Chapter-6

A new testing procedure for sequence of step length in an unpredictable manner of nonlinear conjugate gradient method

The general nonlinear conjugate gradient method always considers a critical initial search direction to find a convergent solution. In the present work we have found a new procedure to find the sequence of step lengths which tend to vary in an unpredictable manner. The convergence of the C.G method with this proposed step length has been discussed.

Introduction

Conjugate Gradient (C.G) methods comprise a class of unconstrained optimization algorithms which is characterized by low memory requirements and strong local and global convergence properties. In the seminal 1952 paper [1] of Hestenes and Stiefel, the algorithm is presented as an approach to solve symmetric, positive-definite linear systems.

In this survey, we focus on conjugate gradient methods applied to the linear unconstrained optimization problem

$$\min\{ f(x) : x \in R^n \} \qquad (5.13)$$

Where $f : R^n \to R$ is a confunction, especiallyiable function especially if the dimension n is large. They are on the form

$$x_{k+1} = x_k + \alpha_k d_k \qquad (5.14)$$

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

$$d_k = \begin{cases} -g_k, k = 1 \\ -g_k + \beta_k d_{k-1}, k \ge 2 \end{cases} \qquad (5.14)$$

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 .In this case when f is a convex quadratic function

$$f(x) = g^T x + \frac{1}{2} x^T H_x \qquad (5.15)$$

It is known from (5.13) and (5.14) that only the step size α_k and the parameter β_k remain to be determined in the definition of the Conjugate Gradient method. In this case that if f is a convex quadratic, the choice of β_k should be such that the method (5.13)-(5.14) reduces to the linear Conjugate Gradient method if the line search is exact namely

$$\alpha_k = \arg\min\{f(x_k + \alpha d_k); \alpha > 0\} \tag{5.16}$$

The conjugate gradient method is such that the conjugacy conditions hold, namely

$$d_i^T H d_j = 0, \quad \forall i \neq j \tag{5.17}$$

Denote y_{k-1} to be the gradient change,

$$y_{k-1} = g_k - g_{k-1} \tag{5.18}$$

For general nonlinear functions, we know by the mean value theorem that there exists some $t \in (0,1)$ such that

$$\alpha_{k-1}^{-1} d_k^T y_{K-1} = d_k^T \nabla^2 f(x_{k-1} + t\alpha_{K-1} d_{K-1}) d_{K-1} \tag{5.19}$$

Therefore, it is reasonable to replace (1.6) with the following conjugacy condition

$$d_k^T y_{k-1} = 0 \tag{5.20}$$

Multiplying y_{k-1} in (5.14) and using (5.20) we can deduce a formula for the scalar β_k

$$\beta_k^{HS} = \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}} \tag{5.21}$$

This is so called HS formula, which was given by Hestenes and Stiefel [1]. In practical computation, the HS method resembles the PRP method (see [16] and [17] for the PRP method); both methods are generally believed to be two of the most efficient conjugate gradient methods.

However, both the conjugacy conditions (5.17) and (5.20) depend on exact line searches. In practical computation one normally carries out inexact line searches instead of exact line searches. In the case when $g_{k+1}^T d_k \neq 0$,the conjugacy conditions (5.17) and (5.20) may have some disadvantages(for instance see

[18]).Suppose we minimize the convex quadratic function (5.15) on a subspace spanned by a set of mutually conjugate directions $\{d_1, d_2, ..., d_k\}$ .Suppose that the line search along d_1 is not exact,ie $\alpha_1 \neq \alpha_1^*$ where α_1^* is the step length that solves (5.16).Then no matter what line searches are used in the subsequent iterations, it is always true that

$$(x_{k+1} - x^*)^T H (x_{k+1} - x^*) \geq \alpha_{k+1} - \alpha_1^* d_1^T H d_1 \qquad (5.22)$$

Where $x^* = -H^{-1}g$ is the minimum of the objective function(5.15).Hence we see that the error left in the current iteration cannot be eliminated in the subsequent iterations as long as the subsequent search directions conjugate to the current search condition.

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 the nonlinear Conjugate Gradient method. Meanwhile the parameter β_k is called the Conjugate Gradient parameter. The equivalence of the linear system to the minimization problem of $\frac{1}{2} x^T A x - b^T x$ 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[2]) while n-step quadratic convergence rate can be established if the method is restarted along the negative gradient every n-step.(See Cohen [3] and McCormick and Ritter[4])

In 1964 the method has been extended to nonlinear problems by Fletcher and Reeves [5], 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[6].Since the exact line search is usually expensive and impractical, the strong Wolfe line search is often considered the implementation of the nonlinear Conjugate Gradient methods .It aims to find a step size satisfying the strong Wolfe conditions.

$$f(x_k + \alpha_k d_k) - f(x_k) \le \rho \alpha_k g_k^T d_k$$
$$\left| g(x_k + \alpha_k d_k)^T \right| \le -\sigma g_k^T d_k$$
$$where \ \ 0 < \rho < \sigma < 1.$$

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 satisfied with a step size satisfying the standard wolf conditions and

$$g(x_k + \alpha_k d_k)^T d_k \ge \sigma g_k^T d_k \ where \ 0 < \rho < \sigma < 1$$

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

$$g_k^T d_k < 0$$

For Conjugate Gradient method, by multiplying (1.3) 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 paper we say that a Conjugate Gradient method is descent if (5.19) 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>0.However we have to point out that the borderlines between these Conjugate Gradient methods are not strict.

If $s_k = x_{k+1} - x_k$ and in the following $y_k = g_{k+1} - g_k$ .Different Conjugate Gradient algorithms corresponds to different choices for the parameter β_k.Therefore a crucial element in any Conjugate Gradient algorithm in the formula definition of any Conjugate Gradient algorithm has very simple general structure as illustrated below.

The prototype of conjugate gradient algorithm

Step1: Select the initial starting point $x_0 \in dom$ f and compute:

$f_0 = f(x_0)$ and $g_0 = \nabla f(x_0)$ .Set for example $d_0 = -g_0$ and k=0

Step2: Test a criterion for stopping the iteration. For example, ifs$\| g_k \|_\infty \le \varepsilon$, then stop; otherwise continue with step 3.

Step 3: Determine the Step length α_k

Step 4: 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 $s_k = x_{k+1} - x_k$

Step 5: Determine β_k

Step 6: Compute the search direction as: $d_{k+1} = -g_{k+1} + \beta_k s_k$.

Step 7: 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 $d_{k+1} = -g_{k+1}$

Step 8: Compute the initial guess $\frac{\alpha_{k-1} \| d_{k-1} \|}{\| d_k \|}$,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 and those variants focus on parameter β_k computation and on the step length determination.

The main objective of this work is to investigate the sequence of step length $\{x_k\}$ tend to vary in a totally unpredictable manner taking an arbitrary initial search direction although the convergence for general linear as well as nonlinear functions with an initial search direction already exist.

In this section we have discussed, how the sequence of step length $\{x_k\}$ tend to vary in a totally unpredictable manner taking an arbitrary initial search direction as discussed in Chapter-4.Here we have taken the initial search direction as $d_0 = -g_o + \gamma g_o$ Instead of $d_0 = -g_o$ *where* $\gamma \in (0,1)$

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

$$x_1 = x_0 + \alpha_0 d_0$$

$$\Rightarrow \alpha_0 = \frac{x_1 - x_0}{d_0} = \frac{x_1 - x_0}{g_0(\gamma - 1)}$$

$$\alpha_1 = \frac{x_2 - x_1}{g_1(\gamma - 1)}$$

$$\alpha_2 = \frac{x_3 - x_2}{g_2(\gamma - 1)}$$

$$\alpha_3 = \frac{x_4 - x_3}{g_3(\gamma - 1)}$$

Continuining this process we get

$$\alpha_k = \frac{x_{k+1} - x_k}{g_k(\gamma - 1)}$$

$$\alpha_{k+1} = \frac{x_{k+2} - x_{k+1}}{g_{k+1}(\gamma - 1)}$$

Now if we investigate the differences of the consecutive step lengths, we get

$$\alpha_1 - \alpha_0 = \frac{x_2 - x_1}{g_1(\gamma - 1)} - \frac{x_1 - x_0}{g_0(\gamma - 1)}$$

$$= \frac{g_1 x_1 - g_1 x_0 - g_0 x_2 + g_0 x_1}{g_0 g_1(\gamma - 1)}$$

$$\alpha_2 - \alpha_1 = \frac{x_3 - x_2}{g_2(\gamma - 1)} - \frac{x_2 - x_1}{g_1(\gamma - 1)}$$

$$= \frac{g_1 x_3 - g_1 x_2 - g_2 x_2 + g_2 x_1}{g_1 g_2(\gamma - 1)}$$

$$\alpha_3 - \alpha_2 = \frac{x_4 - x_3}{g_3(\gamma - 1)} - \frac{x_3 - x_2}{g_2(\gamma - 1)}$$
$$= \frac{g_2 x_4 - g_2 x_3 - g_3 x_3 + g_3 x_2}{g_2 g_3(\gamma - 1)}$$

and so on.

The difference of the k+1 th step length and the k th step length is

$$\alpha_{k+1} - \alpha_k = \frac{x_{k+2} - x_{k+1}}{g_{k+1}(\gamma - 1)} - \frac{x_{k+1} - x_k}{g_k(\gamma - 1)}$$
$$= \frac{g_k x_{k+2} - g_k x_{k+1} - g_{k+1} x_{k+1} + g_{k+1} x_k}{g_k g_{k+1}(\gamma - 1)}$$

The convergence of the sequence of step lengths $\{\alpha_k\}$ can be obtained from the difference of k+1 th term and k th term of the sequence.

$$\alpha_{k+1} - \alpha_k = \frac{g_k x_{k+2} - g_k x_{k+1} - g_{k+1} x_{k+1} + g_{k+1} x_k}{g_k g_{k+1}(\gamma - 1)}$$
$$= \frac{g_k (x_{k+2} - x_{k+1}) - g_{k+1}(x_{k+1} - x_k)}{g_k g_{k+1}(\gamma - 1)}$$
$$= \frac{g_k s_{k+1} - g_{k+1} s_k}{g_k g_{k+1}(\gamma - 1)}$$

Taking norm on both side

$$\|\alpha_{k+1} - \alpha_k\| = \frac{\|g_k s_{k+1} - g_{k+1} s_k\|}{\|g_k\|\|g_{k+1}\|(\gamma - 1)}$$
$$\leq \frac{d}{\|g_k\|\|g_{k+1}\|(\gamma - 1)} = \in (say)$$

$$\Rightarrow \|\alpha_{k+1} - \alpha_k\| < \in$$

Which implies that the sequence $\{\alpha_k\}$ converge

Again from, Again from Wolfe conditions,

$$x_{k+1} = x_k + \alpha_k d_k$$
$$d_k = -g_k + \beta_k d_{k-1}$$

Applying Taylor's series, we have

$$f(x_k + \alpha_k d_k) = f(x_k) + \frac{\alpha_k d_k}{1!} \nabla f(x_k) + \frac{\alpha_k^2 d_k d_k^T}{2!} \nabla^2 f(x_k) + \text{........} to \ \infty$$

$$f(x_k + \alpha_k d_k) - f(x_k) = \alpha_k d_k g_k^T + R_n$$
$$\Rightarrow f(x_k + \alpha_k d_k) - f(x_k) \leq \alpha_k d_k g_k^T$$
, Where R_n is the remainder after n terms

Therefore the given function satisfies the Wolfe conditions and it converges even with the unpredictable values of α_k .