NONLINEAR CONJUGATE GRADIENT METHODS

YU-HONG DAI
State Key Laboratory of Scientific and Engineering Computing, Institute of Computational Mathematics and Scientific/Engineering Computing Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, P.R. China

INTRODUCTION

Conjugate gradient methods are a class of important methods for solving unconstrained optimization problems

$$\min f(x), \quad x \in R^n, \tag{1}$$

especially if the dimension n is large. They are of the form

$$x_{k+1} = x_k + \alpha_k d_k, \tag{2}$$

where α_k is a stepsize obtained by a line search, and d_k is the search direction defined by

$$d_k = \begin{cases} -g_k, & \text{for } k = 1; \\ -g_k + \beta_k d_{k-1}, & \text{for } k \geq 2, \end{cases} \tag{3}$$

where β_k is a parameter, and g_k denotes $\nabla f(x_k)$.

It is known from Equations (2) and (3) that only the stepsize α_k and the parameter β_k remain to be determined in the definition of conjugate gradient methods. In the case that f is a convex quadratic, the choice of β_k should be such that the methods (2) and (3) reduces to the *linear* conjugate gradient method if the line search is exact, namely,

$$\alpha_k = \arg\min \left\{ f(x_k + \alpha d_k); \alpha > 0 \right\}. \tag{4}$$

For nonlinear functions, however, 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 as *nonlinear* conjugate gradient method. Meanwhile, the parameter β_k is called *conjugate gradient parameter*.

The linear conjugate gradient method can be dated back to a seminal paper by Hestenes and Stiefel [1] in 1952 for solving a symmetric positive definite linear system $Ax = b$, where $A \in R^{n\times n}$ and $b \in R^n$. An easy and geometrical interpretation of the linear conjugate gradient method can be founded in Shewchuk [2]. The equivalence of the linear system to the minimization problem of $\frac{1}{2} x^{\mathrm{T}} A x - b^{\mathrm{T}} x$ motivated Fletcher and Reeves [3] to extend the linear conjugate gradient method for nonlinear optimization. This work of Fletcher and Reeves in 1964 not only opened the door to the nonlinear conjugate gradient field but greatly stimulated the study of nonlinear optimization. In general, the nonlinear conjugate gradient method without restarts is only linearly convergent [4], while n-step quadratic convergence rate can be established if the method is restarted along the negative gradient every n-step [5,6]. Some recent reviews on nonlinear conjugate gradient methods can be found in Hager and Zhang [7], Nazareth [8,9], and Nocedal [10,11]. This article aims to provide a perspective view on the methods from the angle of descent property and global convergence.

Since exact line search is usually expensive and impractical, the strong Wolfe line search is often considered in the implementation of nonlinear conjugate gradient methods. It aims to find a stepsize satisfying the strong Wolfe conditions

$$f(x_k + \alpha_k d_k) - f(x_k) \leq \rho\, \alpha_k g_k^{\mathrm{T}} d_k, \tag{5}$$

$$|g(x_k + \alpha_k d_k)^{\mathrm{T}} d_k| \leq -\sigma\, g_k^{\mathrm{T}} d_k, \tag{6}$$

Wiley Encyclopedia of Operations Research and Management Science, edited by James J. Cochran

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 computations, a typical choice for σ that controls the inexactness of the line search is $\sigma = 0.1$.

On the other hand, for a general nonlinear function, one may be satisfied with a stepsize satisfying the standard Wolfe conditions, namely, Equation (5) and

$$g(x_k + \alpha_k d_k)^{\mathrm{T}} d_k \geq \sigma\, g_k^{\mathrm{T}} d_k, \tag{7}$$

where again $0 < \rho < \sigma < 1$. As is well known, the standard Wolfe 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 [12,13] indicates that the use of standard Wolfe line searches is possible in the nonlinear conjugate gradient field. Besides this, there are quite a few Refs 14,15,16,17 that deal with Armijo-type line searches.

A requirement for an optimization method to use the above line searches is that, the search direction d_k must have the descent property, namely,

$$g_k^{\mathrm{T}} d_k < 0. \tag{8}$$

For conjugate gradient methods, by multiplying Equation (3) with g_k^{T}, we have

$$g_k^{\mathrm{T}} d_k = -\|g_k\|^2 + \beta_k g_k^{\mathrm{T}} d_{k-1}. \tag{9}$$

Thus if the line search is exact, we have $g_k^{\mathrm{T}} d_k = -\|g_k\|^2$ since $g_k^{\mathrm{T}} d_{k-1} = 0$. Consequently, d_k is descent provided $g_k \neq 0$. However, this may not be true in case of inexact line searches for early conjugate gradient methods. A simple restart with $d_k = -g_k$ may remedy these bad situations, but will probably degrade the numerical performance since the second-derivative information along the previous direction d_{k-1} is discarded [18]. Assume that no restarts are used. In this article we say that, a conjugate gradient method is *descent* if Equation (8) holds for all k, and is *sufficient descent* if the sufficient descent condition

$$g_k^{\mathrm{T}} d_k \leq -c\, \|g_k\|^2, \tag{10}$$

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 (see the discussion at the beginning of the section titled "Sufficient Descent Conjugate Gradient Methods").

This survey is organized in the following way. In the next section, we will address two general convergence theorems for the methods of the form (2) and (3) assuming the descent property of each search direction. Afterwards, we divide conjugate gradient methods into three categories: early conjugate gradient methods, descent conjugate gradient methods, and sufficient descent conjugate gradient methods. They will be discussed in sections titled "Early Conjugate Gradient Methods", "Descent Conjugate Gradient Methods", and "Sufficient Descent Conjugate Gradient Methods", respectively, with the emphases on the Fletcher–Reeves (FR) method, the Polak–Ribière–Polyak (PRP) method, the Hestenes–Stiefel (HS) method, the Dai–Yuan method, and the CG_DESCENT method by Hager and Zhang. Some research issues on conjugate gradient methods are mentioned in the last section.

GENERAL CONVERGENCE THEOREMS

In this section, we give two global convergence theorems for any methods of the form (2) and (3) assuming the descent condition (8) for all k. The first one deals with the strong Wolfe line search, while the second treats the standard Wolfe line search.

At first, we give the following basic assumptions on the objective function. Throughout this article, the symbol $\|\cdot\|$ denotes the two norm.

Assumption 1. (i) The level set $\mathcal{L} = \{x \in R^n : f(x) \leq f(x_1)\}$ is bounded, where x_1 is the starting point; (ii) in some neighborhood $\mathcal{N}$ of $\mathcal{L}$, f is continuously differentiable, and its gradient is Lipschitz continuous; namely, there exists a constant $L > 0$ such that

$$\|g(x) - g(y)\| \leq L\, \|x - y\|, \quad \text{for all } x, y \in \mathcal{N}. \tag{11}$$

Sometimes, the boundedness assumption for $\mathcal{L}$ in item (i) is unnecessary and we only require that f is bounded below in $\mathcal{L}$. However, we will just use Assumption 1 for the convergence results in this survey. Under Assumption 1 on f, we state a very useful result, which was obtained by Zoutendijk [19] and Wolfe [20,21]. The relation (12) is usually called as the *Zoutendijk condition*.

Lemma 1. *Suppose that Assumption 1 holds. Consider any iterative method of the form (2), where d_k satisfies $g_k^T d_k < 0$ and α_k is obtained by the standard Wolfe line search. Then we have that*

$$\sum_{k=1}^{\infty} \frac{(g_k^T d_k)^2}{||d_k||^2} < +\infty. \tag{12}$$

To simplify the statements of the following results, we assume that $g_k \neq 0$, for all k, for otherwise a stationary point has been found. Assume also that $\beta_k \neq 0$, for all k. This is because if $\beta_k = 0$, the direction in Equation (3) reduces to the negative gradient direction. Thus, either the method converges globally if $\beta_k = 0$ for infinite number of k, or one can take some x_k as the new initial point. In addition, we say that a method is *globally convergent* if

$$\liminf_{k\to\infty} \|g_k\| = 0, \tag{13}$$

and is *strongly convergent* if

$$\lim_{k\to\infty} \|g_k\| = 0. \tag{14}$$

If the iterations $\{x_k\}$ stay in a bounded region, Equation (13) means that there exists at least one cluster point, which is a stationary point of f, while Equation (14) indicates that every cluster point of $\{x_k\}$ will be a stationary point of f.

To analyze the methods of the form (2) and (3), besides Equation (9), we derive another basic relation. By Equation (3), we have $d_k + g_k = \beta_k d_{k-1}$ for all $k \geq 2$. Squaring both sides of this relation yields

$$\|d_k\|^2 = -2g_k^{\mathrm{T}} d_k - \|g_k\|^2 + \beta_k^2 \|d_{k-1}\|^2. \tag{15}$$

The following theorem gives a general convergence result for any descent methods of the form (2) and (3) under the strong Wolfe line search. It indicates that, if $\|d_k\|^2$ is at most linearly increasing, namely, $\|d_k\|^2 \leq c_1 k + c_2$, for all k, a descent conjugate gradient method with strong Wolfe line search is globally convergent.

Theorem 1 [22]. *Suppose that Assumption 1 holds. Consider some methods of the form (2) and (3) with d_k satisfying $g_k^T d_k < 0$ and with the strong Wolfe line search (5) and (6). Then the method is globally convergent if*

$$\sum_{k\geq 1} \frac{1}{\|d_k\|^2} = +\infty. \tag{16}$$

Proof. It follows from Equations (9) and (6) that $|g_k^{\mathrm{T}} d_k| + \sigma|\beta_k||g_{k-1}^{\mathrm{T}} d_{k-1}| \geq \|g_k\|^2$, which with the Cauchy–Schwarz inequality gives

$$(g_k^{\mathrm{T}} d_k)^2 + \beta_k^2 (g_{k-1}^{\mathrm{T}} d_{k-1})^2 \geq c_1 \|g_k\|^4, \tag{17}$$

where $c_1 = (1+\sigma^2)^{-1}$ is constant. By Equation (15), $g_k^{\mathrm{T}} d_k < 0$ and Equation (17), we have

$$\begin{aligned}
&\frac{(g_k^{\mathrm{T}} d_k)^2}{\|d_k\|^2} + \frac{(g_{k-1}^{\mathrm{T}} d_{k-1})^2}{\|d_{k-1}\|^2} \\
&= \frac{1}{\|d_k\|^2}\left[(g_k^{\mathrm{T}} d_k)^2 + \frac{\|d_k\|^2}{\|d_{k-1}\|^2}(g_{k-1}^{\mathrm{T}} d_{k-1})^2\right] \\
&\geq \frac{1}{\|d_k\|^2}\left[(g_k^{\mathrm{T}} d_k)^2 + \beta_k^2 (g_{k-1}^{\mathrm{T}} d_{k-1})^2 \right. \\
&\qquad \left. - \frac{(g_{k-1}^{\mathrm{T}} d_{k-1})^2}{\|d_{k-1}\|^2}\|g_k\|^2\right] \\
&\geq \frac{1}{\|d_k\|^2}\left[c_1\|g_k\|^4 - \frac{(g_{k-1}^{\mathrm{T}} d_{k-1})^2}{\|d_{k-1}\|^2}\|g_k\|^2\right].
\end{aligned} \tag{18}$$

Assume that Equation (13) is false and there exists some constant $\gamma > 0$ such that

$$\|g_k\| \geq \gamma, \text{for all } k \geq 1. \tag{19}$$

Notice that the Zoutendijk condition (12) implies that $g_k^{\mathrm{T}} d_k / \|d_k\|$ tends to zero. By this,

Equations (18) and (19), we have for sufficiently large k,

$$\frac{(g_k^{\mathrm{T}}d_k)^2}{\|d_k\|^2}+\frac{(g_{k-1}^{\mathrm{T}}d_{k-1})^2}{\|d_{k-1}\|^2}\geq\frac{c_1}{2}\frac{\|g_k\|^2}{\|d_k\|^2}. \quad (20)$$

Thus, by the Zoutendijk condition and Equation (19), we must have that

$$\sum_{k\geq1}\frac{1}{\|d_k\|^2}\leq\frac{1}{\gamma^2}\sum_{k\geq1}\frac{\|g_k\|^2}{\|d_k\|^2}<+\infty, \quad (21)$$

which is a contradiction to the assumption (16). Therefore, we must have that the convergence relation (13) holds.

We are now going to provide another general global convergence theorem for any descent methods (2) and (3) with the standard Wolfe line search. To this aim, we define

$$t_k=\frac{\|d_k\|^2}{\phi_k^2},\quad \phi_k=\begin{cases}\|g_1\|^2, & \text{for k=1 ;}\\ \prod_{j=2}^k\beta_j^2, & \text{for}k\geq2.\end{cases} \quad (22)$$

By dividing Equation (15) by ϕ_k^2 and noticing that $d_1=-g_1$, we can obtain [23] that for all $k\geq1$

$$t_k=-2\sum_{i=1}^k\frac{g_i^{\mathrm{T}}d_i}{\phi_i^2}-\sum_{i=1}^k\frac{\|g_i\|^2}{\phi_i^2}. \quad (23)$$

Theorem 2 [17]. *Suppose that Assumption 1 holds. Consider any methods of the form (2) and (3) with d_k satisfying $g_k^Td_k<0$ and with the standard Wolfe line search (5) and (7). Then the method is globally convergent if the scalar β_k is such that*

$$\sum_{k\geq1}\prod_{j=2}^k\beta_j^{-2}=+\infty. \quad (24)$$

Proof. Define ϕ_k as in Equation (22). The condition (24) is equivalent to

$$\sum_{k\geq1}\frac{1}{\phi_k^2}=+\infty. \quad (25)$$

Noting that $-2g_i^{\mathrm{T}}d_i-\|g_i\|^2\leq(g_i^{\mathrm{T}}d_i)^2/\|g_i\|^2$, it follows from Equation (23) that

$$t_k\leq\sum_{i=1}^k\frac{(g_i^{\mathrm{T}}d_i)^2}{\|g_i\|^2\phi_i^2}. \quad (26)$$

Since $t_k\geq0$, the relation (23) also gives

$$-2\sum_{i=1}^k\frac{g_i^{\mathrm{T}}d_i}{\phi_i^2}\geq\sum_{i=1}^k\frac{\|g_i\|^2}{\phi_i^2}. \quad (27)$$

Noting that $-4g_i^{\mathrm{T}}d_i-\|g_i\|^2\leq4(g_i^{\mathrm{T}}d_i)^2/\|g_i\|^2$, we get by this and Equation (27) that

$$\begin{aligned}4\sum_{i=1}^k\frac{(g_i^{\mathrm{T}}d_i)^2}{\|g_i\|^2\phi_i^2}&\geq-4\sum_{i=1}^k\frac{g_i^{\mathrm{T}}d_i}{\phi_i^2}-\sum_{i=1}^k\frac{\|g_i\|^2}{\phi_i^2}\\&\geq\sum_{i=1}^k\frac{\|g_i\|^2}{\phi_i^2}.\end{aligned} \quad (28)$$

Now we proceed by contradiction and assume that Equation (19) holds. Then by Equations (28), (25), and (19), we have that

$$\sum_{k\geq1}\frac{(g_k^{\mathrm{T}}d_k)^2}{\|g_k\|^2\phi_k^2}\geq\frac{\gamma^2}{4}\sum_{k\geq1}\frac{1}{\phi_k^2}=+\infty, \quad (29)$$

which means that the sum series in the right-hand side of Equation (26) is divergent. By Lemma 6 in Pu and Yu [24], we then know that

$$\begin{aligned}+\infty=\sum_{k\geq1}\frac{(g_k^{\mathrm{T}}d_k)^2}{\|g_k\|^2\phi_k^2}\frac{1}{t_k}&=\sum_{k\geq1}\frac{(g_k^{\mathrm{T}}d_k)^2}{\|g_k\|^2\|d_k\|^2}\\&\leq\frac{1}{\gamma^2}\sum_{k\geq1}\frac{(g_k^{\mathrm{T}}d_k)^2}{\|d_k\|^2},\end{aligned} \quad (30)$$

which contradicts the Zoutendijk condition (12). The contradiction shows the truth of Equation (13).

Theorem 2 provides a condition on β_k, which is sufficient for the global convergence of a conjugate gradient method using the standard Wolfe line search. Instead of the sufficient descent condition (10), only the descent condition $d_k^{\mathrm{T}}g_k<0$ is used here. An easy understanding between Theorems 1

and 2 is given in Dai [23] under the strong Wolfe line search, in which situation we have the estimate $d_k \approx \beta_k d_{k-1}$, if there is no convergence. Since different nonlinear conjugate gradient methods only vary with the scalar β_k, we believe that the condition (24) in Theorem 2 is very powerful in the convergence analysis of conjugate gradient methods. See Dai [23] for some further uses of Equation (24).

EARLY CONJUGATE GRADIENT METHODS

The Fletcher–Reeves Method

In 1964, Fletcher and Reeves [3] proposed the first nonlinear conjugate gradient method and used the following conjugate gradient parameter

$$\beta_k^{\mathrm{FR}} = \frac{||g_k||^2}{||g_{k-1}||^2}. \tag{31}$$

The introduction of the FR method is a milestone in the field of large-scale nonlinear optimization.

Early analysis with the FR method is based on the exact line search. Zoutendijk [19] proved that the FR method with line search is globally convergent for nonlinear functions. Al-Baali [25] first analyzed the FR method with strong Wolfe inexact line searches (5) and (6). He showed that if $\sigma < 1/2$, the sufficient condition (10) holds and there is global convergence. Liu *et al*. [26] extended Al-Baali's result to the case that $\sigma = 1/2$. Dai and Yuan [27] presented a simpler proof to this result by showing that the sufficient condition (10) holds for at least one of any two neighboring iterations. Here it is worth noting that after the descent condition (8) has been verified, we can establish the global convergence easily by Theorem 2. More exactly, assuming that there is no convergence and Equation (19) holds, we can see that $\prod_{j=2}^{k} \beta_j^2$ is at most linearly increasing and hence Equation (24) holds. Consequently, there will be global convergence by Theorem 2, leading to a contradiction.

Further, if $\sigma > 1/2$, Dai and Yuan [27] proved that even for the one-dimensional quadratic function

$$f(x) = \frac{1}{2}x^2, \quad x \in R,$$

the FR method may fail due to generating an uphill search direction. Interestingly enough, if we continue the FR method by searching its opposite direction once an uphill direction is generated and keeping $x_{k+1} = x_k$ if g_{k+1} is orthogonal to d_{k+1}, it is still possible to establish the global convergence of the method. Dai and Yuan [28] considered this idea and showed the global convergence of the FR method under a generalized Wolfe line search.

Powell [18] analyzed the global efficiency of the FR method with the exact line search. Denote by θ_k, the angle between d_k and $-g_k$. The exact line search implies that g_{k+1} is orthogonal to d_k for all k and hence

$$\|d_{k+1}\| = \sec\theta_k \, \|g_k\| \tag{32}$$

and

$$\beta_{k+1}\|d_k\| = \tan\theta_k \, \|g_{k+1}\|. \tag{33}$$

By using the above two relations and substituting the formula (31), we can obtain

$$\tan\theta_{k+1} = \sec\theta_k \, \|g_{k+1}\|/\|g_k\| > \tan\theta_k \, \|g_{k+1}\|/\|g_k\|. \tag{34}$$

Now, if θ_k is close to $\frac{1}{2}\pi$, the iteration may take a very small step, in which case both the step $s_k = x_{k+1} - x_k$ and the change $y_k = g_{k+1} - g_k$ are small. Thus the ratio $\|g_{k+1}\|/\|g_k\|$ is close to one. Consequently, by Equation (34), θ_{k+1} is close to $\frac{1}{2}\pi$, which indicates that slow progress may occur again on the next iteration. The drawback that the FR method may fall into some circle of tiny steps was extended by Gilbert and Nocedal [29] to the strong Wolfe line search, and was observed by many researchers in the community. It explains why the FR method is sometimes very slow in practical computations.

Suppose that after some iterations, the FR method enters a region in the space of the

variables where f is the quadratic function

$$f(x) = \frac{1}{2}x^{\mathrm{T}}x, \quad x \in R^n. \tag{35}$$

In this case, the exact line search along d_k and $g_k = x_k$ implies that

$$\|g_{k+1}\| = \|g_k\| \sin\theta_k. \tag{36}$$

By this and the equality in (34), we obtain $\theta_{k+1} = \theta_k$. Thus the angle between the search direction and the steepest descent direction remains constant for all consecutive iterations, which makes the method very slow if θ_k is close to $\frac{1}{2}\pi$. This example was addressed by Powell [18] for the two-dimensional case and is actually valid for any dimension.

As will be seen in the section titled "The Polak–Ribière–Polyak Method", unlike the FR method, the PRP method can generate a search direction close to the steepest descent direction once a small step occurs and hence can avoid cycles of tiny steps. On the other hand, the PRP method need not converge even with the exact line search. This motivated Touati-Ahmed and Storey [30] to extend Al-Baali's convergence result on the FR method to the general methods (2) and (3) with

$$\beta_k \in \left[0,\ \beta_k^{\mathrm{FR}}\right] \tag{37}$$

and suggested the formula

$$\beta_k^{\mathrm{TS}} = \max\left\{0, \min\left\{\beta_k^{\mathrm{PRP}}, \beta_k^{\mathrm{FR}}\right\}\right\}. \tag{38}$$

Gilbert and Nocedal [29] further extended Equation (37) to the interval

$$\beta_k \in \left[-\beta_k^{\mathrm{FR}},\ \beta_k^{\mathrm{FR}}\right] \tag{39}$$

and proposed the formula

$$\beta_k^{\mathrm{GN}} = \max\left\{-\beta_k^{\mathrm{FR}}, \min\left\{\beta_k^{\mathrm{PRP}}, \beta_k^{\mathrm{FR}}\right\}\right\}. \tag{40}$$

However, the numerical results in Gilbert and Nocedal [29] show that the Gilbert and Nocedal (GN) method is not as good as the PRP method, although it indeed performs better than the FR method.

The Polak–Ribière–Polyak Method

In 1969, Polak and Ribière [31] and Polyak [32] proposed another conjugate gradient parameter, independently, that is

$$\beta_k^{\mathrm{PRP}} = \frac{g_k^{\mathrm{T}} y_{k-1}}{||g_{k-1}||^2}, \tag{41}$$

where $y_{k-1} = g_k - g_{k-1}$. In practical computations, the PRP method performs much better than the FR method for many optimization problems because it can automatically recover once a small step is generated. For this, we still consider the exact line search. It follows from Equation (41) that

$$|\beta_{k+1}^{\mathrm{PRP}}| \le \|g_{k+1}\|\,\|g_{k+1} - g_k\|/||g_k||^2. \tag{42}$$

By using the relations (32), (33), and (42), we can obtain

$$\tan\theta_{k+1} \le \sec\theta_k \|g_{k+1} - g_k\|/\|g_k\|. \tag{43}$$

Assume that the angle θ_k between $-g_k$ and d_k is close to $\frac{1}{2}\pi$ and $\|s_k\| = \|x_{k+1} - x_k\| \approx 0$. Then we have that $\|g_{k+1} - g_k\| \ll \|g_k\|$ and hence

$$\tan\theta_{k+1} \ll \sec\theta_k. \tag{44}$$

Consequently, the next search direction d_{k+1} will tend to $-g_{k+1}$ and avoid the occurrence of continuous tiny steps. The PRP method was believed to be one of the efficient conjugate gradient methods in the last century.

Nevertheless, the global convergence of the PRP method only proves for strictly convex functions [33]; for general functions, Powell [34] showed that the PRP method can cycle infinitely without approaching a solution even if the stepsize α_k is chosen to the least positive minimizer of the line search function. To change this unbalanced state, Gilbert and Nocedal [29] considered Powell [35]'s suggestion of modifying the PRP method by setting

$$\beta_k^{\mathrm{PRP+}} = \max\{\beta_k^{\mathrm{PRP}}, 0\}, \tag{45}$$

and showed that this modification of the PRP method, called PRP$^+$, is globally convergent both for exact and inexact line searches. More

exactly, Gilbert and Nocedal established the following result.

Theorem 3. *Suppose that Assumption 1 holds. Consider the PRP$^+$ method, namely, Equations (2) and (3), where β_k is given by Equation (45). If the line search satisfies the standard Wolfe conditions (5) and (7), and the sufficient descent condition (10) for some constant $c > 0$, the method is globally convergent.*

The technique of their proof is quite sophisticated. First, they define the so-called Property $(*)$, which is a mathematical lifting of the property of avoiding cycles of tiny steps.

***Property* $(*)$.** Consider the methods (2) and (3) and assume that $0 < \gamma < \|g_k\| \le \overline{\gamma}$. Then we say that the method has Property $(*)$, if there exist constants $b > 1$ and $\zeta > 0$ such that for all k,

$$|\beta_k| \le b, \tag{46}$$

and

$$\|s_{k-1}\| \le \zeta \Longrightarrow |\beta_k| \le \frac{1}{2b}. \tag{47}$$

It is not difficult to see that both PRP and PRP$^+$ possess such a property. Secondly, defining $u_k = d_k/\|d_k\|$, Gilbert and Nocedal observed that if $\beta_k \ge 0$ and if $\|d_k\| \to \infty$, u_k and u_{k-1} will tend to be the same, namely, $\|u_k - u_{k-1}\| \to 0$. Their proof then proceeds by contradiction. If there is no convergence, then $\|d_k\|$ must tend to infinity. Consequently, by Property $(*)$, the method has to take big steps for at least half of the iterations, otherwise $\|d_k\|$ becomes finite. However, since d_k tends to be the same direction for sufficiently large k, the iterations will lie out of the bounded level set $\mathcal{L}$ if there are many big steps. A contradiction is then obtained.

The convergence result of Gilbert and Nocedal requires the sufficient descent condition (10). If the strong Wolfe line search is used instead of the standard Wolfe line search, the sufficient descent condition can be relaxed to the descent condition of the search direction [22]. However, the following one-dimensional quadratic example shows that the PRP method with the strong Wolfe line search may generate an uphill search direction [36]. Consider

$$f(x) = \frac{1}{2}\lambda x^2, \qquad x \in R^1, \tag{48}$$

where $\lambda = \min\{1 + \sigma, 2 - 2\delta\}$ and suppose that the initial point is $x_1 = 1$. Then for any constant δ and σ satisfying $0 < \delta < \sigma < 1/2$, direct calculations show that the unit stepsize satisfies the strong Wolfe conditions (5) and (6). Consequently, $x_2 = 1 - \lambda$ and

$$g_2^{\mathrm{T}} d_2 = \lambda^2(\lambda - 1)^3 > 0, \tag{49}$$

which means that d_2 is uphill. Thus, for any small $\sigma \in (0, 1)$, the strong Wolfe line search cannot guarantee the descent property of the PRP method even for convex quadratic functions.

To ensure the sufficient descent condition, required by Theorem 3 for the PRP$^+$ method, in practical computations, Gilbert and Nocedal [29] designed a dynamic inexact line search strategy. As a matter of fact, their strategy applies to any of the methods (2) and (3) with nonnegative β_k's. Let us look at Equation (9). If $g_k^{\mathrm{T}} d_{k-1} \le 0$, we already have Equation (10) since $\beta_k \ge 0$. On the other hand, If $g_k^{\mathrm{T}} d_k > 0$, it must be the case that $g_k^{\mathrm{T}} d_{k-1} > 0$, which means that a one-dimensional minimizer has been bracketed. Then $g_k^{\mathrm{T}} d_{k-1}$ can be reduced and Equation (10) holds by applying a line search algorithm, such as that given by Lemaréchal [37], Fletcher [38], or Moré and Thuente [39]. Comparing with the PRP method, however, no significant improvement is reported in Gilbert and Nocedal [29] for the PRP$^+$ method.

Is there any clever inexact line search that can guarantee the global convergence of the original PRP method? Grippo and Lucidi [15] answered this question positively by generalizing an Armijo-type line search in De Leone *et al*. [40]. Given constants $\tau > 0$, $\sigma \in (0, 1)$, $\delta > 0$, and $0 < c_1 < 1 < c_2$, their line search

aims to find

$$\alpha_k = \max\left\{\sigma^j \frac{\tau|g_k^{\mathrm{T}} d_k|}{||d_k||^2}; j = 0, 1, \ldots, \right\} \quad (50)$$

such that $x_{k+1} = x_k + \alpha_k d_k$ and $d_{k+1} = -g_{k+1} + \beta_{k+1}^{\mathrm{PRP}} d_k$ satisfy

$$f(x_{k+1}) \leq f(x_k) - \delta\, \alpha_k^2\, ||d_k||^2 \quad (51)$$

and

$$-c_2||g_{k+1}||^2 \leq g_{k+1}^T d_{k+1} \leq -c_1||g_{k+1}||^2, \quad (52)$$

Such a stepsize must exist because of the following observations. If α_k or, equivalently, $\|s_k\| = \|x_{k+1} - x_k\| = \alpha_k\|d_k\|$ is small, $\beta_{k+1}^{\mathrm{PRP}}$ tends to zero and hence d_{k+1} gets close to $-g_{k+1}$. On the other hand, the difference in the objective function, $f(x_{k+1}) - f(x_k)$, is $O(-g_k^{\mathrm{T}} s_k)$ or $O(\|s_k\|)$, whereas the expected decrease is only of the second order $O(\|s_k\|^2)$. Therefore, Equations (51) and (52) must hold provided that α_k is sufficiently small. Furthermore, since the total reduction of the objective function is finite, the line search condition (51) enforces $\lim_{k\to\infty} \|s_k\| = 0$. By this property, the strong global convergence relation (14) can be achieved for the PRP algorithm of Grippo and Lucidi. From the view point of computations, Grippo and Lucidi [15] refined their line search algorithm so that the first one-dimensional minimizer of the line search function can be accepted. Again, the numerical experience [41] does not suggest a significant improvement of their algorithm over the PRP method.

Along the line of Grippo and Lucidi [15], Dai and Yuan [36] builds the strong convergence of the PRP method with constant stepsizes

$$\alpha_k \equiv \eta, \text{ where } \eta \in (0, \frac{1}{4L}) \text{ is constant,} \quad (53)$$

where L is the Lipschitz constant in Assumption 1. This result was extended in Dai [42] for the case that $\alpha_k \equiv \frac{1}{4L}$. Chen and Sun [43] further studied the PRP method together with other conjugate gradient methods using fixed stepsizes of the form

$$\alpha_k = \frac{-\delta\, g_k^{\mathrm{T}} d_k}{d_k^{\mathrm{T}} Q_k d_k}, \quad (54)$$

where $\delta > 0$ is constant and $\{Q_k\}$ is a sequence of positive definite matrices determined in some way.

The Hestenes–Stiefel Method

In this section, we briefly discuss the HS conjugate gradient methods, namely, (2) and (3) where β_k is calculated by

$$\beta_k^{\mathrm{HS}} = \frac{g_k^{\mathrm{T}} y_{k-1}}{d_{k-1}^{\mathrm{T}} y_{k-1}}. \quad (55)$$

Such a formula is first used by Hestenes and Stiefel in the proposition of the linear conjugate gradient method in 1952.

A remarkable property of the HS method is that, no matter whether the line search is exact or not, by multiplying Equation (3) with y_{k-1} and using Equation (55), we always have that

$$d_k^{\mathrm{T}} y_{k-1} = 0. \quad (56)$$

In the quadratic case, y_{k-1} is parallel to $A\, d_{k-1}$, where A is the Hessian of the function. Then Equation (56) implies $d_k^{\mathrm{T}} A d_{k-1} = 0$, that is, d_k is conjugate to d_{k-1}. For this reason, the relation (56) is often called *conjugacy condition*. If the line search is exact, we have by Equation (9) that $g_k^{\mathrm{T}} d_k = -\|g_k\|^2$ since $g_k^{\mathrm{T}} d_{k-1} = 0$. It follows that $d_{k-1}^{\mathrm{T}} y_{k-1} = \|g_{k-1}\|^2$ and $\beta_k^{\mathrm{HS}} = \beta_k^{\mathrm{PRP}}$. Therefore the HS method is identical to the PRP method in the case of exact line searches. As a result, Powell [34]'s counterexample for the PRP method also applies to the HS method, showing the nonconvergence of the HS method with the exact line search. Unlike the PRP method, whose convergence can be guaranteed by the line search of Grippo and Lucidi [15], it is still not known whether there exists a clever line search such that the (unmodified) HS method is well defined at each iteration and converges globally. The answer is perhaps negative. One major observation is that, when $\|s_{k-1}\|$

is small, both the nominator and denominator of β_k^{HS} become small so that β_k^{HS} might be unbounded. Another observation is that, for any one-dimensional function, we always have

$$\begin{aligned} d_2 &= -g_2 + \beta_2^{\mathrm{HS}} d_1 \\ &= -g_2 + \frac{g_2 \cdot y_1}{d_1 \cdot y_1} d_1 \\ &= -g_2 + g_2 = 0 \end{aligned} \tag{57}$$

independent of the line search. Consequently, there is some special difficulty to ensure the descent property of the HS method with inexact line searches.

Similar to the PRP$^+$ method, we can consider the HS$^+$ method, where

$$\beta_k^{\mathrm{HS}+} = \max\left\{\beta_k^{\mathrm{HS}}, 0\right\}. \tag{58}$$

In case of the sufficient descent condition (10), it is easy to verify that both HS and HS$^+$ have Property $(*)$. Further, we can similarly modify the standard Wolfe line search to ensure the sufficient descent condition and global convergence for the HS$^+$ method. If the sufficient descent condition (10) is relaxed to the descent condition, Qi *et al*. [44] established the global convergence of a modified HS method, where β_k takes the form

$$\beta_k^{\mathrm{QHL}} = \max\left\{0, \min\left\{\beta_k^{\mathrm{HS}}, \frac{1}{\|g_k\|}\right\}\right\}. \tag{59}$$

Early in 1977, Perry [45] observed that the search direction in the HS method can be written as

$$d_k = -P_k g_k, \tag{60}$$

where

$$P_k = I - \frac{d_{k-1} y_{k-1}^{\mathrm{T}}}{d_{k-1}^{\mathrm{T}} y_{k-1}}. \tag{61}$$

Noting that $P_k^{\mathrm{T}} y_{k-1} = 0$, P_k is an affine transformation that transforms R^n into the null space of y_{k-1}. To ensure the descent property of d_k, however, we may wish the matrix P_k is positive definite. It is obvious that there is no positive definite matrix P_k such that $P_k^{\mathrm{T}} y_{k-1} = 0$. Instead, we look for a positive definite matrix P_k such that the conjugacy condition (56) holds. In case of exact line searches, it is sufficient to require P_k to satisfy

$$P_k^{\mathrm{T}} y_{k-1} = s_{k-1}, \tag{62}$$

which is exactly the quasi-Newton equation (see Yuan [33] or the article on ***Quasi-Newton Methods***). Following this line, we can consider to generate P_k by using the BFGS update from $\gamma_{k-1} I$, where γ_{k-1} is some scaling factor. This yields

$$\begin{aligned} P_k(\gamma_{k-1}) &= \gamma_{k-1}\left(I - \frac{s_{k-1} y_{k-1}^{\mathrm{T}} + y_{k-1} s_{k-1}^{\mathrm{T}}}{s_{k-1}^{\mathrm{T}} y_{k-1}}\right) \\ &\quad + \left(1 + \frac{\gamma_{k-1}\|y_{k-1}\|^2}{s_{k-1}^{\mathrm{T}} y_{k-1}}\right)\frac{s_{k-1} s_{k-1}^{\mathrm{T}}}{s_{k-1}^{\mathrm{T}} y_{k-1}}. \end{aligned} \tag{63}$$

Shanno [46] explored this idea with $\gamma_{k-1} = 1$ (namely, no scaling is considered in the BFGS update) and obtained the search direction

$$\begin{aligned} d_k &= -g_k + \left[\frac{g_k^{\mathrm{T}} y_{k-1}}{s_{k-1}^{\mathrm{T}} y_{k-1}} - \left(1 + \frac{\|y_{k-1}\|^2}{s_{k-1}^{\mathrm{T}} y_{k-1}}\right)\right. \\ &\quad \left.\frac{g_k^{\mathrm{T}} s_{k-1}}{s_{k-1}^{\mathrm{T}} y_{k-1}}\right] s_{k-1} + \frac{g_k^{\mathrm{T}} s_{k-1}}{s_{k-1}^{\mathrm{T}} y_{k-1}} y_{k-1}. \end{aligned} \tag{64}$$

The methods (2) and (64) are called *memoryless BFGS methods* by Buckley [47]. It is easy to see that the memoryless BFGS method reduces to the HS method if the line search is exact. Without much more calculations and storage at each iteration, the memoryless BFGS method performs much better than the HS method in practical computations.

In case of inexact line searches, Dai and Liao [48] derived the following relation directly from Equations (60) and (62),

$$\begin{aligned} d_k^{\mathrm{T}} y_{k-1} = -(P_k g_k)^{\mathrm{T}} y_{k-1} &= -g_k^{\mathrm{T}}(P_k^{\mathrm{T}} y_{k-1}) \\ &= -g_k^{\mathrm{T}} s_{k-1}. \end{aligned} \tag{65}$$

By introducing a scaling factor t, Dai and Liao considered a generalized conjugacy condition,

$$d_k^{\mathrm{T}} y_{k-1} = -t g_k^{\mathrm{T}} s_{k-1}, \tag{66}$$

and proposed the following choice for β_k,

$$\beta_k^{\mathrm{DL}}(t) = \frac{g_k^{\mathrm{T}} y_{k-1} - t g_k^{\mathrm{T}} s_{k-1}}{d_{k-1}^{\mathrm{T}} y_{k-1}}. \tag{67}$$

Clearly, if the line search is exact, namely, $g_k^{\mathrm{T}} s_{k-1} = 0$, the DL direction is identical to the HS direction. If $g_k^{\mathrm{T}} s_{k-1} \neq 0$, an analysis for quadratic functions is presented in Dai and Liao [48], showing that for small values of t, the DL direction can bring a bigger descent in the objective function than the HS direction if an exact line search is done at the kth iteration. The numerical experiments in Dai and Liao [48] showed that the DL method with $t = 0.1$ is a significant improvement of the HS method. In addition, similar to PRP$^+$ and HS$^+$, Dai and Liao [48] established the global convergence of a modified DL method, where

$$\beta_k^{\mathrm{DL}^+}(t) = \max\left\{\frac{g_k^{\mathrm{T}} y_{k-1}}{d_{k-1}^{\mathrm{T}} y_{k-1}}, 0\right\} - t\,\frac{g_k^{\mathrm{T}} s_{k-1}}{d_{k-1}^{\mathrm{T}} y_{k-1}}, \tag{68}$$

that allows negative values.

Two further developments of the DL method are made by Yabe and Takano [49] and Li *et al*. [50]. Specifically, based on a modified secant condition given by Zhang *et al*. [51,52], Yabe and Takano [49] suggested the variants of Equations (67) and (68) with the vector y_{k-1} replaced with

$$z_{k-1} = y_{k-1} + \left(\frac{\rho\,\lambda_k}{s_{k-1}^{\mathrm{T}} u_{k-1}}\right) u_{k-1}, \tag{69}$$

where $\lambda_k = 6(f_{k-1} - f_k) + 3(g_{k-1} + g_k)^{\mathrm{T}} s_{k-1}$, $\rho \geq 0$ is a constant, and $u_{k-1} \in R^n$ satisfies $s_{k-1}^{\mathrm{T}} u_{k-1} \neq 0$ (for example, $u_{k-1} = d_{k-1}$). Li *et al*. [50] considered the modified secant condition in Wei *et al*. [53] and suggested the following replacement of y_{k-1} in Equations (67) and (68):

$$y_{k-1}^* = y_{k-1} + \frac{\nu_{k-1}}{\|s_{k-1}\|^2}\, s_{k-1}, \tag{70}$$

where $\nu_{k-1} = 2(f_{k-1} - f_k) + (g_{k-1} + g_k)^{\mathrm{T}} s_{k-1}$. Owing to the use of precise modified secant conditions, certain numerical improvements are expected for these variants over the DL and DL$^+$ methods.

DESCENT CONJUGATE GRADIENT METHODS

From the previous section, we can see that none of the FR, PRP, and HS methods can ensure the descent property of the search direction even if the strong Wolfe conditions (5) and (6) with arbitrary $\sigma \in (0, 1)$. For the FR method, the descent condition can be guaranteed by restricting $\sigma \leq 1/2$. However, this is not true any more for $\sigma > 1/2$. For any constant value of $\sigma \in (0, 1)$, there is always some possibility for the PRP and HS methods not to generate a descent search direction.

If a descent search direction is not produced, a practical remedy is to restart the method along $-g_k$. However, this might degrade the efficiency of the method since the second-derivative information achieved along the previous search direction is discarded [18]. From the previous section, we see that many efforts have been made for early conjugate gradient methods to guarantee a descent direction and hence avoid the use of the remedy, including modifying the conjugate gradient parameter β_k or designing some special line search. In this section, we will first address the conjugate descent method and then emphasize the Dai–Yuan method, both of which can ensure the descent condition under strong Wolfe conditions and standard Wolfe conditions, respectively. A hybrid of the two methods is briefly mentioned at the end, which can ensure a descent direction at every iteration without line searches.

The Conjugate Descent Method

In his monograph [38], Fletcher proposed the conjugate descent (CD) methods, namely, (2) and (3) with β_k given by

$$\beta_k^{\mathrm{CD}} = \frac{\|g_k\|^2}{-d_{k-1}^{\mathrm{T}} g_{k-1}}. \tag{71}$$

Other than the FR, PRP, and HS methods, the CD method can ensure the descent property of each search condition, provided that the strong Wolfe conditions (5) and (6) are used. To see this, we first introduce the

following variants of the strong Wolfe conditions, namely, (5) and

$$\sigma_1 g_k^{\mathrm{T}} d_k \le g(x_k + \alpha_k d_k)^{\mathrm{T}} d_k \le -\sigma_2 g_k^{\mathrm{T}} d_k, \quad (72)$$

where $0 < \delta < \sigma_1 < 1$ and $0 \le \sigma_2 < 1$. If $\sigma_1 = \sigma_2 = \sigma$, the above conditions reduce to the strong Wolfe conditions (5) and (6). Now, by Equations (9) and (71), we have

$$-g_k^{\mathrm{T}} d_k = \|g_k\|^2 \left[1 + g_k^{\mathrm{T}} d_{k-1} / g_{k-1}^{\mathrm{T}} d_{k-1}\right]. \quad (73)$$

The above relation and Equation (72) indicate that

$$1 - \sigma_2 \le -g_k^{\mathrm{T}} d_k / \|g_k\|^2 \le 1 + \sigma_1. \quad (74)$$

Since $\sigma_2 < 1$, the left inequality in Equation (74) means that Equation (10) holds with $c = 1 - \sigma_2$ and hence the descent condition holds.

Global convergence analysis of the CD method is made in Dai and Yuan [54] using the generalized strong Wolfe conditions (5) and (72). Specifically, if $\sigma_1 < 1$ and $\sigma_2 = 0$, it follows from Equations (71),(74) and (31) that

$$0 \le \beta_k^{\mathrm{CD}} \le \beta_k^{\mathrm{FR}}. \quad (75)$$

Therefore, by the result of Touati-Ahmed and Storey [30] related to the relation (37), there is global convergence of the CD method. However, for any $\sigma_2 > 0$, it is possible that the square norm $\|d_k\|^2$ in the method increases to infinity at an exponential rate. Specifically, Dai and Yuan [54] considered the following two-dimensional function

$$f(x, y) = \xi x^2 - y, \quad \text{where } \xi \in (1,\ 9/8), \quad (76)$$

and showed that the CD method with the generalized strong Wolfe line search may not solve Equation (76). In real computations, the CD method is even inferior to the FR method.

The Dai–Yuan Method

To enforce a descent direction in case of the standard Wolfe line search, Dai and Yuan [12] proposed a new conjugate gradient method, where

$$\beta_k^{\mathrm{DY}} = \frac{\|g_k\|^2}{d_{k-1}^{\mathrm{T}} y_{k-1}}. \quad (77)$$

For the DY method, it follows by Equations (3),(77) and direct calculations that

$$g_k^{\mathrm{T}} d_k = \frac{\|g_k\|^2}{d_{k-1}^{\mathrm{T}} y_{k-1}} g_{k-1}^{\mathrm{T}} d_{k-1}. \quad (78)$$

The fraction in Equation (78) is exactly the DY formula (77). With this observation, we can get an equivalent expression of β_k^{DY} from Equation (78),

$$\beta_k^{\mathrm{DY}} = \frac{g_k^{\mathrm{T}} d_k}{g_{k-1}^{\mathrm{T}} d_{k-1}}. \quad (79)$$

The following theorem establishes the descent property and global convergence of the DY method with the standard Wolfe line search.

Theorem 4. *Suppose that Assumption 1 holds. Consider the DY methods, namely, (2) and (3) where β_k is given by Equation (77). If the line search satisfies the standard Wolfe conditions (5) and (7), we have that $g_k^{\mathrm{T}} d_k < 0$ for all $k \ge 1$. Further, the method converges in the sense that $\liminf_{k\to\infty} \|g_k\| = 0$.*

Proof. It is obvious that $d_1^{\mathrm{T}} g_1 < 0$ since $d_1 = -g_1$. Assume that $g_{k-1}^{\mathrm{T}} d_{k-1} < 0$. It follows by this and Equation (7) that $d_{k-1}^{\mathrm{T}} y_{k-1} > 0$. Thus by Equation (78), we also have that $g_k^{\mathrm{T}} d_k < 0$. Therefore by induction, $g_k^{\mathrm{T}} d_k < 0$ for all $k \ge 1$.

Now, let us denote

$$q_k = \frac{\|d_k\|^2}{(g_k^{\mathrm{T}} d_k)^2}, \qquad r_k = -\frac{g_k^{\mathrm{T}} d_k}{\|g_k\|^2}. \quad (80)$$

Dividing Equation (15) by $(g_k^{\mathrm{T}} d_k)^2$ and using Equations (79) and (80), we obtain

$$q_k = q_{k-1} + \frac{1}{\|g_k\|^2} \frac{2}{r_k} - \frac{1}{\|g_k\|^2} \frac{1}{r_k^2}. \quad (81)$$

Noting that $\frac{2}{r_k} - \frac{1}{r_k^2} \leq 1$, an immediate corollary of Equation (81) is

$$q_k \leq q_{k-1} + \|g_k\|^{-2}. \tag{82}$$

Assume that $\liminf_{k\to\infty} \|g_k\| \neq 0$ and Equation (19) holds. By Equations (19),(82) and $d_1 = -g_1$, we have $q_k \leq k/\gamma^2$ and hence $\sum_{k\geq 1} q_k^{-1} = +\infty$, which contradicts the Zoutendijk condition (12). Therefore the statement is true.

By Equation (81), we can further exploit the self-adjusting property of the DY method [55]. To this aim, we first notice that the r_k defined in Equation (80) is a quantity that reflects the descent degree of the search direction d_k, since the descent condition (8) is equivalent to $r_k > 0$ and the sufficient condition (10) is the same as $r_k \geq c$. Now let us focus on the relation (81). The second term on the right side of Equation (81) increases the value of q_{k-1}, whereas the third term decreases the value of q_{k-1}. Considering the two terms together, we see that q_{k-1} increases if and only if $r_k \geq 1/2$. If r_k is close to zero, then q_{k-1} will be significantly reduced, since the order of $1/r_k$ in the second term is only one, but its order in the third term is two. This and the fact that $q_k \geq 0$ for all k imply that, in the case when q_{k-1} is very small, r_k must be relatively large. Further investigations along the observations can lead to the following result of the DY method independent of the line search.

Theorem 5. *Consider the DY methods (2),(3) and (77), where d_k is a descent direction. Assume that $0 < \gamma \leq \|g_k\| \leq \overline{\gamma}$ holds for all $k \geq 1$. There must exist positive constants δ_1, δ_2 and δ_3 such that the relations*

$$-g_k^T d_k \geq \frac{\delta_1}{\sqrt{k}}, \qquad \|d_k\|^2 \geq \frac{\delta_2}{k}, \qquad r_k \geq \frac{\delta_3}{\sqrt{k}} \tag{83}$$

hold for all $k \geq 1$. Further, for any $p \in (0, 1)$, there must positive constants δ_4, δ_5, δ_6 such that, for any k, the relations

$$-g_i^T d_i \geq \delta_4, \qquad \|d_i\|^2 \geq \delta_5, \qquad r_i \geq \delta_6 \tag{84}$$

holds for at least $\lceil pk \rceil$ values of $i \in [1, k]$.

The above theorem enables us to establish global convergence for the DY method provided that the line search is such that

$$f_k - f_{k+1} \geq c \min\left\{-g_k^{\mathrm{T}} d_k, \|d_k\|^2, q_k^{-1}\right\}, \tag{85}$$

for all $k \geq 1$ and some $c > 0$. Consequently, we can analyze the convergence properties of the DY method using the standard Wolfe line search, the Armijo line search [56], and the line search proposed in De Leone *et al*. [40] and Grippo *et al*. [57] for no-derivative methods.

In general, once some optimization method fails to generate a descent direction, a usual remedy is to do a restart along $-g_k$. As shown in Dai [55], the DY direction can act the role of the negative gradient and meanwhile guarantee the global convergence. A numerical experiment with the memoryless BFGS method in Dai [55] demonstrated this finding.

Since the DY method has the same drawback as the FR method, namely, it cannot recover from cycles of tiny steps, it is natural to consider the hybrid of the DY and HS methods like those for the FR and PRP methods in Touati-Ahmed and Storey [30] and Gilbert and Nocedal [29]. Under the standard Wolfe line search, Dai and Yuan [13] extended Theorem 4 to any method (2),(3) with

$$\beta_k \in \left[-\frac{1-\sigma}{1+\sigma}\,\beta_k^{\mathrm{DY}},\ \beta_k^{\mathrm{DY}}\right], \tag{86}$$

where σ is the parameter in the Wolfe condition (7). In spite of a large admissible interval, the numerical results of Dai and Yuan [13] indicated that the following hybrid is preferable in real computations

$$\beta_k^{\mathrm{DYHS}} = \max\left\{0, \min\left\{\beta_k^{\mathrm{HS}}, \beta_k^{\mathrm{DY}}\right\}\right\}. \tag{87}$$

Unlike the Touati-Ahmed and Storey (TS) and GN hybrid methods, the DYHS method using standard Wolfe line searches performs much better than the PRP method using

strong Wolfe line searches [13]. The latter was generally believed to be one of the most efficient conjugate gradient algorithms.

It is well known that some quasi-Newton methods can be expressed in a unified way and their properties can be analyzed uniformly [58,59]. On the contrary, nonlinear conjugate gradient methods were often analyzed individually. To change the situation, Dai and Yuan [60] proposed a family of conjugate gradient methods, in which

$$\beta_k(\lambda) = \frac{||g_k||^2}{\lambda||g_{k-1}||^2 + (1-\lambda)d_{k-1}^{\mathrm{T}}y_{k-1}}, \quad \lambda \in [0,1]. \tag{88}$$

This family can be regarded as some kind of convex combination of the FR and DY methods. Dai and Yuan [61] further extended the family to the case $\lambda \in (-\infty, +\infty)$ and presented some unified convergence results. Independently, Nazareth [8] regarded the FR, PRP, HS, and DY formulas as the four leading contenders for the conjugate gradient parameter and proposed a two-parameter family:

$$\beta_k(\lambda_k, \mu_k) = \frac{\lambda_k||g_k||^2 + (1-\lambda_k)g_k^{\mathrm{T}}y_{k-1}}{\mu_k||g_{k-1}||^2 + (1-\mu_k)d_{k-1}^{\mathrm{T}}y_{k-1}}, \quad \lambda_k, \mu_k \in [0,1]. \tag{89}$$

The methods that take the convex combination $\lambda_k\beta_k^{\mathrm{HS}} + (1-\lambda_k)\beta_k^{\mathrm{DY}}$, considered in Andrei [62], can be regarded as a subfamily of Equation (89) with $\mu_k = 0$. Several efficient choices for λ_k in this subfamily are also studied in Andrei [62] based on different secant conditions.

Later, based on FR, PRP, HS, DY, CD, and the formula

$$\beta_k^{LS} = \frac{g_k^{\mathrm{T}}y_{k-1}}{-d_{k-1}^{\mathrm{T}}g_{k-1}} \tag{90}$$

by Liu and Storey [63], Dai and Yuan [64] proposed a three-parameter family

$$\beta_k(\lambda_k, \mu_k, \omega_k) = \frac{||g_k||^2 - \lambda_k g_k^{\mathrm{T}}g_{k-1}}{||g_{k-1}||^2 + \mu_k g_k^{\mathrm{T}}d_{k-1} - \omega_k\beta_{k-1}g_{k-1}^{\mathrm{T}}d_{k-2}}, \tag{91}$$

where $\lambda_k \in [0,1]$, $\mu_k \in [0,1]$, and $\omega_k \in [0, 1-\mu_k]$ are parameters. One subfamily of the methods (91) with $\lambda_k = 1$, $\mu_k = 0$, and $\omega_k = u$ is studied in Shi and Guo [65] with an efficient nonmonotone line search. Further, Dai [66] studied a family of hybrid conjugate gradient methods, in which

$$\beta_k(\mu_k, \omega_k, \tau_k) = \frac{\max\left\{0, \min\left\{g_k^{\mathrm{T}}y_{k-1}, \tau_k\|g_k\|^2\right\}\right\}}{(\tau_k + \omega_k)g_k^{\mathrm{T}}d_{k-1} + \mu_k\|g_{k-1}\|^2 + (1-\mu_k)(-d_{k-1}^{\mathrm{T}}g_{k-1})} \tag{92}$$

where $\mu_k \in [0,1]$, $\omega_k \in [0, 1-\mu_k]$ and $\tau_k \in [1, +\infty)$ are parameters.

The DYCD Method

Suppose that M is some fixed positive integer, and λ and δ are constants in $(0,1)$. Given an initial guess $\overline{\alpha}_k$ at the kth iteration, the nonmonotone line search by Grippo *et al*. [67] is to compute the least nonnegative integer m such that the steplength $\alpha_k = \overline{\alpha}_k\lambda^m$ satisfies the following relation:

$$f(x_k + \alpha_k d_k) \le \max\{f_k, \ldots, f_{k-M(k)}\} + \delta\alpha_k g_k^{\mathrm{T}}d_k, \tag{93}$$

where $M(k) = \min(M, k-1)$. To enforce a descent search direction at every iteration in this situation, Dai [14] considered a hybrid of the DY and CD methods, namely, (2) and (3) with

$$\beta_k^{\mathrm{DYCD}} = \frac{\|g_k\|^2}{\max\left\{d_{k-1}^{\mathrm{T}}y_{k-1}, -d_{k-1}^{\mathrm{T}}g_{k-1}\right\}}. \tag{94}$$

It is proved in Dai [14] that the DYCD method possesses the descent property without line searches. Further, there is the global convergence if the DYCD method is combined with

the above nonmonotone line search. Surprisingly, a variant of the DYCD method tested in Dai [14] was able to solve all the 18 test problems in Moré *et al* [68].

SUFFICIENT DESCENT CONJUGATE GRADIENT METHODS

In this section, we summarize several nonlinear conjugate gradient methods that can guarantee the sufficient descent condition (10), especially the CG_Descent method by Hager and Zhang [7,69].

Though the sufficient descent condition (10) is not scale invariant, there is, however, some difficulty to differentiate *descent* conjugate gradient methods and *sufficient descent* conjugate gradient methods. More exactly, if d_k satisfies $g_k^{\mathrm{T}} d_k < 0$, we can define another method whose search direction is $\overline{d}_k = \left(-c\,\|g_k\|^2 / g_k^{\mathrm{T}} d_k\right) d_k$, such that $g_k^{\mathrm{T}} \overline{d}_k = -c\|g_k\|^2$.

Let us take the DY method as an illustrative example. A variant of the DY method is given in Dai [55], where d_k takes the form

$$d_k = -\frac{d_{k-1}^{\mathrm{T}} y_{k-1}}{\|g_k\|^2} g_k + d_{k-1}. \tag{95}$$

Since $d_1 = -g_1$, we can get by the induction principle that

$$g_k^{\mathrm{T}} d_k = -\|g_1\|^2, \quad \text{for all } k \geq 1. \tag{96}$$

Further, if a scaling factor $\|g_k\|^2 / \|g_{k-1}\|^2$, that is, the formula β_k^{FR} exactly, is introduced for each search direction d_k (except d_1), we obtain the scheme

$$d_k = -\frac{d_{k-1}^{\mathrm{T}} y_{k-1}}{\|g_{k-1}\|^2} g_k + \beta_k^{\mathrm{FR}} d_{k-1}. \tag{97}$$

In this case, we have that $-g_k^{\mathrm{T}} d_k = \|g_k\|^2$ for all k, which implies that the sufficient descent condition (10) holds with $c = 1$. It is worth mentioning that the above scheme (97) is obtained by Zhang *et al*. [70] (see also the section titled "Several New Methods That Guarantee Sufficient Descent") with the motivation of modifying the FR method. They found that, the numerical performance of this scheme is very promising for a large collection of test problems in the CUTEr library [71].

The CG_Descent Method

To ensure the sufficient descent condition (10), Hager and Zhang [7] proposed a family of conjugate gradient methods, where

$$\beta_k^{\mathrm{HZ}}(\lambda_k) = \frac{g_k^{\mathrm{T}} y_{k-1}}{d_{k-1}^{\mathrm{T}} y_{k-1}} - \lambda_k \left(\frac{\|y_{k-1}\|^2 g_k^{\mathrm{T}} d_{k-1}}{(d_{k-1}^{\mathrm{T}} y_{k-1})^2} \right), \tag{98}$$

where $\lambda_k \geq \overline{\lambda} > 1/4$ controls the relative weight placed on the conjugacy degree versus the descent degree of the search direction. This family is clearly related to the DL method (67) with

$$t = \lambda_k \frac{\|y_{k-1}\|^2}{s_{k-1}^{\mathrm{T}} y_{k-1}}. \tag{99}$$

To verify the sufficient descent condition for the Hager and Zhang (HZ) method, we have by Equations (3) and (98) that

$$g_k^{\mathrm{T}} d_k = -\|g_k\|^2 + \left(\frac{g_k^{\mathrm{T}} y_{k-1} (g_k^{\mathrm{T}} d_{k-1})}{d_{k-1}^{\mathrm{T}} y_{k-1}} \right) - \lambda_k \left(\frac{\|y_{k-1}\|^2 (g_k^{\mathrm{T}} d_{k-1})^2}{(d_{k-1}^{\mathrm{T}} y_{k-1})^2} \right). \tag{100}$$

Now, by applying

$$u_k = \frac{1}{\sqrt{2\lambda_k}} (d_{k-1}^{\mathrm{T}} y_{k-1}) g_k, \qquad v_k = \sqrt{2\lambda_k} (g_k^{\mathrm{T}} d_{k-1}) y_{k-1} \tag{101}$$

into the inequality

$$u_k^{\mathrm{T}} v_k \leq \frac{1}{2} \left(\|u_k\|^2 + \|v_k\|^2 \right), \tag{102}$$

we can obtain

$$\frac{g_k^{\mathrm{T}} y_{k-1} (g_k^{\mathrm{T}} d_{k-1})}{d_{k-1}^{\mathrm{T}} y_{k-1}} \leq \frac{1}{4\lambda_k} \|g_k\|^2 + \lambda_k \left(\frac{\|y_{k-1}\|^2 g_k^{\mathrm{T}} d_{k-1}}{(d_{k-1}^{\mathrm{T}} y_{k-1})^2} \right). \tag{103}$$

Therefore by Equations (100) and (103), we have that

$$g_k^{\mathrm{T}} d_k \leq -\left(1 - \frac{1}{4\lambda_k}\right) \|g_k\|^2, \qquad (104)$$

which with the restriction of λ_k means that the sufficient descent condition (10) holds with $c = 1 - (4\overline{\lambda})^{-1}$.

In order to obtain global convergence for general nonlinear functions, Hager and Zhang truncated their conjugate gradient parameter similar to the PRP$^+$ method. More exactly, they suggested to choose

$$\beta_k^{\mathrm{HZ}+}(\lambda_k) = \max\left\{\beta_k^{\mathrm{HZ}}(\lambda_k),\, \eta_k\right\},$$
$$\eta_k = \frac{-1}{\|d_{k-1}\|^2 \min\{\eta,\, \|g_{k-1}\|\}}, \qquad (105)$$

where $\eta > 0$ is a constant. With this truncation, they established the global convergence of the modified method (105) with the standard Wolfe line search for general functions.

Hager and Zhang [69,72] tested the value of $\lambda_k = 2$ for the family with a precisely developed efficient line search. For a large collection of large-scale test problems in the CUTEr library [71], the new method, called CG_DESCENT, performs better than both PRP$^+$ of Gilbert and Nocedal and L-BFGS of Liu and Nocedal.

More efficient choices of λ_k, however, have been found in Kou and Dai [73] by projecting the scaled memoryless BFGS direction defined in Equations (60) and (63) into the one-dimensional manifold $\{-g_k + \beta\, d_k : \beta \in R\}$. By taking the scaling factors $\gamma_{k-1} = \frac{s_{k-1}^{\mathrm{T}} y_{k-1}}{\|y_{k-1}\|^2}$ and $\gamma_{k-1} = \frac{\|s_{k-1}\|^2}{s_{k-1}^{\mathrm{T}} y_{k-1}}$, they suggest the uses of $\lambda_k = 2 - \frac{d_{k-1}^{\mathrm{T}} y_{k-1}}{\|d_{k-1}\|^2 \|y_{k-1}\|^2}$ and $\lambda_k = 1$, respectively. A simple and efficient nonmonotone line search criterion is also designed in Kou and Dai [73], that can guarantee the global convergence of the new methods.

Several New Methods That Guarantee Sufficient Descent

The remarkable property of the HZ method (98) that can guarantee the sufficient descent condition (10) for general functions have attracted several further investigations.

A direct generalization of Equation (98) is given in Yu and Guan [74] (see also Yu *et al*. [75]). They found that, for any β_k of the form

$$\beta_k = \frac{g_k^{\mathrm{T}} v_k}{\Delta_k}, \quad \text{for some } v_k \in R^n \text{and} \Delta_k \in R, \qquad (106)$$

there is a corresponding formula

$$\beta_k^{\mathrm{YG}}(C) = \frac{g_k^{\mathrm{T}} v_k}{\Delta_k} - \frac{C\,\|v_k\|^2}{\Delta_k^2} g_k^{\mathrm{T}} d_{k-1}, \quad (107)$$

where $C > 1/4$, such that Equation (10) holds with $c = 1 - (4C)^{-1}$. Since almost all of the conjugate gradient parameters can be written into Equation (106), we can obtain various extensions that can guarantee sufficient descent. It is obvious that the HZ formula (98) is corresponding to Equation (107) with the HS formula (55), where $v_k = y_{k-1}$ and $\Delta_k = d_{k-1}^{\mathrm{T}} y_{k-1}$. The extensions of β_K^{FR}, β_k^{PRP}, β_k^{DY}, β_k^{CD}, and β_k^{LS} are also provided in Yu and Guan [74]. A further generalization of this framework on the spectral conjugate gradient method (see Birgin and Martinez [76] or the section titled "Several Topics on Conjugate Gradient Methods") is given in Yu *et al*. [75].

Another general way of producing sufficient descent conjugate gradient methods is provided in Cheng [77] and Cheng and Liu [78]. Its basic is as follows. For any search direction $-g_k + \beta_k\, d_{k-1}$, which need not be descent, an orthogonal projection to the null space of g_k leads to the vector

$$d_k^{\perp} = \left(I - \frac{g_k g_k^{\mathrm{T}}}{\|g_k\|^2}\right)(-g_k + \beta_k\, d_{k-1}). \quad (108)$$

The search direction defined by

$$d_k = -g_k + d_k^{\perp} = -\left(1 + \beta_k \frac{g_k^{\mathrm{T}} d_{k-1}}{\|g_k\|^2}\right) g_k + \beta_k\, d_{k-1} \qquad (109)$$

then always satisfies $g_k^{\mathrm{T}} d_k = -\|g_k\|^2$. If the line search is exact, the second term in the paranthesis of Equation (109) is missing

since $g_k^{\mathrm{T}} d_{k-1} = 0$. Hence the above scheme reduces to the linear conjugate gradient method in the ideal case. The above procedure with $\beta_k = \beta_k^{\mathrm{PRP}}$ is studied in Cheng [77]. As shown in Cheng and Liu [78], setting $\beta_k = \beta_k^{\mathrm{FR}}$ in Equation (109) leads to the scheme (97). Another variant corresponding to Yabe and Takano [49] (see the end of the section titled "The Hestenes–Stiefel Method") is also investigated in Cheng and Liu [78].

Observing that the search direction (64) defined by the memoryless BFGS method is formed by the vectors $-g_k$, d_{k-1}, and y_{k-1}, Zhang *et al*. [17] proposed the following modification of the PRP method

$$d_k = -g_k + \frac{g_k^{\mathrm{T}} y_{k-1}}{\|g_{k-1}\|^2} d_{k-1} - \frac{g_k^{\mathrm{T}} d_{k-1}}{\|g_{k-1}\|^2} y_{k-1}. \tag{110}$$

By multiplying the above d_k with g_k^{T}, one can see that the corresponding values of the last two terms have opposite signs and hence $g_k^{\mathrm{T}} d_k = -\|g_k\|^2$. The above scheme was implemented in Zhang *et al*. [17] with an Armijo-type line search in relation to De Leone *et al*. [40], yielding comparable numerical results with CG_DESCENT. Although Equation (110) reduces to Equation (3) in case of exact line searches, this scheme is not a standard conjugate gradient method of the form (3) any more.

SEVERAL TOPICS ON CONJUGATE GRADIENT METHODS

As shown in the previous sections, various formulas have been proposed for the nonlinear conjugate gradient parameter β_k, whereas there is not much to do with the choice of this parameter in the linear conjugate gradient method (it is a consensus to use the FR formula (31) there). While some of the existing conjugate gradient algorithms, like the DYHS method (87) and the CG_DESCENT method (98) among others, have proved more efficient than the PRP method, we feel there is still much more room to seek the best nonlinear conjugate gradient algorithms.

As a lot of attention has been paid to the choice of β_k, it is actually also important how to choose the stepsize α_k. Some joint consideration is given by Yuan and Stoer [79], which aims to find the best points of the function over the two-dimensional manifold

$$x_k + \mathrm{Span}\left\{-\alpha g_k + \beta d_{k-1} : \alpha, \beta \in R^2\right\} \tag{111}$$

as the next iterates. Motivated by the success of the Barzilai–Borwein stepsize in the steepest descent method [80,81], Birgin and Martinez [76] proposed the so-called spectral conjugate gradient method that takes the search direction

$$d_k = -\frac{1}{\delta_k} g_k + \frac{g_k^{\mathrm{T}}(y_{k-1} - \delta_k s_{k-1})}{\delta_k d_{k-1}^{\mathrm{T}} y_{k-1}} d_{k-1}. \tag{112}$$

The efficient combination of the Barzilai–Bor wein method and the conjugate gradient method, however, is still not known to us. Specifically, the study of Dai and Liao [48] indicates that when the iterate gets close to the solution, the Barzilai–Borwein stepsize can always be accepted by the often-employed nonmonotone line search. We wonder whether there is a similar result for the spectral conjugate gradient method or some of its suitable alternatives.

In addition to the standard conjugate gradient methods of the form (2) and (3), another class of two-term conjugate gradient methods is called *method of shortest residuals* (SR), that was first presented by Hestenes [82] and studied by Pytlak and Tarnawski [83], Dai and Yuan [84], and the references therein. The SR method defines the search direction by

$$d_k = -Nr\{g_k,\ -\beta_k d_{k-1}\}, \tag{113}$$

where β_k is a scalar and $Nr\{a,\ b\}$ is defined as the point from a line segment spanned by the vectors a and b which has the smallest

norm, namely,

$$\|Nr\{a,\ b\}\| = \min\left\{\|\lambda a + (1-\lambda) b\| : 0 \le \lambda \le 1\right\}. \quad (114)$$

If $\beta_k \equiv 1$, the corresponding variant of the SR method generates the same iterations as the FR method does in case of exact line searches. The formula of β_k corresponding to the PRP method is also given in Pytlak [85] and modified in Dai and Yuan [84]. If, further, the function is quadratic, these variants of the SR method are equivalent and the direction d_k proves to be the SR in the $(k-1)$-simplex whose vertices are $-g_1, \ldots, -g_k$. For the SR method, the descent property of d_k is naturally implied by its definition (see Pytlak [85] and Dai and Yuan [84] for details). In contrast to the standard conjugate gradient method, where the size of d_k may become very large, the SR method has the trend of pushing $\|d_k\|$ very small. Therefore, we wonder whether there exists some family of methods that includes the standard conjugate gradient method and the SR method as its members. If this is the case, it might be possible to find more efficient methods that monitor the size of $\|d_k\|$ in a more efficient way.

If the storage of more vectors is admissible, one may consider to choose, for example, three-term conjugate gradient methods [17,18,86] and limited-memory quasi-Newton methods [87,88] for solving large-scale optimization problems, other than the two-term conjugate gradient methods. As an alternative, one may think of forming some preconditioner for conjugate gradient methods through the information already achieved in the previous fewer iterations [89–91]. Unlike the linear conjugate gradient method, where a constant preconditioner is usually satisfactory, a robust and efficient conjugate gradient method for highly nonlinear functions requires to be dynamically preconditioned. Therefore, it remains to study how to precondition the nonlinear conjugate gradient method in more effective ways.

Acknowledgments

The author thanks Professor Ya-xiang Yuan and the anonymous editor very much for their useful suggestions and comments. This work was partially supported by the Chinese NSF grants 10831106 and the CAS grant kjcx-yw-s7-03.

REFERENCES

1. Hestenes MR, Stiefel E. Method of conjugate gradient for solving linear system. J Res Natl Bur Stand 1952;49:409–436.
2. Shewchuk JR. An introduction to the conjugate gradient method without the agonizing pain. Technical Report CS-94-125. Pittsburgh (PA): Carnegie Mellon University; 1994.
3. Fletcher R, Reeves CM. Function minimization by conjugate gradients. Comput J 1964;7:149–154.
4. Crowder HP, Wolfe P. Linear convergence of the conjugate gradient method. IBM J Res Dev 1969;16:431–433.
5. Cohen A. Rate of convergence of several conjugate gradient method algorithms. SIAM J Numer Anal 1972;9:248–259.
6. McCormick P, Ritter K. Alternative proofs of the convergence properties of the conjugate-gradient method. J Optim Theory Appl 1975;13(5):497–518.
7. Hager WW, Zhang H. A survey of nonlinear conjugate gradient methods. Pac J Optim 2006;2(1):335–358.
8. Nazareth L. Conjugate-gradient methods. In: Floudas C, Pardalos P, editors. Encyclopedia of optimization. Boston (MA), Dordrecht, The Netherlands: Kluwer Academic Publishers; 2001. pp. 319–323.
9. Nazareth JL. Conjugate gradient method. Comput Stat 2009;1(3):348–353. Wiley Interdisciplinary Reviews.
10. Nocedal J. Theory of algorithms for unconstrained optimization. Acta Numerica 1991;1:199–242.
11. Nocedal J. Conjugate gradient methods and nonlinear optimization. In: Adams L, Nazareth JL, editors. Linear and nonlinear conjugate gradient-related methods. Philadelphia (PA): SIAM; 1996. pp. 9–23.
12. Dai YH, Yuan Y. A nonlinear conjugate gradient method with a strong global convergence property. SIAM J Optim 1999;10(1):177–182.

13. Dai YH, Yuan Y. An efficient hybrid conjugate gradient method for unconstrained optimization. Ann Oper Res 2001;103:33–47.
14. Dai YH. A nonmonotone conjugate gradient algorithm for unconstrained optimization. J Syst Sci Complex 2002;15(2):139–145.
15. Grippo L, Lucidi S. A globally convergent version of the Polak-Ribière conjugate gradient method. Math Program 1997;78:375–391.
16. Wang CY, Zhang YZ. Global convergence properties of *s*-related conjugate gradient methods. Chin Sci Bull 1998;43(23):1959–1965.
17. Zhang L, Zhou W, Li D. A descent modified Polak-Ribière-Polyak conjugate gradient method and its global convergence. IMA J Numer Anal 2006;26(4):629–640.
18. Powell MJD. Restart procedures of the conjugate gradient method. Math Program 1977;12:241–254.
19. Zoutendijk G. Nonlinear programming, computational methods. In: Abadie J, editor. Integer and nonlinear programming. Amsterdam: North-Holland Publishing Co.; 1970. pp. 37–86.
20. Wolfe P. Convergence conditions for ascent methods. SIAM Rev 1969;11:226–235.
21. Wolfe P. Convergence conditions for ascent methods II: some corrections. SIAM Rev 1971;13:185–188.
22. Dai YH, Han J, Liu G, et al. Convergence properties of nonlinear conjugate gradient methods. SIAM J Optim 1999;10(2): 345–358.
23. Dai YH. Convergence analysis of nonlinear conjugate gradient methods. In: Wang Y, Yagola AG, Yang C, editor. Optimization and regularization for computational inverse problems and applications. Berlin Heidelberg: Springer; 2010. pp. 157–181.
24. Pu D, Yu W. On the convergence properties of the DFP algorithms. Ann Oper Res 1990;24:175–184.
25. Al-Baali M. Descent property and global convergence of the Fletcher-Reeves method with inexact linesearch. IMA J Numer Anal 1985;5:121–124.
26. Liu GH, Han JY, Yin HX. Global convergence of the Fletcher-Reeves algorithm with an inexact line search. Appl Math J Chin Univ Ser B 1995;10:75–82.
27. Dai YH, Yuan Y. Convergence properties of the Fletcher-Reeves method. IMA J Numer Anal 1996;16:155–164.
28. Dai YH, Yuan Y. Convergence of the Fletcher-Reeves method under a generalized Wolfe search. J Comput Math Chin Univ 1996;2:142–148.
29. Gilbert JC, Nocedal J. Global convergence properties of conjugate gradient methods for optimization. SIAM J Optim 1992;2:21–42.
30. Touati-Ahmed D, Storey C. Global convergent hybrid conjugate gradient methods. J Optim Theory Appl 1990;64:379–397.
31. Polak E, Ribière G. Note sur la convergence de méthods de directions conjugées. Rev Fr Inform Rech Oper 1969;16:35–43.
32. Polyak BT. The conjugate gradient method in extremum problems. USSR Comput Math Math Phys 1969;9:94–112.
33. Yuan Y. Numerical methods for nonlinear programming. Shanghai: Shanghai Scientific & Technical Publishers; 1993 (in Chinese).
34. Powell MJD. Nonconvex minimization calculations and the conjugate gradient method. In: Griffiths DF, editor. Numerical analysis, Lecture notes in mathematics 1066. Berlin: Springer; 1984. pp. 122–141.
35. Powell MJD. Convergence properties of algorithms for nonlinear optimization. SIAM Rev 1986;28:487–500.
36. Dai YH, Yuan Y. Nonlinear conjugate gradient methods. Shanghai: Shanghai Scientific & Technical Publishers; 2000 (in Chinese).
37. Lemaréchal C. A view of line searches. In: Auslander A, Oetti W, Stoer J, editors. Optimization and optimal control, Lecture notes in control and information 30. Berlin: Springer; 1981. pp. 59–78.
38. Fletcher R. Practical methods of optimization. Volume 1, Unconstrained Optimization. New York: John Wiley & Sons, Inc.; 1987.
39. Moré J, Thuente DJ. On line search algorithms with guaranteed sufficient decrease. ACM Trans Math Softw 1994;20:286–307.
40. De Leone R, Gaudioso M, Grippo L. Stopping criteria for linesearch methods without derivatives. Math Program 1984;30:285–300.
41. Nocedal J. Large scale unconstrained optimization. In: Watson A, Duff I, editors. The state of the art in numerical analysis. Oxford: Oxford University Press; 1997. pp. 311–338.
42. Dai YH. Convergence of conjugate gradient methods with constant stepsizes. Optim Meth Software 2010; DOI: 10.1080/10556781003721042.

43. Chen XD, Sun J. Global convergence of two-parameter family of conjugate gradient methods without line search. J Comput Appl Math 2002;146:37–45.
44. Qi HD, Han JY, Liu GH. A modified Hestenes-Stiefel conjugate gradient algorithm. Chin Ann Math Ser A 1996;17(3):177–184.
45. Perry JM. A class of conjugate gradient algorithms with a two-step variable-metric memory. Discussion Paper 269. Evanston (IL): Center for Mathematical Studies in Economics and Management Sciences, Northwestern University; 1977.
46. Shanno DF. Conjugate gradient methods with inexact searches. Math Oper Res 1978;3:244–256.
47. Buckley A. Conjugate gradient methods. In: Powell MJD, editor. Nonlinear optimization 1981. London: Academic Press; 1982. pp. 17–22.
48. Dai YH, Liao LZ. New conjugacy conditions and related nonlinear conjugate gradient methods. Appl Math Optim 2001;43(1):87–101.
49. Yabe H, Takano M. Global convergence properties of nonlinear conjugate gradient methods with modified secant condition. Comput Optim Appl 2004;28:203–225.
50. Li GY, Tang CM, Wei ZX. New conjugacy condition and related new conjugate gradient methods for unconstrained optimization. J Comput Appl Math 2007;202:523–539.
51. Zhang JZ, Deng NY, Chen LH. New quasi-Newton equation and related methods for unconstrained optimization. J Optim Theory Appl 1999;102:147–167.
52. Zhang JZ, Xu CX. Properties and numerical performance of quasi-Newton methods with modified quasi-Newton equations. J Comput Appl Math 2001;137:269–278.
53. Wei Z, Yu G, Yuan G, et al. The superlinear convergence of a modified BFGS-type method for unconstrained optimization. Comput Optim Appl 2004;29(3):315–332.
54. Dai YH, Yuan Y. Convergence properties of the conjugate descent method. Adv Math 1996;26(6):552–562.
55. Dai YH. New properties of a nonlinear conjugate gradient method. Numer Math 2001;89(1):83–98.
56. Armijo L. Minimization of functions having Lipschitz continuous partial derivatives. Pac J Math 1966;16:1–3.
57. Grippo L, Lampariello F, Lucidi S. Global convergence and stabilization of unconstrained minimization methods without derivatives. J Optim Theory Appl 1988;56:385–406.
58. Broyden CG. The convergence of a class of double-rank minimization algorithms 1. General considerations. J Inst Math Appl 1970;6:76–90.
59. Byrd R, Nocedal J, Yuan Y. Global convergence of a class of variable metric algorithms. SIAM J Numer Anal 1987;4:1171–1190.
60. Dai YH, Yuan Y. A class of globally convergent conjugate gradient methods. Research report ICM-98-030. Beijing: Institute of Computational Mathematics and Scientific/Engineering Computing, Chinese Academy of Sciences; 1998.
61. Dai YH, Yuan Y. An extension class of nonlinear conjugate gradient methods. In: Li D, editors. Proceedings of the 5th International Conference on Optimization: Techniques and Applications. Hongkong: Institute of Computational Mathematics and Scientific/Engineering Computing, Chinese Academy of Sciences; 2001. pp. 778–785.
62. Andrei N. Accelerated hybrid conjugate gradient algorithm with modified secant condition for unconstrained optimization. Numer Algorithms 2010;54(1):1017–1398.
63. Liu Y, Storey C. Efficient generalized conjugate gradient algorithms, part 1: theory. J Optim Theory Appl 1991;69:129–137.
64. Dai YH, Yuan Y. A three-parameter family of conjugate gradient methods. Math Comput 2001;70:1155–1167.
65. Shi ZJ, Guo J. A new family of conjugate gradient methods. J Comput Appl Math 2009;224:444–457.
66. Dai YH. A family of hybrid conjugate gradient methods for unconstrained optimization. Math Comput 2003;72:1317–1328.
67. Grippo L, Lamparillo F, Lucidi S. A nonmonotone line search technique for Newton's method. SIAM J Numer Anal 1986;23:707–716.
68. Moré JJ, Garbow BS, Hillstrom KE. Testing unconstrained optimization software. ACM Trans Math Softw 1981;7:17–41.
69. Hager WW, Zhang H. A new conjugate gradient method with guaranteed descent and an efficient line search. SIAM J Optim 2005;16(1):170–192.
70. Zhang L, Zhou W, Li D. Global convergence of a modified Fletcher-Reeves conjugate gradient method with Armijo-type line search. Numer Math 2006;104(4):561–572.

71. Bongartz I, Conn AR, Gould NIM, et al. CUTE: constrained and unconstrained testing environments. ACM Trans Math Softw 1995;21:123–160.
72. Hager WW, Zhang H. Algorithm 851: CG_DESCENT, a conjugate gradient method with guaranteed descent. ACM Trans Math Softw 2006;32(1):113–137.
73. Kou CX, Dai YH. A new conjugate gradient algorithm with nonmonotone line search, Research report. Beijing: LSEC, ICMSEC, Academy of Mathematics and Systems Science, Chinese Academy of Sciences; 2010.
74. Yu GH, Guan LT. New descent nonlinear conjugate gradient methods for large-scale optimization. Technical Report. Department of Scientific Computation and Computer Applications, Sun Yat-Sen University, Guangzhou, P. R. China; 2005.
75. Yu GH, Guan LT, Chen WF. Spectral conjugate gradient methods with sufficient descent property for large-scale unconstrained optimization. Optim Methods Softw 2008;23(2):275–293.
76. Birgin EG, Martinez JM. A spectral conjugate gradient method for unconstrained optimization. Appl Math Optim 2001;43:117–128.
77. Cheng WY. A two-term PRP-based descent method. Numer Funct Anal Optim 2007;28:1217–1230.
78. Cheng WY, Liu QF. Sufficient descent nonlinear conjugate gradient methods with conjugacy conditions. Numer Algorithms 2010;53:113–131.
79. Yuan Y, Stoer J. A subspace study on conjugate gradient algorithms. Z Angew Math Mech 1995;75(1):69–77.
80. Barzilai J, Borwein JM. Two-point step size gradient methods. IMA J Numer Anal 1988;8:141–148.
81. Raydan M. The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem. SIAM J Optim 1997;7(1):26–33.
82. Hestenes MR. Conjugate direction methods in optimization. New York, Heidelberg, Berlin: Springer; 1980.
83. Pytlak R, Tarnawski T. On the method of shortest residuals for unconstrained optimization. J Optim Theory Appl 2007;133(1):99–110.
84. Dai YH, Yuan Y. Global convergence of the method of shortest residuals. Numer Math 1999;83:581–598.
85. Pytlak R. On the convergence of conjugate gradient algorithm. IMA J Numer Anal 1989;14:443–460.
86. Nazareth JL. A conjugate direction algorithm without line searches. J Optim Theory Appl 1977;23(3):373–387.
87. Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. Math Program 1989;45:503–528.
88. Sun LP. Updating the self-scaling symmetric rank one algorithm with limited memory for large-scale unconstrained optimization. Comput Optim Appl 2004;27:23–29.
89. Buckley A. A combined conjugate gradient quasi-Newton minimization algorithms. Math Prog 1978;15:200–210.
90. Morales JL, Nocedal J. Automatic preconditioning by limited memory quasi-newton updating. SIAM J Optim 2000; 10(4):1079–1096.
91. Andrei N. Accelerated scaled memoryless BFGS preconditioned conjugate gradient algorithm for unconstrained optimization. Eur J Oper Res 2010;204:410–420.