On Use of Statistical Models in Global Optimization

Antanas ŽILINSKAS, Gražina GIMBUTIENĖ

Vilnius University
Institute of Mathematics and Informatics, Lithuania
Research Council of Lithuania, grant No. MIP-051/2014.
A PLENARY LECTURE AT
INTERNATIONAL CONFERENCE AND
SCHOOL
HIGH-PERFORMANCE COMPUTING,
OPTIMIZATION AND APPLICATIONS

Nizhnij Novgorod, November 16-20, 2015.
Abstract. Statistical models for global optimization are substantiated and constructed. Global optimization algorithms are constructed implementing principles of the theory of rational decisions under uncertainty. Global convergence of proposed algorithms is analyzed. Statistical and deterministic models are generalized to the problems of non convex multiobjective optimization. Corresponding algorithms are developed implementing principles of the rational decision theory.
Contents

• Introductory remarks.
• Single-objective global optimization.
• Multi-objective global optimization.
• Open problems.
Introductory remarks
In global optimization prevail heuristic methods: genetic algorithms, ant colony optimization, honey bees dancing, cuckoo search, harmony search, etc. In some cases heuristic algorithms are substantiated theoretically, e.g. simulated annealing.

In local optimization, contrary, prevail a major paradigm of descent which is combined with the use of an appropriate model of objective functions in question.
Single-objective global optimization
Black box optimization

• Mathematically the problem of optimization is formulated as

\[ f^* = \min_{x \in A} f(x), \]

where \( f(x) \) is a nonlinear objective function of continuous variables \( f : \mathbb{R}^d \to \mathbb{R}, \ A \subset \mathbb{R}^d \). Global minimum and minimum point should be found.

• Frequently the objective functions of practical problems are available as computer programs, and the properties of the objective function are difficult to elicit. We assume that the objective function values are given by a "black box'.
We are interested in the construction of an algorithm which would be well substantiated theoretically including estimate of its convergence rate.
Statistical model
A standard model of functions under uncertainty considered in the probability theory is a stochastic function. Frequently the terms "stochastic process" and "random field" are used for stochastic functions of one and several variables correspondingly. Since theoretical properties of stochastic functions, explicitly presented in probabilistic textbooks, normally are not sufficient to substantiate the selection of a suitable statistical model, the visual analysis of sample functions of a candidate model can be helpful.
An objective function intended to minimize normally is not a completely black box, since some information on $f(\cdot)$ is available, e.g. from the similar problems solved in the past. Let values of the considered function be observed at some points $x_i$, but $f(x)$ be not yet observed. The weakest assumption on available information seems to be the comparability of likelihood of inclusions $f(x) \in Y_1$ and $f(x) \in Y_2$ where $Y_1$, $Y_2$ are arbitrary intervals. If it seems more likely that $f(x) \in Y_1$ than $f(x) \in Y_2$, such a subjective assessment will be denoted $Y_1 \succ_x Y_2$. 
The axioms on rationality of comparative (subjective) likelihood imply the existence of a probability density compatible with the axioms, i.e. the existence of a probability density \( p_x(t) \) such that

\[
Y_1 \succeq_x Y_2 \iff \int_{Y_1} p_x(t) dt \geq \int_{Y_1} p_x(t) dt.
\]

Correspondingly, the family of random variables \( \xi(x), x \in A \), with probability densities \( p_x(\cdot) \) is acceptable as a statistical model of objective functions. The distribution of \( \xi(x) \) is frequently accepted Gaussian because of computational reasons.
Let $k$ values of $f(x)$ are known: $y_i = f(x_i)$, $i = 1, ..., n$, and uncertainty with respect to unknown values of $f(x)$ correspond to the postulated above. Then the family of random variables $\xi(x)$, $x \in A$ is acceptable as a statistical model of the considered function. By means of similar axioms the parameters of $\xi(x)$ can be defined implying formulas for their computing:

$$E\{\xi(x)\} = m(x|x_i, y_i), \quad D\{\xi(x)\} = s^2(x|x_i, y_i).$$
Assumptions on rationality of search
The algorithm is defined by functions $d_i(\cdot)$ for computation of the point of current computation of an objective function value

$$x_i = d_i(x_j, y_j, j = 1, \ldots, i - 1), i = 2, \ldots, n,$$

(1)

and the functions for computing the point of the first computation $x_1$ which depends on the considered model and on $n$, and for computing the approximation of the global minimizer $x^{on}$, which depends on the complete information extracted during the search.
An optimal algorithm with respect to a stochastic function selected for a model of objective functions can be defined similarly as an optimal algorithm with respect to a deterministic model replacing the criterion of worst case error with the criterion of average error
\[ \tilde{D} = \arg \min_D E(\xi(x^{on}) - \min_{x \in \mathbf{A}} \xi(x)), \]
(1)where expectation is defined with respect to the probability measure defining the stochastic function in question, \( x^{on} \) the random point in \( \mathbf{A} \) found as an approximation to the minimizer of \( \xi(x) \) by the algorithm \( D \) after the fixed in advance number \( n \) of computations of function values.
The implementation of the optimal Bayesian algorithm is difficult, therefore various simplified versions were proposed. One of the most popular simplifications is named a one-step Bayesian algorithm \[ \text{[?]} \], which at the current step is defined by the following formula

\[
x_i = \arg \min_{x \in \mathbf{A}} E(\min\{\xi(x), y_{o,i-1}\}|x_j, y_j, j = 1, \ldots, i - 1),
\]

(2)

where \( E\{\cdot | \cdot \} \) denotes the condition average, and

\[
y_{o,i} = \min_{j=1,\ldots,i} y_j.
\]
The first global optimization algorithm based on a statistical model selects a current point for computation of function value according to the following formula

\[
x_i = \arg \max_{x \in A} P(\xi(x) \leq y^{o,i-1} - \varepsilon|x_j, y_j, j = 1, \ldots, i - 1),
\]

\[
y^{o,j} = \min_{i=1,\ldots,j} y_i, \varepsilon > 0,
\]

meaning that it is aimed at maximal probability of the improvement of the current estimate of global minimum. Later this algorithm was substantiated axiomatically and was named the P-algorithm.
The rationality of the P-algorithm was defined by the formalization of the following informal assumptions

- The choice of the current point is not rational in the close vicinities of the points of previous computations of values of objective functions.

- The priority of points where large function values are expected with large probabilities is low.

- Aspirations for elicitation of global information and for local improvement do not dominate each other.
The algorithm which satisfies such axioms, called the P-algorithm, computes at the current optimization step the objective function at the point

\[ x_{n+1} = \arg \max_{x \in A} P_n(x), \quad P_n(x) = P\{\xi(x) < y_{on}^n - \varepsilon | \xi(x_i) = y_i, i = 1, \ldots, n\}, \]

where \( y_i \) are values of the objective function computed at previous optimization steps, and \( f(x_i) = y_i \).
Any characterization of a random variable normally includes a location parameter (e.g., mean) and a spread parameter (e.g., standard deviation); we use a minimal description of $\xi_x$ by these two parameters which will be denoted by $m(x)$ and $s(x)$. The dependence of both parameters on the information available at the current optimization step ($x_i, y_i, i = 1, \ldots, n$) will be included into the notations where needed.
Let us assume that the utility $u_{n+1}(x)$ of computation of the current objective function value at the point $x$ depends on $x$ via $m(x)$ and $s(x)$. A desired target value $y^{'on}$, $y^{'on} < \min_{1 \leq i \leq n} y_i$, is also assumed as a parameter which defines $u_{n+1}(x)$:

$$u_{n+1}(x) = U(m(x), s(x), y^{'on}),$$  \hspace{1cm} (5)$$

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

$$U(m(x) + c, s(x), y^{on} + c) = U(m(x), s(x), y^{on}),$$
$$U(m(x) \cdot C, s(x) \cdot C, y^{on} \cdot C) = U(m(x), s(x), y^{on}), \quad C > 0.$$
Since the utility function is used for the construction of a minimization algorithm, a smaller function value obtained at the current optimization step is considered more desirable than a larger one; therefore we postulate that

\[ m < \mu \text{ implies } U(m, s, y) > U(\mu, s, y). \]  

(6)
The analysis above suggests the construction of a global optimization algorithm, the \((n + 1)\)st iteration of which is defined by the solution of the following maximization problem:

\[
x_{n+1} = \max_{x \in A} P \left( \frac{y^o - m(x|x_i, y_i, i = 1, \ldots, n)}{s(x|x_i, y_i, i = 1, \ldots, n)} \right).
\]  

(7)
Implementation
We consider the planning of $n + 1$-th computing of the objective function value. The feasible region is partitioned using the Delaunay triangulation. The optimal triangle is selected, where optimality was defined by the assumptions of the rationality of search. The value of the objective function is computed at the center of the selected triangle.
Some results needed for the further development are presented in A. Žilinskas, G. Gimbutiene *On an Asymptotic Property of a Simplicial Statistical Model of Global Optimization*

\[ \rho_i = \frac{y_{on} - \hat{Y}_i - \varepsilon_n}{|T_i|}, \]  

(8)
The main theorem basically says that for large enough \( n \), we have

\[
\Delta_n \leq \exp(-\sqrt{nC}),
\]

where \( C \) depends on the condition number for \( f(x) \) at global minimum point, but not on \( n \). Thus the convergence rate is exponential in \( n \), roughly speaking.
Examples
• A quadratic function

\[ f(x_1, x_2) = (x_1 + x_2 - 1)^2 + 10(x_1 - x_2)^2, \quad -1 \leq x_1, x_2 \leq 1. \]

• The Rastrigin function

\[ f(x_1, x_2) = x_1^2 + x_2^2 - \cos(18x_1) - \cos(18x_2), \quad -2 \leq x_1, x_2 \leq 2. \]

• The Branin function

\[ f(x_1, x_2) = \left( x_2 - \frac{5.1}{4\pi^2} x_1^2 + 5\frac{x_1}{\pi} - 6 \right)^2 + 10 \left( 1 - \frac{1}{8\pi} \right) \cos(x_1) + 10, \]
\[ -5 \leq x_1 \leq 10, 0 \leq x_2 \leq 15. \]

• The six-hump camel back function

\[ f(x_1, x_2) = (4 - 2.1x_1^2 + x_1^4/3) x_1^2 + x_1 x_2 + (-4 + 4x_2^2) x_2^2, \]
\[ -3 \leq x_1 \leq 3, \quad -2 \leq x_2 \leq 2. \]
To exclude the influence of the coincidence of specific properties of the algorithm and a test problem each problem has been optimized 100 times with bounds perturbed by random numbers uniformly distributed in the interval of 10% of the length of the corresponding ranges of variables.
The dependence of the averaged errors (in the base-10 logarithmic scale, $\Delta_n = \log_{10}\{(y^{on} - f_{min})\}$) on the number of function evaluations performed is presented in the next slide. For comparison these test functions have also been minimized by the well-known global optimization code DIRECT, and by the simplicial partition algorithm based on a statistical model.
Figure 1: Comparison of normalized errors for different test functions; the quadratic function (top left), the Rastrigin function (top right), the Branin function (lower left), and the camel back function (lower right).
The figures corroborate the high convergence rate of the proposed algorithm predicted by the theoretical analysis.
Multi-objective global optimization
Nonlinear multi-objective optimization is very active research area. Depending on the properties of a multi-objective optimization problem, different approaches to its solution can be applied. The best direction developed is optimization of convex problems; for the latter problems the methods that generalize the ideas of classical mathematical programming suit well. For the problems with the objectives not satisfying the assumption of convexity, metaheuristic methods are frequently favorable. However, there remains a class of important problems without sufficient attention of researchers: the problems with black-box, multimodal, and expensive objectives.
Black box multi objective optimization

- Mathematically the problem of optimization is formulated as

\[ f^* = \min_{x \in A} f(x), \]

where \( f(x) \) is a nonlinear objective function of continuous variables \( f : \mathbb{R}^n \to \mathbb{R}^m, \ A \subset \mathbb{R}^n \). Pareto optimal solutions should be found.

- Frequently the objective functions of practical problems are available as computer programs, and the properties of the objective function are difficult to elicit. We assume that the objective function values are given by a ”black box”.
Figure 3  Mapping a decision variable from $\mathbb{R}^d$ into $\mathbb{R}^n$
The mathematical model of the weighted sum method takes the form of:

$$\min f(x), \quad f(x) = \sum_{i=1}^{m} \omega_i f_i(x),$$  \hspace{1cm} (10)$$

where $\omega_i$ is the weight of $i$-th criterion,

$0 \leq \omega_i \leq 1, \ i = 1, \ldots, m,$ and $\sum_{i=1}^{m} \omega_i = 1.$
Weighted sum scalarization theoretically is validated for convex problems. The other versions of scalarization, e.g., lexicographic goal programming algorithm are developed to generate solutions in the Pareto set of non-convex problems. For the problems with the objectives not satisfying the assumption of convexity, metaheuristic methods are frequently favorable.
**Theorem.** The worst-case optimal (passive/adaptive) multi-objective optimization algorithm selects the points $x_1, \ldots, x_n$, at the centers of $n$ balls of minimum radius which cover the feasible region $A$. The minimum worst-case error is equal to the radius $r$ of the balls of the optimal cover. The worst-case objective functions $\varphi(\cdot)$ and $\phi(\cdot)$ are defined by the following formulae:

\[
\varphi_k(x) = c_k, \ x \in A, \ k = 1, \ldots, m,
\]

\[
\phi_k(x) = \max_{1 \leq i \leq n} \phi_k(x_i) - (\|x - x_i\|), \ k = 1, \ldots, m.
\]
A surprising result is that the components of the *worst-case* vector objective function coincide up to linear scaling. Such a coincidence means that the *worst-case* model of multi-objective global optimization degrades to the single objective model.
Similarly to the case of single-objective optimization we assume that the utility of choice of the point for the current computation of the vector value $f(x)$ has the following structure

$$u_{n+1}(x) = U(m(x), \Sigma(x), y^{on}),$$

(11)

where $m(x) = (m_1(x), \ldots, m_m(x))^T$, and $y^{on}$ denotes a vector desired to improve.
At the current optimization step a point for computing the value of $f(x)$ is sought by an optimization algorithm which maximizes $u_{n+1}(x)$ which should be invariant with respect to the scales of data. Such a rationality assumption can be expresses by the following properties of $U(\cdot)$:

\[
U(m(x) + c, \Sigma(x), y^{on} + c) = U(m(x), \Sigma(x), y^{on}), \quad c = (c_1, \ldots, c_m)^T,
\]

\[
U(C \cdot m(x), C \cdot \Sigma(x), C \cdot y^{on}) = U(m(x), \Sigma(x), y^{on}), \quad C_i > 0,
\]

\[
C = \begin{pmatrix}
C_1 & 0 & \ldots & 0 \\
0 & \ldots & \ldots & \ldots \\
0 & 0 & \ldots & C_m
\end{pmatrix}.
\]
Theorem A function which satisfies assumptions (12) has the following structure

$$U(m(x), \Sigma(x), y_1^{on}) = \pi \left( \frac{y_1^{on} - m_1(x)}{s_1(x)}, \ldots, \frac{y_m^{on} - m_m(x)}{s_m(x)} \right).$$

(11) If moreover, assumption of monotonicity is satisfied then $\pi(\cdot)$ is an increasing function of all its variables.
As an example, the bi-objective $\pi$-algorithm has been implemented. A product of two arctangents was used for $\pi(\cdot)$. Then the $n + 1$ step of the $\pi$-algorithm is defined as the following optimization problem

$$x_{n+1} = \arg\max_{x \in A} \left( \arctan\left( \frac{y_{1n} - m_1(x)}{s_1(x)} \right) + \frac{\pi}{2} \right) \cdot \left( \arctan\left( \frac{y_{2n} - m_2(x)}{s_2(x)} \right) + \frac{\pi}{2} \right),$$

(11)
Experimental results
Test problem proposed by Schaffer (1984)

\[ f_1(x) = \begin{cases} 
-x, & \text{if } x \leq 1, \\
 x - 2, & \text{if } 1 < x \leq 3, \\
 4 - x, & \text{if } 3 < x \leq 4, \\
 x - 4, & \text{if } x > 4,
\end{cases} \]

\[ f_2(x) = (x - 5)^2, \quad -1 \leq x \leq 8. \]
The feasible objective region of the Schaffer problem
The points generated by RUS
The points generated by P-algorithm
The vector objective function is composed of the Rastrigin function

\[ f_1(x) = (x + 0.5)^2 - \cos(18(x + 0.5)), \]
\[ f_2(x) = (x - 0.5)^2 - \cos(18(x - 0.5)), \quad -1 \leq x \leq 1. \]

Rastrigin function is frequently used for testing global optimization algorithms.
The feasible objective region of the Rastrigin problem, and the points generated by MC
Rastrigin problem. The points generated by the P-algorithm and the $\pi$-algorithm.
Table 1: Performance criteria of the P-algorithm and the RUS method for problems Rastrigin, and Schaffer

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>P-algorithm</th>
<th>RUS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Problem Ras</td>
<td>Problem SCH</td>
</tr>
<tr>
<td>NP</td>
<td>31.4</td>
<td>41.9</td>
</tr>
<tr>
<td>NN</td>
<td>34.3</td>
<td>43.5</td>
</tr>
<tr>
<td>MD</td>
<td>0.095</td>
<td>0.099</td>
</tr>
<tr>
<td>DP</td>
<td>0.038</td>
<td>0.013</td>
</tr>
</tbody>
</table>
The vector objective function is composed of two Shekel functions

\[ f(x) = (f_1(x), f_2(x))^T, \quad x = (x_1, x_2)^T, \quad 0 \leq x_1 \leq 1, \quad 0 \leq x_2 \leq 1, \]

\[ f_1(x) = -\frac{1}{0.1 + (x_1 - 0.1)^2 + 2(x_2 - 0.1)^2} - \frac{1}{0.14 + 20(x_1 - 0.45)^2 + (x_2 - 0.55)^2}, \]

\[ f_2(x) = -\frac{1}{0.15 + 40(x_1 - 0.55)^2 + (x_2 - 0.45)^2} - \frac{1}{0.1 + (x_1 - 0.3)^2 + (x_2 - 0.95)^2}. \]

Such test functions are frequently used for testing global optimization algorithms.
Contours of the objective functions
Feasible objective region (Shekel), and points generated by MC
Results of search by $\pi$-algorithm $z_\circ = (-9, -9)$
Results of search by $\pi$-algorithm $z_o = (-6, -6)$
Test problem proposed by Fonseca and Fleming (1995)

\[
f_1(X) = 1 - \exp\left(-\sum_{i=1}^{d} (x_i - 1/\sqrt{d})^2\right),
\]

\[
f_2(X) = 1 - \exp\left(-\sum_{i=1}^{d} (x_i + 1/\sqrt{d})^2\right),
\]

\[-4 \leq x_i \leq 4, \ i = 1, \ldots, d.\]
Figure 2: The feasible region of problem Fon.
Figure 3: The points generated by the RUS in the objective feasible region of the problem Fon.
Figure 4: The points generated by the $\pi$-algorithm in the objective feasible region of problem Fon.
Open problems
Choice of a stochastic function adequate to the available information on the problem in question.

Estimation of parameters of the selected (Gaussian) stochastic function.

Difficulties in numerical computations (ill defined covariance matrices).

Unexpected behavior of one-step Bayesian algorithm.
Figure 5: The points generated by the one-step Bayesian algorithm for true (Wiener process) model.
Figure 6: The points generated by the one-step Bayesian algorithm for true (stationary process) model.
Absence of well substantiated testing methodology of algorithms aimed at expensive problems.