1981

A Bayesian based method for global optimization

Groch
Carnegie Mellon University

Luis M. Vidigal

Stephen W. Director

Follow this and additional works at: http://repository.cmu.edu/ece
NOTICE WARNING CONCERNING COPYRIGHT RESTRICTIONS:
The copyright law of the United States (title 17, U.S. Code) governs the making
of photocopies or other reproductions of copyrighted material. Any copying of this
document without permission of its author may be prohibited by law.
A BAYESIAN BASED METHOD FOR GLOBAL OPTIMIZATION

by

A. Groch, L. M. Vidigal, and S. W. Director

DRC-18-33-81

September 1981
A Bayesian Based Method for Global Optimization

A. Groch, L.M. Vidigal, and S.W. Director

Department of Electrical Engineering
Carnegie-Mellon University
Pittsburgh, Pa. 15213

Abstract

In this paper a new method for global optimization is presented. The method based on a one-step one-dimensional Bayesian technique seems to be very efficient, as a neighbourhood of the global optimum is attained in a relatively small number of function evaluations. The method is then extended to higher dimensional cases. A number of numerical examples is presented to illustrate the behaviour of the algorithm.

1. Introduction

In electronic circuit optimization it is often desirable to determine the global minimum (maximum) of some multieextremal function which characterizes the performance of the circuit. To this juncture designers had to settle for results which were obtainable using local optimization techniques. However, in the past several years methods for global optimization have been developed [1]. These methods differ from each other with regard to accuracy and accompanying computational effort. To be useful for circuit optimization it is important that the method not require too many function evaluations since each function evaluation entails a computationally expensive circuit simulation. Because of this need the recently developed probabilistic methods, and mainly Bayesian methods, which involve a relatively small number of function evaluations seem most appropriate. Unfortunately, Bayesian methods in their original form require the solution of complex systems of recurrent optimization equations expressing conditional probabilities [3], and even the simplest one-dimensional case involves a significant amount of computational overhead.

In this paper an effort to simplify the one-dimensional one-stage Bayesian method is presented. The approach used for solution of the one-dimensional problem is then generalized to second and higher dimensional cases. However, for this simple and compact algorithm only purely deterministic interpretation is possible. The method proposed quickly determines the sequence of so called observation points and locates a neighbourhood of the global optimum. We believe, based on these results, that this simplified method warrants further investigation.

2. One-Dimensional One-Stage Bayesian Method

We begin our discussion of Bayesian methods by considering a real stochastic process \( f(x, w), x \in \mathbb{R}, w \in Q \) where \( Q \) is a set of random events. This process is statistically determined by its \( k \)th order distribution functions

\[
F_{\mathbb{R}}^{x_{1},...x_{k}}(y_{1},...y_{k}) = P(f(x_{1},wK_{y_{1}}),...,f(x_{k},wK_{y_{k}}))
\]

for any value of \( k \) and for any set of points \( x_{1},...,x_{k} \). Here \( P(f(x_{1},wK_{y_{1}}),...,f(x_{k},wK_{y_{k}})) \) denotes the probability of an event

\[
(f(x_{1},wK_{y_{1}}),...,f(x_{k},wK_{y_{k}}))
\]

The stochastic function \( f(x, w) \) may be considered to be a family of functions, one for each value of \( w \). We assume that the function to be minimized \( f(x) \) is a particular element of that family, i.e. \( f(x) = f(x, w_{1}) \). Moreover, we will assume later an a priori distribution function (1) of the stochastic process.

Assume a sequence of \( k \) observation points \( x_{1},...,x_{k} \) and corresponding function evaluations \( f(x_{1}),...,f(x_{k}) \), called observations. Then the information known about \( f(x) \) can be summarized by the vector \( z_{k} \)

\[
z_{k} = (f(x_{1}),...,f(x_{k}))
\]

The probabilistic method for seeking the minimum of \( f(x) \) is called Bayesian if given the a priori decision of making a total of \( N \) observations, the next observation point \( x_{N+1} \) is chosen so that it minimizes the expected deviation of \( f(x_{N+1}) \) from the minimum \( f^{*} = \min f(x) \). Determining \( x_{N+1} \) requires the solution of a complex system of \( (N+1) \) recurrent equations containing conditional probabilities with respect to \( z_{k} \). \( k = 1,...,N \). To avoid this computational cost a simplified form of the algorithm called the "one-stage" method is often used [4]. In this method, solution of the system of recurrent equations is avoided by assuming that the next observation will be the last one.

Given \( k \) observation points, the one-stage Bayesian method chooses the next point \( x_{N+1} \) to be the solution of the problem

\[
x_{N+1} = \arg \min f(x)
\]
\[ \max \phi_k(x) = f(x^{opt_k}) - E\left( \left\{ f(x^{opt_i}) \cdot f(x^{*}) \right\} | x^k \right) \] (4)

where \( E \) denotes expected value.

Unfortunately, as it will be shown later, the computational overhead of the one-stage method is still significant, even for the one-dimensional problem. The convergence and the efficiency of the computation (4) strongly depends on the selection of a priori distribution function (1). Among the possible choices, the most convenient for the one-dimensional case has proven to be a Wiener process. Features of the Wiener process such as the normal distribution of every increment \( f(x \cdot Ax_i) - f(x,u) \), and the independence of these increments for every pair of not overlapping subintervals of \( A \), make the computation of conditional expectations (see eq.(4)) relatively simple.

Notice that in (4) the function \( \phi_k(x) \) denotes the average improvement towards optimum. Considering a one-dimensional case and assuming that \( f(x,u) \) represents a Wiener process, it can be shown [5] that

\[ \psi_k(x) = E \left( \frac{f(x^{opt_k})}{\sum \psi_k(x)} \left( \frac{1}{2\pi} \right)^{1/2} \exp \left( -\frac{x^2}{2} \right) \right) du \] (5)

where

\[ \psi_k(x) = E(f(x)|x^k) \quad \varphi_k(x) = \text{Var}(f(x)|x^k) \]

\[ m_k(x) = 0 \quad \alpha_k(x) = a^2 x \quad f(x^k,u) = 0 \]

It is convenient to order the observation points in a monotonically increasing sequence. Assuming \( A \) is the unit interval \([0,1]\) we will represent this sequence of points, at the \( k \)-th stage, by

\[ x_0 \leq x_1 \leq \cdots \leq x^* = x_{k+1} \] (6)

(Notice that \( x^* \) of eq.(3) does not correspond to \( x^{**} \) here.)

Using well known properties of the Wiener process [2], the conditional expected value and variance in each subinterval can be written

\[ m_k(x) = x_0 \leq x_1 \leq \cdots \leq x^k \leq x^* \leq x_{k+1} \]

\[ \sigma_k(x) = a \left( \frac{x^k - x^{k+1}}{x^k - x^*} \right) \]

where \( a \) is a parameter characteristic of the Wiener process whose value must be estimated for each subinterval. For this purpose an unbiased maximum likelihood estimator is used [4]. Estimation of \( a \) requires \( M \) initial, arbitrarily chosen, observations (e.g. \( M = 6 \)).

To determine the next observation point \( x^{k+1} \) the maximum values of \( \theta_{k+1}(x) \) in each subinterval have to be found and then compared. This task is made easier by the following properties of \( \theta_{k+1}(x) \):

1. It is unimodal in each subinterval
2. It is a decreasing function of \( a \) and a nondecreasing function of the subinterval length
3. It is a nonincreasing function of the difference \( m_k(x) - f(x^{opt_k}) \).

In any one-stage Bayesian method [5], the \( (k+1)^{th} \) approach to the optimum \( x^{opt_k} \), denoted as \( x^{opt_{k+1}} \), is the point at which the conditional expectation \( E(f(x)|x^{k+1}) \) is minimized. This is a piecewise linear function of \( x \) (see eq.(7)), so \( x^{opt_{k+1}} \) is the point at which \( f(x^{opt_{k+1}}) \), \( i = 0,1,...,k+1 \) is of the lowest value. Although the proof of convergence to the global optimum requires in the limit that \( k \to \infty \) [5], usually no more than 10 to 20 observations are needed to obtain a point in the neighbourhood of the minimum (at least local), even for highly oscillatory functions as will be shown in the examples.

The efficiency of the algorithm depends on a constant \( c \geq 1 \), which is associated to the variance of the Wiener process. For large values of \( c \) the method becomes more "global" but usually more observations are needed to locate the neighbourhood of the global minimum. In Fig.1 the typical shape of \( \phi_k(x) \) is shown. Note that if \( N \) denotes the total number of observations , for small values of \( c \) we increase the risk that the final solution may correspond to a local minimum only.

### 3. Simplified Algorithm

A careful study of the behaviour of \( \phi_k(x) \) suggests a method to simplify the algorithm described by equation (5). Instead of considering the function \( \phi_k(x) \), a new function \( w_k(x) \) is introduced:

\[ w_k(x) = \text{var}(f(x)|x^k) \] (8)

Observe that for \( c \leq 2 \), we can say with 97.7% confidence, that for any point \( x \), \( f(x) \geq w_k(x) \), and therefore it is reasonable to search for the minimum of \( f(x) \) at the points where \( w_k(x) \) is minimal. Further, notice that the function \( \phi_k(x) \) has all the properties listed for \( w_k(x) \) in Section 2.

To determine the minimum value of \( w_k(x) \) on the interval \([0,1]\) we follow a similar procedure, namely we find the minima of \( w_k(x) \) in all subintervals defined by (6) and select the one that corresponds to the minimum value of \( w_k(x) \). The computational effort associated with this procedure is clearly much smaller than the work required to evaluate \( \theta_{k+1}(x) \) as the integral in (5) (error function) can not be evaluated analytically, and some numerical technique must be used.

In all numerical examples (see Section 5) the results obtained, using (5) and (8) are in close agreement. For example the distances between the points determined by both algorithms do not exceed 1.5% of the subinterval length (see Table 1 in Section 5). The differences of 20% relate only to the subintervals on which \( m_k(x) \) is much greater than \( f(x^k,u) \), however the chance of selecting \( x^{opt_k} \) from such subintervals is rather small. The simplified algorithm (8), has proven to be a convenient and reliable tool in all examples considered. Using (8) in place of (5) decreases CPU time by about a factor of 50 while maintaining almost the same quality as the original Bayesian algorithm.

Unfortunately the simplified algorithm (8) introduces the risk of missing the global minimum even if \( k \to \infty \). This will happen if the minimum of \( w_k(x) \) in some subinterval is greater than \( f(x^{opt_k}) \). While this risk will decrease for larger value of \( c \), increasing \( c \) usually will increase the number of observations required.

Despite the similarities of both algorithms, due to the lack of appropriate mathematical proof for their equivalence, the algorithm (8) has to be considered separately. Our simplified approach is in fact a new deterministic way of subinterval division and selection of the next observation point. We can make a physical interpretation of this procedure. Specifically, the selected value \( x^{k+1} \) is the point
Section 5.1 example i;k«5 of "the lowest potential energy" between successive pairs of observations \( f(x^{i-1}) \) and \( f(x^i) \) (see Fig 1).

4. Extension to Higher Dimensions

Because of the dose agreement between both algorithms we are motivated to extend the method proposed in Section 3. For higher dimensions we replace the interval division defined by (6), by a simplicial division of the region \( A \). Initially some points are selected and used to partition this region. As a new point is determined, a new simplicial division can be found.

In the one-dimensional case \( m(x) \) is a piecewise linear function connecting successive \( f(x^i) \), \( i=0,1,...,k \). In the \( n \)-dimensional case we assume that \( x \in \mathbb{R}^n \), for each simplex a hyperplane joining the vertices (observation values). Similarly the quadratic equation for \( a_i(x) \) is generalized to a \( n \)-dimensional quadratic surface for each simplex. The coordinates of the optimal point \( x_{0j}^* \) for any simplex \( s \) are given by

\[
x_{0j}^* = x_{j}^* - \sum_{i=1}^{n} a_i \cdot \left( \begin{array}{c}
x_i^* - x_j^* - x_{ij}^* \\
\end{array} \right)
\]

where

- \( x_j^* \) - coordinate of the center of hypersphere described on considered simplex \( s \)
- \( r^* \) - radius of this hypersphere
- \( a_i \) - coefficient in the expression for \( m_{ij}^*(x) \) (10).

\[
m_{ij}^*(x) = \sum_{i=1}^{n} a_i \cdot x_i^* \\
\]

As in the one-dimensional case, \( m_j(x) \) contains all the observation values calculated for vertices of a given simplex. Similarly \( a_{ij}(x) \) becomes zero at the observation points. Using (10), the algorithm (9) maintains all the properties mentioned in Section 3 regardless of the dimension \( n \). It may happen that the optimum \( x_{k+1}^* \) obtained from (9) will lie outside the subregion under consideration. Then a solution in an adjacent face of this subregion is sought.

As in the one-dimensional case the next observation point \( x_{k+1}^* \) is the point of "the lowest potential energy" among all the points \( x_{k+1}^* \) (one for each simplicial division), i.e., \( x_{k+1}^* \) is the point at which "\( k^{th} \) is minimum in whole hypercube \([0,1]^n\)".

This algorithm was used with several two-dimensional functions. The results as discussed in the next Section seem very promising. The number of observations required increases only slightly in comparison with one-dimensional case.

As a final note, the problem of estimating \( a \) is stressed again. To be consistent we have used classical formula for the unbiased maximum likelihood estimator [4]. The meaning of the constant \( c \) remains unchanged in higher dimensions.

<table>
<thead>
<tr>
<th>Function f(x)</th>
<th>( e )</th>
<th>Bayesian Method</th>
<th>Simplified Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>f(x)</td>
<td>1</td>
<td>.7789</td>
<td>.7624</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>.7570</td>
<td>.7599</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>.7430</td>
<td>.7437</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>.7227</td>
<td>.7204</td>
</tr>
<tr>
<td>f(x) Oaxsi</td>
<td>1</td>
<td>3680</td>
<td>3720</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3461</td>
<td>3461</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3336</td>
<td>3327</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>3164</td>
<td>3147</td>
</tr>
<tr>
<td>f(x) Oaffi</td>
<td>1</td>
<td>6316</td>
<td>6451</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6364</td>
<td>6366</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>7334</td>
<td>7346</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>7.609</td>
<td>7356</td>
</tr>
</tbody>
</table>

Table 1: Comparison of \( x^{k+1} \) points determined by the Bayesian method and the simplified method.
5. Numerical Examples

5.1 One-dimensional Case

Three different one-dimensional multiextremal functions were considered:

1. \( f_1(x) = 2(x-0.75)^2 \cdot \sin(5wx-0.4x) - 0.125 \)
   
   with global minimum at \( x = 0.7795 \) and \( f_1 = -1.1232 \)

2. \( f_2(x) = 2(x-0.75)^2 \cdot \sin(8wx-0.5x) - 0.125 \)
   
   with global minimum at \( x = 0.17 \) with \( f_2 = -1.1232 \) and two
   significant local minima at \( x = 0.75 \) where \( f(x) = -1.125 \), and
   \( x = 0.99842 \) where \( f(x) = -1.0007867 \)

3. \( f_3(x) = 2x \cdot \sin(9wx-0.5x) - 0.125 \)

   with equal minimum at \( x = 0.1 \) with \( f_3 = -1.125 \) and two
   significant local minima at \( x = 0.75 \) where \( f(x) = -1.125 \), and
   \( x = 0.99842 \) where \( f(x) = -1.0007867 \)

An three functions were optimized by both methods. It was
decided a priori to fix the number of observations at 32. Table 2
shows the two best results for different values of c and appropriate
observation numbers at which those values were achieved.

5.2 Two-dimensional Case

Two different two-dimensional multiextremal functions were
considered:

1. Branin’s function.
   
   \( y(x,y) = a(y-bx^2 + c)^2 + d \cdot I(\cos(x) + 1) \)
   
   where \( a = 1 \), \( b = 5.1/8 \), \( c = 5/\pi \), \( d = 6 \). \( I = 1/8 \)
   
   \( -5 < x < 10 \), \( 0 < y < 15 \)

   This function has three equal global minima \( f_{\text{min}} = 0.397887 \)
   at points \((-3.14159, 1.275), (3.14159, 1.275), (9.42478, 1.275)\).

2. Goldstein’s and Price’s function.

   \( \sum (x - y)^2 + 10(1 - x)^2 \cdot \sin(x) \cdot \sin(y) \)

   This function has global minimum \( f_{\text{min}} = 3 \) at point \((0,0,0, -1.000)\)
   and local minima at below:

   \[
   \begin{array}{cccc}
   x & y & K & f(x,y) \\
   \hline
   -0.60 & -0.60 & 600 & 600 \\
   -0.60 & -0.60 & 600 & 600 \\
   -0.60 & -0.60 & 600 & 600 \\
   \end{array}
   \]
Those functions were optimized by algorithm (8) and fixed number of 36 observations was assumed in advance. Two best results for different values of \( c \) are shown in Table 3.

### Table 3: Results for the optimization of functions of two variables

<table>
<thead>
<tr>
<th>Function</th>
<th>( c )</th>
<th>( x )</th>
<th>( y )</th>
<th>( f(x,y) )</th>
<th>obs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f(x,y) )</td>
<td>1</td>
<td>9.4179</td>
<td>2.2902</td>
<td>4.301</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>9.3651</td>
<td>2.7011</td>
<td>4.811</td>
<td>18</td>
</tr>
<tr>
<td>(-5x+10)</td>
<td>3</td>
<td>9.3230</td>
<td>2.0587</td>
<td>5.573</td>
<td>36</td>
</tr>
<tr>
<td>( G = 119.396 )</td>
<td>7</td>
<td>-3.2242</td>
<td>13.2242</td>
<td>9.928</td>
<td>36</td>
</tr>
<tr>
<td>( f(x,y) )</td>
<td>1</td>
<td>.1303</td>
<td>.1098</td>
<td>14.782</td>
<td>24</td>
</tr>
<tr>
<td>( \frac{G}{2} )</td>
<td>2</td>
<td>.1703</td>
<td>.9254</td>
<td>10.132</td>
<td>17</td>
</tr>
<tr>
<td>( 3x^2 )</td>
<td>3</td>
<td>.1601</td>
<td>.9070</td>
<td>11.159</td>
<td>18</td>
</tr>
<tr>
<td>( 3y^2 )</td>
<td>7</td>
<td>.1512</td>
<td>.9381</td>
<td>8.623</td>
<td>18</td>
</tr>
<tr>
<td>( G = 444434 )</td>
<td>7</td>
<td>.1593</td>
<td>.9103</td>
<td>8.674</td>
<td>17</td>
</tr>
</tbody>
</table>

6. Conclusions

A new method for seeking the global minimum, which is based on the probabilistic one-dimensional one-stage Bayesian method, has been developed. This method has features similar to the original one but is much simpler and faster and the new formulation may be easily extended to higher dimensions.

Experimental results confirm the efficiency of this method for global optimization, as the number of function evaluations required grows only slightly with the dimensionality. While the method presented here appears promising for use in global optimization, additional mathematical investigations are still needed.

### References

[1] Gomulka, J.
Towards Global Optimization. : Deterministic vs Probabilistic Approaches to Global Optimization.

A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise.

On Bayesian Methods to Seek the Extremum.

Towards Global Optimization. : The Application of Bayesian Methods for Seeking the Extremum.

The One-Step Bayesian Method for Seeking the Extremum of the Function of One Variable.
Cybernetics (Russian) 1:139-144, 1975.