Noname manuscript No.
(will be inserted by the editor)

On Global Bayesian Optimization and Paralell Computing

Jonas Mockus

Received: date / Accepted: date

Abstract New versions of global optimization using the Bayesian Approach (BA) are described.
The first one called ”Exkor” is some method of multi-dimensional optimization by applying a sequence of one-dimensional global optimizers starting from the best points obtained by previous one-dimensional optimizers. The new element is that observation points are defined by explicit formulas.The results depend on the initial point.
Therefore, in parallel computing, uniformely distributed random initial points are generated and the best final result is accepted.
The second version is multi-dimensional where the sequence of observation points is generated in some hyper-recangule by minimization the risk function.
The minimization of risk functions is performed comparing S randomly generated points. Thus, the parameter S controls the relation of auxiliary and main calculations.
In the both versions, The globality of search is controled by the parameter ε.

Institute of Mathematics and Informatics
Tel.: +3705 2109338
E-mail: jmockus@gmail.com

1 Outline

The traditional numerical analysis considers optimization algorithms that guarantee some accuracy for all functions to be optimized. This includes the exact algorithms. That is the worst case analysis. To limit maximal errors one needs computational efforts that often increase exponentially with the size of the problem. This is the main disadvantage of the worst case analysis.

The alternative is the average case analysis. Here an average error is not limited but is made as small as possible. The average is taken over a set of functions to be optimized. The average case analysis is the Bayesian Approach (BA) [2,5].

The Bayesian Approach (BA) is defined by fixing a prior distribution P on a set of functions $f(x)$ and by minimizing the Bayesian risk function [1,5]. The risk function $R_0(x)$ is the expected deviation from the global minimum at a fixed point x. The distribution P is considered as a stochastic model of $f(x),\ x \in R^m$, where $f(x)$ might be a deterministic or a stochastic function. In the Gaussian case, assuming [5] that the $(n+1)$th observation is the last one

$$R_0(x) = \frac{1}{\sqrt{2}\pi s_n(x)} \times \int_{-\infty}^{+\infty} \min(c_n, z) e^{-\frac{1}{2}\left(\frac{y-m_n(x)}{s_n(x)}\right)^2} dz. \quad (1)$$

Here $c_n = \min_i z_i - \varepsilon,\ z_i = f(x_i)$. $m_n(x)$ is a conditional expectation with respect to observed values $z_i,\ i = 1,\ldots,n$. $s_n^2(x)$ is a conditional variance, and $\varepsilon > 0$ is a correction parameter. The minimum of the risk function $R_0(x)$ is obtained [5] at the point

$$x_{n+1} = \arg\max_x \frac{s_n(x)}{m_n(x) - c_n}. \quad (2)$$

The objective of BA, used mainly in continuous cases, is to provide as small average error as possible while keeping convergence conditions.

2 Bayesian Approach (BA)

The Wiener process is a common [4,8,9] stochastic model in the one-dimensional case $m = 1$.

Figure 1 shows the conditional expectation, the conditional standard, and the risk function with respect to available evaluations. The Wiener model implies continuity of almost all sample functions $f(x)$. The model assumes that increments $f(x_4) - f(x_3)$ and $f(x_2) - f(x_1)$, $x_1 < x_2 < x_3 < x_4$, are stochastically independent. Here $f(x)$ is Gaussian $(0, \sigma x)$ at any fixed $x > 0$. Note, that the Wiener process originally provided a mathematical model of a particle in the Brownian motion.

The Wiener model is extended to multidimensional case, too [5]. However, simplified stochastic models are preferable if $m > 1$. They can be preferable in one-dimensional cases, too, since the Wiener model is just an approximation of actual functions $f(x)$. The simplified models are designed by replacing the traditional Kolmogorov consistency conditions. The conistency conditions require the inversion of matrices of nth order for computing the conditional expectation $m_n(x)$ and variance $s_n^2(x)$ [1].

[1] The favorable exceptions are the Markovian processes including the Wiener one. Extending the Wiener process to $m > 1$ the Markovian property disappears.

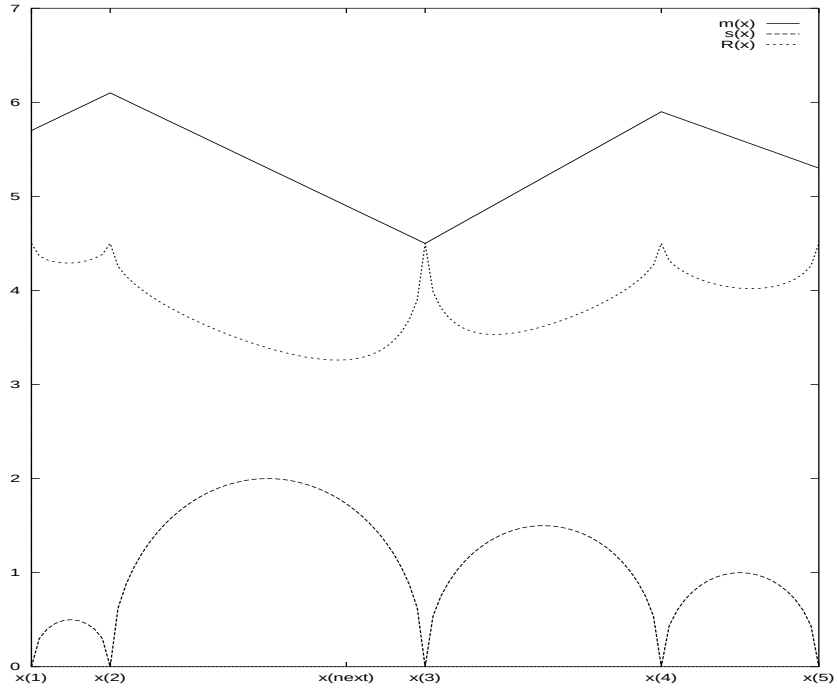


Fig. 1 The Wiener model. The conditional expectation $m(x)$, the conditional standard $s(x)$, and the risk function $R(x)$ with respect to observed values $x(1), y(1), x(2), y(2), ...$

Replacing the regular consistency conditions by:
- continuity of the risk function $R(x)$
- convergence of x_n to the global minimum
- simplicity of expressions of $m_n(x)$ and $s_n(x)$

the following simple expression of $R(x)$ is obtained using the results of [5]

$$R(x) = \min_{1 \le i \le n} z_i - \min_{1 \le i \le n} \frac{\|x - x_i\|^2}{z_i - c_n}. \tag{3}$$

The minimum of the risk function $R(x)$ is obtained at the point

$$x_{n+1} = \arg\max_{x} \min_{1 \le i \le n} \frac{||x - x_i||^2}{z_i - c_n}. \tag{4}$$

The aim of BA is to reduce the expected deviation. In addition, BA has some good asymptotic properties, too. It is shown [5] that

$$d^*/d_a = \left(\frac{f_a - f^* + \varepsilon}{\varepsilon} \right)^{1/2}, \; n \to \infty. \tag{5}$$

Here d^* is a density of points x_i around the global optimum. d_a is an average density of x_i in the feasible area. f_a is an average value of $f(x)$ in this area. f^* is an average value of $f(x)$ around the global minimum. ε is the correction parameter in expression (1). That means that BA provides convergence to the global minimum for any continuous $f(x)$ and a greater density of observations x_i around the global optimum, if n is large.

Note, that the correction parameter ε has a similar influence as the temperature in simulated annealing. However, that is a superficial similarity. The good asymptotic behavior is just some "by-product" of BA. The reason is that Bayesian decisions are applied for small size samples where asymptotic properties are not noticeable.
General expressions (3) and (4) are for K-dimensional functions, where $x=(x^k,\ k=1,...,K)$. In one-dimensional case there exists explicit solution (8) - (14). If $K>1$, then some numeric procedure should be used. The simplest one is the Monte Carlo search when the best point $x(s^*)$ of S random uniformly distributed points $x(s),\ s=1,...,S$ is selected by (25) - (26).

To choose the next point x_{n+1} by BA, one minimizes the expected deviation $R(x)$ from the global optimum (see Figure 1). The minimization of $R(x)$ is a complicated auxiliary optimization problem. That means that BA is useful for expensive, in terms of computing times, functions of a few ($m<20$) continuous variables. This happens in wide variety of problems.

Some examples are described in [5]. These include maximization of the yield of differential amplifiers, optimization of mechanical systems of a shock absorber, optimization of composite laminates, evaluation of parameters of immunological models and nonlinear time series, planning of extremal experiments on thermostable polymeric compositions. A large set of test problems in global optimization is considered in [3].

3 One-dimensional Optimization

If $m=1$ then from (4) follows

$$R(x)=\min_{1\le i\le n} z_i-\min_{1\le i\le n-1}\frac{(x-x_i)^2}{z_i-c_n}. \tag{6}$$

The minimum of the risk function $R(x)$ is obtained at the point

$$x_{n+1}=\arg\max_{a\le x\le b}\min_{1\le i\le n-1}\frac{(x-x_i)^2}{z_i-c_n}. \tag{7}$$

Here the maximum is reached at some point $x=x_i^*,\ i=1,...,n-1$ satisfying conditions

$$\frac{(x-x_i)^2}{z_i-c_n}=\frac{(x-x_{i+1})^2}{z_{i+1}-c_n}, \tag{8}$$

$$x_i\le x\le x_{i+1},\ x_1=a,\ x_n=b,\ i=1,...,n-2. \tag{9}$$

This way condition (6) is reduced to

$$x_{n+1}=\arg\min_{1\le i\le n-1} R(x_i^*). \tag{10}$$

Here

$$R(x_i^*)=\min_{1\le i\le n-1} z_i-\frac{(x_i^*-x_i)^2}{z_i-c_n}. \tag{11}$$

Fig. 2 . illustrates the simple case when $n=5$ and $x_6=\min_i R(x_i^*),\ i=1,2,3,4$

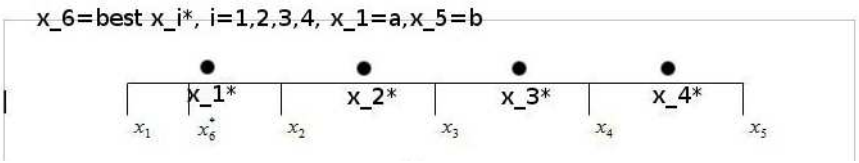


Fig. 2 Illustrates the observations $x_i, i = 1,2,3,4,5$ and the solutions $x_i, i = 1,2,3,4$.

The points satisfying (7) are defined by conditions

$$x_i^+ = \frac{(x2-x1)\sqrt{(v1*v2)} - v1*x2 + v2*x1}{v2-v1} \tag{12}$$

$$x_i^- = \frac{(x2-x1)\sqrt{(v1*v2)} + v1*x2 - v2*x1}{v2-v1} \tag{13}$$

Here $v1 = z_i - c_n,\ v2 = z_{i+1} - c_n,\ x1 = x_i,\ x2 = x_{I+1},\ i = 1,...,n-1$
The points satisfying both (7) and (8) are determined as follows

$$x_i^* = \begin{cases} x_i^+, & \text{if } x_i \le x_i^+ \le x_{i+1} \\ x_i^-, & \text{if } x_i \le x_i^- \le x_{i+1}, \end{cases} \tag{14}$$

4 Coordinate Optimization (CO)

The one-dimensional model is applied for coordinate optimization when $x = (x^j,\ j = 1,...,m)$
We start from some initial point $x = x(0)$ and apply one-dimensional search by this sequential procedure:

$$x(1) = arg \min_{a_1 \le x^1 \le b_1} f(x(0)) \tag{15}$$

$$x(2) = arg \min_{a_2 \le x^2 \le b_2} f(x(1)) \tag{16}$$

.....

$$x(m) = arg \min_{a_m \le x^m \le b_m} f(x(m-1)) \tag{17}$$

$$x(m+1) = arg \min_{a_1 \le x^1 \le b_1} f(x(m)) \tag{18}$$

.....

We stop when $m = M$ and denote the best point as $x* = x(M)$. Here results depend on initial points $x(0)$.
However the coordinate optimization is a convenient tool of visualization by one-dimensional projections. Coordinate optimization converges to global minimum if

$$f(x) = \sum_i f_i(x_i)\ or\ f(x) = \prod_i f_i(x_i) \tag{19}$$

Otherwise, it provides some approximation.

5 Global Coordinate Optimization (GCO)

We start from random initial points $x = x(0)_t,\ t = 1,...,T$ and apply one-dimensional search by repeating this this sequential procedure T times, preferably by different processors

$$x(1)_t = arg \min_{a_1 \le x^1 \le b_1} f(x(0)_r) \tag{20}$$

$$x(2)_t = arg \min_{a_2 \le x^2 \le b_2} f(x(1)_r) \tag{21}$$

.....

$$x(m)_t = arg \min_{a_m \le x^m \le b_m} f(x(m-1)_r) \tag{22}$$

$$x(m+1)_t = arg \min_{a_1 \le x^1 \le b_1} f(x(m)_r) \tag{23}$$

.....

Then we select the best point using results of all T processors (and all M iterations, if $f(x)$ is stochastic)

$$x^* = arg \min_{t=1,...,T,\ m=1,...,M} f(x(M_t)) \tag{24}$$

Here results converge to the global minimum of any continuous function when $T \to \infty$
GCO is intended for parallel computing by multi-processor systems. Different processors generate different starting points. The final result is obtained by the master processor using (24).

6 Parallel Bayesian Optimization (PBO)

Here different processors perform separate Bayesian optimization using (25) and (26).
To minimize the risk function (25) S random points $x(s)$ are generated and the best point $x(s^*)$ is selected to define the next observation point $x_{n+1} = x(s*),\ n = 1,...,M-1$. Therefore, different processors $t = 1,...,T$ produce different results denoted by index t.
Here

$$R(x(s)_t) = \min_{1 \le i \le n} z_{it} - \min_{1 \le i \le n} \frac{\|x(s)_t - x_i\|^2}{z_{it} - c_{nt}}. \tag{25}$$

The minimum of the risk function $R(x)$ is obtained at the point

$$x(s*)_{nt} = \arg\max_{x(s)_t} \min_{1 \le i \le n} \frac{||x(s)_t - x_i||^2}{z_{it} - c_{nt}}. \tag{26}$$

where

$$||x(s)_t - x_i||^2 = \Sigma_{k=1}^{K} (x^k(s)_t - x_i^k)^2. \tag{27}$$

Then the best point using results of all T processors and all M observations is selected by a master processor.

$$x^* = arg \min_{t=1,...T,\ n=1,...,M} f(x(s^*)_{nt}) \tag{28}$$

Here results converge to the global minimum of any continuous function when $T \to \infty$

7 Parameter selection

Globality of search is controlled by the parameter ε. Large ε defines global, nearly uniform search. Small ε performs most of the observations around local minima. The global convergence is provided at any $\varepsilon > 0$. However, since the convergence rate depend on ε, some experimentation with different ε could be useful.

The minimization of the risk function in one-dimensional case is performed by exact expressions (9) and (10). If $K > 1$ the accuracy of risk function minimization depends on the parameter S which controls the balance of computing resources between basic calculations defining values of function $f(x)$ and auxiliary calculations needed to minimize the risk function $R(x)$.

The simple rule is to select large S if $\tau_S < \tau_F$, otherwise select small S. Here τ_S is estimated time of risk minimization using parameter S, and τ_F is estimated time to calculate the function $f(x)$ at some fixed x.

The parameters ε and S are in the grafical users interface (GUI) for conv enient adaprtation to the problem to be solved.

References

1. M. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, New York, 1970.
2. P. Diaconis. Bayesian numerical analysis. In *Statistical Decision Theory and Related Topics*, pages 163–175. Springer Verlag, 1988.
3. C.A. Floudas, P.M. Pardalos, C.S. Adjiman, W.R. Esposito, us Z.H. Gu, S.T. Harding, J.L. Klepeis, C.A. Meyer, and C.A. Schweiger. *Handbook of test problems in local and global optimization*. Kluwer Academic Publishers, Dordrecht-Boston-London, 1999.
4. H.J. Kushner. A new method of locating the maximum point of an arbitrary multi-peak curve in the presence of noise. *J. of Basic Engineering*, 86:97–100, 1964.
5. J. Mockus. *Bayesian approach to global optimization*. Kluwer Academic Publishers, Dordrecht-London-Boston, 1989.
6. J. Mockus. *A Set of Examples of Global and Discrete Optimization: Application of Bayesian Heuristic Approach*. Kluwer Academic Publishers, Dordrecht-Boston-London., 2000.
7. J. Mockus, W. Eddy, A. Mockus, L. Mockus, and G. Reklaitis. *Bayesian Heuristic Approach to Discrete and Global Optimization*. Kluwer Academic Publishers, ISBN 0-7923-4327-1, Dordrecht-London-Boston, 1997.
8. V.R. Saltenis. On a method of multi-extremal optimization. *Automatics and Computers (Avtomatika i Vychislitelnayya Tekchnika)*, (3):33–38, 1971. (in Russian).
9. A. Torn and A. Zilinskas. *Global optimization*. Springer-Verlag, Berlin, 1989.