LILIANA SPIRCU
Department of Economic Cybernetics
Academy of Economic Studies

Cluster Analysis in Global Optimization

One of the most important problem in mathematical programming is to find the global minimum of a function of several variables. The paper presents several algorithms for approximating the global minimum, emphasis being laid on the algorithms using cluster methods. Thus Törn's algorithm [3] *is presented and another original algorithm is proposed for approximating the global minimum,based on the theory of statistical estimations of a density function and on the properties of the mode* [5].

1. INTRODUCTION

Any productive system achieving a number of n products by consuming a number of p limited resources and intended to perform the qualitative and quantitative plan indices with a minimum consumption of resources can be described with the aid of a mathematical programming model. If the objective-function of the productive system representing the profit or the production cost is a nonlinear (nonconvex) function, the usual techniques of finding the optimum do not always lead to the global optimal solution of the model, but to a local optimum [8]. If we are interested in approximating the global optimum of the model, therefore of the structure of production which gives a maximum global profit, respectively, which requires a global minimum production cost, we must use special techniques in order to solve the mathematical programming problems.

The obtaining of a global optimum is necessary also in electronics (for example in circuit design), where the optimizing function expresses the deviation from the accepted tolerances and is usually nonconvex [10].

Problem SG: Find a global solution x_* for which

$$f(x_*) = \min_{x \in S} f(x) \tag{1.1}$$

where f is a nonconvex function and $S \subset R^n$, assuming that the local optima of f are interior points of the set S (not on the boundary of S). Under these circumstances, problem SG may be considered to be without constraints on variables $X = (x^1, ..., x^n)$. Naturally, S is supposed to be a bounded set. We may reduce the general case to a simpler situation in which S is a parallelipiped in R^n, by specifying for each variable x^i a variation interval

$$[k^i, K^i], \ i = 1, ..., n.$$

This sort of problem frequently appears in the sequential minimizing techniques of a function (SUMT) developed by Fiacco and McCormick [8], [9] supposing the solution of a number of subproblems of the SG type whose solutions converge to the searched minimum of the function. Obviously, the known optimizing techniques can be used in determining or approximating a local optimum of the problem (1.1) [2], [8].

The global minimizing algorithms based on the generation of random points in the domain S require the generation of a big number of points in S in order to increase the probability of obtaining a global minimum. The difficulties appearing hence are linked with a great computation effort which is most often useless ,because the random points generated in an uncontrolled manner in S can be in the neighbourhood of a local minimum already obtained after a previous generation. Several algorithms, which are intended to conduct the generation of random points in S so that the search of the global minimum be as efficient as possible, are presented in [1].

Further, we shall present a global minimizing algorithm based on the iterative definition of a marginal density function φ^i, for each component of vector x, on the variation interval $[k^i, K^i]$. For the beginning, we suppose that each φ^i is uniformly distributed in the variation interval.

Algorithm GA. Step 1. We generate N points $X_k \in S$ and calculate the values $f(X_k)$. (Points X_k are generated after density function φ, that is each component x_k^i of vector X_k is generated after φ^i).

Step 2. The marginal density functions φ^i are modified as follows: (a) We say that a value $f(X_k)$ is a „success" if in the increasingly ordered sequence of the values $f(X_j)$, $j = 1, ..., N$, $f(X_k)$ is among the first aN of the values, a being a fixed threshold (0.25 can be taken).

b) We suppose each interval $[k^i, K^i]$ divided into $m(\geq 2)$ subintervals of equal length. We denote by M_{ij} the rate of success for subinterval j of component i; it is obtained by dividing the number of "success" points which have component i belonging to subinterval j, by the total number of points which have this property, in case there exist such points; otherwise take $M_{ij} = 0$.

c) We redefine the marginal density function φ^i by putting

$$p_{ij} = M_{ij} / \sum_{j=1}^{m} M_{ij}$$

where p_{ij} is the probability that component i of a generated point should belong to subinterval j.

Step 3. If for an index i there is an index $j(i)$ for which we have

$$p_{ij} = \begin{cases} 0 \text{ for } j \neq j(i), \\ 1 \text{ otherwise,} \end{cases}$$

then we reduce the set S by eliminating the subintervals $\neq j(i)$ of component i, and repeat with step 1; otherwise, we continue directly with step 1.

It is pointed out in [3] that in global optimization, the cluster analysis can be used successfully, by considerably reducing the computation effort.

Before presenting several algorithms using cluster analysis we shall give a few introductory elements about the cluster analysis.

2. CLUSTER ANALYSIS IN GLOBAL OPTIMIZATION.

2.1. Let $I = (i_1, i_2, ..., i_N)$ be a set of N individuals belonging to a po pulation P. Let also consider a set of n characteristics $C = (c_1, c_2, ..., c_n$ noticeable on the individuals of I, characteristics which in reality may be qualitative or quantitative, which we shall suppose to be only quantitative and measurable. Thus, let X_{ij} be the value of characteristic c_i measured on the individual i_j and let X be the matrix of measurements, matrix of size $n \times N$.

Problem GC. Let m be a positive integer (given before or which is to be determined), smaller than N. Determine m groups (or clusters) of individuals of I, be they $\pi_1, \pi_2, ..., \pi_m$ so that each individual $i \in I$ belongs only to a single cluster π_j, and the individuals clustered into the same class π_j "are similar" between them, at the same time being "different" from the individuals of any other class π_k, $k \neq j$.

A solution of the problem GC (called a cluster problem) consists, in fact, in determining that partition of set I which, of the set of all possible partitions, optimizes a certain objective function which is most often a distance function. A number of algorithms for clustering based on the distance functions is given in [6].

Wishart [7] has developed another cluster method based on the estimation of the density function and on the analysis of modes, a cluster being associated to a mode.

Given a random vector x with n components, let $\{X_1, X_2, ..., X_N\}$ be a selection of volume N on it. For value $\psi(x)$ of the density function of X, the following estimation based on the selection can be considered:

$$\psi_N(X) = \frac{1}{N} \sum_{j=1}^{N} \delta_N(X - X_j) \tag{2.1.}$$

where δ_N is a δ—generating sequence [5].

We know that if $K(X)$ is a density function defined on R^n which satisfies the following two conditions:

1) $|K(X)| < \infty$ for every $X \in R^n$;

2) $lim\, |x^1 \ldots x^n| \, |K(X)| = 0$ when $\|X\| \to \infty$,

then the sequence:

$$\delta_N(X) = \frac{1}{h_1(N) \ldots h_n(N)} K\left(\frac{x^1}{h_1(N)}, \ldots, \frac{x^n}{h_n(N)}\right)$$

where $lim\, \|h(N)\| = 0$ when $N \to \infty$ is a δ-generating sequence. K is called a *weight function.*

It is pointed out in [5] that (2.1.) is a consistent, asymptotically normal, unbiased estimation for density ψ. In the same paper, the following weight-functions are suggested for estimating the density function in the n-dimensional case:

$$K(y) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left(-\frac{1}{2} y^T \Sigma^{-1} y\right) \text{ the } n - \text{dimensional normal;}$$

$$K(y) = \begin{cases} 1/2^n, & \text{for } y \in U \\ 0, & \text{otherwise;} \end{cases}$$

$$K(y) = \begin{cases} \prod_{j=1}^{n} (1 - |y^j|) & \text{for } y \in U, \\ 0 & \text{otherwise;} \end{cases}$$

where U is the n-dimensional cube $[-1,1] \times \ldots \times [-1,1]$;

$$K(y) = \frac{1}{2^n} \exp\left(-\sum_{j=1}^{n} |y^j|\right);$$

$$K(y) = \left(\frac{2}{n}\right)^n \prod_{j=1}^{n} \left(\frac{\sin \frac{y^i}{2}}{y^i}\right)^2.$$

The *mode* of a random vector X with n components is called any point $\theta \in R^n$ which is a local maximum of the density function of vector X. The point $\theta_N \in R^n$, which is a maximum point of estimation ψ_N given by (2.1) is called *estimation* of the mode θ.

It is pointed out in [5] that in certain hypotheses imposed on function ψ, θ_N is a consistent and asymptotically normal estimation for θ.

2.2. We now present the algorithm proposed by Tőrn in [3] based on the classification into clusters using spheres with fixed centres and variable radii.

Algorithm CSG. Step 1. Generate uniformly N points $X_1, X_2, ..., X_N$ in domain S.

Step 2. Consider X_i $(1 \leqslant i \leqslant N)$ as initial points in an algorithm of classical minimization, and execute one or several iterations, obtaining the points y_i, $i = 1, ..., N$.

Step 3. Apply an algorithm for clustering the points $y_1, y_2, ..., y_N$ and consider NC the number of clusters obtained; if $NC = 1$ continue with step 5, otherwise with step 4.

Step 4. Choose a representative of each cluster and repeat form step, 2 with $N = NC$.

Step 5. We are in the neighbourhood of the global minimum; choose a representative of the only cluster and apply a rapid algorithm for finding the minimum (efficient in the neighborhood of the optimum).

We shall further specify some aspects linked to the cluster technique used by Tőrn at step 3 of this algorithm CSG. Two facts are important, namely: the choice of the grouping centres and the choice of the stopping threshold for extending the respective cluster. The adequate choosing of a threshold-distance r is important because if it is too big all points will be clustered around a single centre and if r is too small each point will form a cluster ,hence the grouping does not lead to any reduction of the volume of calculi.

The following way of choosing a threshold for limiting the growth of a cluster is proposed. Let S be the reference domain and let $X_1, X_2, ..., X_N$ be the N points uniformly generated in S. We consider a parallelipipedic envelope of the generated points, be it A, having the volume V. We divide N by V and obtain the density

$$\rho = N/V$$

for the uniform distribution of these points in A. A cluster is obtained by grouping the points belonging to a sphere of centre y and radius r, a radius which will successively be increased $(r_1 < r_2 < ...)$ as long as the density of the points belonging to the sphere is bigger than the uniform density ρ.

If $S(n, r_i, X)$ is the sphere of radius r_i and centre X in the space R^n and $W(n, r_i)$ is its volume, we have the following relation from which we can obtain the radius r_i:

$$\rho W(n, r_i) = i, \quad i = 1, 2, ...$$

Since

$$W(n, r) = \begin{cases} \dfrac{\pi r^n}{\left(\dfrac{n}{2}\right)!} & \text{for } n \text{ even,} \\ \dfrac{2\sqrt{(2\pi)^{n-1}}\, r^n}{1 \cdot 3 \cdot ... \cdot n} & \text{for } n \text{ odd,} \end{cases}$$

we have:

$$r_i = Q(n, \rho)\sqrt[n]{i},$$

where:

$$Q(n, \rho) = \begin{cases} \dfrac{\sqrt[n]{\dfrac{(n/2)!}{\varphi}}}{\sqrt{\pi}} & \text{for } n \text{ even,} \\ \dfrac{1 \cdot 3 \cdot \ldots \cdot n}{2\rho\sqrt{(2^{\pi})^{n-1}}} & \text{for } n \text{ odd.} \end{cases}$$

Proceed as follows: calculate the values $f(X_k)$ for all generated points X_k, then choose as a grouping centre X_0 that generated point which has the minimum value $f(X_k)$. Calculate the distances from this grouping centre to all other generated points. Iteratively, as far as it is possible, calculate radii $r_1, r_2, \ldots$ and cluster the points having the distances up to the grouping centre between r_i and r_{i+1}. When the growth of the cluster is stopped, the method is repeated by choosing another grouping centre, up to the clustering of all generated points.

As volume V is a decisive element in clustering, to achieve the independence of the method from the selection of the parallelipipedic envelope, we shall consider that it was selected having the sides parallel with the eigen vectors of the covariation matrix, calculated with the aid of the formula:

$$s_{ij} = \frac{1}{N-1}\left(\sum_{k=1}^{N} x_{ik}x_{jk} - \frac{1}{N}\sum_{k=1}^{N} x_{jk}\sum_{k=1}^{N} x_{ik}\right).$$

In this case

$$V = V(m, L, \lambda) = L\prod_{i=1}^{m}\sqrt{\lambda_i}$$

where $\lambda_i (i = 1, \ldots, n)$ are the eigenvalues of the covariation matrix. Appearing in the formula are only the m eigenvalues which are bigger than a certain threshold β. The constant value L is calculated depending on the density of the points in S [3].

The selection of the grouping centre remains to be explained . The first grouping point is that which gives the lowest value for the objective function. Then, the growth of the cluster around this centre is achieved until the stopping state is obtained. The next grouping centre will be given by that point of the yet ungrouped set which has the lowest value of the objective function, etc.

The number of generated points is:

$$N = \frac{\ln \varepsilon}{\ln(1-p)}$$

where p is the probability that a random generated point in S should belong to the influence region of the global minimum; thus, we shall choose as many points in S so that the probability of having at least a point in the influence region of the global minimum should be at least $1 - \varepsilon$.

In the following, we shall present a global minimizing algorithm based on the estimation of the density function and of its modes.

Algorithm EMG. Step 1. Generate uniformly N points $x_1, x_2, \ldots x_N$ in S.

Step 2. Consider $x_i (1 \leqslant i \leqslant N)$ as initial points in a classical minimizing algorithm (for example the gradient algorithm), execute one or several iterations, obtaining points $y_1, y_2, \ldots, y_N$ in S and values $f(y_1), \ldots, f(y_N)$ for the objective function.

Step 3. We increasingly order values $f(y_i)$, $i = 1, \ldots, N$; we say that point y_j is a "success" if the value $f(y_j)$ is among the first αN of the values $f(y_i)$ with $i = 1, \ldots, N$ (α can be taken for example 0.25). Let $z_1, \ldots, z_M$ be the success points.

Step 4. Consider $z_1, \ldots, z_M$ as a selection of volume M and estimate the density function using this selection by formula (2.1.); be it ψ.

Step 5. Calculate the modes of density ψ; let $m_1, m_2, \ldots, m_K$. If $K = 1$ go to step 7, otherwise go to step 6.

Step 6. Generate uniformly $N' = (N - K)/2$ points in S. Modify $N = K + N'$ and denote by x_i $(1 \leqslant u \leqslant N)$ the modes m_j $(1 \leqslant j \leqslant K)$ and the new generated points. Continue with step 2.

Step 7. We are in the neighbourhood of the searched minimum. Consider m_1 as being the initial point in a rapid minimizing algorithm, apply this algorithm until the minimum is sufficiently well approximated.

In comparison with Tőrn's algorithm CSG, the advantages of algorithm EMG are the following:

a) It requires a smaller number of point generations in S for the same risk factor.

b) It greatly uses the data related only to the global minimum, neglecting the information related to other optima. (Tőrn's algorithm tends towards the identification of all local minima, whereas the EMG algorithm loses the information referring to the "unimportant" local minima.

The difficulty of the algorithm EMG consists in the calculation of the modes of the density ψ at step 5. But, using as weight function a n-dimensional normal, this calculus can be made by applying rapid algorithms for optimizing function ψ, by starting from the initial points $z_1, \ldots, z_M$.

REFERENCES

1. HARTMAN, J.K.: *Some Experiments in Global Optimization.* Naval Research Logistics Quarterly, 20 (1973), 3.
2. LARCHEV, O.I.: *New Approach to Comparison of Serach Methods Used in Nonlinear Programming Problems.* I. Optimization Theory Appl., 13 (1974), 6.
3. TŐRN A.: *Cluster Analysis Using Seed Points and Densily Determined Hyperspheres as an Aid to Global Optimization.* IEEE Transactions on Systems, Man and Cybernetics (SMC) 7 (1977), 8.

4. TÖRN A.: *Cluster Analysis as a Tool in a Global Optimization Model*, in Modern Trends in Cybernetics and Systems, Springer Verlag, Berlin-Heidelberg-New York, 1977.
5. VĂDUVA I.: *Contribuţii la teoria estimaţiilor statistice ale densităţilor de repartiţie şi aplicaţii* (Contributions to the Theory of Statistical Estimations of Allocation Densities and Applications). Studii şi cercetări matematice ,20, 1968.
6. DURAN B.S., ODELL P.L.: *Cluster Analysis: A Survey*. Springer Verlag Berlin-Heidelberg-New York, 1977.
7. FU K.S. (Ed.): *Digital Pattern Recognition*. Springer Verlag, Berlin-Heidelberg-New York, 1976.
8. DESCHENEAU J.CH.: *Mathematical Programming for Parameter Optimization*. Technical Report TR 73—12, 1973.
9. FIACCO A.V.: *Computational Algorithm for Sequential Unconstrained Minimization Techniques for Nonlinear Programming*. Management Science, 10, 601—607, 1964.
10. SRIDHARA RAO G., SASTRY V. V. RAO VENKATA G.: *Comparison of Nonlinear Programming Techniques for the Optimal Design of Transforms*. Proc. of the Institute of Electrical Engineers, 124, 1977, 12.

TECHNIQUES DE CLASSIFICATION DANS L'OPTIMISATION GLOBALE

L'un des problèmes importants de la programmation mathématique consiste dans l'obtention du minimum global d'une fonction dépendant de plusieurs variables. On présente quelques algorithmes d'approximation du minimum global, insistant sur les algorithmes qui utilisent ses techniques de classification. On présente ainsi l'algorithme de Törn [3] et on propose un algorithme propre d'approximation du minimum global, ayant pour base la théorie des estimations statistiques de la densité de répartition et les propriétés du mode [5].

КЛАССИФИКАЦИОННЫЕ МЕТОДЫ В ГЛОБАЛЬНОЙ ОПТИМИЗАЦИИ

Однои из важных проблем математического программирования является нахождение глобального минимума функции от нескольких переменных. В работе приведено несколько алгоритмов, апроксимирующих глобальный минимум, настаивая на алгоритмы, использующие классификационные методы. Таким образом, представлен алгоритм тэрн [3] и предлогается новый алгоритм апроксимации глобального минимума, на основе теории статистических оценок плотности распределения и своиств модуля [5].

14. 10. 1978