Journal of the Nigerian Association of Mathematical Physics
Volume **12 (May, 2008), 353 - 358**

General convergence analysis of a modified steepest descent method.

[1]Dickson E. Ativie Omorogbe and [2]A. A Osagiede,
[1]Institute of Education
University of Benin, Benin City, Nigeria.
[2]Department of Mathematics
Faculty of Physical Sciences
University of Benin, Benin City, Nigeria.

Abstract

In this paper, we consider a modified steepest method and its convergence Analysis. The application of the eigen vectors and eigenvalues were properly used in the analysis. The more ill condition the matrix (i.e, the larger its condition number), the slower the convergence rate of the modified steepest descent method. In the same vein, the smaller the condition number K, the faster the convergence rate. Convergence of the modified steepest descent method is instantaneous if the eigenvalues are equal, are some of the results drawn from this work. However, it is concluded that if the condition number k is small convergence of the modified steepest descent method is quick irrespective of the starting point unlike the conventional method which is not amenable to such analysis.

1.0 Introduction

Steepest Descent method searches in the direction of the negative gradient for a minimum. The reason for this search direction is not far fetched since the, error function decreases most rapidly in this direction. This is logical as the search must continue for a non-zero distance. The search will only be in the direction of the negative gradient for a small distance. According to Ibiejugba (1999 [1]), the steepest descent method basically consists of two interlocking component parts, first a choice of direction in which to move followed by a minimization over a line in the selected direction. It is conceivable when $X \in R^n$ the application of the steepest descent method necessitates an infinite number of steps to actually achieve the desired minimum, though only a finite number of steps are required to solve:

$$Ax = f,\ x \in X \in R^n$$

where *A* is a densely defined symmetric operator on a Hilbert space *H.* According to Omorogbe et al (2006 [2]), this phenomenon is attributable to an asymptotic restriction of the steepest descent directions to only a two-dimensional subspace; thus, residuals are said to have failed in searching the whole space, H. In this work a modified steepest descent method which attempt to obviate the computational deficiency of the conventional steepest descent methods is considered together with its convergence analysis.

2.0 The conventional steepest descent method

The conventional steepest descent method is contained in the following algorithm.

e-mail: dickomorogbe@yahoo.com

2.1 Algorithm

It is assumed that an estimate $x^{(0)}$ of a minimized x^* of f is known.

a. Set $k = 0$ (2.1)

b. Compute $P^{(k)}$ from $P^{(k)} = -g(X^{(k)})$ (2.2)

c. Compute $\lambda^{(k)}$ such that $F(X^{(k)} + \lambda^{(k)} P^{(k)}) = \min f(X^{(k)} + \lambda P^{(k)})$ (2.3)

d. Compute $X^{(k+1)}$ from $X^{(k+1)} = X^{(k)} + \lambda^{(k)} P^{(k)}$ (2.4)

e. set $k = k+1$ and Go To 2 (2.5)

According Wolfe (1978 [5]), to understand, intuitively how this algorithm works, we consider an objective function $f : R^2 \to R^1$ with contours near a minimizer x^*, this is illustrated in Figure 2.1 below. For such an object function it is clear that there are only two descent directions, namely $-g^{(0)}$ and $-g^{(1)}$ is orthogonal to $g^{(0)}$. All other descent directions are parallel either to $-g^{(0)}$ or to $-g^{(1)}$. For an objective function with nearly circular contours near x^*, it would seen that Algorithm 2.1 would make rapid progress. For an object function with contours as shown in Figure 2.2 however it is clear that Algorithm 2.1 would soon cease to make effective progress after the first little iteration. The kind of subsequent progress to be expected is illustrated in Figure 2.2 by the zig-zag dotted lines. Very many short steps are needed for convergence to x^*. The type of contours in Figure 2.2 is more common in practice than the type shown in Figure 2.1, so it would seem that Algorithm 2.1 although appealingly simple, is not very efficient for practical problems. It is used in conjunction with certain more efficient methods because it is the precursor of all gradient methods and provides a useful insight into the nature of descent methods. (Shmuel, 1966 [4])

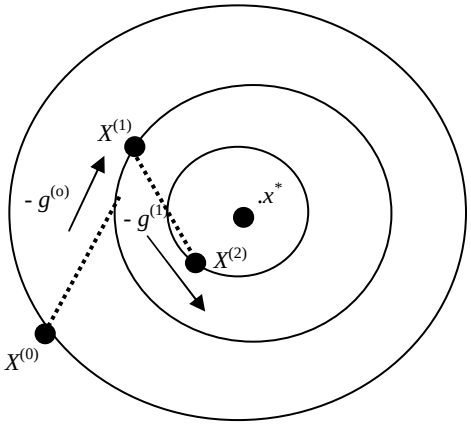

Figure 2.1

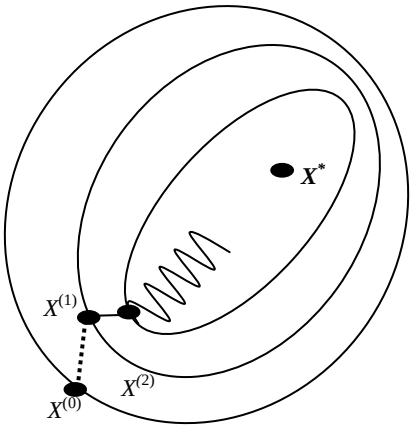

Figure 2.2

3.0 The modified steepest descent method

The modified steepest descent method is contained in the following algorithm

3.1 Algorithm

a. $$r_{(i)} = b - Ax_{(i)} \tag{3.1}$$

where $-f^1(X_{(i)}) = r_{(i)}$ is the direction opposite $f^1(X_{(i)})$

b. $$\alpha_{(i)} = \frac{r_{(i)}^T r_{(i)}}{r_{(i)}^T A r_{(i)}} \tag{3.2}$$

c. $$X_{(i+1)} = X_{(i)} + \alpha_{(i)} r_{(i)} \tag{3.3}$$

The algorithm, as written above, requires two matrix-vector multiplications per iteration. The computational cost of the modified steepest descent is dominated by matrix-vector products; fortunately, this can be eliminated, by premultiplying both sides of Equation (8) by –A and adding b, we have:

$$r_{(i+1)} = r_{(i)} - \alpha_{(i)} A r_{(i)} \tag{3.4}$$

Although equation (3.1) is still needed to compute $r_{(0)}$, then equation (3.4) can be used for every iteration thereafter. The product Ar, which occurs in both equations (3.2) and (3.4) need only be computed once. The disadvantage of using this recurrence is that the sequence defined by equation (3.4) is generated without any feedback from the value of $x_{(i)}$, so that accumulation of floating point round off error may cause $x_{(i)}$ to converge to some point near *x*.(Shewchuk 1994 [3]).

However, this effect can be avoided by periodically using equation (3.1) to recompute the correct residual.

To determine α as used in equation (3.2), note that

$$f^1(x_{(i)}) = -r_{(i)} \tag{3.5}$$

and we have:

$$r_{(I)}^T r_{(0)} = 0 \tag{3.6}$$

$$(b - A(x_{(i)}{}^T r_{(0)}) = 0 \tag{3.7}$$

$$(b - Ax_{(0)} + \alpha r_{(0)})^T r_{(0)} = 0 \tag{3.8}$$

$$(b - Ax_{(0)})^T r_{(0)} - \alpha(A r_{(0}^T r_{(0)} = 0 \tag{3.9}$$

$$(b - Ax_{(0)}{}^T r_{(0)} = \alpha(A r_{(0)}^T r_{(0)} \tag{3.10}$$

$$r_{(0)}^T r_{(0)} = \alpha r_{(0)}^T (A r_{(0)}) \tag{3.11}$$

$$\alpha = \frac{r_{(0}^T r_{(0)}}{r_{(0)}^T A r_{(0)}} \tag{3.12}$$

4.0 Convergence analysis of the modified steepest descent method

Let us consider the case where $e_{(i)}$ is an eigenvector with eigenvalue λ_e. Then, the residual

$$r_{(i)} = -Ae_{(i)} = -\lambda_{e(i)} \tag{4.1}$$

is also aneigenvector. Equation (3.3) gives

$$e_{(i+1)} = e_{(i)} + \frac{r_{(i)}^T r_{(i)}}{r_{(i)}^T A r_{(i)}} \tag{4.2}$$

$$= e_{(i)} + \frac{r_{(i)}^T r_{(i)}(-\lambda_e e_{(i)})}{\lambda_e r_{(i)}^T r_{(i)}} = 0 \qquad (4.3)$$

i.e. choosing $\alpha_{(i)} = \lambda_e^{-1}$ gives us instant convergence.

For a more general analysis, we must express $e_{(i)}$ as a linear combination of eigenvectors, and we shall further require these eigenvectors to be orthonormal. Symmetrically, there exists a set of n orthogonal eigenvectors of A, such that we can express the eigenvectors arbitrarily. Let us choose $\alpha_{(i)} = \lambda_e^{-1}$ so that each eigenvector is of unit length. This choice gives us the property that

$$V_j^T V_k = \begin{cases} 1, & j \neq k \\ 0, & j \neq k \end{cases}$$

Then we express the error term $e_{(i)}$ as a linear combination of the eigenvectors

$$e_{(i)} = \sum_{j=1}^{n} \xi_j V_j \qquad (4.4)$$

where ξ is the length of each component of $e_{(i)}$. From equations (4.3) and (4.4) we have the following identities:

$$r_{(i)} = -Ae_{(i)} - \sum_{j=1}^{n} \xi_j \lambda_j V_j \qquad (4.5)$$

$$\left\| e_{(i)} \right\|^2 = e_{(i)}^T e_{(i)} - \sum \xi_j^2 \qquad (4.6)$$

$$e_{(i)}^T A e_{(i)} = \left(\sum \xi_j V_j^T\right)\left(\sum \xi_j \lambda_j V_j\right) = \sum_j \xi_j^2 \lambda \qquad (4.7)$$

$$\left\| r_{(i)} \right\|^2 = r_{(i)}^T r_{(i)} = \sum \xi_j^2 \lambda_j^2 \qquad (4.8)$$

$$r_{(i)}^T A r_{(i)} = \sum \xi_j^2 \lambda_j^3 \qquad (4.9)$$

Equation (4.5) shows that $r_{(i)}$ can also be expressed as the sum of eigenvector Components, and the length of these components are $-\xi_j \lambda_j$. Equations (4.6) and (4.8) are just Pythagoras law.

We can proceed with the analysis as follows: equation (3.3) gives

$$e_{(i+1)} = e_{(i)} + \frac{r_{(i)}^T r_{(i)}}{r_{(i)}^T A r_{(i)}}(r_{(i)}) = e_{(i)} + \frac{\sum_j \xi_j^2 \lambda_j^2}{\sum_j \xi_j^2 \lambda_j^3} = (r_{(i)}) \qquad (4.10)$$

We saw earlier in this work that, If $e_{(i)}$ has only one eigenvector component, then convergences is achieved in one step by choosing $\alpha_{(i)} = \lambda_e^{-1}$. Let us consider the case where $e_{(i)}$ is arbitrary, but all the eigenvector have a common eigenvalue λ. Then, equation (4.10) becomes:

$$e_{(i+1)} = e_{(i)} + \frac{\lambda^2 \sum_j \xi_j^2}{\lambda^3 \sum_j \xi_j^2}(-\lambda e_{(i)}) = 0$$

Once again, there is instant convergence. This is because all the eigenvalues are equal, the ellipsoid is spherical; hence, no matter what point we start at, the residual must point to the centre of the sphere, this is demonstrated in Figure 4.1. However, if there are several unequal, nonzero eigenvalues, then no choice of $\alpha_{(i)}$ will eliminate all the eigenvector components, and our choice becomes a sort of compromising fact, the fraction in Equation (4.10) is best thought of as a weighted average of the values of λ_e^{-1}. The weight ξ_j^2 ensure that longer components of $e_{(i)}$ are given precedence. As a result, on any given iteration, some of the shorter components of $e_{(i)}$ might actually increase in length (though never for long).

5.0 General convergence.

To bound the convergence of the modified steepest descent method in the general case, we shall

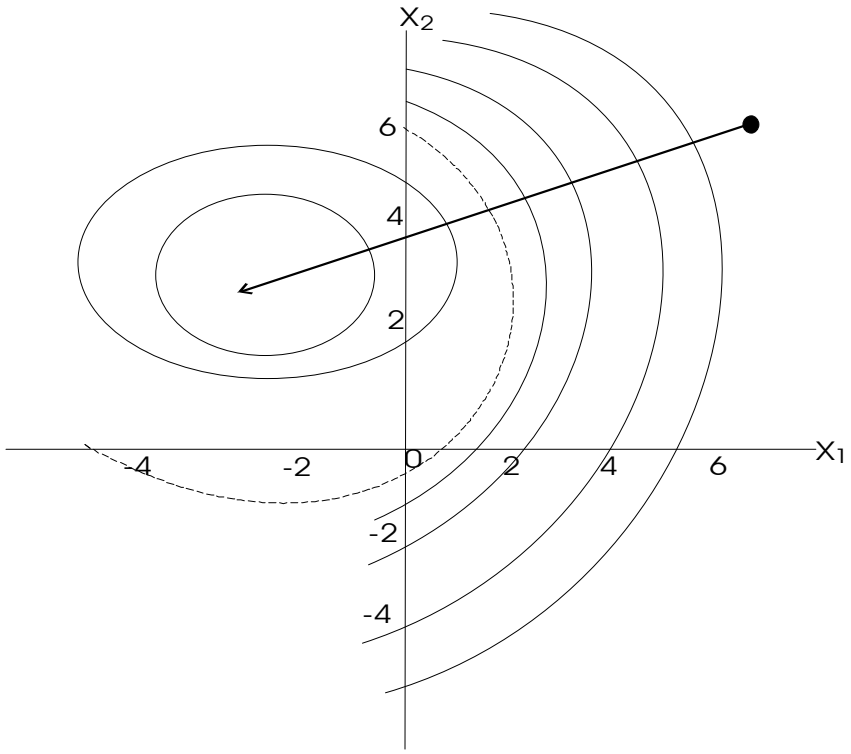

Figure 4.1: Steepest descent method converges to exact solution on the first iteration if the eigenvalues are all equal.

define the energy norm $\|e\|_A = (e^T Ae)^{1/2}$. This norm is easier to work with than the Euclidean norm, and is in some sense a more natural norm. Let us consider the slope because if at some arbitrary point *p* and at the solution point $x = A^{-1}b$ then,

$$F(p) = f(x) + \tfrac{1}{2}(p-x)^T A(p-x) \qquad (5.1)$$

If *A* is symmetric as well as positive definite, it follows that *x* is a global minimum of f. suppose we examine equation (5.1) shows that minimizing $\|e_{(i)}\|_A$ is equivalent to minimizing $f(x_{(i)})$ with this norm, we have

$$\|e_{(i+1)}\|_A^2 - e_{(i+1)}^T Ae_{(i+1)} = (e_{(i)}^T Ae_{(i)} + 2\alpha_{(i)} r^T A(e_{(i)} + \alpha_{(i)} r_{(i)} = e_{(i)}^T Ae_{(i)} + 2\alpha_{(i)} r_{(i)}^T Ae_{(i)} + \alpha_{(i)}^2 r_{(i)}^T Ar_{(i)}$$

(by symmetry of A) $$= \|e_{(i)}\|_A^2 + 2\frac{r_{(i)}^T r_{(i)}}{r_{(i)}^T Ar_{(i)}}(-r_{(i)}^T r_{(i)}) + \left(\frac{r_{(i)}^T r_{(i)}}{r_{(i)}^{T\,Ar_{(i)}}}\right)^2 r_{(i)}^T Ar_{(i)}$$

$$= \|e_{(i)}\|_A^2 - \frac{(r_{(i)}^T r_{(i)})^2}{r_{(i)}^T Ar_{(i)}} = \|e_{(i)}\|_A^2\left[\frac{1-(r_{(i)}^T r_{(i)})^2}{(r_{(i)}^T Ar_{(i)})(e_{(i)}^T Ae_{(i)})}\right] - \|e_{(i)}\|_A^2\left(\frac{1-\left(\sum_j \xi^2\lambda_j^2\right)^2}{\left(\sum_j \xi^2\lambda_j^3\right)\left(\sum_j \xi_j^2\lambda_j\right)}\right) = \|e_{(i)}\|_A^2$$

(by identities 4.7, 4.8 and 4.9). When $\omega^2 = 1 - \frac{\left(\sum_j \xi_j^2 \lambda_j^2\right)^2}{\left(\sum_j \xi_j^2 \lambda_j^3\right)\left(\sum_j \xi_j^2 \lambda_j\right)}$ (5.2)

The analysis depends on finding an upper bound for ω to demonstrate how the weight and eigenvalues affect convergence. We shall derive a result for $n = 2$, assume that $\lambda_1 > \lambda_2$. The spectral condition number of A is defined to be $K = \lambda_1 / \lambda_2 \geq 1$. The slope of $e_{(i)}$ which depends on the starting points, is denoted $\mu = \xi_2 / \xi_1$. We have $\omega = 1 - \frac{\left(\xi_1^2 \lambda_1^2 + \xi_2^2 \lambda_2^2\right)}{\left(\xi_1^2 \lambda_1 + \xi_2^2 \lambda_2\right)\left(\xi_1^2 \lambda_1^3 + \xi_2^2 \lambda_2^3\right)}$

$$= 1 - \frac{\left(k^2 + \mu^2\right)^2}{(k + -\mu^2)(k^3 + \mu^2)} \tag{5.3}$$

The value of ω, which determines the rate of convergence of the modified steepest descent is expressed as a function of μ and k. if $e_{(0)}$ is an eigenvector, then the slope μ is zero or infinite, when ω is zero then convergence is instant. If the eigenvalues are equal, then the condition number *k* is one, implied ω equals zero. The quadratic forms with a large condition number, the modified steepest descent can converge quickly if a fortunate starting point is chosen, but it is usually worst when *k* is large. However, if the condition number is small, the quadratic form is nearly spherical, and convergence is quick regardless of the starting point. Holding *k* constant (because *A* is fixed), a little basic calculus reveals that equation (5.3) is maximized when $\mu = \pm k$. An upper bound for ω which may be the worst starting point is found by setting $\mu^2 = k^2$, Such that $\omega^2 \leq 1 - \frac{4k^4}{k^5 + 2h^4 + k^3} = \frac{k^5 - 2k^4 + k^3}{k^5 + 2k^4 + k^3} - \frac{(k-1)^2}{(k+1)^2}$,

it follows that $$\omega \leq \frac{k-1}{k+1} \tag{5.4}$$

The more ill-condition the matrix (i.e, the larger its condition number *k*), the slower the convergence of the modified steepest descent method. It is proven by using Chebyshev polynomial that equation (5.4) is also valid for *n*>2, if the condition number of a symmetric, positive-definite matrix is defined to be:

$$K = \lambda_{\max} / \lambda_{\min}.$$

The ratio of the largest and smallest eigenvalues of *A*. The convergence results for steepest descent

are $$\left\|e_{(i)}\right\|_A \leq \left(\frac{k-1}{k+1}\right)^i \left\|e_{(0)}\right\|_A \tag{5.5}$$

and $$\frac{f(x_{(i)}) - f(x)}{f(x_{(0)}) - f(x)} = \frac{\frac{1}{2} e_{(i)}^T A e_{(i)}}{\frac{1}{2} e_{(0)}^T A e_{(0)}} \text{ (by equation 5.1)} \leq \left(\frac{k-1}{k+1}\right)^{2i}$$

6.0 Discussion of results and conclusion

Convergence of the modified steepest descent method (per iteration) worsens as the condition number of the matrix (i.e the larger its condition number *k*), the slower the convergence rate of the algorithm. Convergence is instant if eigenvalues are equal. However, if the condition number *k* is small, convergence is quick irrespective of the starting point. However, the analysis depends on finding an upper bound for ω. The value of ω is expressed as a function of μ and k, if $e_{(0)}$ is an eigen vector, then the slope μ is zero or infinite, when ω is zero convergence is instant. If the eigen values are equal, then the condition number *k* is one, implies ω equals zero. In the quadratic form with large condition number, the modified steepest descent method can converge quickly if a good starting point is chosen, but it is usually worst when *k* is large. If *k* is small the quadratic form is nearly spherical and convergence is quick regardless of starting point. From the foregoing, it is worthy of note that the conventional steepest descent method is not amenable to this analysis.

References

[1] Ibiejugba, M. A et al (1999) "Foundation postgraduate course on computational methods and application in optimization", *National mathematical centre, Abuja, Nigeria.*

[2] Omorogbe, Dickson A, Chukwendu, C. R, and Okonta, S. D (2006) "A super convergence of A modified Non-Classical conjugate Gradient method" *knowledge review* Vol. 13, No. 7 p. 100 - 106

[3] Shewchuk, J. R (1994) "A introduction to conjugate gradient method without Agonizing pain" carnegie mellon University Pittsburgh.

[4] Shmuel, kaniel (1996) "Estimate for some computational Techniques in Linear Algebra" *mathematics of computation 20 p. 369 – 378*

[5] Wolfe M. A (1978) "Numerical method for unconstrained optimization" van Nostrond Reinhold co. Ltd Bershire, England.