KYBERNETIKA — VOLUME 24 (1988), NUMBER 5

QUASI-NEWTON GRADIENT METHOD WITH ANALYTICAL DETERMINATION OF THE DIRECTION AND LENGTH OF STEP

PETER HUDZOVIČ

The paper presents an algorithm of the gradient method generating a sequence of matrices approximating inverse of a Hessian matrix in such a way that not only the proper direction of the gradient but even the step-length is determined in every step. So no (time-consuming) one-dimensional search procedure is required during iteration steps and the total time for finding an extreme point is significantly reduced.

1. INTRODUCTION

Quasi-Newton gradient methods are proved to be very useful for solving nonlinear problems of multidimensional static optimization since they provide fast convergency of the iteration procedure. Basically, these methods are modifications of the well-known Newton method, but (instead of inverting a Hessian matrix) they generate a sequence of matrices determining the proper direction of the gradient and approximating the inverse of the Hessian matrix.

Comparing with the Newton method a class of quasi-Newton algorithms has the imperfection that no step-length in the desired direction is obtained. The step-length must be subsequently found by applying one-dimensional search procedure.

In this paper a method for generating an approximating sequence of a Hessian matrix is suggested in such a way that the product of the elements of this sequence with the gradient determines not only the step-direction but also the step-length (in general, the step-length does not correspond to the distance of the local extreme).

2. BASIC RELATIONS

Let the objective function $f(x)$ be a scalar function of a vector argument $x \in \mathbb{R}^n$ and let us assume existence of continuous derivatives of $f(x)$ up to the second order. The gradient $\nabla_x f(x)$ and the Hessian $\nabla_x \nabla_x^T f(x)$ of $f(x)$ will be denoted by $g(x)$

and $H(x)$ respectively. Let k be an integer denoting the iteration step. In order to abbreviate the notations the value of function $f(x_k)$, its gradient $g(x_k)$ and the Hessian $H(x_k)$ at x_k will be denoted by f_k, g_k and H_k respectively.

We shall assume a strict convexity of the objective function, so that

$$[x_{k+1} - x_k]^{\mathrm{T}} . [g_{k+1} - g_k] = r_k^{\mathrm{T}} y_k > 0 \tag{1}$$

where

$$r_k = x_{k+1} - x_k \tag{2}$$

$$y_k = g_{k+1} - g_k . \tag{3}$$

Let the function $f(x)$ take its extreme at the point x^*. Without loosing generality we can assume that the extreme is the minimum taking the value $f(x^*) = f^*$.

Let us express the function $f(x_k + r)$ about the point x_k by the quadratic approximation

$$f(x_k + r) \cong f_k + r^{\mathrm{T}} g_k + \tfrac{1}{2} r^{\mathrm{T}} H_k r \tag{4}$$

and let us search for such a point $x_{k+1} = x_k + r_k$ at which the quadratic approximation reaches its extreme

$$\nabla_{x_k + r} f(x_k + r) = \nabla_r f(x_k + r) \cong g_k + H_k r_k = g_k + H_k(x_{k+1} - x_k) . \tag{5}$$

This condition results directly in the Newtonian algorithm

$$x_{k+1} = x_k - H_k^{-1} g_k \tag{6}$$

requiring calculation of the inverse of theHessian matrix H_k. The advantage of the Newtonian algorithm (6) is a rapid convergency of the iteration process. In case of a quadratic objective function the extreme $x^* = x_{k+1}$ is obtained within a single step from an arbitrary point $x_0 = x_k$.

In accordance with (4) the quadratic approximation of the function $f(x_{k+1} - r)$ about the point x_{k+1} yields

$$f(x_{k+1} - r) \cong f_{k+1} - r^{\mathrm{T}} g_{k+1} + \tfrac{1}{2} r^{\mathrm{T}} H_{k+1} r \tag{7}$$

with the gradient at the point $x_k = x_{k+1} - r_k$ given by

$$g_k = \nabla_{x_{k+1} - r} f(x_{k+1} - r) = -\nabla_r f(x_{k+1} - r) \cong g_{k+1} - H_{k+1} r_k . \tag{8}$$

Combining this relation with equation (3) the so-called quasi-Newton condition is obtained which is the basis of a whole group of gradient methods called ,,variable metric methods"

$$y_k = H_{k+1} r_k . \tag{9}$$

Using the quasi-Newton methods the recurrent relation (6) is modified to

$$x_{k+1} = x_k - t_k M_k g_k , \tag{10}$$

thus the matrix H_k^{-1} is substituted by a positively definite symmetric matrix $t_k M_k$ which deflects the gradient g_k to the required direction r. The step-size t_k, however, needs to be determined additionally by the one-dimensional search

$$f_{k+1} = \min_t f(x_k - tM_k g_k) = f(x_k - t_k M_k g_k) = f(x_k + r_k) \tag{11}$$

After the gradient g_{k+1} at the point x_{k+1} has been calculated the following procedure in the quasi-Newton methods is to derive the matrix M_{k+1} from the M_k matrix by a procedure based on condition (9). And it is the nature of the formula relating the matrices M_k and M_{k+1} which differentiates respective methods of the variable metric from each other. We present three of these methods which are referred to most often.

$$M_{k+1} = M_k + \frac{(r_k - M_k y_k)(r_k - M_k y_k)^{\mathrm{T}}}{(r_k - M_k y_k)^{\mathrm{T}} y_k} \tag{12}$$

belongs to Broyden's method.

The Davidon-Fletcher-Powell method (thereafter only the DFP method) follows with

$$M_{k+1} = M_k + \frac{r_k r_k^{\mathrm{T}}}{r_k^{\mathrm{T}} y_k} - \frac{M_k y_k y_k^{\mathrm{T}} M_k}{y_k^{\mathrm{T}} M_k y_k}. \tag{13}$$

Finally, there is the formula designed by Broyden, Fletcher and Shanno (abbreviated to the BFS method)

$$M_{k+1} = \left[I - \frac{r_k y_k^{\mathrm{T}}}{r_k^{\mathrm{T}} y_k}\right] M_k \left[I - \frac{y_k r_k^{\mathrm{T}}}{r_k^{\mathrm{T}} y_k}\right] + \frac{r_k r_k^{\mathrm{T}}}{r_k^{\mathrm{T}} y_k} \tag{14}$$

where I denotes the identity matrix. The given formulas conclude one step of the variable metric methods and the repetition begins with one-dimensional search.

3. CONVERGENCE OF QUASI-NEWTON ALGORITHMS

In case of the quadratic objective function $f(x)$ the gradient (5) vanishes at the extreme $x^* = x_{k+1}$. Under this assumption the approximate relations (4) and (7) became equations. The Hessian H_k will be a constant matrix H, therefore (6) can be written in the following way

$$r_k = -H^{-1} g_k . \tag{15}$$

Substituting (15) into (4) we get

$$f^* = f_k - \tfrac{1}{2} g_k^{\mathrm{T}} H^{-1} g_k . \tag{16}$$

Combining (4) with (10) and (11) and considering

$$r = -tM_k g_k \tag{17}$$

yields

$$f(x_k - tM_k g_k) = f_k - t g_k^{\mathrm{T}} M_k g_k + \tfrac{1}{2} t^2 g_k^{\mathrm{T}} M_k H M_k g_k . \tag{18}$$

From the zero value of its derivation w.r.t. t the optimal step-length can be determined

$$t_k = \frac{g_k^T M_k g_k}{g_k^T M_k H M_k g_k} \,. \tag{19}$$

Substituting this formula into (18) and modifying it using (16) yields

$$f_{k+1} = f^* + \tfrac{1}{2} g_k^T H^{-1} g_k - \frac{1}{2} \frac{(g_k^T M_k g_k)^2}{g_k^T M_k H M_k g_k} \,. \tag{20}$$

Since in quasi-Newton methods the matrix H^{-1} at x_{k+1} is substituted by the matrix $t_{k+1} M_{k+1}$, the quotient of (20) and (16) results in

$$\varrho_k = \frac{f_{k+1} - f^*}{f_k - f^*} = 1 - \frac{(g_k^T M_k g_k)^2}{g_k^T M_{k+1} g_k g_k^T M_k M_{k+1}^{-1} M_k g_k} \tag{21}$$

which may be considered as the iteration process convergence rate. Using (2) and (10) we get

$$\varrho_k = 1 - \frac{(r_k^T M_k^{-1} r_k)^2}{r_k^T M_k^{-1} M_{k+1} M_k^{-1} r_k r_k^T M_{k+1}^{-1} r_k} \,. \tag{22}$$

The positive definite matrix M_k is expressed as the product

$$M_k = G_k^T G_k \tag{23}$$

where the matrix G_k is called the square root of the matrix M_k. Moreover, we shall introduce the vector

$$w_k = (G_k^T)^{-1} r_k \triangleq G_k^{-T} r_k \,. \tag{24}$$

Employing (24), (22) can be written as

$$\varrho_k = 1 - \frac{(w_k^T w_k)^2}{w_k^T R_k w_k w_k^T R_k^{-1} w_k} \tag{25}$$

where R_k is a symmetric and positive definite matrix

$$R_k = G_k^{-T} M_{k+1} G_k^{-1} \,. \tag{26}$$

Denoting the smallest and largest eigenvalue of this matrix λ_{mk} and λ_{Mk} respectively, then according to the Kantorovich lemma [1], [3] it can be written

$$\varrho_k \leqq 1 - 4 \frac{\lambda_{Mk} \lambda_{mk}}{(\lambda_{Mk} + \lambda_{mk})^2} = \left(\frac{\lambda_{Mk} - \lambda_{mk}}{\lambda_{Mk} + \lambda_{mk}} \right)^2 \,. \tag{27}$$

Substituting (27) into (21) the following inequality is obtained

$$\frac{f_{k+1} - f^*}{f_k - f^*} \leqq \left(\frac{\lambda_{Mk} - \lambda_{mk}}{\lambda_{Mk} - \lambda_{mk}} \right)^2 \triangleq \varkappa_k^2 \tag{28}$$

which can be expressed also by means of the condition number the matrix R_k

$$1 \leqq \operatorname{cond} R_k = \frac{\lambda_{Mk}}{\lambda_{mk}} \,, \tag{29}$$

hence

$$\frac{f_{k+1} - f^*}{f_k - f^*} \leqq \left(\frac{\operatorname{cond} R_k - 1}{\operatorname{cond} R_k + 1}\right)^2 = \varkappa_k^2 . \tag{30}$$

The convergence rate depends proportionally on the inverse value of cond R_k. In accordance to (26) the minimization of cond R_k requires to minimize the condition number change of the matrix M_{k+1} with regard to the condition number of the matrix M_k. This was employed for improving the quasi-Newton methods so that in the relation binding the matrices M_{k+1} and M_k a variable parameter was introduced the value of which was determined so as to ensure a minimum of condition number change of these two matrices. This also gave the name to the present procedure, thereafter referred to only as the "MCC method".

4. MINIMUM CONDITIONALITY CHANGE METHOD

In order to avoid the one-dimensional search procedure, the scalar t will not be explicitly introduced in the algorithm but the inverse of the Hessian H_k^{-1} will be substituted directly by the matrix M_k. Therefore, instead of (6) and (9) we use the equations

$$r_k = -M_k g_k \tag{31}$$

$$r_k = M_{k+1} y_k . \tag{32}$$

Simultaneously, we shall assume that between the matrices M_k and M_{k+1} the following recurrent relation holds

$$M_{k+1} = a_k M_k + A_k \tag{33}$$

where a_k is a positive scalar and A_k is a symmetric matrix. Combination of the last two equation yields

$$r_k = a_k M_k y_k + A_k y_k . \tag{34}$$

For the sake of simplicity we begin with the matrix A_k of rank one

$$A_k = u_k v_k^{\mathrm{T}} . \tag{35}$$

Under the assumption that vectors v_k and y_k are not be orthogonal on substituting the matrix (35) into equation (34) the vector u_k can be written as follows

$$u_k = \frac{1}{v_k^{\mathrm{T}} y_k} (r_k - a_k M_k y_k) . \tag{36}$$

Combination of (33), (35) and (36) gives the matrix

$$M_{k+1} = a_k M_k + \frac{1}{v_k^{\mathrm{T}} y_k} (r_k - a_k M_k y_k) v_k^{\mathrm{T}} \tag{37}$$

the symmetry of which can be provided by putting $v_k = r_k - a_k M_k y_k$ which results in

$$(38) \qquad M_{k+1} = a_k M_k + \frac{(r_k - a_k M_k y_k)(r_k - a_k M_k y_k)^{\mathrm{T}}}{(r_k - a_k M_k y_k)^{\mathrm{T}} y_k}.$$

This is the formula which transforms to Broyden's relation (12) if $a_k = 1$. For our purpose it will be more convenient to provide the symmetry of the matrix M_{k+1} so that the following matrix will be added to the R.H.S. of equation (37)

$$(39) \qquad N_k = \frac{v_k(r_k - a_k M_k y_k)^{\mathrm{T}}}{v_k^{\mathrm{T}} y_k} - \frac{(r_k - a_k M_k y_k)^{\mathrm{T}} y_k}{(v_k^{\mathrm{T}} y_k)^2} v_k v_k^{\mathrm{T}}.$$

The matrix N_k does not break the condition (32) because the product $N_k y_k$ is a zero vector

(40)

$$M_{k+1} = a_k M_k + \frac{(r_k - a_k M_k y_k) v_k^{\mathrm{T}}}{v_k^{\mathrm{T}} y_k} + \frac{v_k(r_k - a_k M_k y_k)^{\mathrm{T}}}{v_k^{\mathrm{T}} y_k} - \frac{(r_k - a_k M_k y_k)^{\mathrm{T}} y_k}{(v_k^{\mathrm{T}} y_k)^2} v_k v_k^{\mathrm{T}}.$$

If the rank of the matrix A_k is two, we must use either $v_k = r_k$, or $v_k = M_k y_k$. Since both the alternatives lead then to the same relations, we choose the first one. After some algebraic manipulation we get by (40)

$$(41) \qquad M_{k+1} = a_k \left[M_k - \frac{M_k y_k r_k^{\mathrm{T}}}{r_k^{\mathrm{T}} y_k} - \frac{r_k y_k^{\mathrm{T}} M_k}{r_k^{\mathrm{T}} y_k} + \frac{y_k^{\mathrm{T}} M_k y_k}{(r_k^{\mathrm{T}} y_k)^2} r_k r_k^{\mathrm{T}} \right] + \frac{r_k r_k^{\mathrm{T}}}{r_k^{\mathrm{T}} y_k}.$$

For further modification of this matrix a vector orthogonal to y_k is used

$$(42) \qquad l_k = \frac{M_k y_k}{y_k^{\mathrm{T}} M_k y_k} - \frac{r_k}{r_k^{\mathrm{T}} y_k}.$$

Hence condition (32) will not be infringed if the matrix of rank one $(b_k r_k^{\mathrm{T}} y_k - a_k y_k^{\mathrm{T}} M_k y_k) l_k l_k^{\mathrm{T}}$ is added to the R.H.S. of (41) where b_k is a scalar parameter. The modification yields

(43)

$$M_{k+1} = a_k \left[M_k - \frac{M_k y_k y_k^{\mathrm{T}} M_k}{y_k^{\mathrm{T}} M_k y_k} \right] + \frac{r_k r_k^{\mathrm{T}}}{r_k^{\mathrm{T}} y_k} + b_k r_k^{\mathrm{T}} y_k \left[\frac{M_k y_k}{y_k^{\mathrm{T}} M_k y_k} - \frac{r_k}{r_k^{\mathrm{T}} y_k} \right] \left[\frac{M_k y_k}{y_k^{\mathrm{T}} M_k y_k} - \frac{r_k}{r_k^{\mathrm{T}} y_k} \right]^{\mathrm{T}}.$$

In accordance to (26) this formula is multiplied from the left by the matrix $G_k^{-\mathrm{T}}$ and from the right by the matrix G_k^{-1}. Moreover, we denote

$$(44) \qquad z_k = G_k y_k.$$

Then, using transformation (24) we obtain

$$(45) \quad R_k = a_k \left[I - \frac{z_k z_k^{\mathrm{T}}}{z_k^{\mathrm{T}} z_k} \right] + \frac{w_k w_k^{\mathrm{T}}}{w_k^{\mathrm{T}} z_k} + b_k w_k^{\mathrm{T}} z_k \left[\frac{z_k}{z_k^{\mathrm{T}} z_k} - \frac{w_k}{w_k^{\mathrm{T}} z_k} \right] \left[\frac{z_k^{\mathrm{T}}}{z_k^{\mathrm{T}} z_k} - \frac{w_k^{\mathrm{T}}}{w_k^{\mathrm{T}} z_k} \right].$$

The matrix R_k being of dimension $n \times n$ has then eigenvalues $\lambda_k = a_k$ being of multiplicity $n - 2$. To calculate the remaining two eigenvalues λ_{mk} and λ_{Mk} satisfying the inequality $\lambda_{mk} \leqq a_k \leqq \lambda_{Mk}$ we employ the fact that the eigenvectors corresponding to λ_{mk} and λ_{Mk} must be linear combinations of vectors w_k and z_k

(46) $$(\alpha w_k + \beta z_k) \lambda_{ik} = R_k(\alpha w_k + \beta z_k), \quad i = m, M.$$

To simplify what follows let us denote

(47) $$c_k = \frac{w_k^T w_k}{w_k^T z_k} = \frac{g_k^T M_k g_k}{r_k^T y_k} = -\frac{r_k^T g_k}{r_k^T y_k}$$

(48) $$d_k = \frac{w_k^T z_k}{z_k^T z_k} = \frac{r_k^T y_k}{y_k^T M_k y_k}.$$

Then the combination of (45) and (46) yields the following system of two linear equations

(49) $$\alpha \lambda_{ik} = \alpha a_k + \alpha c_k + \beta + \alpha b_k(c_k - d_k), \quad i = m, M,$$

(50) $$\beta \lambda_{ik} = \alpha b_k d_k(d_k - c_k) - \alpha a_k d_k, \qquad i = m, M.$$

Thus finding the eigenvalues λ_{mk} and λ_{Mk} results in a solution of the quadratic equation

(51) $$\lambda_{ik}^2 - [a_k + c_k + b_k(c_k - d_k)] \lambda_{ik} + d_k[a_k + b_k(c_k - d_k)] = 0.$$

To receive the maximum possible convergence rate of the iteration process, parameters a_k and b_k will be selected so as to minimize the function $\varkappa_k^2 = \varkappa_k^2(a_k, b_k)$ from the equation (27). For this purpose the quadratic equation (51) need not be solved because it is sufficient to know the product and the sum of the extreme eigenvalues. This leads to the conditions

(52) $$\frac{\partial}{\partial s_k} \frac{d_k[a_k + b_k(c_k - d_k)]}{[a_k + c_k + b_k(c_k - d_k)]^2} = 0, \quad s_k = a_k, b_k,$$

which result in a single relation binding the scalars a_k and b_k

(53) $$a_k = c_k - b_k(c_k - d_k).$$

Substitution in (51) will yields this simple form

(54) $$\lambda_{ik}^2 - 2c_k \lambda_{ik} + c_k d_k = 0, \quad i = m, M.$$

The combination of (27). (47) and (54) gives

(55) $$\varkappa_k = \sqrt{\left(1 - 4\frac{c_k d_k}{(2c_k)^2}\right)} = \sqrt{\left(1 - \frac{d_k}{c_k}\right)} = \sqrt{\left(1 - \frac{(w_k^T z_k)^2}{w_k^T w_k z_k^T z_k}\right)}.$$

In accordance to the assumption (1) on the strict convexity of the objective function it holds $r_k^T y_k = w_k^T z_k > 0$. Therefore, using the Schwarz inequality the expression under the square root sign can be considered non-negative which results in the relation $0 < d_k < c_k$.

Solving the equation (54) we obtain a pair of extreme eigenvalues of the matrix R_k

(56) $$\lambda_{ik} = c_k\left[1 \pm \sqrt{\left(1 - \frac{d_k}{c_k}\right)}\right] = c_k[1 \pm \varkappa_k]$$

which, supposing the assumption (1) holds, will be positive. The same applies also to their converted values which correspond to the extreme eigenvalues of the matrix R_k^{-1}

$$(57) \qquad \frac{1}{\lambda_{ik}} = \frac{1}{d_k}\left[1 \pm \sqrt{\left(1 - \frac{d_k}{c_k}\right)}\right] = \frac{1}{d_k}\left[1 \pm \varkappa_k\right].$$

Positive definiteness of the matrix M_{k+1} requires $a_k > 0$, hence according to (53) the following relation must hold

$$(58) \qquad b_k < \frac{c_k}{c_k - d_k} = \frac{1}{\varkappa_k^2}.$$

Besides, if λ_{Mk} and λ_{mk} are the largest resp. smallest eigenvalues of the matrix R_k and the remaining n-2 eigenvalues equal $\lambda_k = a_k$, the following inequality holds

$$(59) \qquad c_k(1 - \varkappa_k) \leqq c_k - b_k(c_k - d_k) \leqq c_k(1 + \varkappa_k)$$

which in virtue of (58) yields an interval of feasible values of the parameter b_k

$$(60) \qquad -\frac{1}{\varkappa_k} \leqq -1 \leqq b_k \leqq 1 \leqq \frac{1}{\varkappa_k}.$$

Finally, substituting equation (53) into the matrix (43) and modifying it, we obtain

$$(61) \qquad M_{k+1} = \left[c_k - b_k(c_k - d_k)\right] M_k + c_k(b_k - 1)\frac{M_k y_k y_k^{\mathrm{T}} M_k}{y_k^{\mathrm{T}} M_k y_k} - b_k \frac{M_k y_k r_k^{\mathrm{T}} + r_k y_k^{\mathrm{T}} M_k}{y_k^{\mathrm{T}} M_k y_k} + (b_k + 1)\frac{r_k r_k^{\mathrm{T}}}{r_k^{\mathrm{T}} y_k}.$$

The last two relations indicate that in general there is an infinite number of alternatives for calculating the matrix M_{k+1} from the matrix M_k. Anyhow, if the selection of the parameter b_k will depends on the number of operations connected with the evaluation of formula (61), only a finite number of the following alternatives

$$(62) \qquad M_{k+1} = (M_k, r_k, y_k, b_k)$$

is to be considered. Due to the fact that the first element on the R.H.S. of equation (61) must be positive the first alternative of the MCC method will be the one in which the choice of $b_k = 1$ leads to the zero value of the second element. After some algebraic manipulations we have

$$(63) \qquad M_{k+1} = \frac{r_k r_k^{\mathrm{T}}}{r_k^{\mathrm{T}} y_k} + \frac{r_k^{\mathrm{T}} y_k}{y_k^{\mathrm{T}} M_k y_k}\left[I - \frac{r_k y_k^{\mathrm{T}}}{r_k^{\mathrm{T}} y_k}\right] M_k \left[I - \frac{y_k r_k^{\mathrm{T}}}{r_k^{\mathrm{T}} y_k}\right].$$

Comparing (63) with (14) it is clear that the BFS method is a special case of the first alternative of the MCC method.

In the second alternative of the MCC method the third element on the R.H.S.

of (61) vanishes by choosing $b_k = 0$ which results in the formula

$$(64) \qquad M_{k+1} = \frac{r_k r_k^{\mathrm{T}}}{r_k^{\mathrm{T}} y_k} - \frac{r_k^{\mathrm{T}} g_k}{r_k^{\mathrm{T}} y_k} \left[M_k - \frac{M_k y_k y_k^{\mathrm{T}} M_k}{y_k^{\mathrm{T}} M_k y_k} \right].$$

When compared to formula (13), it indicates that the DFP method is a special case of the second alternative of the MCC method.

The vanishing last element on the R.H.S. of equation (61) corresponds to the third alternative of the MCC method where $b_k = -1$

$$(65) \quad M_{k+1} = 2 \frac{r_k^{\mathrm{T}} g_k}{r_k^{\mathrm{T}} y_k} \frac{M_k y_k y_k^{\mathrm{T}} M_k}{y_k^{\mathrm{T}} M_k y_k} - \left[2 \frac{r_k^{\mathrm{T}} g_k}{r_k^{\mathrm{T}} y_k} + \frac{r_k^{\mathrm{T}} y_k}{y_k^{\mathrm{T}} M_k y_k} \right] M_k + \frac{M_k y_k r_k^{\mathrm{T}} + r_k y_k^{\mathrm{T}} M_k}{y_k^{\mathrm{T}} M_k y_k}.$$

Two additional alternatives of the MCC method could be utilized. In the fourth one $b_k = -1/\varkappa_k$, i.e. the smallest possible value from interval (60). With the help of (54) and (56) we get

$$(66) \qquad a_k = c_k(1 + \varkappa_k) = \lambda_{Mk}$$

therefore the matrix R_k will have only one eigenvalue λ_{mk} different from $\lambda_k = a_k$. This means that the matrix A_k occurring in (33) will be of rank one. Formula (38) corresponds to this case if we substitute

$$(67) \qquad a_k = -\frac{r_k^{\mathrm{T}} g_k}{r_k^{\mathrm{T}} y_k} \left[1 + \sqrt{\left(1 + \frac{(r_k^{\mathrm{T}} y_k)^2}{r_k^{\mathrm{T}} g_k y_k^{\mathrm{T}} M_k y_k} \right)} \right]$$

which, in fact, is a generalized Broyden's method represented by (12). The same can be said about the fifth alternative of the MCC method in which the largest possible value $b_k = 1/\varkappa_k$ from interval (60) is chosen. When substituted in (54) it yields

$$(68) \qquad a_k = -\frac{r_k^{\mathrm{T}} g_k}{r_k^{\mathrm{T}} y_k} \left[1 - \sqrt{\left(1 + \frac{(r_k^{\mathrm{T}} y_k)^2}{r_k^{\mathrm{T}} g_k y_k^{\mathrm{T}} M_k y_k} \right)} \right]$$

and this value is substituted in formula (38).

Let us remark that in the kth step the current metric is $(x^{\mathrm{T}} M_k^{-1} x)^{1/2} = (x^{\mathrm{T}} G_k^{-1} G_k^{-\mathrm{T}} x)^{1/2} \triangleq (u^{\mathrm{T}} u)^{1/2}$, while in the step $k + 1$ we use the metric $(x^{\mathrm{T}} M_{k+1}^{-1} x)^{1/2} = (u^{\mathrm{T}} G_k M_{k+1}^{-1} G_k^{\mathrm{T}} u)^{1/2} = (u^{\mathrm{T}} R_k^{-1} u)^{1/2}$.

The relevant aspect for the choice of the parameter b_k from the interval (60) is the convergence rate of the iteration procedure from the probabilistic point of view. The convergence rate will be highest provided the equiscalar levels of the quadratic approximation of the objective function in the metric valid in the kth step become hyperspheres in the $(k + 1)$st step. This corresponds to the case when all the eigenvalues $\lambda_k^{-1} = a_k^{-1}$ of the matrix R_k^{-1} are equal. A pair of the extreme eigenvalues λ_{mk}^{-1} and λ_{Mk}^{-1} will, however, cause a flattening of the above mentioned equiscalar levels into hyperelipsoidal ones. Therefore, by a proper choice of the parameter b_k we shall strive, above all, for the minimum "flattening" of these hyperelipsoids.

In accordance with this reasoning the best results should be achieved with the first alternative of the MCC method at which the matrix R_k^{-1} has $n - 2$ eigenvalues $\lambda_k^{-1} = a_k^{-1} = d_k^{-1}$, while according to formula (57) the other two are located symmetrically with regard to it. In the remaining cases the above mentioned flattening is always higher, hence the probability of reaching the extremal point is smaller. This applies also to the second alternative of the MCC method in which the matrix R_k has $n - 2$ eigenvalues $\lambda_k = a_k = c_k$ and the two extreme values are in accordance to (56) located symmetrically with regard to c_k which means that their reciprocal values with regard to $a_k^{-1} = c_k^{-1}$ cannot be symmetrically located. As will be shown further, numerical results prove validity of these statements.

5. ALGORITHM OF THE MCC METHOD

Similarly as at the other iteration algorithms the most important as well as the most problematic step of the suggested method is the beginning of the iteration procedure with a heuristic choice of certain data, e.g. the position of the "support" point x_0 and values of the initial matrix M_0 elements. It is because the aforementioned data effect essentially the number of iteration cycles N and herewith also the overall time for searching the extreme T which can be considered the quality rate of specific algorithms.

A key moment in the entire procedure is to determine the length of the first step t_0 at the support point x_0 in the direction of the antigradient $-g_0$ for the objective function

$$\varphi(t) \triangleq f(x_0 - tg_0) . \tag{69}$$

An optimal step-length $t = t_0$ can be found by a one-dimensional search according to the relation

$$\left[\frac{\mathrm{d}}{\mathrm{d}t}\varphi(t)\right]_{t=t_0} = -g_0^{\mathrm{T}} g(x_0 - t_0 g_0) = -g_0^{\mathrm{T}} g_1 \triangleq 0 . \tag{70}$$

To obtain this optimum length it is advantageous to employ the procedure suggested by Davidon. This procedure is based on cubic extrapolation described e.g. in [5].

The first step for a gradual updating of value t_0 which satisfies the condition

$$|g_0^{\mathrm{T}} g(x_0 - t_0 g_0)| \leqq \varepsilon \tag{71}$$

where ε is a given small positive constant, may correspond to a fixed value t, or it can be estimated in the following way

$$t_c = t_a - \frac{\varphi'(t_a)}{\varphi''(t_a)} \doteq t_a - \frac{\varphi(t_a) - \varphi(t_b)}{\varphi'(t_a) - \varphi'(t_b)} \triangleq t . \tag{72}$$

If $t_a = 0$, then obviously $\varphi'(t_a) = -g_0^{\mathrm{T}} g_0$. Furthermore, we shall assume that $\varphi'(t_b) = 0$, thus necessarily $\varphi(t_a) > \varphi(t_b)$. This justifies to put $\varphi(t_a) - \varphi(t_b) = v|\varphi(t_a)| = v|f(x_0)| \triangleq v|f_0|$, where $v > 0$ and in the case of an objective function with non-negative function values $v \leqq 1$. This leads to the value of the first step

$$t = v \frac{|f_0|}{g_0^T g_0} \tag{73}$$

which starts up the procedure of one-dimensional search, the output of which is the step-length t_0. If at the start of the iteration process an identity matrix $J_0 = I$ is chosen, then the inverse of the Hessian matrix is substituted by the first estimate

$$M_0 = t_0 I . \tag{74}$$

To finish the entire process of searching the extremal point the following condition is used

$$0 \leqq \sqrt{(g_{k+1}^T g_{k+1})} \leqq \eta \tag{75}$$

where η is a preselected small positive constant determining the accuracy requirements on the locating of the extremal point $x^* \cong x_{k+1}$.

The derivation of the MCC method is based on the assumption that the objective function satisfies condition (1) of a strict convexity. If, however, this assumption is not satisfied the matrix M_{k+1} looses the property of positive definiteness. In order to eliminate such a situation, if inequality $r_k^T y_k > 0$ doesn't hold, the entire cycle is cancelled by putting $J_0 = M_k$. Then follows the turn to the beginning of the algorithm together with the search for the optimum step-length t_0 at $x_0 \triangleq x_k$.

The determination of the position of an extreme in accordance with the MCC method has the following steps:

1. Starting points are: determination of the matrix $J_0 = I$, heuristic determination of the parameters ε, v and η, as well as the choice of the support point x_0 in which the function value f_0 and gradient g_0 are calculated.

2. The determination of the step-length t_0 follows using a cubic extrapolation which begins either with a fixed chosen value t, or formula (73) can be used for this purpose. It stops by satisfying the condition (71). The matrix M_0 is determined according to (74).

3. Next step is the calculation of the point

$$x_{k+1} = x_k - M_k g_k \tag{76}$$

in which the function value f_{k+1} and the gradient g_{k+1} are calculated.

4. This is followed by testing the condition for stopping the iteration process (75). If this condition is satisfied, the process stops with the result $x^* \cong x_{k+1}$. If not, step No. 5 is continued.

5. A pair of vectors r_k and y_k is calculated in accordance with (2) resp. (3).

6. The condition of a strict convexity of the objective function (1) is verified. If this condition is not fulfilled the cycle is cancelled by putting at the same time $J_0 = M_k$ (preferably choose $J_0 = I$ – the identity matrix) and $x_0 = x_k$, and the return to the second step is accomplished. In case of a strict convex objective function the next step follows.

7. One of the alternatives of the MCC method is chosen and we calculate

$$M_{k+1} = F(M_k, r_k, y_k)\,. \tag{77}$$

We recommend to use the first alternative. The entire cycle is concluded by the return to the third step after the variable k was increased by one.

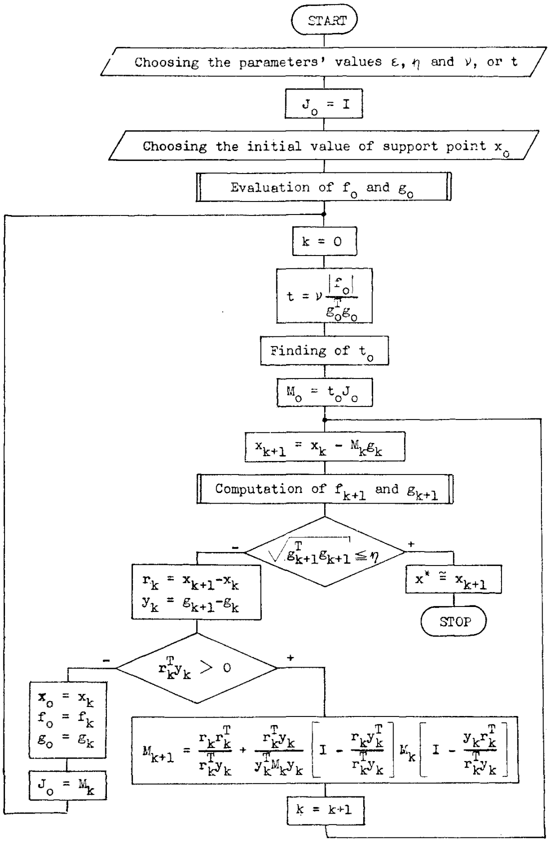

Fig. 1. Flow chart diagram of the MCC method.

Fig. 1. shows a flow chart diagram of the algorithm for the first alternative of the MCC method using (63).

6. EXAMPLES

The properties of individual alternatives of the MCC method were compared to the BFS and DFP methods, at first on the case of traditional testing Rosenbrock's "banana" function

$$f(x_1, x_2) = 100\,(x_1^2 - x_2)^2 + (x_1 - 1)^2 \tag{78}$$

the minimum of which is at the point $x^* = (1, 1)^{\mathrm{T}}$. Moreover, it is necessary to note that this function does not satisfy the requirement of strict convexity. Since its function values are non-negative, $\nu \leqq 1$. We choose $\nu = 0{\cdot}1$ and besides, $\varepsilon = 10^{-6}$ and also $\eta = 10^{-6}$.

The calculations for all five alternatives of the MCC method were done using a pocket minicalculator SHARP PC 1500 A, as well as for the BFS and DFP methods in nine positions of the support point x_0. Table 1 shows the results, where the T line indicates the calculation of time T, N line the number of steps N.

Table 1.

Method		Position of support point x_0									Total
	x_{02}	1	1	1	2	2	2	3	3	33	
	x_{02}	1	2	3	1	2	3	1	2	3	
DFP	T	27	479	40	28	32	27	49	34	27	743
	N	4	46	8	4	5	4	8	6	5	90
	i	1	3	1	1	1	1	2	1	1	x_i^*
BFS	T	27	106	37	29	50	38	40	31	30	388
	N	4	16	7	4	6	7	7	5	5	61
	i	1	3	1	1	3	1	1	1	1	x_i^*
MCC 1	T	18	15	38	18	12	15	16	16	24	172
	N	6	6	24	8	4	7	6	6	15	82
	i	1	1	1	1	1	1	1	1	1	x_i^*
MCC 2	T	18	17	35	20	19	17	17	16	20	179
	N	6	7	21	9	9	8	7	5	11	83
	i	1	1	1	1	1	1	1	1	1	x_i^*

In the second example, for the same values of constants ε, ν and η and again from nine support points positions the extremization of Eason and Fenton function [5] was done using the BFS and DFP methods and the first two alternatives of the MCC method

$$f(x_1, x_2) = \frac{1}{10}\left[12 + x_1^2 + \frac{1 + x_2^2}{x_1^2} + \frac{x_1^2 x_2^2 + 100}{x_1^4 x_2^4}\right] \tag{79}$$

with the following four local minima: $x_1^* = (1{\cdot}74345,\ 2.02969)^{\mathrm{T}}$, $x_2^* = (1{\cdot}74345,$

$-2{\cdot}02969)^{\mathrm{T}}$, $x_3^* = (-1{\cdot}74345, 2{\cdot}02969)^{\mathrm{T}}$ and $x_4^* = (-1{\cdot}74345, -2{\cdot}02969)^{\mathrm{T}}$. Results are given in Table 2, which in addition to T and N contain also a minimum's index to which the process converged.

Table 2.

Method		Position of support point x_0									Total
	x_{01}	−4	−4	−4	0	0	0	4	4	4	
	x_{02}	−4	0	4	−4	0	4	−4	0	4	
DFP	T	30	46	21	43	64	26	23	27	37	317
	N	3	5	2	5	7	3	2	3	4	34
BFS	T	30	43	21	42	62	25	24	26	36	309
	N	3	5	2	5	7	3	2	3	4	34
MCC 1	T	11	34	10	30	31	20	11	19	16	182
	N	5	15	4	18	12	10	5	9	9	77
MCC 2	T	10	39	10	34	28	20	12	23	17	193
	N	4	13	4	22	11	8	5	12	9	88
Extreme	i	1	1	2	1	3	2	4	4	3	x_i^*

The third objective function used to compare properties of the first two alternatives of the MCC method with the BFS and DFP methods is the function designed by Himmelblau [5]

$$f(x_1, x_2) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2\,. \tag{80}$$

This function also has four minima at the points: $x_1^* = (-3{\cdot}77931, -3{\cdot}28319)^{\mathrm{T}}$, $x_2^* = (-2{\cdot}80512, 3{\cdot}13131)^{\mathrm{T}}$, $x_3^* = (3, 2)^{\mathrm{T}}$ and $x_4^* = (3{\cdot}58443, -1{\cdot}84813)^{\mathrm{T}}$. The same as above applies to this example with the only exception: in accordance with all four methods the iteration process led from one support point to the same extreme point. That is why Table 3 differs in this sense from Table 2.

7. EVALUATION OF RESULTS AND CONCLUSION

Since it is impossible to make any generalizing conclusions on the basis of 27 examples and their numerical results, consequently, the properties of the MCC method compared to the BFS and DFP methods cannot be evaluated objectively, either.

But the results presented indicate certain relations between the calculation time T and number of iteration cycles N for individual algorithms. Though in comparison with the BFS and DFP methods the MCC method shows approximately a double

Table 3.

Formula		Method									Total
		Position of support point x_0									
	x_{01}	0	0	2	2	−1	−1·2	−1	−1·2	−1·1	
	x_{02}	0	2	0	2	1	1	1·2	1·2	1·1	
		DFP									
(13)	T	91	106	96	72	144	146	145	159	171	1130
	N	12	13	11	9	19	19	19	20	20	142
		BFS									
(14)	T	92	98	85	73	133	144	139	143	136	1043
	N	12	13	11	9	19	20	19	20	19	142
		MCC 1									
(63)	T	40	43	35	30	64	62	68	70	70	482
	N	36	39	27	25	55	50	56	57	64	409
		MCC 2									
(64)	T	43	39	35	35	80	109	77	81	62	561
	N	38	33	25	29	72	100	70	75	53	495
		MCC 3									
(65)	T	53	52	39	41	72	128	86	71	81	623
	N	47	42	32	28	65	106	73	58	74	525
		MCC 4									
(12) + (67)	T	44	51	38	26	78	75	85	68	66	559
	N	36	42	29	19	68	63	72	57	55	441
		MCC 5									
(12) + (68)	T	41	46	36	41	68	88	87	73	81	561
	N	31	37	27	33	52	69	68	53	63	433

amount of iteration cycles, it is important that the overall calculation time is reduced even to a half.

The first of the five alternatives of the MCC methods proves to have the best properties. The second one follows with some distance and then the remaining alternative follows not only from the point of view of calculations requirements but also from the viewpoint of the overall number of steps and last but not least from aspect of the smallest sensitivity to the parameter v value.

Finally, it is useful to remark that if the one-dimensional search begins with a fixed choice of the parameter t so that formula (73) is omitted, then with an analytically specified gradient of the objective function the MCC method in course of the entire iteration process does not require the evaluation of the objective function.

(Received January 23, 1987.)

REFERENCES

[1] M. Aoki: Introduction to Optimization Techniques—Fundamentals and Applications of Nonlinear Programming. Macmillan, New York 1971. Russian translation: Nauka, Moscow 1977.

[2] P. E. Gill, W. Murray and M. H. Wright: Practical Optimization. Academic Press, New York 1981. Russian translation: Mir, Moscow 1985.

[3] M. Maňas: Optimization Methods (in Czech). SNTL, Praha 1979.

[4] B. T. Polyak: Vvedenie v optimizaciju. Nauka, Moscow 1983.

[5] G. V. Reklaitis, A. Ravindran and K. M. Ragsdell: Engineering Optimization—Methods and Applications. J. Wiley and Sons, New York 1983. Russian translation: Mir, Moscow 1986.

[6] A. G. Sukharev, A. V. Timokhov and V. V. Fedorov: Kurs metodov optimizacii. Nauka, Moscow 1986.

Doc. Ing. Peter Hudzovič, CSc., Katedra automatizovaných systémov riadenia technologických procesov, Elektrotechnická fakulta SVŠT (Department of Automated Control Systems for Technological Processes, Faculty of Electrical Engineering—Slovak Technical University), 812 19 Bratislava, Mlynská dolina. Czechoslovakia.