Journal of the Nigerian Association of Mathematical Physics
Volume 10 (November 2006), **363 - 370**
© J of NAMP

Jacobian approach to optimal determination of perturbation parameter for gradient method

J. O. Omolehin*, K. Rauf*, B. Opawoye*, and W. B. Yahya†
•Mathematics Department, University of Ilorin, Ilorin, Nigeria
†Statistics Department University of Ilorin, Ilorin, Nigeria
.

Abstract

In this work, the optimal determination of the perturbation factor (λ) or perturbation parameter for gradient method is considered. The spectrum analysis of the associated Jacobian of the associated matrix has laid the basis for the judicious selection of the perturbation factor. Numerical work is carried out to prove our hypothesis.

Keywords: Hessian; Matrix; Gradient; Spectrum; Minimize
C.R. Categories: G.1.7

1.0 Introduction

Let us consider the quadratic function $f : IR^n \to R$ which is continuously differentiable in some domain $D\underline{C}IR^n$ and it is assumed that f assumes a local minimum value in D at a point $x \in D_o$ where D_o is the interior of D .Now considering the Taylor's series expansion [3]. $f\left(x-\lambda\frac{\partial f}{\partial x}(x)^T\right)= f(x)+\frac{\partial f}{\partial x}(x)\left(-\lambda\frac{\partial f}{\partial x}(x)^T\right)+\theta(\lambda)$

$= f(x)-\lambda\left\|\frac{\partial f}{\partial x}(x)\right\|_e^2+\theta(\lambda)$. Where e is the usual Euclidian space, $\|\cdot\|$ is a norm in e. It is assumed that $\frac{\partial f}{\partial x}(x)\neq 0,$ then for sufficiently small $\lambda > 0$ we have $f\left(x-\lambda\frac{\partial f}{\partial x}(x)^T\right) < f(x)$. Hence, the gradient method (GM) is defined as the construction of sequence x_k of points in IR^n by the recursion equation.

$x_{k+1} = x_k - \lambda\frac{\partial f}{\partial x}(x)^T, k = 0,1,2,\cdots$ We make an initial guessed value for $x_o[3, pg,345-293]$. The convergence rate of GM has been extensively considered but the associated problem with the convergence rate of GM is not stable for the problems considered in [3]. Different values of λ show different types of convergence profiles. There could be a situation where by one of the components of the vector might be converging consistently and rapidly, the other components might not show any pattern of convergence at all. These are the problems we are set to solve with a view to finding the optimal parameter λ that will make the convergence rate of the method more stable. The spectrum analysis of the control operator A in the Conjugate Gradient Method (CGM) algorithm due to Ibiejugba [2] will be employed in this study.

2.0 Conjugate Gradient method (CGM)

The conventional CGM was originally developed by Hestenes and Stiefel [4] and it was used for quadratic minimization. To this end we define quadratic functional as:

$f(x)= f_o + \langle a, x\rangle_H + \frac{1}{2}\langle x, Ax\rangle_H$ Where A is an nxn symmetric, positive definite operator on the Hilbert space H , and a is vector in H . The steps involved in CGM algorithm are described as follows if $H \equiv IR^n$

Step 1:

The first element $x_o \in H$ of the sequence descent sequence is guessed, while the remaining members of the sequence are computed with the aid of the formulae.

Step 2
$p_o = -g_o = -(a + Ax_o)$
(p_o is the descent direction;
g_o is the gradient of $f(x)$ when $x = x_o$)
Step 3:
$x_{i+1} = x_i + \alpha_i p_i$,
$\alpha_i = \dfrac{\langle g_i, g_i \rangle_H}{\langle p_i, AP_i \rangle_H}$,
α is the step length.
$g_{i+1} = g_i + \alpha_i AP_i$,
$P_{i+1} = -g_{i+1} + \beta_i P_i$, $\beta i = \dfrac{\langle g_{i+1}, g_{i+1} \rangle_H}{\langle g_i, g_i \rangle_H}$
Step 4: if $g_i = 0$, for some i , terminate the sequence, else set
$i = i + 1$
And go to step 3.

3.0 Spectrum analysis of GM for quadratic functional

***Theorem*: 3.1**

The convergence rate of GM algorithm for quadratic functional remains stable if $\lambda = \dfrac{m}{M}$ *where* m *and* M *are the smallest and largest eigen values of the control operator A respectively.*

Proof:

Recall the problem of the minimization of
$f : IR^n \to R$ Given by

$$f(x) = f_o + \langle a, x \rangle_H + \frac{1}{2}\langle x, Ax \rangle_H \qquad (3.1)$$

where A is an nxn symmetric positive definite matrix and H is a Hilbert space. In our own case here $H \equiv R^n$. Differentiate (1.1) to obtain

$$\frac{\partial f(x)}{\partial x} = a + Ax \qquad (3.2)$$

Consider

$$x_{k+1} = K_k - \lambda \frac{\partial f(x)^T}{\partial x} \qquad (3.3)$$

Substitute equation (3.2) in (3.3), to obtain

$$x_{k+1} = x_k - \lambda A x_k - \lambda a = (I - \lambda A)x_k - \overline{\lambda} \text{ (} \lambda a = \overline{\lambda} \text{ Constant).} \qquad (3.4)$$

It has been established [6] that there exists an orthogonal matrix p which diagonalizes A, i.e.

$$P^{-1}AP = P_T \quad AP = \text{diag}(\lambda, \lambda, ..., \lambda_n) \qquad (3.5)$$

where P^{-1} and P^T are inverse and transpose of P respectively.

$$x_k = p_{yk} , k = 0,1,2,3.\cdots \qquad (3.6)$$

Therefore

$$P_{k+1} = (I - \lambda A)p_{yk} \qquad (3.7)$$

(Without loss of generality the constant $\overline{\lambda}$ can be dropped.

$$y_{k+1} = \left(I - \lambda p A P^{-1}\right) yk = \text{ diag}\left(1-\lambda\lambda_1, -\lambda\lambda_2, 1-\lambda\lambda_3, \cdots 1-\lambda\lambda_n y_n\right) \qquad (3.8)$$

$$y^j{}_{k+1} = \left(1-\lambda\lambda_j\right)^k y_o{}^j, \ j = 1,2,\cdots n \qquad (3.9)$$

Thus
$$y^j{}_k = \left(1-\lambda\lambda_j\right)^k y_o{}^j \quad j = 1,2,\cdots.n \qquad (3.10)$$

Therefore $y_k{}^j$ form a geometric progression $y_o{}^j, \left(1-\lambda\lambda_j\right) y_o{}^j, \left(1-\lambda\lambda_j\right)^2 y_o{}^j, \cdots$
The rate at which the numbers approach the minimum is dependent on

$$\left|1-\lambda\lambda_j\right| \qquad (3.11)$$

Ibiejugba et al. established the convergence rate of CGM as $\dfrac{E(x_n)}{E(x_o)} \le \dfrac{1-\dfrac{m}{M}}{1+\dfrac{m}{M}},$ (3.12)

Where m and M smallest and largest eigen values of matrix A respectively. Equate RHS of (12) to (3.11) and simplify. Therefore, we have

$$\frac{1-\dfrac{m}{M}}{1+\dfrac{m}{M}} = 1-\lambda\lambda_j$$

$$\frac{M-m}{M+m} - 1 = -\lambda\lambda_j \qquad (3.13)$$

$$\frac{2m}{\lambda_j} = \lambda \qquad (3.14)$$

We take $\lambda_j = 2M$ (for some j) and equation (3.14) becomes

$$\lambda = \frac{m}{M} \qquad (3.15)$$

Therefore our theorem has been proved.

4.0 Convergence rate of CGM algorithm

To fully understand our numerical work reported in the next section it will be necessary to show the convergence rate of CGM algorithm [2].

Recall the quadratic functional

$f(X) = f_o + \langle a, X\rangle_H + \frac{1}{2}\langle X, AX\rangle_H$ Where f_o is constant, H is a Hilbert space, X is a nxn dimensional vector in *H*, a positive definite constant matrix operator.

***Theorem* 4.1**

The law of convergence of CGM algorithm is given as

$$E(X_n) = \left\{\frac{1-\dfrac{m}{M}}{1+\dfrac{m}{M}}\right\}^{2n} E(X_o),$$

where m and M are the smallest and largest eigen value of A respectively..

Proof

Define $E(X) = \frac{1}{2}\left\langle \left(X - X^*\right), A\left(X - X^*\right)\right\rangle_H$ therefore,

$$E(X)=\frac{1}{2}\langle (X-X^{*})A(X-X^{*})\rangle_{H}$$

$$=\frac{1}{2}\langle X+A^{-1}{}_{a},A(X+A^{-1}a)\rangle_{H}=\frac{1}{2}\langle X+A^{-1}a,AX+AA^{-1}a\rangle$$

$$+\frac{1}{2}\langle A^{-1}a,AX\rangle_{H}+\frac{1}{2}\langle A^{-1}a,a\rangle_{H}$$

$$=\frac{1}{2}\langle X+A^{-1}a,AX+a,AX+a\rangle_{H}$$

$$=\frac{1}{2}\langle X,AX\rangle_{H}+\frac{1}{2}\langle X,a\rangle_{H}$$

$$=\frac{1}{2}\langle X,AX\rangle_{H}+\frac{1}{2}\langle X,a\rangle_{H}+\frac{1}{2}\langle A^{-1}a,AX\rangle_{H}+\frac{1}{2}\langle -X^{*},a\rangle_{H}$$

$$=\frac{1}{2}\langle X,AX\rangle_{H}+\frac{1}{2}\langle X,a\rangle_{H}+\frac{1}{2}\langle X^{*}AX^{*},\rangle_{H}+\frac{1}{2}\langle A^{-1}a,AX\rangle_{H}$$

$$E(x)=F(X)-F_{o}+\frac{1}{2}\langle X^{*},AX^{*}\rangle_{H}$$

$$=F(X)-\tilde{F}(o)$$

Therefore, $E(X)$ is $F(X)$ plus a constant term, hence the convergence of $E(X)$ is considered instead of that of $F(X)$ as from now Recall that,

$$E(X)=\frac{1}{2}\langle X+A^{-1}a,AX+a\rangle_{H}=\frac{1}{2}\langle A^{-1}(AX+a),AX+a\rangle_{H}=\frac{1}{2}\langle A^{-1}g(X),g(X)\rangle_{H}$$

Hence, $E(Xi)-E(X_{i+1})=\frac{1}{2}\langle X_{i}-X^{*},A(X_{i}-X^{*})\rangle_{H}-\frac{1}{2}\langle X_{i+1}-X^{*},A(X_{i+1}-X^{*})\rangle_{H}$

But

$X_{i+1}=X_{i}+\alpha_{i}p_{i}$, therefore

$$E(X_{i})-E(X_{i+1})=\frac{1}{2}\langle X_{i}-X^{*},A(X_{i}-X^{*})\rangle_{H}-\frac{1}{2}\langle X_{i}+X_{i}p_{i}-X^{*},A(Xi+\alpha_{i}p_{i}-X^{*})\rangle_{H}$$

$$=\frac{1}{2}\langle X_{i}-X^{*},A(X_{i}-X^{*})\rangle_{H}-\frac{1}{2}\langle X_{i}-X^{*},A(X_{i}-X^{*})\rangle_{H}$$

$$-\frac{1}{2}\alpha_{i}\langle p_{i},A(X_{i}+\alpha_{i}p_{i}-X^{*})\rangle_{H}-\frac{1}{2}\alpha_{i}\langle X_{i}-X^{*},Ap_{i}\rangle_{H}$$

$$=-\frac{\alpha_{i}}{2}\langle p_{i},A(X_{i}-X^{*})\rangle_{H}-\frac{1}{2}\alpha_{i}\langle X_{i}-X^{*},Ap_{i}\rangle_{H}-\frac{1}{2}\alpha_{i}\langle p_{i},A\alpha_{o}p_{o}\rangle_{H}$$

$$=-\alpha_{i}\langle p_{i},Ax_{i}+a\rangle_{H}-\frac{1}{2}\alpha_{i}{}^{2}\langle p_{i},Ap_{i}\rangle_{H}-\frac{1}{2}\alpha_{i}^{2}\langle p_{i},Ap_{i}\rangle_{H}$$

$$=-\alpha_{i}\langle p_{i},g_{i}\rangle_{H}-\frac{1}{2}\alpha_{i}\langle g_{i},g_{i}\rangle \quad \text{since } \alpha_{i}=\frac{\langle g_{i},g_{i}\rangle_{H}}{\langle p_{i},Ap_{i}\rangle_{H}}$$

$$=-\alpha_{i}\langle -g_{i}+\beta_{i-1}p_{i-1},g_{i}\rangle_{H}-\frac{1}{2}\alpha_{i}\langle g_{i},g_{i}\rangle_{H}=\alpha_{i}\langle g_{i},g_{i}\rangle_{H}-\alpha_{i}\beta_{i-1}\langle p_{i-1},g_{i}\rangle_{H}-\frac{1}{2}\alpha_{i}\langle g_{i},g_{i}\rangle_{H}$$

$=\alpha_{i}\langle g_{i},g_{i}\rangle-\frac{1}{2}\alpha_{i}\langle g_{i},g_{i}\rangle_{H}$, (since $\langle p_{i-1},g_{i}\rangle_{H}=0$, orthogonality of p_{i-1} and g_{i})

$$=\alpha_{i}\langle g_{i},g_{i}\rangle_{H}-\frac{1}{2}\alpha_{i}\langle g_{i},g_{i}\rangle_{H}=\frac{1}{2}\alpha_{i}\langle g_{i},g_{i}\rangle_{H}$$

$=\dfrac{\frac{1}{2}\langle g_{i},g_{i}\rangle^{2}{}_{H}}{\langle p_{i},Ap_{i}\rangle_{H}}$ because $\alpha_{i}=\dfrac{\langle g_{i},g_{i}\rangle}{\langle p_{i},Ap_{i}\rangle}$. Hence

$$E(X_i)-E(X_{i+1})=\frac{\langle g_i,g_i\rangle^2{}_H\,E(X_i)}{\langle p_i,Ap_i\rangle_H\langle g_i,A^{-1}g_i\rangle_H}$$

Using the fact that $g_i=\beta_{i-1}p_{i-1}-p_i$ we get

$$\langle g_i,Ag_i\rangle_H=\langle\beta_{i-1}p_{i-1},A(\beta_{i-1}p_{i-1}-p_i)\rangle_H$$

$$=\beta_{i-1}{}^2\langle p_{i-1},Ap_{i-1}\rangle_H+\langle p_i,Ap_i\rangle_H$$

$\geq\langle p_i,Ap_i\rangle_H$, since $\langle p_{i-1},Ap_{i-1}\rangle_H\geq 0$ (due to the positive definiteness of operator A),

$$\langle g_i,Ag_i\rangle_H\geq\langle p_i,Ap_i\rangle_H$$

Therefore

$$E(X_i)-E(X_{i+1})\geq\frac{\langle g_i,g_i\rangle^2{}_H\,E(X_i)}{\langle g_i,Ag_i\rangle_H\langle g_i,A^{-1}g_i\rangle_H}$$

But for a bounded self-adjoint operator in a Hilbert space H, Kantorovich established the following inequality

$\frac{\langle X,X\rangle^2{}_H}{\langle X,AX\rangle_H\langle X,A^{-1}\rangle_H}\geq\frac{4mM}{(m+M)^2}$, where m and M are respectively the greatest lower and least upper bounds of the spectrum of operator A. using Kantorovich's inequality we obtain

$$E(X_{i+1})\leq\left\{\frac{1-\frac{m}{M}}{1+\frac{m}{M}}\right\}^{2n}E(X_o)$$

This shows the convergence rate of the CGM. {N.B $m\leq M$ }

In this case A is a matrix operator where m is the smallest eigen value and M is the greatest eigen value. Note that the convergence rate is just reported here for proper understanding of the paper, it is not our work. Our own work is reported in main result.

Having reported the convergence rate of CGM algorithm it is the values of m and M we are going to consider. Observe that the convergence rate of CGM algorithm shows that the method will converge in at most n iterations. However, this convergence rate depends on the initial value. In our work, we determine the Hessian matrix of the problem and subsequently determined the eigen values of the matrix operator. These eigen values together with the initial values will be used in our numerical example for gradient method.

The numerical work and the analysis of our results will be reported in the next section.

5.0 Main result

Let us consider a particular problem. Observe that maximization of f is the minimization of $-f$.

Problem:

Maximize $f(x,y)=\frac{x}{1+x+x^2}+\frac{\left(y-\frac{y^2}{20}\right)\left(x+\frac{1}{2}\right)}{1+x+x^2}$

(The exact solution to this problem is $x=0.5$ and $y=10$)

Solution: let us compute the Hessian matrix as follows:

$$H=\begin{pmatrix}\frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x\partial y}\\ \frac{\partial^2 f}{\partial y\partial x} & \frac{\partial^2 f}{\partial y^2}\end{pmatrix}=\begin{pmatrix}H_{11} & H_{12}\\ H_{21} & H_{22}\end{pmatrix},\ \frac{\partial f}{\partial y}=\frac{\left(x+\frac{1}{2}\right)\left(1-\frac{y}{10}\right)}{1+x+x^2},$$

$$\frac{\partial f}{\partial x}=\frac{(1-x^2)+\left(y-\frac{y^2}{20}\right)\left(\frac{1}{2}-x-x^2\right)}{(1+x+x^2)^2}$$

$$H_{12} = \frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x} = \frac{1}{1+x+x^2}\left[1-\frac{y}{10}\right] - \frac{2x+1}{\left(1+x+x^2\right)^2}\left[\left(x+\frac{1}{2}\right)\left(1-\frac{y}{10}\right)\right] = H_{21}\,, \quad \frac{\partial^2 f}{\partial y^2} = -\frac{1}{10}\left[\frac{x+\frac{1}{2}}{1+x+x^2}\right] = H_{22}$$

$$\frac{\partial^2 f}{\partial x^2} = -\frac{(1+2x)}{1+x+x^2}\left[1+y-\frac{y^2}{10}\right] - \frac{2x+1}{\left[1+x+x^2\right]^2}\left[1+\left(y-\frac{y^2}{20}\right)\right] - \left[x+\left(y-\frac{y^2}{20}\right)\left(x+\frac{1}{2}\right)\right]\left[\frac{-6\left(x+x^2\right)}{\left(1+x+x^2\right)^3}\right] = H_{11}$$

To determine the eigen value we proceed as follows:

Let $H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}$,

$$H_\lambda = \begin{vmatrix} H_{11}-\lambda & H_{12} \\ H_{21} & H_{22}-\lambda \end{vmatrix} = 0$$

$$\Rightarrow (H1_1 - \lambda)(H_{22} - \lambda) - H_{21}H_{12} = 0$$

$$H_{11}H_{22} - \lambda(H_{11} + H_{22}) + \lambda^2 - H_{21}H_{12} = 0, \ \lambda^2 - \lambda(H_{11} + H_{22}) + H_{11}H_{22} - H_{21}H_{12} = 0$$

Let a =1, b $= -(H_{11} + H_{22})$,

c $= H_{11}H_{22} - H_{21}H_{12}$

Therefore $a\lambda^2 + b\lambda + c = 0$

We can now apply quadratic formula to solve for λ. .A program was written in FORTRAN for that purpose.

In the next section we are going to report and analyze of our numerical result.

6.0 Numerical Example

We now consider some numerical examples. We are going to focus on the role played by the choice of the initial values and the value of λ . The convergence profile of our minimizing vectors will be given in the table that follows.

Problem:

Maximize $f(x,y) = \dfrac{x}{1+x+x^2} + \dfrac{\left(y-\frac{y^2}{20}\right)\left(x+\frac{1}{2}\right)}{1+x+x^2}$ (Analytic solution is given as $x = 0.5, y = 10$)

The numerical results are tabulated in table.

7.0 Analysis of numerical results

Since the convergence rate of the Gradient Method depends on the parameter λ and the initial values for updating of our minimizing vector, it is imperative that research should focus emphasis on the optimal selection of these parameters.

In Tables 1 and 2 in the Table of results, we have the same initial values but different values of λ. When λ = 1.0, the values for x show no sign of convergence but the exact value of y was obtained at the 155th iteration. We now make the value of λ = 0.5 and the exact value of y was obtained at the 261st iteration and the approximate value for x is 0.453 instead of 0.5.

Table of result: Summary of convergence profile

Table	**Initial values**	**Minimizing vector**	**Iteration number**
1	$x_o = 1.0$ $y_o = 0.5$ $\lambda = 1.0$	$x =\ -$ $y = 10.0$	155 155

2	$x_o = 1.0$ $y_o = 0.5$ $\lambda = 0.5$	$x = 0.453$ $y = 9.99$	261 261

Table	Initial values	Minimizing vector	Iteration number
3	$x_o = 1.0$ $y_o = 0.5$ variable $\lambda =$ 48.9514	$x = -$ $y = 10.0$	160 160
4	$x_o = 0.6$ $y_o = 1.045499$ variable $\lambda = 0.1228$	$x = 0.46$ $y = 10.0$	91 91
5	$x_o = 0.6$ $y_o = 1.045499$,variable $\lambda = 0.5$	$x = 0.453$ $y = 9.99$	106 221

We now calculated the eigen values of the associated Hessian matrix of the problem considered. At each circle of iteration, the value of λ is re calculated through the Hessian matrix operator. In Table 3 of Table of results with initial values $x_o = 1.0$ and $y_o = 0.5$ with judiciously varying λ at each iteration, the exact value of y was obtained at 160th iteration with $\lambda = 48.9514$. However, the convergence of x is discouragingly slow.

When our initial guess is $x_o = 0.6$ and $y_o = 1.045499$, the exact value of $x = 0.5$ and approximate value of $y = 9.99$ was obtained at the 75th iteration. The exact value of $y = 10$ was obtained at the 91st iteration and the approximate value for $x = 0.46$ they are as shown in Table 4. If we look at Table 5 in the Table of results, it is observed that both x and y converge but none of them give the exact value. Their values are $x = 0.453$ and $y = 9.99$ they are acceptable. The value of λ is 0.5 and the values of the initial values are $x_o = 0.6$ and $y_0 = 1.045499$. This is an interesting result and it is better than Russell's [3] results.

8.0 Conclusion

We have succeeded in making the convergence rate of our method consistent than Rushell [3] approach using algorithm generated by Ibiejugba in [2]. We observed here that the calculation of our perturbation term (λ) is still a constant at every iteration. It does not take into consideration the individual component of the vector. Also it does not address the minimizing vector at each iteration. All these identified shortcomings will be addressed in our next work.

Bu and large, this is an interesting result and it is better than Rushell's [3] results because the consistency of the convergence rate is more stable.

References

[1] Omolehin, J.O. - Computational convergence rate of Gradient Method, NJMA. (Accepted), (1996).

[2] Ibiejugba, M.A. – Computational methods in optimization, Ph.D. Thesis, University of Leeds, Leeds, U.K., (1985).

[3] Russel, David, L. – Optimization theory, New York, W.A. Benjamin, Inc., (1970).

[4] Omolehin, J.O. – On the control of reaction diffusion equation, PhD. Thesis, University of Ilorin, Ilorin, (1991).

[5] Omolehin, J.O. – Experiment with extended conjugate gradient method algorithm, M.sc.

[6] Omolehin, J.O. – Eigen value perturbation for gradient method, Anaele Stiintifice Ale Universitata “al,I.Cuza,Tomul L1,s.I.Matematical,f.1,(2005).