Chapter-4

A new initial search direction for nonlinear conjugate gradient method

Introduction:

The general nonlinear Conjugate Gradient method always considers critical initial search direction to find a convergent solution. In the present work our aim is to establish the fact that the convergent solution can also be obtained for nonlinear function, even if the initial search direction is different from the initial search direction taken in most of the algorithms of the nonlinear Conjugate Gradient method. In the present work we try to established the above mentioned fact theoretically.

Conjugate Gradient(C.G) methods comprise a class of unconstrained optimization algorithms which are characterised by low memory requirements and strong local and global convergence properties. In the seminal 1952 paper of Hestenes and Stiefel, the algorithm is presented as an approach to solve symmetric, positive-definite linear systems.

A non-linear unconstrained optimization problem can be stated as

$$\min\{ f(x) : x \in R^n \} \tag{4.1}$$

Where $f : R^n \to R$ is a continuously differentiable function especially if the dimension is very large.

The Conjugate Gradient method to solve the general nonlinear problem defined by (4.1) is of the form

$$x_{k+1} = x_k + \alpha_k d_k \tag{4.2}$$

where α_k is a step size obtained by a line search and d_k is the search direction obtained by

$$d_k = \begin{cases} -g_k, k = 1 \\ -g_k + \beta_k d_{k-1}, k \geq 2 \end{cases} \tag{4.3}$$

Where β_k is a parameter and g_k denotes $\nabla f(x_k)$ where the gradient $\nabla f(x_k)$ of f at x_k is a row vector and g_k is a column vector. Different C.G methods correspond to different choices for the scalar β_k.

It is known from (4.2) and (4.3) that only the step size α_k and the parameter β_k remain to be determined in the definition of Conjugate Gradient method. In this case that if f is a convex quadratic, the choice of β_k should be such that the method (4.2)-(4.3) reduces to the linear Conjugate Gradient method if the line search is exact namely

$$\alpha_k = \arg\min\{ f(x_k + \alpha d_k); \alpha > 0\} \tag{4.4}$$

For non linear functions, different formulae for the parameter β_k result in different Conjugate Gradient methods and their properties can be significantly different. To differentiate the linear Conjugate Gradient method, sometimes we call the Conjugate Gradient method for unconstrained optimization by nonlinear Conjugate Gradient method. Meanwhile the parameter β_k is called Conjugate Gradient parameter. The equivalence of the linear system to the minimization problem of $\frac{1}{2} x^T Ax - b^T x$ Motivated Fletcher and Reeves to extend the linear Conjugate Gradient method for nonlinear optimization. This work of Fletcher and Reeves in 1964 not only opened the door of nonlinear C.G Field but greatly stimulated the study of nonlinear optimization.

In general the nonlinear Conjugate Gradient method without restarts is only linearly convergent(See Crowder and Wolfe[54]) while n-step quadratic convergence rate can be established if the method is restarted along the negative gradient every n-step.(See Cohen [55] and McCormick and Ritter[56])

In 1964 the method has been extended to nonlinear problems by Fletcher and Reeves [44], which is usually considered as the first nonlinear Conjugate Gradient algorithm. Since then a large number of variations of Conjugate Gradient algorithms have been suggested. A survey on their definition including 40 nonlinear Conjugate Gradient algorithms for unconstrained optimization is given by Andrei[73].Since the exact line search is usually expensive and impractical, the strong Wolfe line search is often consider the implementation of the nonlinear Conjugate Gradient methods .It aims to find a step size satisfying the strong Wolfe conditions.

$$f(x_k+\alpha_k d_k)-f(x_k)\le \rho\alpha_k g_k^T d_k \qquad (4.5)$$

$$\left|g(x_k+\alpha_k d_k)^T\right|\le -\sigma g_k^T d_k \qquad (4.6)$$

Where $0<\rho<\sigma<1$

The strong Wolfe line search is often regarded as a suitable extension of the exact line search since it reduces to the latter. If σ is equal to zero, in practical computation a typical choice for σ that controls the inexactness of the line search is σ=0.1.On the other hand general non linear function ,one may be satisfy with a step size satisfying the standard wolf conditions ,namely and

$$g(x_k+\alpha_k d_k)^T d_k \ge \sigma g_k^T d_k \qquad (4.7)$$

$$0<\rho<\sigma<1.$$

As is well known the standard Wolf line search is normally used in the implementation of Quasi-Newton methods, another important class of methods for unconstrained optimization. The work of Dai and Yuan indicates that the use of standard Wolfe line search is possible in the nonlinear Conjugate Gradient field. A requirement for an optimization method to use the above line searches is that, the search direction d_k must have descent property namely

$$g_k^T d_k<0 \qquad (4.8)$$

For Conjugate Gradient method, by multiplying (4.3) with g_k^T , we have

$$g_k^T d_k=-\| g_k \|^2+\beta_k g_k^T d_{k-1}$$

Thus if the line search is exact, we have $g_k^T d_k=-\| g_k \|^2$ since $g_k^T d_{k-1}=0$.

Consequently d_k is descent provided $g_k\ne 0$.

In this work we assume that a Conjugate Gradient method is descent if (4..8) holds for all k and is sufficient descent if the sufficient descent condition

$$g_k^T d_k\le -c\| g_k \|^2$$

Holds for all k and some constant c>1.However we have to point out that the borderlines between these Conjugate Gradient methods are not strict.

Any Conjugate Gradient algorithm has very simple and the prototype of the general nonlinear Conjugate Gradient algorithm can illustrated as follows:

Step1: Select the initial stating point

$x_0 \in dom\, f$ and compute f_0 and $g_0 = \nabla f(x_0)$

$d_0 = -g_0 \; and \; k = 0$

Step2: Test a criterion for stopping the iteration.

For example, if $s \| g_k \|_\infty \leq \in$, then stop otherwise continue with step 3

Step3: Determine the step length $\alpha_k = \dfrac{g_k^T d_k}{d_k^T d_k}$

Step4: Update the variables as: $x_{k+1} = x_k + \alpha_k d_k$

Compute f_{k+1} and g_{k+1}

Compute $y_k = g_{k+1} - g_k$ and $s_k = x_{k+1} - x_k$

Step5: Determine β_k

Step6: Compute the search direction as: $d_{k+1} = -g_{k+1} + \beta_k s_k$

Step7: Restart criterion.

For example if the restart criterion of Powell $\left| g_{k+1}^T g_k \right| > 0.2 \| g_{k+1} \|^2$ is satisfied ,

then set $d_{k+1} = -g_{k+1}$

Step8: Compute the initial guess $\dfrac{\alpha_{k-1} \| d_{k-1} \|}{d_k}$, set $k = k+1$ and continue with step 2.

This is a prototype of the conjugate gradient algorithm but some more sophisticated variants are also known

These variants focus on parameter β_k computation and on the step length determination.

The objective of this present work is to search an initial search direction other than $d_0 = -g_0$ for general unconstrained nonlinear Conjugate Gradient method and to establish the fact that the general unconstrained nonlinear conjugate gradient algorithm can be used with this new search direction. The other objective of the present paper is to test the convergence of the Conjugate Gradient algorithm for this new search direction.

Already the Conjugate Gradient method has been devised by Magnus Hestenes (1906-1991) and Eduard Stiefel (1909-1978) in their seminal paper where an algorithm for solving symmetric, positive-definite linear algebraic systems has been presented and after a relatively short period of stagnation ,the paper by Reid brought the Conjugate Gradient method as a main active area of research in unconstrained optimization and later in 1964 the method has been extended to nonlinear problems by Fletcher and Reeves.

In the present work of this thesis, we will assume that the initial search direction for the nonlinear conjugate gradient algorithm is slightly deflect from the direction - g_0.

In our work, instead of $d_0 = -g_0$ we have taken the initial search direction as

$$d_0 = -g_0 + \gamma g_0 \qquad (4.9)$$

Where γ is very small and $0 < \gamma < 1$.The values of α_k can be obtained by (4.9) in $\alpha_k = \frac{g_k^T d_k}{d_k^T d_k}$

Therefore for $k = 0$ we have

$$\begin{aligned}\alpha_0 &= \frac{g_0^T d_0}{d_0^T d_0} \\ &= \frac{-g_0^T g_0}{(1-\gamma) g_0^T g_0} \\ &= -\frac{1}{1-\gamma}\end{aligned} \qquad (4.10)$$

Putting the value of α_k in (1.2) (chapter 1)

for $k = 1$

$$\begin{aligned} x &= x_0 + \frac{1}{1-\gamma}(1-\gamma) g_0 \\ &\Rightarrow x_1 - x_0 = g_0 \end{aligned} \qquad (4.11)$$

Therefore

$$d_1 = -g_1 + \beta_1 d_0$$
$$= -g_1 + \beta_1(1-\gamma)g_0$$
$$= -g_1 - \frac{\|g_1\|^2}{\|g_0\|^2}(1-\gamma)g_0$$
$$= -g_1 - Lg_0 \qquad (4.12)$$

Where $L = \frac{\|g_1\|^2}{\|g_0\|^2}(1-\gamma)$ (4.13)

Which gives

$$g_1^T d_1 = -\|g_1\|^2 - \frac{\|g_1\|^2}{\|g_0\|^2}(1-\gamma)g_1^T g_0 \le -c\|g_1\|^2 \qquad (4.14)$$

Therefore

$$\alpha_1 = \frac{g_1^T d_1}{d_1^T d_1}$$
$$= \frac{\|g_1\|^2 - Lg_1^T g_0}{\|g_1\|^2 + L(g_1^T g_0 + g_0^T g_1) + (1-\lambda)L\|g_1\|^2} \qquad (4.15)$$

Neglecting the remaining terms .

Again for $k = 2$

$$x_2 = x_1 + \alpha_1 d_1$$

$$= x_1 - \frac{\|g_1\|^2 - Lg_1^T g_0}{\|g_0\|^2 + L(g_1^T g_0 + g_0^T g_1) + (1-\lambda)L\|g_1\|^2}(g_1 + Lg_0)$$

$$\Rightarrow |x_2 - x_1| = \frac{\left|\|g_1\|^2 - Lg_1^T g_0\right|\left|(g_1 + Lg_0)\right|}{\left|\|g_0\|^2 + L(g_1^T g_0 + g_0^T g_1) + (1-\lambda)L\|g_1\|^2\right|} \qquad (4.16)$$

Continuing as above we have

$$x_{k+1} - x_k = \frac{(\|g_k\|^2 - Lg_k^T g_{k-1})(g_k + Lg_{k-1})}{\|g_{k-1}\|^2 + L(g_k^T g_{k-1} + g_{k-1}^T g_k) + (1-\gamma)L\|g_0\|^2}$$
$$\le \frac{\frac{1}{c}}{\|g_{k-1}\|^2 + (1-\gamma)^2 g_{k-1}\|g_k\|^4}$$

$$\le \frac{\frac{1}{c}}{c+(1-\gamma)^2 g_{k-1}} \tag{4.17}$$

$$\Rightarrow \| x_{k+1}-x_k \| \le \frac{c}{\| c+(1-\gamma)^2 g_{k-1} \|} \qquad [c>1, \frac{1}{c}<1]$$

$$\le \frac{c}{c-(1-\gamma)^2 \|g_{k-1}\|}$$

$$< \frac{c}{c-(1-\gamma)^2 \frac{1}{\sqrt{c}}} = \in \text{ (say)} \qquad [\ \|g_{k-1}\| \le \frac{1}{\sqrt{c}}\ ] \tag{4.18}$$

Therefore x_k converges again from (4.2) and (4.3)

$$x_{k+1} = x_k + \alpha_k d_k$$
$$d_k = -g_k + \beta_k d_{k-1}$$
$$f(x_k + \alpha_k d_k) - f(x_k) = \alpha_k d_k g_k^T + R_n$$

Here R_n is the remainder after n terms.

Therefore,

$$f(x_k + \alpha_k d_k) - f(x_k) \le \alpha_k d_k g_k^T$$

Applying Taylor's series ,we have

$$x_{k+1} = x_k + \alpha_k d_k$$
$$d_k = -g_k + \beta_k d_{k-1}$$
$$f(x_k + \alpha_k d_k) - f(x_k) \le \alpha_k d_k g_k^T$$
$$f(x_k + \alpha_k d_k) = f(x_x) + \frac{\alpha_k d_k}{1} \nabla f(x_k) + \frac{\alpha_k^2 d_k d_k^T}{2} \nabla^2 f(x_k) + ...\infty$$

Again from Wolfe conditions,

$$x_{k+1} = x_k + \alpha_k d_k$$
$$d_k = -g_k + \beta_k d_{k-1}$$
$$f(x_k + \alpha_k d_k) - f(x_k) \le \rho \alpha_k d_k g_k^T$$
$$= \rho \alpha_k g_k^T \{(-g_k) + \beta_k d_{k-1}\}$$
$$= -\rho \alpha_k \| g_k \|^2 + \rho \alpha_k g_k^T \beta_k d_{k-1}$$
$$\Rightarrow f(x_k + \alpha_k d_k) - f(x_k) < -\rho \alpha_k \| g_k \|^2$$
$$\Rightarrow f(x_k + \alpha_k d_k) - f(x_k) < 0$$

Therefore

$$f(x_k + \alpha_k d_k) < f(x_k) < f(x_{k-1}) < \ldots\ldots\ldots\ldots\ldots < f(x_1) < f(x_0)$$

From above we can observe that the function satisfies the Wolfe Conditions and so it is convergent.