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[1]

A.Groch[2],
L.M. Vidigal[3]
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-dimensionai 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 destreable to' determine the global minimum (maximum) of some muitiextremal 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 \n their original form require the solution of complex systems of recurrent 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 € A C R^n. *u* € Q where Q is a set of random events. This process is statistically determined by its k^m order distribution functions

$$F_{x^1,\dots,x^k}(y_1,\dots,y_k) = P\{f(x^1,\omega)<y_1,\dots,f(x^k,\omega)<y_k\} \quad (1)$$

for any value of k and for any set of points $x^1,\dots,x^k$, where $P\{f(x^1,uKy_r-..,f(x^k.uKy_k\}$ denotes the probability of an event

$$\{f(x^1_f«Ky_1.-..f(x^k_twKy_k\} \quad (2)$$

The stochastic function f(x,*>) may be considered to be a family of functions, one for each value of *u.* We assume that the function to be minimized f(x) is a particular element of that family, i.e. f(x)«ffcuj). Moreover, we will assume later an a priori distribution function (1) of the stochastic process.

Assume a sequence of k <u>observation points</u> x\...,x^k and corresponding function evaluations $f(x^1),\dots,f(x^k)$, called <u>observations</u>. Then the information known about f(x) can be summarized by the vector z^k

$$z^k * (f(x^1);\dots_i f(x^k)_t x^1,\dots_t x^k) \quad (3)$$

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 $x^{0***"*^1}$ is chosen so that it minimizes the expected -deviation of $fCx^{001,*^1*^1},^)$ from the minimum f^ » $ffx^{0*}) = \min_{x\in A} f(x)$. Determining $x^{optN>1}$ 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 always assuming that the next observation will be the last one.

Given k observation points, the one-stage Bayesian method chooses the next point x^{K*^1} to be the solution of the problem

[1] Th* w:m r c h was supported in part by the National Science Foundation under Grant ECS 79-23191

[2] *Qn* leave from Technical Urttversity off Gdansk. Poland, as Fulbright-Hays visiting scholar

[3] Dn leave from institute Superior Tecmco.L'cbon.Portugal

UNfVERSITY LIBRARIES
CARNE6fE-MELL0N UNIVERSITY
Pi n S B U R G K f SYLVAN 15213

$$\max \varphi_{k+1}(x) = f(x^{opt,k}) - E\{\min (f(x^{optk}).f(x.)) \mid z^k\} \quad (4)$$

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 • Ax,u>)-f(x,u), and the independence of these increments for every pair of not overlapping subintervais of A, make the computation of conditional expectations (see eq.(4)) relatively simple.

Notice that in (4) the function $\varphi_{k+1}(x)$ 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

$$\varphi_{k+1}(x) = c\,\sigma_k(x) \int_{-\infty}^{\frac{f(x^{optk}) - m_k(x)}{c\,\sigma_k(x)}} ((1/(2\pi))^{1/2} \int_{-\infty}^{u} e^{-t^2/2} dt)du \quad (5)$$

where

$$m_k(x) = E\{f(x)|z^k\} \qquad \sigma_k^2(x) = Var\{f(x)|z^k\}$$

$$m_0(x) = 0 \qquad \sigma_0^2(x) = \sigma^2 x \qquad f(x^{0,0}) = 0$$

It is convenient to order the observation points into a monotonicaUy increasing sequence. Assuming A is the unit interval [0,1] we wiH represent this sequence of points, at the k^{m} stage, by

$$0 = x^{0,k} \le x^{1,k} \le \ldots \le x^{k,k} = 1 \quad (6)$$

(Notice that x* of eq.(3) does not correspond to $x^{i,k}$ here.)

Using well known properties of the Wiener process [2], the conditional expected value and variance in each subintervai ry-i.k xi,k j (i m 1 2 . . . k) can be written

$$m_k(x) = \frac{f(x^{i,k})(x - x^{i-1,k}) + f(x^{i-1,k})(x^{i,k} - x)}{x^{i,k} - x^{i-1,k}}$$

and (7)

$$\sigma_k(x) = \sigma \left(\frac{(x - x^{i-1,k})(x^{i,k} - x)}{x^{i,k} - x^{i-1,k}}\right)^{1/2}$$

where a is a parameter characteristic of the Wiener process whose value must be estimated for each problem. For this purpose an unbiased maximum likelihood estimator is used [4]. Estimation of 9 requires M initial, arbitraly chosen, observations (e.g. M » 6).

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

1. it is unimodal in each subintervai
2. it is an increasing function of *a* and a nondecreasing function of the subinterval length
3. it is a nonincreasing function of the difference $(m_k(x)-f(x^{optk}))$.

In any one-stage Bayesian method [5], the (k + 1)* approach to the optimum x^{opt}. denoted as x^{k+1}, is the point at which the conditional expectation $E\{f(x)|z^{k+1}\}$ is minimzed. This is a piecewise linear function of x (see eq.(7)), so x^{optk+1} is the point at which $f(x^{i,k+1})$, (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 $c \ge 1$, 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 $\varphi_{k+1}(x)$ 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 $\varphi_{k+1}(x)$ suggests a method to simplify the algorithm described by equation (5). Instead of considering the function $\varphi_{k+1}(x)$ a new function $w_{k+1}(x)$ is introduced:

$$w_{k+1}(x) = m_k(x) - c\sigma_k(x) \quad (8)$$

Observe that for c = 2. we can say with 97.7% confidence, that for any point x. $f(x) > w_{k+1}(x)$, and therefore it is reasonable to search for the minimum of f(x) at the points where $w_{k+1}(x)$ is minimal. Further, notice that the function $(-w_{k+1}(x))$ has all the properties listed for for $\varphi_{k+1}(x)$ in Section 2.

To determine the minimum value of $w_{k+1}(x)$ on the interval [0,1] we follow a similar procedure, namely we find the minima of $w_{k+1}(x)$ in all subintervais defined by (6) and select the one that corresponds to the minimum value of $w_{k+1}(x)$. The computational effort associated with this procedure is clearly much smaller than the work required to evaluate $\varphi_{k+1}(x)$ 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 subintervais on which $m_k(x)$ is much greater than $f(x^{optk})$, however the chance of selecting x^{k+1} from such subintervais 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 $w_{k+1}(x)$ in some subinterval is greater than $f(x^{optk})$. 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 seperately. 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

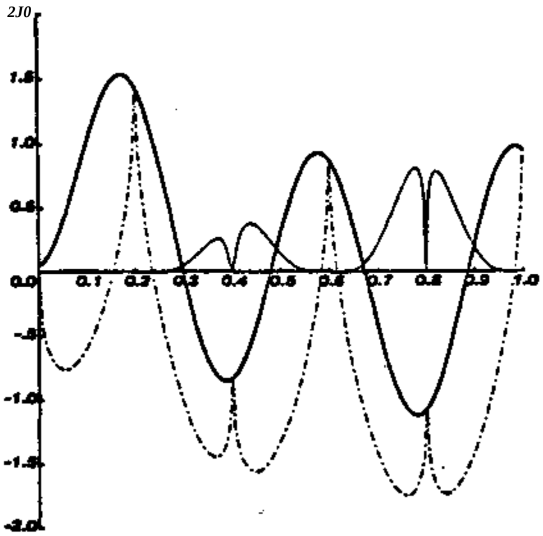

Figure 1: Functions f(x)-bold line, m_k(x)-solid and w_k(x)-dotdash; Section 5.1 example i; k«5

of "the lowest potential energy" between successive pairs of observations $f(x^{k,1J})$ and $f(x^{k})$ (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 $m_k(x)$ is a piecewise linear function connecting successive $f(x^{ik})$, (i«0,1....k). In the n-dimensional case we assume that %(x), x€ACR^n, is for each simplex a hyperpJane joining the vertices (observation values). Similarly the quadratic'equation for $a_k(x)$ is generalized to a n-dimensional quadratic surface for each simplex. The coordinates of the optimal point x_j^{0*}, (j = 1.2.....n) for any simplex s are given by

$$x_i^{o3} - x_i^{f**} - a_i V \{\frac{c^2 2}{2^{s}} + \sum_{j=1}^{n} (a_j^s)^2\}^{-1/2} \quad i = 1....n \qquad (9)$$

where

- x_i^{f3} • j^{91} coordinate of the center of hypersphere described on considered simplex s
- r* - radius of this hypersphere
- $a_i^?$ - j* coefficient in the expression for m_k^* (x) (10).

$$mj(x) = \sum_{j=1}^{n} a_j^s x_j + a_{n+1}^s$$

$$\sigma_k^s(x) = \sigma (2r^s)^{-1/2} \times \{(r^s)^2 - \sum_{i=1}^{n} (x-x_i^{(s)})^2\}^{-1/2} \qquad (10)$$

As in the one-dimensional case, mj (x) contains all the observation values calculated for vertices of a given simplex. Similarly σ_k^* (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^{0**} 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^{**1} is the point of "the lowest potential energy" among all the points x^{03} (one for each simplicial division), i.e., x^{k*1} is the point at which ^{w}k* 1^ is minimum in whole hypercube [0,1]°.

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.

Function	e	Bayesttn Method	Simplified Method
f(x)	i	.7789	.7624
	2	.7570	.7599
	3	.7430	.7437
Oaxsi	7	.7227	.7204
	1	3680	3720
$f_2(x)$	2	3461	3461
	3	3336	3327
Oftti	7	3164	3147
	1	6316	&451
$f_3(x)$	2	6364	6366
	3	7334	7346
-1O*X*1O	7	7.609	7356

Table 1: Comparison of x^{k*1} points determined by the Bayesian method and the simplified method

Function	c	Bayesian Method			Simplified Method		
		x	«*)	ote.	x	Kx>	obs.
' , M	1	.7798	-1.1232	23	.7799	-1.1232	30
		.7789	•1.1232	7	.7790	•1.1232	15
O«x*1	2	.7792	•1.1232	9	.7794	-1.1232	9
		.7775	-1.1227	32	.7816	-1.1226	26
ffm23B2	3	.7787	•1.1232	22	.7797	-1.1232	21
		.7768	•1.1224	32	.7765	-1.1221	29
	7	.7816	-1.1227	20	.7812	•1.1229	20
		.7735	•1.1187	32	.7720	•1.1162	30
Oferfi	1	3963	•1.0008	9	3985	•1.0008	16
		3876	•1.0006	11	3963	•1.0006	22
	2	3883	•1.0008	15	3983	•1.0008	22
		3878	-1.0007	9	3986	•1.0006	12
6*» 1-229	3	.7504	•1.1250	22	3965	•1.0008	21
		.7484	•1.1248	25	3982	-1.0008	32
	7	.7480	-1.1247	17	.7498	-1.1250	16
		.7466	•1.1210	32	.7529	-1.1224	26
•3W	1	5.7656	•113243	8	5.7817	-12.0312	24
		- --	•9.4350	9	5.7929	-12.0311	10
•tttot*O	2	-.4838	-12.0303	16	5.8401	-11.6783	32
		-.4818	-12.0171	15	5.8436	•11.6256	31
6"» 9-308	3	•4855	•12.0287	14	5.3230	-11.8820	32
		-.4840	-12.0236	11	5.6285	•113258	31
	7	4.5587	-9.4841	25	4.5539	•9.4826	21
		4.5800	•8.3486	29	4.5654	•8.4865	26

Table 2: The results of the optimization 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 • sin(5wx-0.4*)-0.125 0≤x≤1

with global minimum at x » 0.7795 and f „_ »-1.1232287
ran

2. fjM.minCf^xtf^x)) 0<x<1

where $f_{21}(x)$ * 2 (x-0.75)^2 • sin(8wx-0.5v) - 0.125

f^M * -125 +75(0.17-x) x≤0.17

«1.25+35(x-0.17) x>0.17

with global minimum at x a 0.17 with f^*-^{1}-^ and two significant local minima at x*0.75 where f(x) =-1.125, and x « 0.99842 where f(x)» -1.0007867

3. $f_3(x) = -\Sigma_{k=1}^{5}[k \sin((k+1)x+k)]$ -10≤x≤10

with three equal minima at the points: -6.77457, -0.49139 and 5.79179 with t_{mm} = -12.03125, and one signifcant local minimum at x« 4.5577 with f(x)«-9.4947. Table 1 shows the comparison of x^{11}*^{1} points determined for those three functions by original Bayesian method and the simplified one. for different values of c. The results correspond to the case when the distances between the observation points x'^{k} (i « 0.1....k; k * 5) were equal.

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:

LBranfn's function.

yx.yj-aiy-bx^cx-d)^2 • l(if)cos(x) + I

where a*1. b*5.1/4*^{2}, c»5/v, d»6. I*10. f»1/8«

-5≤x≤10$_t$ 0≤y£15

This function has three equal global minima t_{nm} • 0.397887 at points (-3.14159.12.275). (3.14159, £275), (9.42478T2.475).

2. Goldstein's and Price's function.

f_2(x,y) = [1 + (x + y 1 ^{2} 19-14xy+ 3x^2-14x 3+ y +)]^2 x
[30 • (2x-3y)^2(18-32x • 12x^2 • 48y-36xy • 27y^2))

This function has global minimum f^ » 3 at point (0.000, -1.000) and local minima as below:

x	•600	•396	1800	1.200
y	-.400	•602	.200	£00
K«.y)	30	35	84	840

Function	c	x	y	f(x,y)	obs.
$f_1(x,y)$	1	9.4179	2.2902	.4301	22
		9.2918	2.6383	.5573	33
	2	9.3651	2.7011	.4911	18
-5≤x≤10		9.3230	2.0587	.5573	36
0≤y≤15	3	9.3814	3.2025	.9903	25
		-3.2242	13.2242	.9928	36
δ = 119.996	7	-3.2882	12.9992	.6371	27
		2.6962	2.2525	1.4908	24
$f_2(x,y)$	1	.1303	-1.0692	14.7923	24
		.2085	-1.0429	32.7279	13
	2	.1703	-.9254	10.1322	17
-2≤x≤2		.1601	-.9070	11.1597	18
-2≤y≤2	3	.1512	-.9381	8.6230	18
		.1562	-.9103	8.8741	17
δ = 444434	7	.1371	-.9021	8.2519	24
		.1386	-.8909	9.0065	23

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

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.

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 Approachs to Global Optimization.
North-Holland Publishing Company, 1978, pages 19-29.

[2] Kushner, H.J.
A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise.
Journal of Basic Engineering :97-106, March, 1964.

[3] Mockus,I.B.
On Bayesian Methods to Seek the Extremum.
Automatics and Computers (russian) 3:53-62, 1972.

[4] Mockus, J., Tiesis, V. and Zilinskas, A.
Towards Global Optimization. : The Application of Bayesian Methods for Seeking the Extremum.
North-Holland Publishing Company, 1978, pages 118-129.

[5] Zilinskas,A.G.
The One-Step Bayesian Method for Seeking the Extremum of the Function of One Variable.
Cybernetics (russian) 1:139-144, 1975.