)$ corresponding conditional variance. It defines a degree of uncertainty about $f(x)$ after n observations.

Denote by $\mu(x)$ an a priori expected value of $f(x)$ and by $\sigma^2(x)$ an initial uncertainty represented as an a priori variance.

Denote by σ_{ij} the a priori covariance between $f(x_i)$ and $f(x_j)$. By σ_{xi} denote a priori covariance between $f(x)$ and $f(x_i)$. Here the corresponding covariance matrices can be expressed as $\Sigma = (\sigma_{ij})$ and $\Sigma_x = (\sigma_{xi})$.

Now we can express a conditional expectation:

$$\mu_n(x) = \mu(x) + \Sigma_x \Sigma^{-1} (y - \mu(x)) \tag{5}$$

and a conditional variance:

$$\sigma_n^2(x) = \sigma^2(x) - \Sigma_x \Sigma^{-1} \Sigma_x^T \tag{6}$$

If a priori distribution corresponds to (3) and $m = 1$ then expression (5) is a linear interpolation between the observed values y_i. Expression (6) defines a piecewise quadratic positive function with zero values at observed points x_i. In this special case we need no inversion of covariance matrix. The reason is that the a priori distribution corresponding to (3) is Markovian. It means that conditional expectation and conditional variance depends only on the two closest observations. One observation on the left and one on the right. It helps to design simple and efficient Bayesian methods for one-dimensional global optimization, see Zilinskas (1989). Unfortunately the Markovian property holds only in one-dimensional case.

6. How to Define and Minimize a Risk Function

It follows from (2) that a Bayesian method can be expressed by condition:

$$d^* = arg \min_d \int_{C_A} (f(x_{N+1}(d)) - f(x^*)) \, \mathrm{d}P(f)$$

This condition can be made simpler omitting a second component of the integral which does not depend on d. So we have:

$$d^* = arg \min_d \int_{C_A} f(x_{N+1}(x)) \, \mathrm{d}P(f) \tag{7}$$

Equation (7) can be reduced to N-dimensional dynamic programing problem, see Mockus (1972). However, this problem is too difficult. So one-step approximation is used. At any n we suppose that the $n+1$-th observation is the last one. And so on until we reach $n = N$.

Assume that P is such, that the minimum of conditional expectation $\mu_{0n} =$

$\min_{x \in A} \mu_n(x)$ is equal to the minimal observed value $y_{0n} = \min_{1 \leqslant i \leqslant n} y_i$ like in the Wiener process.

Then the point of the next observation $x = x_{n+1}$ should minimize a conditional expectation of the least of two values: $f(x)$ and c_n, where $f(x)$ is a predicted value at a point x and $c_n = y_{0n} - \epsilon_n$.

Here ϵ_n is a correction parameter. This parameter can control influence of remaining observations. It should be large if there are many of them. It should be small otherwise.

In the one-step Gaussian case a risk function can be expressed as:

$$R_n(x) = \frac{1}{\sqrt{2\pi}\sigma_n(x)} \int_{-\infty}^{+\infty} \min(y, c_n) e^{-\frac{(y-\mu_n(x))^2}{2\sigma_n^2(x)}} \mathrm{d}y \tag{8}$$

and a Bayesian method can be defined as:

$$x_{n+1} = arg \min_{x \in A} R_n(x), \quad n = 1, \dots, N. \tag{9}$$

We can see from (5), (6) and (8) that to define the risk function in cases when $m > 1$ we need to inverse the covariance matrix of n-th order. It is too difficult when n is large. So we have no choice but to abandon the classical framework of probability theory. Some of its basic assumptions we shall replace by other conditions more convenient for calculations.

The inversion of covariance matrix Σ corresponds to solution of the system of linear equations which represent the Kolmogorov's consistency conditions. So the only way to get rid of expensive inversion is to omit those conditions.

It seems reasonable to replace Kolmogorov's consistency conditions by the following three assumptions:

(i) Continuity of risk function (8)

(ii) Convergence of method (9) to a global minimum of any continuous function

(iii) Simplicity of expressions for $\mu_n(x)$ and $\sigma_n(x)$

Then method (9) can be expressed, see Mockus (1989), as:

$$x_{n+1} = arg \max_{x \in A} \min_{1 \leqslant i \leqslant n} \frac{\|x - x_i\|^2}{y_i - y_{0n} + \epsilon_n}, \quad n = 1, \dots, N. \tag{10}$$

7. Convergence of Bayesian Methods

The main motivation to introduce the Bayesian search was not asymptotic at all. We wanted to minimize average deviation after finite (usually small) number of observations. So we defined the Bayesian methods (9) and (10).

Now let us consider those methods also from the asymptotic point of view. Usual convergence conditions we already included as a condition (ii) defining Bayesian method (10). This condition is weak. It does not show an efficiency of search. Efficiency of search can be defined as a relation of density of observations in an area of global minimum to corresponding average density. We shall denote it by K_n. Then asymptotic of K_n for method (10) can be expressed, see Mockus (1989), as:

$$K = \lim_{n \to \infty} K_n = \frac{f_a - f_0 + \epsilon}{\epsilon} \tag{11}$$

Here

$$\epsilon = \lim_{n \to \infty} \epsilon_n$$

where $\epsilon_n > 0$ is a correction parameter, see (10).

We can see from (11) that the search will be nearly asymptotic uniform if the correction parameter ϵ is large or if the objective function $f(x)$ is flat, meaning that $f_a - f_0$ is small.

If the correction parameter is small then most of the observations will be placed asymptotically around a global minimum. However, it will take more observations to reach this asymptotic condition.

8. Software

The global optimization software was developed considering the results of national (USSR) and international "competition" of different algorithms of global optimization, see Dixon and Szego (1978). Some experience in real life optimization problems also was used selecting the set of optimization algorithms, see the "applications" part of this paper. The set of algorithms of global optimization includes four versions of Bayesian search, one version of clustering, a version of uniform deterministic grid and pure Monte Carlo search.

Usually it is reasonable to start optimization by a global method and to finish it by some local method. Two global methods, the clustering and the Zilinskas version of Bayesian technique contains some simple algorithms of local search. It means that the local search is not necessary for those two methods, but it may be useful.

There are three local optimization methods. One method is of variable metric type, with Lagrangian multipliers and penalty functions, for constrained optimization of smooth functions, see Schittkowski (1985). The second method is of simplex type of Nelder and Mead, with penalty functions for constrained optimization of nondifferentiable functions, see Himmelblau (1972). The third is of stochastic approximation with Bayesian step size control, for functions with "noise," see Mockus (1989).

Each subroutine represents a global or a local method. The choice of method has to follow the idea that the computational complexity of the method roughly corresponds to that of the objective function.

For "expensive" functions the Bayesian methods could be recommended. Those methods need a lot of auxiliary calculations, but each observation is very efficient.

For “cheap” functions simple grid methods, such as Monte Carlo or uniform deterministic grid, see Sobolj (1969), can happen to be better. Here observations are not so efficient, but auxiliary calculations are negligible. It explains relative efficiency of simple methods optimizing simple functions.

If we expect the number of local minima to be small, then clustering techniques, see Torn (1989), may be the best choice.

For global optimization of one-dimensional functions a relatively simple Bayesian technique is available.

There are optimization problems where objective functions can be roughly represented as a sum of components depending on different variables. Here the Bayesian method of coordinate search usually shows very good results. This method globally optimizes one variable at a time using one-dimensional Bayesian search. The difference of this method from other methods of global optimization is that it depends on a starting point. So the deviation from global minimum can be made as small as we want by applying a multi-start procedure from different uniformly distributed starting points.

All global methods optimize in a rectangular region. Linear and nonlinear inequality constraints is represented as some penalty function. The same applies also for local method of stochastic approximation type.

In local methods of simplex and variable metrics type linear and nonlinear constraints can be defined directly. It may be done by subroutines for constraints, supplied by the user.

The global optimization software is in three versions. The first one is a library of portable Fortran subroutines. The users guide and source codes are in a book by Mockus (1989), including a floppy disk.

The portable Fortran version can run on any computer with standard Fortran compiler. Users should represent objective functions in a form FUNCTION FI(X,N), where X is an array of variables, N is its dimension.

Rectangular constraints are given by arrays of lower bounds A and upper bounds B. For local methods of simplex and variable metrics type we may represent constraints by subroutine CONSTR(X,N,G,MC). Here G is an array of length MC which contains the values of constraints at the point X and MC is the number of constraints.

The second version is an interactive system. Here the objective function can be written in one of two forms: Fortran or C. This system is for users which prefer to represent objective functions both in C and in Fortran. This version needs Microsoft C and Fortran compilers, see Mockus (1990).

In this version, besides the regular scalar optimization, there is also a possibility of vector optimization applying the idea of Pareto optimality. The Pareto set can be approximately defined using a Bayesian or a grid method.

The latest interactive version is for objective functions represented in C. It needs a Turbo C compiler. Here a user represents objective function $f(x)$ as some C subroutine.

In both interactive systems users can select the global or local method by a menu system. The current results can be observed in table and graphical forms. There are two graphical forms:

1. GRAPH: the dependence of best value of objective function on observation number.

2. PROJECTION: the dependence of observed objective function values on one of the variables.

The main advantage of the first version, the Fortran library, is its portability. Disadvantage is that interactive possibilities are very limited. It means that this version can be easily used for some well defined optimization problems, but not for preliminary investigations, where interaction is essential.

The advantages of the second (Microsoft) and the third (turbo C) versions are reasonably good interactive facilities. Disadvantage is that both those versions can be used only on PC. It is all right for preliminary investigation and for solution of small scale problems. For real life global optimization problems the computational power of PC is usually not sufficient.

So the fourth version is developed. It is a global optimization software designed for UNIX and X Window systems. It has excellent interactive facilities plus portability to almost any computer, including super computers and some parallel computers.

9. Examples of Applications

Many examples of applications are about optimization of parameters of mathematical models represented as some systems of nonlinear differential equations. The objective function $f(x)$ here depends on a solution of the equations. Variables x represent the parameters of system which we can control. To such family of problems belong the three following examples:

- Maximization of general yield of differential amplifiers.
- Optimization of mechanical system of shock-absorber.
- Estimation of parameters of nonlinear regression of immunological model.

The last example suggest a broad area for the applications of global optimization. It is well known that in nonlinear regression the square deviation and also the likelihood functions could happen to be multimodal for some data. The number of local minima can be very large, even in simple cases. An example:

- Estimation of unknown parameters of bilinear time series.

The large source of difficult global optimization problems is engineering design. Here we are optimizing parameters of some mathematical models, usually nonlinear. An example:

- Optimization of composite laminates.

Many laws of nature could be defined in the terms of global optimization. An example:

- The "Disk" problem: minimization of potential energy of organic molecule.

We often cannot describe the behavior of new materials and technologies by mathematical models, because the corresponding information and knowledge is not available. Here optimization can be done by direct experiments, changing the control variables and observing the results. An example:

- The planning of extremal experiments of thermostable polymeric composition.

Let us now to consider those examples separately.

10. Maximization of General Yield of Differential Amplifiers

In the production of LSI (Large Scale Integration) electronic circuits there are some inevitable deviations of dimensions of LSI elements such as transistors and resistors from the ratings set by designer. As a result some of the circuits will not meet the standards. The reason is that such parameters as delay time τ, lower level of output voltage U or a bias of zero u may get out of the feasible region.

Those parameters we can define by a system of nonlinear differential equations, depending on dimensions of transistors and resistors. Usually we solve the system of differential equations using specific numerical techniques. Deviations of dimensions from the fixed ratings can be simulated using some Monte Carlo techniques assuming multivariate Gaussian distribution.

In addition there are also so called "catastrophic" rejects, due to the defects of silicon crystal.

So the yield function can be expressed as a product of three components:

- The number of circuits from a crystal. It is a decreasing function of dimensions.
- The unit minus the percentage of rejects, attributable to deviation of parameters. It is an increasing function of dimensions.
- The unit minus the part of rejects, due to crystal defects. It is a decreasing function of dimensions.

The product of monotonous functions is not necessarily unimodal. For differential amplifier the yield function can happen to be two-modal for each variable representing width or length of transistor. It means the possibility of 2^{2m} modality, where m is the number of transistors.

The multimodality of yield function together with the presence of noise which is inevitable in any Monte Carlo simulation makes the problem very difficult. So only coordinate optimization (global line search along each variable one by one) seems convenient enough.

Here the noise was filtered by Wiener smoothing. It means that the yield function is assumed to be a Wiener process and that the simulation error is a Gaussian noise with standard deviation σ, see Baskis and Mockus (1988).

If σ is large then we shall get a horizontal line corresponding to average value of observations. It means complete filtering and no optimization. If σ is zero then we shall see piecewise linear line connecting the observed values. It means no filtering at all and a large number of pseudo local minima. Good value of smoothing parameter σ for differential amplifier was about 10.

In one-dimensional global optimization Wiener smoothing seems more convenient comparing with the well known techniques of smoothing of scatterplots, see Friedman *et al.* (1980). In Wiener smoothing there is only one parameter controlling the smoothing level, and this parameter has clear statistical meaning. It can be regarded as a variance of Gaussian noise and can be estimated using the results of observations.

The idea of Wiener smoothing can be regarded as a spline smoothing under some conditions, see Craven and Wahba (1979). The meaning of standard deviation of noise σ is similar to that of smoothing parameter of spline λ.

11. Optimization of the Mechanical System of Shock-Absorber

Let us consider the mechanical object of two masses. Suppose that the shock is instantaneous impulse and that the shock-absorber can be represented by a system of linear differential equations. The parameters of the shock-absorber should minimize the maximal deviation during the transitional process.

$$f(x) = \max_{0 \leqslant t \leqslant T} | \nu(t)|$$

where $\nu(t)$ denotes a trajectory of lower mass and $f(x)$ means the maximal deviation during the transitional process. The four components of vector $x \in B$ represent different relations of elasticities, masses and dampers. The feasible set B we define by nonlinear constraints. Using penalty functions we can reduce the problem to the optimization in a rectangular set A, where $B \subset A$.

Here we see two multimodal problems: one-dimensional one $\max_{t\in[0,T]} | \nu(t)|$, and four-dimensional one $\min_{x\in A} f(x)$.

A convenient way to maximize one-dimensional multimodal function $\nu(t)$ is by relatively simple one-dimensional Bayesian method, see Zilinskas (1976). The four-dimensional function is not unimodal and rather expensive. Calculation of this function for a fixed x is defined algorithmically and includes two procedures: one defining the trajectory $\nu(t)$ and the other maximizing $|\nu(t)|$. So minimizing $f(x)$ it is natural to use the global multi-dimensional Bayesian method, see Mockus (1989).

The case of nonlinear shock-absorber (when we represent the object by a system of nonlinear differential equations) can be treated in a similar way. One difference is that numerical instead of analytical integration of nonlinear differential equations should be applied. The other difference is that there appears an additional fifth parameter of "nonlinearity."

12. Estimation of Parameters of Nonlinear Regression of Immunological Model

The well known mathematical model of immune response is the system of nonlinear differential equations with time delay:

$$\begin{aligned}\frac{dv}{dt} &= -\gamma f v\\ \frac{df}{dt} &= \rho c - \eta\gamma f v - \mu_1 f\\ \frac{dc}{dt} &= \alpha f v|_{t-\tau} - \mu_2(c - c_0)\end{aligned}$$

Here $v = v(t)$ is the density of plasma antigen, $f = f(t)$ is the density of specific antibodies, $c = c(t)$ is the density of plasma cells, τ is time delay, $\gamma, \rho, \eta, \mu_1, \mu_2, \alpha$ are the unknown parameters of the model which we shall represent by vector x.

The relation between the parameters x of the model and the trajectories $v(t), f(t)$ and $c(t)$ at the points $t_i, i = 1, \ldots, K$ can be defined by numerical methods. Belykh (1983) developed very efficient numerical methods for integration of differential equations with time delay.

The objective function $f(x)$ is a likelihood function for some given data. Experimental data corresponds to the reaction of a homogeneous sample of mice to the inoculation of a nonpathogenic antigen. The results of experiments are the plasma cell densities at six fixed moments of time $t_i, i = 1, \ldots, 6$.

For this data the likelihood function appears as a unimodal one. However, global optimization happened to be useful selecting a good starting point for local optimization. It is important, because the efficiency of local search depends on a starting point.

Function $f(x)$ is not expensive, due to efficiency of specific numerical integration methods by Belykh (1983). So the methods of uniform deterministic optimization happened to be sufficiently good for preliminary global search.

13. Estimation of Parameters of Bilinear Time Series

The class of bilinear time series models is useful for describing many non-linear phenomena. Let us consider simple example:

$$y_i = x^1 y_{i-1} + x^2 y_{i-2} + x^3 y_{i-1} e_{i-1} + x^4 y_{i-2} e_{i-1} + e_i$$

Here $x = (x^1, \ldots, x^4)$ are unknown parameters, $y_i, i = 1, \ldots, k$ is experimental data, e_i are residuals.

The sort of behavior seen from this model is common in seismological data, see Subba Rao and Gabr (1984). For this type of data, the activity due to an event is of very short duration.

The unknown parameters x can be estimated by minimizing the sum of residual squares:

$$f(x) = \sum_{i=1}^{K} e_i{}^2$$

It is easy to see that when i is large then the residuals e_i are polynomials of high degree, so multimodality of $f(x)$ is almost inevitable. In that sense it is a good example for application of global optimization. The objective function $f(x)$ is simple, so the uniform deterministic search can be recommended for global optimization. The function $f(x)$ is smooth, so a sort of second order techniques such as variable metrics techniques may be used for local optimization.

14. Optimization of Composite Laminates

Optimization is usually most efficient in new developments. There are several reasons for that, such as the lack of experience and the absence of library of ready made designs to choose. There is no "feeling" of the problem, which usually appears after some experience.

An example of new application area is the design of laminated composite materials. Laminated composites are of several thin layers (or plies), which are bound together to form a composite laminate. A single ply consists of long reinforcing fibers (e.g. graphite fibers), embedded within a relatively weak matrix material (e.g., epoxy). All fibers within an individual ply are oriented in one direction. Composite laminates are usually fabricated such, that fiber angles vary from ply-to-ply.

In our studies we have restricted the decision variable x^k to be fiber orientation angle in the k-th play. Here k ranges from 1 to n and n equals the number of plies in the laminate. Angles x^i ranges from -90^0 to $+90^0$.

The objective function $f(x)$ was defined by experts as a sum of "penalty" functions corresponding to different plies $k = 1, \ldots, n$ under different load conditions $j = 1, \ldots, m$. It was supposed that penalties are exponentially increasing functions of calculated ply strains $\epsilon^k_{i(j)}$, $i = 1,2$ and ply shear strain $\gamma^k_{12(j)}$. It was assumed that the penalty functions are multiplied by factor $\delta = m * n$ if the corresponding critical strains ϵ^{cr}_i and γ^{cr}_{12} are exceeded. So:

$$\begin{aligned} f(x) \;&=\; \left(\frac{1}{m}\right) \sum_{j=1}^{m} \left[\sum_{k=1}^{n} \left[\left(\delta \cdot \exp \frac{|\epsilon^k_{1(j)}| - \epsilon^{cr}_1}{\epsilon^{cr}_1} \right)^2 \right.\right. \\ &\left.\left. + \left(\delta \cdot \exp \frac{|\epsilon^k_{2(j)}| - \epsilon^{cr}_2}{\epsilon^{cr}_2} \right)^2 + \left(\delta \cdot \exp \frac{|\gamma^k_{12(j)}| - \gamma^{cr}_{12}}{\gamma^{cr}_{12}} \right)^2 \right] \right] \end{aligned}$$

The ply strains $\epsilon^k_{i(j)}, i = 1,2$ and $\gamma^k_{12(j)}$ are functions of the decision variables x^k and the applied loadings. The summation is over all n-plies and m-loading conditions.

Three methods of global optimization were compared, Bayesian, uniform deterministic and Hit-and-Run, see Zabinsky *et al.* (1990). As expected Bayesian method used the available observations in a most efficient way. The method of Hit-and-Run used less computer time, due to simpler auxiliary calculations. The method of uniform deterministic search was somewhere between Bayesian and Hit-and-Run methods, both in terms of computer time and number of calculations.

This ordering of efficiency of methods is natural for simple functions. We expect that increasing the complexity of $f(x)$, what apparently will happen taking into account many additional important factors, the Bayesian method may turn out to be the best one.

15. The "Disk" Problem: Minimization of Potential Energy of Organic Molecule

Let us assume that the potential energy $f(x)$ of a molecule can be represented as a sum of functions $v_{ij}(r_{ij})$, where r_{ij} is the distance between atom i and atom j. So:

$$f(x) = \sum_{i,j=1,\ldots,m,i<j} v_{ij}$$

where:

$$v_{ij} = \left(\frac{s_{ij}}{r_{ij}}\right)^{12} - \left(\frac{s_{ij}}{r_{ij}}\right)^{6}$$

Here s_{ij} is diameter of atom j and

$$r_{ij} = \sqrt{(x^i - x^j)^2 + (y^i - y^j)^2 + (z^i - z^j)^2}.$$

Vector (x^j, y^j, z^j) represents the center of atom j.

The two-dimensional "Disk" problem was a test to compare different methods of global optimization at the CECAM Workshop on Global Optimization and Molecular Chemistry, Paris, June 1990.

Bayesian algorithm was good in a sense of number of observations. Some other methods such as tunneling, see Levy *et al.* (1982), stochastic equations, see Alufi-Pentini *et al.* (1985), and discrete dynamical system, see Donnelly and Rogers (1988), were more efficient in computer time, due to simpler auxiliary calculations.

The Disk problem can be considered in several stages.

In the first stage the main objective would be to get just some preliminary understanding of the behavior of the objective function. To do it we would like to define some initial approximation to the global minimum, using a very small number of observations.

Most of the methods of optimization at the Workshop were using also gradient values. It was supposed gradients are roughly N+1 (where N is number of variables) times more "expensive" and more "informative" comparing with single observation. It was assumed that the minimal reasonable number of iterations in the seven point (14 variables) disk problem is ten. So the minimal equivalent number of observations can be estimated as $150 = (14 + 1) * 10$.

The advantage of this stage of research is that different methods can be easily compared using different computers, including PC. The disadvantage is that the test problem of just 7 “atoms” is very small comparing with real life organic molecules with 300 and more atoms.

There are good reasons to suppose that the Disk problem belongs to a NP-complete family. It means that any realistic number of observations for say a 100 point case will appear at least as small as 150 observations in the 7 point case. The reason is the exponential complexity of NP-complete family of problems. So the results of comparison of methods obtained on a small problem using small number observations can happen to be useful for larger problems and correspondingly larger numbers of observations.

For the tasks of this stage a useful tool would be some PC compatible global optimization software with good graphics, for example the system GLOBAL MINIMUM. The result after 9 observations of global Bayesian search and 108 observations of local search was –9.45. It deviates significantly from the global minimum which is about –12.6.

One of the reasons of such large deviation is that the number of observations was very small. The other reason is that the global optimization methods usually are most efficient, if objective function can be represented as a sum of two components: one unimodal and one multimodal. Multimodal component makes the sum of two multimodal. The projection of the potential energy function looks differently; like almost constant function with occasional high and narrow spices.

The second stage is to compare the results without restricting observations number. The result after 153952 observations using the clustering algorithm was -12.5396, what is quite near to the global minimum. Some other methods achieved a similar result, but the number of observations was larger.

16. Planning of Extremal Experiments of Thermostable Polymeric Composition

The objective function $f(x)$ is the specific loss of mass kg/m^2 in the flow of high temperature gases during the fixed interval (20 sec.).

There are four variables $x = (x^1, x^2, x^3, x^4)$, where

x^1 is the proportion of carbonized phenol-formaldehyde resin to phenol-formaldehyde.

x^2 is the proportion of urotropine,

x^3 is the specific pressure of moulding kg/cm^2,

x^4 is time of moulding min.

The technological constraints are given in the form $a_i \leqslant x^i \leqslant b_i, i = 1, \ldots, 4$.

No mathematical description of the objective function was available. The opinion was that $f(x)$ perhaps have more than one local minimum. The only way to define values of $f(x)$ at fixed points x_i was by physical experiment. In such a case some noise η usually is present. The presence of noise makes testing of unimodality of $f(x)$ very difficult. It means that applying well known local methods of extremal experimental planning we can get stuck to some local minimum, which is far away from the global one.

Here the application of Bayesian methods seems natural. In agreement with usual Bayesian techniques the first 12 observations were uniformly distributed. The following 14 observations were carried out by one-step Bayesian method with the Gaussian a priori distribution (3).

Table I shows only the results of those observations when the quality of material was increasing. The best value of the specific loss 0.385 was considered to be good for the materials of the given type.

Table I. Minimization of the specific loss of mass

i	x_i^1	x_i^2	x_i^3	x_i^4	$f(x_i)$
1	50	13	4 070	8	2.141
2	55	14.5	6 990	12	0.737
4	65	12.5	5 210	4	0.514
5	70	11	5 909	30	0.495
11	70	14	3 289	3	0.493
18	77	7.7	3 700	20	0.417
26	80	10.5	4 320	17	0.385

References

Al-Khayyal, F.A. and Falk, J.E. (1983), Jointly Constrained Biconvex Programming, *Mathematics of Operations Research* **8**, 273–286.

Alufi-Pentini, F., Parisi, V. and Zirilli, F. (1985), Global Optimization and Stochastic Differential Equations, *J. of Optimization Theory and Applications* **47**, 1–16.

Archetti, F. and Betro, B. (1979), A Probabilistic Algorithm for Global Optimization, *Calcolo* **16**, 335–343.

Baskis, A. and Mockus, L. (1988), Application of Global Optimization Software for the Optimization of Differential Amplifier, *Theory of Optimal Decisions*, 9–16, Vilnius, Lithuania (in Russian).

Belykh, L.N. (1983), On the Computational Methods in Disease Models, *Mathemathical Modeling in Inmunology and Medicine,* ed. G.I. Marchuk and L.N. Belykh, North-Holland Publishing Company, New York, 79–84.

Benson, H.P. (1982), Algorithms for Parametric Nonconvex Programming, *J. of Optimization Theory and Applications* **38**, 316–340.

Boender, G. and Rinnoy Kan, A. (1987), Bayesian Stopping Rules for Multi-Start Global Optimization Methods, *Mathematical Programming* **37**, 59–80.

Craven, P. and Wahba, G. (1979), Smoothing Noisy Data with Spline Functions, *Numerische Mathematik* **31**, 377–403.

De Groot, M. (1970), *Optimal Statistical Decisions*, McGraw-Hill, New York.

Dixon, L.C.W. and Szego, G.P. (1978), *Towards Global Optimization*, North Holland, Amsterdam.

Donnelly, R.A. and Rogers, J.W. (1988), A Discrete Search Technique for Global Optimization, *International Journal of Quantum Chemistry: Quantum Chemistry Symposium* **22**, 507–513.

Ermakov, S.M. and Zigliavski, A.A. (1983), On Random Search of Global Extremum, *Probability Theory and Applications* **83**, 129–136 (in Russian).

Evtushenko, Yu. G. (1985), *Numerical Optimization Techniques*, Optimization Software, Inc., New York.

Floudas, C.A. and Pardalos, P.M. (1987), *A Collection of Test Problems for Constrained Global Optimization Algorithms*, Lecture Notes in Computer Science **455**, Springer-Verlag.

Friedman, J.H., Jacobson, M., and Stuetzle, W. (1980), Projection Pursuit Regression, Technical Report # 146, March, Department of Statistics, Stanford University, 1–27.

Galperin, E. and Zheng, Q. (1987), Nonlinear Observation via Global Optimization Methods: Measure Theory Approach, *J. of Optimization Theory and Applications* **54**, 63–92.

Hansen, E. (1984), Global Optimization with Data Perturbation, *Computational Operations Research* **11**, 97–104.

Hong, Ch.S. and Zheng, Q. (1988), *Integral Global Optimization* Lecture Notes in Economics and Mathematical Systems **298**, Springer-Verlag.

Horst, R. and Tuy, H. (1990), *Global Optimization*, Springer-Verlag.

Kiefer, J. (1953), Sequential Minimax Search for a Maximum, *Proceedings of American Mathematical Society* **4**, 502–506.

Kushner, M.J. (1964), A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise, *J. of Basic Engineering* **86**, 97–106.

Levy, A.V., Montalvo, A., Gomez, S. and Calderon, A. (1982), *Topics in Global Optimization*, Lecture Notes in Mathematics # 909, 18–33.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953), Equations of State Calculation by Fast Computing Machines, *Journal of Chemical Physics* **21**, 1087–1092.

Michalevich, V., Supel, A. and Norkin, V. (1978), *Methods of Nonconvex Optimization*, Nauka, Moscow (in Russian).

Mockus, A. and Mockus, L. (1990), Design of Software for Global Optimization, *Informatica* **1**, 71–88.

Mockus, J. (1972), On Bayesian Methods of Extremum Search, *Automatics and Computer Technics* **72**, 53–62 (in Russian).

Mockus, J. (1989), *Bayesian Approach to Global Optimization*, Kluwer Academic Publishers, Dordrecht-London-Boston.

Mockus, J. and Mockus, L. (1991), Bayesian Approach to Global Optimization and Applications to Multiobjective and Constrained Optimization, *J. of Optimization Theory and Applications* **70**, 155–171.

Pardalos, P.M. and Rosen, J.B. (1987), *Constrained Global Optimization: Algorithms and Applications*, Lecture Notes in Computer Science # 268, Berlin, Springer-Verlag.

Pijavskij, S.A. (1972), An Algorithm for Finding the Absolute Extremum of Function, *Computational Mathematics and Mathematical Physics*, 57–67.

Powell, M.J.D. (1971), On the Convergence Rate of the Variable Metric Algorithm, *J. Inst. of Mathematics and Applications* **7**, 21–36.

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

Ratschek, H. and Rokne, J. (1988) *New Computer Methods for Global Optimization*, John Willey, New York.

Saltenis, V. (1971), On One Method of Multiextremal Optimization, *Automatics and Computer Technics* **71**, 33–38.

Schittkowski, K. (1985/86), NLPQL: A FORTRAN Subroutine Solving Constrained Nonlinear Programming Problems, *Annals of Operations Research* **5**, 485–500.

Schnabel, R.B. (1987), Concurrent Function Evaluations in Local and Global Optimization, *Computer Methods in Applied Mechanics and Engineering* **64**, 537–552.

Shubert, B.O. (1972), A Sequential Method Seeking the Global Maximum of Function, *SIAM Journal on Numerical Analysis* **9**, 379–388.

Sobolj, I.M. (1967), On a Systematic Search in a Hypercube, *SIAM Journal on Numerical Analysis* **16**, 790–793.

Stoyan, Yu.G. and Sokolowskij, V.Z. (1980), *Solution of some Multiextremal Problems by the Method of Narrowing Domains*, Naukova Dumka, Kiev (in Russian).

Strongin, R.G. (1978), *Numerical Methods of Multiextremal Optimization*, Nauka, Moscow.

Subba Rao, T. and Gabr, M.M. (1984), *An Introduction to Bispectral Analysis and Bilinear Time Series Models*, Lecture Notes in Statistics # 24, Berlin, Springer-Verlag.

Sukharev, A.G. (1975), *Optimal Search of Extremum*, Moscow University Press, Moscow.

Torn, A. and Zilinskas, A. (1989), *Global Optimization*, Lecture Notes in Computer Science # 350, Berlin, Springer-Verlag.

Zabinsky, Z.B., Smith, R.L. and McDonald, J.F. (1990), *Improving Hit and Run for Global Optimization*, Working paper, Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI.

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

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