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,
Zhong Guan Cun Donglu 55, Beijing, 100190, P.R. China.
E-mail: dyh@lsec.cc.ac.cn

Abstract

Conjugate gradient methods are a class of important methods for solving linear equations and for solving nonlinear optimization. In this article, a review on conjugate gradient methods for unconstrained optimization is given. They are divided into early conjugate gradient methods, descent conjugate gradient methods and sufficient descent conjugate gradient methods. Two general convergence theorems are provided for the conjugate gradient method assuming the descent property of each search direction. Some research issues on conjugate gradient methods are mentioned.

Key words. conjugate gradient method, line search, descent property, sufficient descent condition, global convergence.

Mathematics Subject Classification: 49M37, 65K05, 90C30.

*This work was partially supported by the Chinese NSF grants 10831106 and the CAS grant kjcx-yw-s7-03.

1 Introduction

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

$$\min f(x), \quad x \in R^n, \tag{1.1}$$

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

$$x_{k+1} = x_k + \alpha_k d_k, \tag{1.2}$$

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

$$d_k = \begin{cases} -g_k, & \text{for } k = 1; \\ -g_k + \beta_k d_{k-1}, & \text{for } k \geq 2, \end{cases} \tag{1.3}$$

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

It is known from (1.2) and (1.3) 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 method (1.2)-(1.3) reduces to the *linear* conjugate gradient method if the line search is exact, namely,

$$\alpha_k = \arg\min\{f(x_k + \alpha d_k); \alpha > 0\}. \tag{1.4}$$

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 by *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 [46] 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 [77]. The equivalence of the linear system to the minimization problem of $\frac{1}{2}x^TAx - b^Tx$ motivated Fletcher and Reeves [37] 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 conjugate gradient 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 [16]), while n-step quadratic convergence rate can be established if the method is restarted along the negative gradient every n-step (see Cohen [15] and MicCormick and Ritter [54]). Some recent reviews on nonlinear conjugate gradient methods can be found in Hager and Zhang [44], Nazareth [60, 61], Nocedal [62, 63], *etc.* This paper aims to provide a perspective view on the methods from the angle of descent property and global convergence.

Since the 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

$$\begin{aligned} f(x_k+\alpha_k d_k)-f(x_k) &\le \rho\,\alpha_k\, g_k^T d_k, && (1.5)\\ |g(x_k+\alpha_k d_k)^T d_k| &\le -\sigma\, g_k^T d_k, && (1.6) \end{aligned}$$

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, (1.5) and

$$g(x_k+\alpha_k d_k)^T d_k \ge \sigma\, g_k^T d_k, \tag{1.7}$$

where again $0<\rho<\sigma<1$. As is well known, the standard Wolfe line search is normly used in the implementation of quasi-Newton methods, another important class of methods for unconstrained optimization. The work of Dai and Yuan [30, 33] indicates that the use of standard Wolfe line searches is possible in the nonlinear conjugate gradient field. Besides this, there are quite a few references (for example, see [19, 41, 81, 93]) 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,

$$g_k^T d_k < 0. \tag{1.8}$$

For conjugate gradient methods, by multiplying (1.3) with g_k^T, we have

$$g_k^T d_k \;=\; -\|g_k\|^2+\beta_k\, g_k^T d_{k-1}. \tag{1.9}$$

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 $g_k\ne 0$. 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 (see [68]). Assume that no restarts are used. In this paper we say that, a conjugate gradient method is *descent* if (1.8) holds for all k, and is *sufficient descent* if the sufficient descent condition

$$g_k^T d_k \leq -c \, \|g_k\|^2, \tag{1.10}$$

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 § 5).

This survey is organized in the following way. In the next section, we will address two general convergence theorems for the method of the form (1.2)-(1.3) 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 3 to 5, respectively, with the emphases on the Fletcher-Reeves method, the Polak-Ribière-Polyak method, the Hestenes-Stiefel 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.

2 General convergence theorems

In this section, we give two global convergence theorems for any method of the form (1.2)-(1.3) assuming the descent condition (1.8) 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 paper, the symbol $\| \cdot \|$ denotes the two norm.

Assumption 2.1. *(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*

$$\|g(x) - g(y)\| \leq L \, \|x - y\|, \quad \textit{for all } x, y \in \mathcal{N}. \tag{2.1}$$

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 2.1 for the convergence results in this survey. Under Assumption 2.1 on f, we state a very useful result, which was obtained by Zoutendijk [94] and Wolfe [83, 84]. The relation (2.2) is usually called as the Zoutendijk condition.

Lemma 2.2. *Suppose that Assumption 2.1 holds. Consider any iterative method of the form (1.2), 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*

$$\sum_{k=1}^{\infty} \frac{(g_k^T d_k)^2}{||d_k||^2} < +\infty. \tag{2.2}$$

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 (1.3) 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

$$\liminf_{k\to\infty} \|g_k\| = 0, \tag{2.3}$$

and is *strongly convergent* if

$$\lim_{k\to\infty} \|g_k\| = 0, \tag{2.4}$$

If the iterations $\{x_k\}$ stay in a bounded region, (2.3) means that there exists at least one cluster point which is a stationary point of f, while (2.4) indicates that every cluster point of $\{x_k\}$ will be a stationary point of f.

To analyze the method of the form (1.2)-(1.3), besides (1.9), we derive another basic relation. By (1.3), we have $d_k + g_k = \beta_k d_{k-1}$ for all $k \geq 2$. Squaring both sides of this relation yields

$$\|d_k\|^2 = -2g_k^T d_k - \|g_k\|^2 + \beta_k^2 \|d_{k-1}\|^2. \tag{2.5}$$

The following theorem gives a general convergence result for any descent method of the form (1.2)-(1.3) 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 2.3. *[22] Suppose that Assumption 2.1 holds. Consider any method of the form (1.2)-(1.3) with d_k satisfying $g_k^T d_k < 0$ and with the strong Wolfe line search (1.5) and (1.6). Then the method is globally convergent if*

$$\sum_{k \geq 1} \frac{1}{\|d_k\|^2} = +\infty. \tag{2.6}$$

Proof. It follows from (1.9) and (1.6) that $|g_k^T d_k| + \sigma|\beta_k||g_{k-1}^T d_{k-1}| \geq \|g_k\|^2$, which with the Cauchy-Schwarz inequality gives

$$(g_k^T d_k)^2 + \beta_k^2 (g_{k-1}^T d_{k-1})^2 \geq c_1 \|g_k\|^4, \tag{2.7}$$

where $c_1 = (1+\sigma^2)^{-1}$ is constant. By (2.5), $g_k^T d_k < 0$ and (2.7), we have

$$\begin{aligned}
& \frac{(g_k^T d_k)^2}{\|d_k\|^2} + \frac{(g_{k-1}^T d_{k-1})^2}{\|d_{k-1}\|^2} \\
= \ & \frac{1}{\|d_k\|^2} \left[(g_k^T d_k)^2 + \frac{\|d_k\|^2}{\|d_{k-1}\|^2} (g_{k-1}^T d_{k-1})^2 \right] \\
\geq \ & \frac{1}{\|d_k\|^2} \left[(g_k^T d_k)^2 + \beta_k^2 (g_{k-1}^T d_{k-1})^2 - \frac{(g_{k-1}^T d_{k-1})^2}{\|d_{k-1}\|^2} \|g_k\|^2 \right] \\
\geq \ & \frac{1}{\|d_k\|^2} \left[c_1 \|g_k\|^4 - \frac{(g_{k-1}^T d_{k-1})^2}{\|d_{k-1}\|^2} \|g_k\|^2 \right].
\end{aligned} \tag{2.8}$$

Assume that (2.3) is false and there exists some constant $\gamma > 0$ such that

$$\|g_k\| \geq \gamma, \quad \text{for all } k \geq 1. \tag{2.9}$$

Notice that the Zoutendijk condition (2.2) implies that $g_k^T d_k / \|d_k\|$ tends to zero. By this, (2.8) and (2.9), we have for sufficiently large k,

$$\frac{(g_k^T d_k)^2}{\|d_k\|^2} + \frac{(g_{k-1}^T d_{k-1})^2}{\|d_{k-1}\|^2} \geq \frac{c_1}{2} \frac{\|g_k\|^2}{\|d_k\|^2}. \tag{2.10}$$

Thus by the Zoutendijk condition and (2.9), we must have that

$$\sum_{k \geq 1} \frac{1}{\|d_k\|^2} \leq \frac{1}{\gamma^2} \sum_{k \geq 1} \frac{\|g_k\|^2}{\|d_k\|^2} < +\infty, \tag{2.11}$$

which is a contradiction to the assumption (2.6). Therefore we must have the convergence relation (2.3) holds. □

We are now going to provide another general global convergence theorem for any descent method (1.2)-(1.3) with the standard Wolfe line search. To this aim, we define

$$t_k = \frac{\|d_k\|^2}{\phi_k^2}, \qquad \phi_k = \begin{cases} \|g_1\|^2, & \text{for } k = 1; \\ \prod_{j=2}^k \beta_j^2, & \text{for } k \geq 2. \end{cases} \tag{2.12}$$

By dividing (2.5) by ϕ_k^2 and noticing that $d_1 = -g_1$, we can obtain ([17]) that for all $k \geq 1$

$$t_k = -2\sum_{i=1}^k \frac{g_i^T d_i}{\phi_i^2} - \sum_{i=1}^k \frac{\|g_i\|^2}{\phi_i^2}. \tag{2.13}$$

Theorem 2.4. *[17] Suppose that Assumption 2.1 holds. Consider any method of the form (1.2)-(1.3) with d_k satisfying $g_k^T d_k < 0$ and with the standard Wolfe line search (1.5) and (1.7). Then the method is globally convergent if the scalar β_k is such that*

$$\sum_{k\geq 1}\prod_{j=2}^k \beta_j^{-2} = +\infty. \tag{2.14}$$

Proof. Define ϕ_k as in (2.12). The condition (2.14) is equivalent to

$$\sum_{k\geq 1}\frac{1}{\phi_k^2} = +\infty. \tag{2.15}$$

Noting that $-2g_i^T d_i - \|g_i\|^2 \leq (g_i^T d_i)^2/\|g_i\|^2$, it follows from (2.13) that

$$t_k \leq \sum_{i=1}^k \frac{(g_i^T d_i)^2}{\|g_i\|^2\,\phi_i^2}. \tag{2.16}$$

Since $t_k \geq 0$, the relation (2.13) also gives

$$-2\sum_{i=1}^k \frac{g_i^T d_i}{\phi_i^2} \geq \sum_{i=1}^k \frac{\|g_i\|^2}{\phi_i^2}. \tag{2.17}$$

Noting that $-4g_i^T d_i - \|g_i\|^2 \leq 4(g_i^T d_i)^2/\|g_i\|^2$, we get by this and (2.17) that

$$4\sum_{i=1}^k \frac{(g_i^T d_i)^2}{\|g_i\|^2\phi_i^2} \geq -4\sum_{i=1}^k \frac{g_i^T d_i}{\phi_i^2} - \sum_{i=1}^k \frac{\|g_i\|^2}{\phi_i^2} \geq \sum_{i=1}^k \frac{\|g_i\|^2}{\phi_i^2}. \tag{2.18}$$

Now we proceed by contradiction and assume that (2.9) holds. Then by (2.18), (2.15) and (2.9), we have that

$$\sum_{k\geq 1} \frac{(g_k^T d_k)^2}{\|g_k\|^2 \phi_k^2} \geq \frac{\gamma^2}{4} \sum_{k\geq 1} \frac{1}{\phi_k^2} = +\infty, \tag{2.19}$$

which means that the sum series in the right hand side of (2.16) is divergent. By Lemma 6 in [71], we then know that

$$+\infty = \sum_{k\geq 1} \frac{(g_k^T d_k)^2}{\|g_k\|^2 \phi_k^2} \frac{1}{t_k} = \sum_{k\geq 1} \frac{(g_k^T d_k)^2}{\|g_k\|^2 \|d_k\|^2} \leq \frac{1}{\gamma^2} \sum_{k\geq 1} \frac{(g_k^T d_k)^2}{\|d_k\|^2}, \tag{2.20}$$

which contradicts the Zoutendijk condition (2.2). The contradiction shows the truth of (2.3). □

Theorem 2.4 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 (1.10), only the descent condition $d_k^T g_k < 0$ is used here. An easy understanding between Theorems 2.3 and 2.4 is given in [17] 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 the condition (2.14) in Theorem 2.4 is very powerful in the convergence analysis of conjugate gradient methods. See [17] for some further uses of (2.14).

3 Early conjugate gradient methods

3.1 The Fletcher-Reeves method

In 1964, Fletcher and Reeves ([37]) proposed the first nonlinear conjugate gradient method and used the following conjugate gradient parameter

$$\beta_k^{\text{FR}} = \frac{||g_k||^2}{||g_{k-1}||^2}. \tag{3.21}$$

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 [94] proved that the FR method with the line search is globally convergent for nonlinear function. Al-Baali [1] first analyzed the FR method with strong Wolfe inexact line searches (1.5)-(1.6). He showed that if $\sigma < 1/2$, the sufficient condition (1.10) holds and there is global convergence. Liu *et al* [51] extended Al-Baali's result to the case that $\sigma = 1/2$. Dai and Yuan [25] presented a simpler proof to this result by showing that the sufficient condition (1.10) holds for at least one of any two neighboring iterations. Here it is worth noting that after the descent condition (1.8) has been verified, we can establish the global convergence easily by Theorem 2.4. More exactly, assuming that there is no convergence and (2.9) holds, we can see that $\prod_{j=2}^{k} \beta_j^2$ is at most linearly increasing and hence (2.14) holds. Consequently, there will be global convergence by Theorem 2.4, leading to a contradiction.

Further, if $\sigma > 1/2$, Dai and Yuan [25] 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. Interesting 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 [27] considered this idea and showed the global convergence of the FR method under a generalized Wolfe line search.

In [68], Powell analyzed the global efficiency of the FR method with the exact line search. Denote θ_k to be 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

$$\|d_{k+1}\| = \sec\theta_k \, \|g_k\| \tag{3.22}$$

and

$$\beta_{k+1}\|d_k\| = \tan\theta_k \, \|g_{k+1}\|. \tag{3.23}$$

By using the above two relations and substituting the formula (3.21), we can obtain

$$\tan\theta_{k+1} = \sec\theta_k \, \|g_{k+1}\|/\|g_k\| > \tan\theta_k \, \|g_{k+1}\|/\|g_k\|. \tag{3.24}$$

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 (3.24), 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 [38] to the strong Wolfe line search, and was observed by many researchers in the community. It explains why the FR method sometimes is 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

$$f(x) = \frac{1}{2} x^T x, \quad x \in R^n. \tag{3.25}$$

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

$$\|g_{k+1}\| = \|g_k\| \sin \theta_k. \tag{3.26}$$

By this and the equality in (3.24), 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 [68] for the two-dimension case and is actually valid for any dimension.

As will be seen in § 3.2, 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 motivates Touati-Ahmed and Storey [80] to extend Al-Baali's convergence result on the FR method to the general method (1.2)-(1.3) with

$$\beta_k \in \left[0,\, \beta_k^{FR}\right] \tag{3.27}$$

and suggested the formula

$$\beta_k^{TS} = \max\left\{0, \min\left\{\beta_k^{PRP}, \beta_k^{FR}\right\}\right\}. \tag{3.28}$$

Gilbert and Nocedal [38] further extended (3.27) to the interval

$$\beta_k \in \left[-\beta_k^{FR},\, \beta_k^{FR}\right] \tag{3.29}$$

and proposed the formula

$$\beta_k^{GN} = \max\left\{-\beta_k^{FR}, \min\left\{\beta_k^{PRP}, \beta_k^{FR}\right\}\right\}. \tag{3.30}$$

However, the numerical results in [38] show that the GN method is not so good as the PRP method although it indeed performs better than the FR method.

3.2 The Polak-Ribière-Polyak method

In 1969, Polak and Ribière [66] and Polyak [67] proposed another conjugate gradient parameter, independently, that is

$$\beta_k^{\mathrm{PRP}} = \frac{g_k^T y_{k-1}}{||g_{k-1}||^2}, \tag{3.31}$$

where $y_{k-1} = g_k - g_{k-1}$. In practical computations, the Polak-Ribière-Polyak (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 (3.31) that

$$|\beta_{k+1}^{\mathrm{PRP}}| \leq \|g_{k+1}\| \, \|g_{k+1} - g_k\| / ||g_k||^2. \tag{3.32}$$

By using the relations (3.22), (3.23) and (3.32), we can obtain

$$\tan \theta_{k+1} \leq \sec \theta_k \|g_{k+1} - g_k\| / \|g_k\|. \tag{3.33}$$

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

$$\tan \theta_{k+1} \ll \sec \theta_k. \tag{3.34}$$

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 [88]; for general functions, Powell [69] 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 [38] considered Powell [70]'s suggestion of modifying the PRP method by setting

$$\beta_k^{PRP+} = \max\{\beta_k^{PRP}, 0\}, \tag{3.35}$$

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 3.1. *Suppose that Assumption 2.1 holds. Consider the PRP$^+$ method, namely, (1.2) and (1.3) where β_k is given by (3.35). If the line search satisfies the standard Wolfe conditions (1.5) and (1.7), and the sufficient descent condition (1.10) for some constant $c > 0$, the method is globally convergent.*

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

Property $(*)$ Consider the method (1.2) and (1.3) and assume that $0 < \gamma < \|g_k\| \leq \bar{\gamma}$. Then we say that the method has Property $(*)$, if there exist constants $b > 1$ and $\zeta > 0$ such that for all k,

$$|\beta_k| \leq b, \tag{3.36}$$

and

$$\|s_{k-1}\| \leq \zeta \Longrightarrow |\beta_k| \leq \frac{1}{2b}. \tag{3.37}$$

It is not difficult to see that both PRP and PRP$^+$ possesses such property. Secondly, defining $u_k = d_k/\|d_k\|$, Gilbert and Nocedal observed that if $\beta_k \geq 0$ and if $\|d_k\| \to \infty$, u_k and u_{k-1} will tend to be the same, namley, $\|u_k - u_{k-1}\| \to 0$. 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 (1.10). 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 (see [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 (see [32]). Consider

$$f(x) = \frac{1}{2}\lambda x^2, \qquad x \in R^1. \tag{3.38}$$

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 (1.5)-(1.6). Consequently, $x_2 = 1 - \lambda$ and

$$g_2^T d_2 = \lambda^2(\lambda - 1)^3 > 0, \tag{3.39}$$

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

To ensure the sufficient descent condition, required by Theorem 3.1 for the PRP$^+$ method, in practical computations, Gilbert and Nocedal [38] designed a dynamic inexact line search strategy. As a matter of fact, their strategy applies to any method (1.2) and (1.3) with nonnegative β_k's. Let us look at (1.9). If $g_k^T d_{k-1} \le 0$, we already have (1.10) since $\beta_k \ge 0$. On the other hand, If $g_k^T d_k > 0$, it must be the case that $g_k^T d_{k-1} > 0$, which means a one-dimensional minimizer has been bracketed. Then $g_k^T d_{k-1}$ can be reduced and (1.10) holds by applying a line search algorithm, such as that given by Lemaréchal [49], Fletcher [36] or Moré and Thuente [57]. Comparing with the PRP method, however, no significant improvement is reported in [38] 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 [41] answered this question positively by generalizing an Armijo-type line search in [35]. Given constants $\tau > 0$, $\sigma \in (0, 1)$, $\delta > 0$ and $0 < c_1 < 1 < c_2$, their line search aims to find

$$\alpha_k = \max \left\{ \sigma^j \, \frac{\tau |g_k^T d_k|}{||d_k||^2}; \; j = 0, 1, \cdots \right\} \tag{3.40}$$

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

$$f(x_{k+1}) \le f(x_k) - \delta \, \alpha_k^2 \, ||d_k||^2 \tag{3.41}$$

and

$$-c_2 ||g_{k+1}||^2 \le g_{k+1}^T d_{k+1} \le -c_1 ||g_{k+1}||^2, \tag{3.42}$$

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, β_{k+1}^{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 $O(-g_k^T s_k)$ or $O(\|s_k\|)$, whereas the expected decrease is only of the second order $O(\|s_k\|^2)$. Therefore (3.41) and (3.42) must hold provided that α_k is sufficiently small. Furthermore, since the total reduction of the objective function is finite, the line search condition (3.41) enforces $\lim_{k \to \infty} \|s_k\| = 0$. By this property, the strong global convergence relation (2.4) can be achieved for the PRP algorithm of Grippo and Lucidi. From the view point of computations, Grippo and Lucidi [41] refined their line search algorithm so that the first one-dimensional minimizer of the line search function can be accepted. Again, the numerical experience (see [64]) does not suggest a significant improvement of their algorithm over the PRP method.

Along the line of [41], Dai and Yuan [32] builds the strong convergence of the PRP method with constant stepsizes

$$\alpha_k \equiv \eta, \quad \text{where } \eta \in (0, \frac{1}{4L}) \text{ is constant,} \tag{3.43}$$

where L is the Lipschitz constant in Assumption 2.1. This result was extended in [21] for the case that $\alpha_k \equiv \frac{1}{4L}$. Chen and Sun [12] further studies the PRP method together with other conjugate gradient methods using fixed stepsizes of the form

$$\alpha_k = \frac{-\delta\, g_k^T d_k}{d_k^T Q_k d_k}, \tag{3.44}$$

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

3.3 The Hestenes-Stiefel method

In this subsection, we briefly discuss the Hestenes-Stiefel (HS) conjugate gradient method, namely, (1.2)-(1.3) where β_k is calculated by

$$\beta_k^{\text{HS}} = \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}}. \tag{3.45}$$

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 (1.3) with y_{k-1} and using (3.45), we always have that

$$d_k^T y_{k-1} = 0. \tag{3.46}$$

In the quadratic case, y_{k-1} is parallel to $A\, d_{k-1}$, where A is the Hessian of the function. Then (3.46) implies $d_k^T A d_{k-1} = 0$, namely, d_k is conjugate to d_{k-1}. For this reason, the relation (3.46) is often called *conjugacy condition.*

If the line search is exact, we have by (1.9) that $g_k^T d_k = -\|g_k\|^2$ since $g_k^T d_{k-1} = 0$. It follows that $d_{k-1}^T y_{k-1} = \|g_{k-1}\|^2$ and $\beta_k^{HS} = \beta_k^{PRP}$. Therefore the HS method is identical to the PRP method in case of exact line searches. As a result, Powell [69]'s counter-example 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 [41], it is still known yet 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

$$d_2 = -g_2 + \beta_2^{HS} d_1 = -g_2 + \frac{g_2 \cdot y_1}{d_1 \cdot y_1}\, d_1 = -g_2 + g_2 = 0 \tag{3.47}$$

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

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

$$\beta_k^{HS+} = \max\{\beta_k^{HS}, 0\}. \tag{3.48}$$

In case of the sufficient descent condition (1.10), 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 (1.10) is relaxed to the descent condition, Qi *et al* [74] established the global convergence of a modified HS method, where β_k takes the form

$$\beta_k^{QHL} = \max\left\{0, \min\left\{\beta_k^{HS}, \frac{1}{\|g_k\|}\right\}\right\}. \tag{3.49}$$

Early in 1977, Perry [65] observed that the search direction in the HS method can be written as

$$d_k = -P_k\, g_k, \tag{3.50}$$

where

$$P_k = I - \frac{d_{k-1} y_{k-1}^T}{d_{k-1}^T y_{k-1}}. \tag{3.51}$$

Noting that $P_k^T y_{k-1} = 0$, 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 $P_k^T y_{k-1} = 0$. Instead, we look for a positive definite matrix P_k such that the conjugacy condition (3.46) holds. In case of exact line searches, it is sufficient to require P_k to satisfy

$$P_k^T y_{k-1} = s_{k-1}, \tag{3.52}$$

which is exactly the quasi-Newton equation (for example, see [88]). 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

$$P_k(\gamma_{k-1}) = \gamma_{k-1}\left(I - \frac{s_{k-1}^T y_{k-1} + y_{k-1} s_{k-1}}{s_{k-1}^T y_{k-1}}\right) + \left(1 + \frac{\gamma_{k-1}\|y_k\|^2}{s_{k-1}^T y_{k-1}}\right)\frac{s_{k-1}s_{k-1}^T}{s_{k-1}^T y_{k-1}}. \tag{3.53}$$

Shanno [76] explored this idea with $\gamma_{k-1} = 1$ (namely, no scaling is considered in the BFGS update) and obtained the search direction

$$d_k = -g_k + \left[\frac{g_k^T y_{k-1}}{s_{k-1}^T y_{k-1}} - \left(1 + \frac{\|y_{k-1}\|^2}{s_{k-1}^T y_{k-1}}\right)\frac{g_k^T s_{k-1}}{s_{k-1}^T y_{k-1}}\right] s_{k-1} + \frac{g_k^T s_{k-1}}{s_{k-1}^T y_{k-1}} y_{k-1}. \tag{3.54}$$

The method (1.2) and (3.54) is called memoryless BFGS method by Buckley [11]. 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 [23] derived the following relation directly from (3.50) and (3.52),

$$d_k^T y_{k-1} = -(P_k g_k)^T y_{k-1} = -g_k^T (P_k^T y_{k-1}) = -g_k^T s_{k-1}. \tag{3.55}$$

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

$$d_k^T y_{k-1} = -t\, g_k^T s_{k-1}, \tag{3.56}$$

and proposed the following choice for β_k,

$$\beta_k^{DL}(t) = \frac{g_k^T y_{k-1} - t\, g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}. \tag{3.57}$$

Clearly, if the line search is exact, namely, $g_k^T s_{k-1} = 0$, the DL direction is identical to the HS direction. If $g_k^T s_{k-1} \neq 0$, an analysis for quadratic functions is presented in [23], 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 k-th iteration. The numerical experiments in [23] showed that the DL method with $t = 0.1$ is a significant improvement of the HS method. In addition, similarly to PRP$^+$ and HS$^+$, Dai and Liao [23] established the global convergence of a modified DL method, where

$$\beta_k^{DL+}(t) = \max\left\{\frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}}, 0\right\} - t\,\frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}, \tag{3.58}$$

that allows negative values.

Two further developments of the DL method are made by Yabe and Takano [85] and Li *et al* [50]. Specifically, based on a modified secant condition given by Zhang *et al* [90, 91], Yabe and Takano [85] suggested the variants of (3.57) and (3.58) with the vector y_{k-1} replaced with

$$z_{k-1} = y_{k-1} + \left(\frac{\rho \, \lambda_k}{s_{k-1}^T u_{k-1}} \right) u_{k-1}, \tag{3.59}$$

where $\lambda_k = 6(f_{k-1} - f_k) + 3(g_{k-1} + g_k)^T s_{k-1}$, $\rho \geq 0$ is a constant and $u_{k-1} \in R^n$ satisfies $s_{k-1}^T u_{k-1} \neq 0$ (for example, $u_{k-1} = d_{k-1}$). Li *et al* [50] considered the modified secant condition in Wei *et al* [82] and suggested the following replacement of y_{k-1} in (3.57) and (3.58):

$$y_{k-1}^* = y_{k-1} + \frac{\nu_{k-1}}{\|s_{k-1}\|^2} \, s_{k-1}, \tag{3.60}$$

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

4 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 (1.5) and (1.6) 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 (see [68]). 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 the strong Wolfe conditions and the standard Wolfe conditions, respectively. A hybrid of the two methods is briefly mentioned at last, which can ensure a descent direction at every iteration without line searches.

4.1 The conjugate descent method

In his monograph [36], Fletcher proposed the conjugate descent (CD) method, namely, (1.2)-(1.3) with β_k is given by

$$\beta_k^{CD} = \frac{\|g_k\|^2}{-d_{k-1}^T g_{k-1}}. \tag{4.1}$$

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 (1.5)-(1.6) are used. To see this, we first introduce the following variants of the strong Wolfe conditions, namely, (1.5) and

$$\sigma_1 \, g_k^T d_k \le g(x_k + \alpha_k d_k)^T d_k \le -\sigma_2 \, g_k^T d_k, \tag{4.2}$$

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 (1.5)-(1.6). Now, by (1.9) and (4.1), we have

$$-g_k^T d_k = \|g_k\|^2 \left[1 + g_k^T d_{k-1}/g_{k-1}^T d_{k-1}\right]. \tag{4.3}$$

The above relation and (4.2) indicates that

$$1 - \sigma_2 \le -g_k^T d_k/\|g_k\|^2 \le 1 + \sigma_1. \tag{4.4}$$

Since $\sigma_2 < 1$, the left inequality in (4.4) means that (1.10) 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 [26] using the generalized strong Wolfe conditions (1.5) and (4.2). Specifically, if $\sigma_1 < 1$ and $\sigma_2 = 0$, it follows from (4.1), (4.4) and (3.21) that

$$0 \le \beta_k^{CD} \le \beta_k^{FR}. \tag{4.5}$$

Therefore by the result of Touati-Ahmed and Storey [80] related to the relation (3.27), 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 [26] considered the following two dimensional function

$$f(x, y) = \xi \, x^2 - y, \quad \text{where } \xi \in (1, \, 9/8), \tag{4.6}$$

and showed that the CD method with the generalized strong Wolfe line search may not solve (4.6). In real computations, the CD method is even inferior to the FR method.

4.2 The Dai-Yuan method

To enforce a descent direction in case of the standard Wolfe line search, Dai and Yuan [30] proposed a new conjugate gradient method, where

$$\beta_k^{DY} = \frac{\|g_k\|^2}{d_{k-1}^T y_{k-1}}. \tag{4.7}$$

For the DY method, it follows by (1.3), (4.7) and direct calculations that

$$g_k^T d_k = \frac{\|g_k\|^2}{d_{k-1}^T y_{k-1}} \, g_{k-1}^T d_{k-1}. \tag{4.8}$$

The fraction in (4.8) is exactly the DY formula (4.7). With this observation, we can get an equivalent expression of β_k^{DY} from (4.8),

$$\beta_k^{DY} = \frac{g_k^T d_k}{g_{k-1}^T d_{k-1}}. \tag{4.9}$$

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

Theorem 4.1. *Suppose that Assumption 2.1 holds. Consider the DY method, namely, (1.2) and (1.3) where β_k is given by (4.7). If the line search satisfies the standard Wolfe conditions (1.5) and (1.7), we have that $g_k^T d_k < 0$ for all $k \geq 1$. Further, the method converges in the sense that $\liminf_{k\to\infty} \|g_k\| = 0$.*

Proof. It is obvious that $d_1^T g_1 < 0$ since $d_1 = -g_1$. Assume that $g_{k-1}^T d_{k-1} < 0$. It follows by this and (1.7) that $d_{k-1}^T y_{k-1} > 0$. Thus by (4.8), we also have that $g_k^T d_k < 0$. Therefore by induction, $g_k^T d_k < 0$ for all $k \geq 1$.

Now, let us denote

$$q_k = \frac{\|d_k\|^2}{(g_k^T d_k)^2}, \qquad r_k = -\frac{g_k^T d_k}{\|g_k\|^2}. \tag{4.10}$$

Dividing (2.5) by $(g_k^T d_k)^2$ and using (4.9) and (4.10), we obtain

$$q_k = q_{k-1} + \frac{1}{\|g_k\|^2}\frac{2}{r_k} - \frac{1}{\|g_k\|^2}\frac{1}{r_k^2}. \tag{4.11}$$

Noting that $\frac{2}{r_k} - \frac{1}{r_k^2} \leq 1$, an immediate corollary of (4.11) is

$$q_k \leq q_{k-1} + \|g_k\|^{-2}. \tag{4.12}$$

Assuming that $\liminf_{k\to\infty} \|g_k\| \neq 0$ and (2.9) holds. By (2.9), (4.12) 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 (2.2). Therefore the statement is true. □

By (4.11), we can further exploit the self-adjusting property of the DY method ([18]). To this aim, we first notice that the r_k defined in (4.10) is a quantity that reflects the descent degree of the search direction d_k, since the descent condition (1.8) is equivalent to $r_k > 0$ and the sufficient condition (1.10) is the same as $r_k \geq c$. Now let us focus on the relation (4.11). The second term on the right side of (4.11) 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 4.2. *Consider the DY method (1.2), (1.3) and (4.7) where d_k is a descent direction. Assume that $0 < \gamma \leq \|g_k\| \leq \bar{\gamma}$ holds for all $k \geq 1$. There must exist positive constants δ_1, δ_2 and δ_3 such that the relations*

$$-g_k^T d_k \geq \frac{\delta_1}{\sqrt{k}}, \qquad \|d_k\|^2 \geq \frac{\delta_2}{k}, \qquad r_k \geq \frac{\delta_3}{\sqrt{k}} \tag{4.13}$$

holds 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

$$-g_i^T d_i \geq \delta_4, \qquad \|d_i\|^2 \geq \delta_5, \qquad r_i \geq \delta_6 \tag{4.14}$$

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

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

$$f_k - f_{k+1} \geq c \min\left\{-g_k^T d_k, \|d_k\|^2, q_k^{-1}\right\}, \tag{4.15}$$

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 [4] and the line search proposed in [35, 40] 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 [18], 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 [18] demonstrated this finding.

Since the DY method has the same drawback as the FR method, namely, it can not 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 [80, 38]. Under the standard Wolfe line search, Dai and Yuan [33] extended Theorem 4.1 to any method (1.2), (1.3) with

$$\beta_k \in \left[-\frac{1-\sigma}{1+\sigma}\, \beta_k^{DY},\ \beta_k^{DY} \right], \tag{4.16}$$

where σ is the parameter in the Wolfe condition (1.7). In spite of a large admissible interval, the numerical results of Dai and Yuan [33] indicated that the following hybrid is preferable in real computations

$$\beta_k^{DYHS} = \max \left\{ 0,\ \min \left\{ \beta_k^{HS}, \beta_k^{DY} \right\} \right\}. \tag{4.17}$$

Unlike the 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 (see [33]). The latter was generally believed as 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 (for example, see [8, 9]). On the contrary, nonlinear conjugate gradient methods were often analyzed individually. To change the situation, Dai and Yuan [28] proposed a family of conjugate gradient methods, in which

$$\beta_k(\lambda) = \frac{||g_k||^2}{\lambda ||g_{k-1}||^2 + (1-\lambda) d_{k-1}^T y_{k-1}}, \quad \lambda \in [0,1]. \tag{4.18}$$

This family can be regarded as some kind of convex combination of the FR, and DY methods. Dai and Yuan [29] further extended the family to the case $\lambda \in (-\infty, +\infty)$ and presented some unified convergence results. Independently, Nazareth [60] regarded the FR, PRP, HS and DY formulas as the four leading contenders for the conjugate gradient parameter and proposed a two-parameter family:

$$\beta_k(\lambda_k, \mu_k) = \frac{\lambda_k ||g_k||^2 + (1-\lambda_k) g_k^T y_{k-1}}{\mu_k ||g_{k-1}||^2 + (1-\mu_k) d_{k-1}^T y_{k-1}}, \quad \lambda_k, \mu_k \in [0,1]. \tag{4.19}$$

The methods that take the convex combination $\lambda_k\beta_k^{HS} + (1-\lambda_k)\beta_k^{DY}$, considered in Andrei [2], can be regarded as a subfamily of (4.19) with $\mu_k = 0$. Several efficient choices for λ_k in this subfamily are also studied in [2] based on different secant conditions.

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

$$\beta_k^{LS} = \frac{g_k^T y_{k-1}}{-d_{k-1}^T g_{k-1}} \tag{4.20}$$

by Liu and Storey [53], Dai and Yuan [34] proposed a three-parameter family:

$$\beta_k(\lambda_k, \mu_k, \omega_k) = \frac{||g_k||^2 - \lambda_k g_k^T g_{k-1}}{||g_{k-1}||^2 + \mu_k g_k^T d_{k-1} - \omega_k \beta_{k-1} g_{k-1}^T d_{k-2}}, \tag{4.21}$$

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 (4.21) with $\lambda_k = 1$, $\mu_k = 0$ and $\omega_k = u$ is studied in Shi and Guo [78] with an efficient nonmonotone line search. Further, Dai [20] studied a family of hybrid conjugate gradient methods, in which

$$\beta_k(\mu_k, \omega_k, \tau_k) = \frac{\max\{0, \min\{g_k^T y_{k-1}, \tau_k \|g_k\|^2\}\}}{(\tau_k + \omega_k) g_k^T d_{k-1} + \mu_k \|g_{k-1}\|^2 + (1-\mu_k)(-d_{k-1}^T g_{k-1})}, \tag{4.22}$$

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

4.3 The DYCD method

Suppose that M is some fixed positive integer, and λ and δ are constants in $(0,1)$. Given an initial guess $\bar{\alpha}_k$ at the k-th iteration, the nonmonotone line search by [39] is to compute the least non-negative integer m such that the steplength $\alpha_k = \bar{\alpha}_k \lambda^m$ satisfies the following relation:

$$f(x_k + \alpha_k d_k) \le \max\{f_k, \ldots, f_{k-M(k)}\} + \delta \alpha_k g_k^T d_k, \tag{4.23}$$

where $M(k) = \min(M, k-1)$. To enforce a descent search direction at every iteration in this situation, Dai [19] considered a hybrid of the DY and CD methods, namely, (1.2)-(1.3) with

$$\beta_k^{DYCD} = \frac{\|g_k\|^2}{\max\left\{d_{k-1}^T y_{k-1},\ -d_{k-1}^T g_{k-1}\right\}}. \tag{4.24}$$

It is proved in [19] 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 [19] was able to solve all the eighteen test problems in Moré *et al* [56].

5 Sufficient descent conjugate gradient methods

In this section, we summarize several nonlinear conjugate gradient methods that can guarantee the sufficient descent condition (1.10), especially the CG_Descent method by Hager and Zhang [42, 44].

Since the sufficient descent condition (1.10) is not scale invariant, however, there is some difficulty to differentiate *descent* conjugate gradient methods and *sufficient descent* conjugate gradient methods. More exactly, if d_k satisfies $g_k^T d_k < 0$, we can define another method whose search direction is $\bar{d}_k = \left(-c\,\|g_k\|^2/g_k^T d_k\right)\, d_k$ such that $g_k^T \bar{d}_k = -c\|g_k\|^2$.

Let us take the DY method as an illustrative example. A variant of the DY method is given in Dai [18], where d_k takes the form

$$d_k = -\frac{d_{k-1}^T y_{k-1}}{\|g_k\|^2} g_k + d_{k-1}. \tag{5.1}$$

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

$$g_k^T d_k = -||g_1||^2, \quad \text{for all } k \geq 1. \tag{5.2}$$

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

$$d_k = -\frac{d_{k-1}^T y_{k-1}}{\|g_{k-1}\|^2}\, g_k + \beta_k^{FR}\, d_{k-1}. \tag{5.3}$$

In this case, we have that $-g_k^T d_k = \|g_k\|^2$ for all k, which implies that the sufficient descent condition (1.10) holds with $c = 1$. It is worth mentioning that the above scheme (5.3) is obtained by Zhang *et al* [92] (see also § 5.2) with the motivation of modifying the Fletcher-Reeves method. They found that, the numerical performance of this scheme is very promising for a large collection of test problems in the CUTEr library [7].

5.1 The CG_Descent method

To ensure the sufficient descent condition (1.10), Hager and Zhang [44] proposed a family of conjugate gradient methods, where

$$\beta_k^{HZ}(\lambda_k) = \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}} - \lambda_k \left(\frac{\|y_{k-1}\|^2\, g_k^T d_{k-1}}{(d_{k-1}^T y_{k-1})^2} \right), \tag{5.4}$$

where $\lambda_k \geq \bar{\lambda} > 1/4$ 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 (3.57) with

$$t = \lambda_k \frac{\|y_{k-1}\|^2}{s_{k-1}^T y_{k-1}}. \tag{5.5}$$

To verify the sufficient descent condition for the HZ method, we have by (1.3) and (5.4) that

$$g_k^T d_k = -\|g_k\|^2 + \left(\frac{g_k^T y_{k-1}(g_k^T d_{k-1})}{d_{k-1}^T y_{k-1}} \right) - \lambda_k \left(\frac{\|y_{k-1}\|^2 \, (g_k^T d_{k-1})^2}{(d_{k-1}^T y_{k-1})^2} \right). \tag{5.6}$$

Now, by applying

$$u_k = \frac{1}{\sqrt{2\lambda_k}} \, (d_{k-1}^T y_{k-1}) \, g_k, \qquad v_k = \sqrt{2\lambda_k} \, (g_k^T d_{k-1}) \, y_{k-1} \tag{5.7}$$

into the inequality

$$u_k^T v_k \leq \frac{1}{2} \, \left(\|u_k\|^2 + \|v_k\|^2 \right), \tag{5.8}$$

we can obtain

$$\frac{g_k^T y_{k-1}(g_k^T d_{k-1})}{d_{k-1}^T y_{k-1}} \leq \frac{1}{4\lambda_k} \|g_k\|^2 + \lambda_k \left(\frac{\|y_{k-1}\|^2 \, g_k^T d_{k-1}}{(d_{k-1}^T y_{k-1})^2} \right). \tag{5.9}$$

Therefore by (5.6) and (5.9), we have that

$$g_k^T d_k \leq - \left(1 - \frac{1}{4\lambda_k} \right) \|g_k\|^2, \tag{5.10}$$

which with the restriction of λ_k means that the sufficient descent condition (1.10) holds with $c = 1 - (4\bar{\lambda})^{-1}$.

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

$$\beta_k^{HZ+}(\lambda_k) = \max \left\{ \beta_k^{HZ}(\lambda_k), \, \eta_k \right\}, \quad \eta_k = \frac{-1}{\|d_{k-1}\|^2 \min \{ \eta, \, \|g_{k-1}\| \}}, \tag{5.11}$$

where $\eta > 0$ is a constant. With this truncation, they established the global convergence of the modified method (5.11) with the standard Wolfe line search for general functions.

Hager and Zhang [42, 43] 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 [7], 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 [48] by projecting the scaled memoryless BFGS direction defined in (3.50) and (3.53) into the one dimensional manifold $\{-g_k + \beta\, d_k : \beta \in R\}$. By taking the scaling factors $\gamma_{k-1} = \frac{s_{k-1}^T y_{k-1}}{\|y_{k-1}\|^2}$ and $\gamma_{k-1} = \frac{\|s_{k-1}\|^2}{s_{k-1}^T y_{k-1}}$, they suggest the uses of $\lambda_k = 2 - \frac{d_{k-1}^T y_{k-1}}{\|d_{k-1}\|^2 \|y_{k-1}\|^2}$ and $\lambda_k = 1$, respectively. A simple and efficient nonmonotone line search criterion is also designed in [48], that can guarantee the global convergence of the new methods.

5.2 Several new methods that guarantee sufficient descent

The remarkable property of the HZ method (5.4) that can guarantee the sufficient descent condition (1.10) for general functions have attracted several further investigations.

A direct generalization of (5.4) is given in Yu and Guan [86] (see also [87]). They found that, for any β_k of the form

$$\beta_k = \frac{g_k^T v_k}{\Delta_k}, \quad \text{for some } v_k \in R^n \text{ and } \Delta_k \in R, \tag{5.12}$$

there is a corresponding formula

$$\beta_k^{YG}(C) = \frac{g_k^T v_k}{\Delta_k} - \frac{C\,\|v_k\|^2}{\Delta_k^2}\, g_k^T d_{k-1}, \tag{5.13}$$

where $C > 1/4$, such that (1.10) holds with $c = 1 - (4C)^{-1}$. Since almost all of the conjugate gradient parameters can be written into (5.12), we can obtain various extensions that can guarantee sufficient descent. It is obvious that the HZ formula (5.4) is corresponding to (5.13) with the HS formula (3.45) where $v_k = y_{k-1}$ and $\Delta_k = d_{k-1}^T y_{k-1}$. The extensions of β_K^{FR}, β_k^{PRP}, β_k^{DY}, β_k^{CD} and β_k^{LS} are also provided in [86]. A further generalization of this framework on the spectral conjugate gradient method (see [6] or § 6) is given in [87].

Another general way of producing sufficient descent conjugate gradient methods is provided in Cheng [13] and Cheng and Liu [14]. 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

$$d_k^{\perp} = \left(I - \frac{g_k g_k^T}{\|g_k\|^2} \right) (-g_k + \beta_k \, d_{k-1}). \tag{5.14}$$

The search direction defined by

$$\begin{aligned} d_k &= -g_k + d_k^{\perp} \\ &= -\left(1 + \beta_k \frac{g_k^T d_{k-1}}{\|g_k\|^2} \right) g_k + \beta_k \, d_{k-1} \end{aligned} \tag{5.15}$$

then always satisfies $g_k^T d_k = -\|g_k\|^2$. If the line search is exact, the second term in the parathesis of (5.15) is missing since $g_k^T d_{k-1} = 0$. Hence the above scheme reduces to the linear conjugate gradient method in the ideal case. The above procedure with $\beta_k = \beta_k^{PRP}$ is studied in [13]. As shown in [14], setting $\beta_k = \beta_k^{FR}$ in (5.15) leads to the scheme (5.3). Another variant corresponding to to Yabe and Takano [85] (see the end of § 3.3) is also investigated in [14].

Observing that the search direction (3.54) defined by the memoryless BFGS method is formed by the vectors $-g_k$, d_{k-1} and y_{k-1}, Zhang *et al* [93] proposes the following modification of the PRP method

$$d_k = -g_k + \frac{g_k^T y_{k-1}}{\|g_{k-1}\|^2} \, d_{k-1} - \frac{g_k^T d_{k-1}}{\|g_{k-1}\|^2} \, y_{k-1}. \tag{5.16}$$

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 $g_k^T d_k = -\|g_k\|^2$. The above scheme was implemented in [93] with an Armijo-type line search in relation to [35], yielding comparable numerical results with CG_DESCENT. Although (5.16) reduces to (1.3) in case of exact line searches, this scheme is not a standard conjugate gradient method of the form (1.3) any more.

6 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 (3.21) there). While some of the existing conjugate gradient algorithms, like the DYHS method (4.17) and the CG_DESCENT method (5.4) among others, have proved more efficient than the PRP method, we feel there is still much 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 [89], which aims to find the best points of the function over the two-dimensional manifold

$$x_k \ + \ \mathrm{Span}\{-\alpha\, g_k + \beta\, d_{k-1} : \, (\alpha,\, \beta) \in R^2\}. \tag{6.1}$$

as the next iterates. Motivated by the success of the Barzilai-Borwein stepsize in the steepest descent method (for example, see [5, 75]), Birgin and Martinez [6] proposed the so-called spectral conjugate gradient method that takes the search direction

$$d_k = -\frac{1}{\delta_k}\, g_k + \frac{g_k^T (y_{k-1} - \delta_k\, s_{k-1})}{\delta_k\, d_{k-1}^T y_{k-1}}\, d_{k-1}. \tag{6.2}$$

The efficient combination of the Barzilai-Borwein method and the conjugate gradient method, however, is still not known to us. Specifically, the study of Dai and Liao [23] 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 method of the form (1.2)-(1.3), another class of two-term conjugate gradient methods is called method of shortest residuals (SR), that was first presented by Hestenes [45] and studied by Pytlak and Tarnawski [73], Dai and Yuan [31], and the references therein. The SR method defines the search direction by

$$d_k = -Nr\{g_k, \ -\beta_k d_{k-1}\}, \tag{6.3}$$

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,

$$\|Nr\{a, \ b\}\| = \min\{\|\, \lambda\, a + (1-\lambda)\, b\,\| : 0 \le \lambda \le 1\}. \tag{6.4}$$

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 [72] and modified in [31]. If, further, the function is quadratic, these variants of the SR method are equivalent and the direction d_k proves to be the shortest residual in the

$(k-1)$-simplex whose vertices are $-g_1, \cdots, -g_k$. For the SR method, the descent property of d_k is naturally implied by its definition (see [72, 31] 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 such as [68, 58, 93] and limited-memory quasi-Newton methods such as [52, 79] 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, for example see [10, 55, 3]. 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.

Acknowledgements. The author thanks Professor Ya-xiang Yuan and the anonymous editor very much for their useful suggestions and comments.

References

[1] M. Al-Baali, *Descent property and global convergence of the Fletcher-Reeves method with inexact linesearch*, IMA J. Numer. Anal., 5 (1985), pp. 121–124.

[2] N. Andrei, *Accelerated hybrid conjugate gradient algorithm with modified secant condition for unconstrained optimization*, Numerical Algorithms 54:1 (2010), pp. 1017-1398.

[3] N. Andrei, *Accelerated scaled memoryless BFGS preconditioned conjugate gradient algorithm for unconstrained optimization*, European Journal of Operational Research 204 (2010), pp. 410-420.

[4] L. Armijo, *Minimization of functions having Lipschitz continuous partial derivatives*, Pacific J. Math. 16 (1966), pp. 1-3.

[5] J. Barzilai and J. M. Borwein, *Two-point step size gradient methods*, IMA J. Numer. Anal. 8 (1988), pp. 141-148.

[6] E. G. Birgin and J. M. Martinez, *A spectral conjugate gradient method for unconstrained optimization*, Appl. Math. Optim. 43 (2001), pp. 117-128.

[7] I. Bongartz, A. R. Conn, N. I. M. Gould and Ph. L. Toint, *CUTE: constrained and unconstrained testing environments*, ACM Trans. Math. Software 21 (1995), pp. 123-160.

[8] C. G. Broyden, *The convergence of a class of double-rank minimization algorithms 1. general considerations* , J. Inst. Math. Appl. 6 (1970) 76–90.

[9] R. Byrd, J. Nocedal and Y. Yuan, *Global convergence of a class of variable metric algorithms*, SIAM J. Numer. Anal. 4 (1987), 1171-1190.

[10] A. Buckley, *A combined conjugate gradient quasi-Newton minimization algorithms*, Math. Prog. 15 (1978), pp. 200-210.

[11] A. Buckley, *Conjugate gradient methods*, in: M. J. D. Powell, ed., Nonlinear Optimization 1981 (Academic Press, London, 1982), pp. 17-22.

[12] X. D. Chen and J. Sun, *Global Convergence of two-parameter family of conjugate gradient methods without line search*, Journal of Computational and Applied Mathematics 146 (2002), pp. 37-45.

[13] W. Y. Cheng, *A two-term PRP-based descent method*, Numerical Functional Analysis and Optimization 28:11-12 (2007), pp. 1217-1230.

[14] W. Y. Cheng and Q. F. Liu, *Sufficient descent nonlinear conjugate gradient methods with conjugacy conditions*, Numerical Algorithms 53 (2010), pp. 113-131.

[15] A. Cohen. *Rate of convergence of several conjugate gradient method algorithms*, SIAM Journal on Numerical Analysis 9 (1972), pp. 248-259.

[16] H. P. Crowder and P. Wolfe, *Linear convergence of the conjugate gradient method*, IBM J. Res. Dec. 16 (1969), pp. 431-433.

[17] Y. H. Dai, *Convergence analysis of nonlinear conjugate gradient methods*, Research report, LSEC, ICMSEC, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, 2000.

[18] Y. H. Dai, *New properties of a nonlinear conjugate gradient method*, Numerische Mathematics 89:1 (2001), pp. 83-98.

[19] Y. H. Dai, *A nonmonotone conjugate gradient algorithm for unconstrained optimization*, Journal of Systems Science and Complexity 15:2 (2002), pp. 139-145.

[20] Y. H. Dai, *A family of hybrid conjugate gradient methods for unconstrained optimization*, Mathematics of Computation 72 (2003) pp. 1317-1328.

[21] Y. H. Dai, *Convergence of conjugate gradient methods with constant stepsizes*, LSEC, ICMSEC, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, 2008 (accepted by Optimizaton Methods and Software).

[22] Y. H. Dai, J. Han, G. Liu, D. Sun, H. Yin, and Y. Yuan, *Convergence properties of nonlinear conjugate gradient methods*, SIAM Journal on Optimization 10: 2 (1999), pp. 345–358.

[23] Y. H. Dai and L.Z. Liao, *New conjugacy conditions and related nonlinear conjugate gradient methods*, Applied Mathematics and Optimization 43:1 (2001), pp. 87-101.

[24] Y. H. Dai and L. Z. Liao, *R-linear convergence of the Barzilai and Borwein gradient method*, IMA Journal of Numerical Analysis 22 (2002), pp. 1-10.

[25] Y. H. Dai and Y. Yuan, *Convergence properties of the Fletcher-Reeves method*, IMA J. Numer. Anal. 16 (1996), pp. 155–164.

[26] Y. H. Dai and Y. Yuan, *Convergence properties of the conjugate descent method*, Advances in Mathematics 26:6 (1996), pp. 552-562.

[27] Y. H. Dai and Y. Yuan, *Convergence of the Fletcher-Reeves method under a generalized Wolfe search*, Journal of Computational Mathematics of Chinese Universities 2 (1996), pp. 142-148.

[28] Y. H. Dai and Y. Yuan, *A class of globally convergent conjugate gradient methods*, Research report ICM-98-030, Institute of Computational Mathematics and Scientific/Engineering Computing, Chinese Academy of Sciences, 1998.

[29] Y. H. Dai and Y. Yuan, *Extension of a class of conjugate gradient methods*, Research report ICM-98-049, Institute of Computational Mathematics and Scientific/Engineering Computing, Chinese Academy of Sciences, 1998.

[30] Y. H. Dai and Y. Yuan, *A nonlinear conjugate gradient method with a strong global convergence property*, SIAM Journal on Optimization 10 : 1 (1999), pp. 177–182.

[31] Y. H. Dai and Y. Yuan, *Global convergence of the method of shortest residuals*, Numerische Mathematik 83 (1999), pp. 581-598.

[32] Y. H. Dai and Y. Yuan, *Nonlinear Conjugate Gradient Methods*, Shanghai Scientific & Technical Publishers, 2000 (in Chinese).

[33] Y. H. Dai and Y. Yuan, *An efficient hybrid conjugate gradient method for unconstrained optimization*, Annals of Operations Research 103 (2001), pp. 33-47.

[34] Y. H. Dai and Y. Yuan, *A three-parameter family of conjugate gradient methods*, Mathematics of Computation 70 (2001), pp. 1155-1167.

[35] R. De Leone, M. Gaudioso, and L. Grippo, *Stopping criteria for line-search methods without derivatives*, Mathematical Programming 30 (1984), pp. 285-300.

[36] R. Fletcher (1987), *Practical Methods of Optimization vol. 1: Unconstrained Optimization*, John Wiley & Sons (New York).

[37] R. Fletcher and C. M. Reeves, *Function minimization by conjugate gradients*, Comput. J. 7 (1964), pp.149–154.

[38] J. C. Gilbert and J. Nocedal, *Global convergence properties of conjugate gradient methods for optimization*, SIAM J. Optimization 2 (1992), pp. 21–42.

[39] L. Grippo, F. Lamparillo, and S. Lucidi, *A nonmonotone line search technique for Newton's method*, SIAM J. Numer. Anal. 23 (1986), pp. 707-716.

[40] L. Grippo, F. Lampariello, and S. Lucidi, *Global convergence and stabilization of unconstrained minimization methods without derivatives*, Journal of Optimization Theory and Applications 56 (1988), pp. 385-406.

[41] L. Grippo and S. Lucidi, *A globally convergent version of the Polak-Ribière conjugate gradient method,* Math. Prog. 78 (1997), pp. 375-391.

[42] W. W. Hager and H. Zhang, *A new conjugate gradient method with guaranteed descent and an efficient line search*, SIAM Journal on Optimization 16:1 (2005), pp. 170 - 192.

[43] W. W. Hager and H. Zhang, *Algorithm 851: CG_DESCENT, a conjugate gradient method with guaranteed descent*, ACM Transactions on Mathematical Software 32:1 (2006), pp. 113 - 137.

[44] W. W. Hager and H. Zhang, *A survey of nonlinear conjugate gradient methods*, Pacific Journal of Optimization 2:1 (2006), pp. 335-58.

[45] M. R. Hestenes, *Conjugate direction methods in optimization*, Springer-Verlag, New York Heidelberg Berlin, 1980.

[46] M. R. Hestenes and E. Stiefel, *Method of conjugate gradient for solving linear system*, J. Res. Nat. Bur. Stand. 49 (1952), pp. 409–436.

[47] Y. F. Hu and C. Storey, *Global convergence result for conjugate gradient methods*, J. Optim. Theory Appl. 71 (1991), pp. 399–405.

[48] C. X. Kou and Y. H. Dai, *New conjugate gradient methods with an efficient nonmonotone line search*, Research report, LSEC, ICMSEC, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, 2010.

[49] C. Lemaréchal, *A view of line searches*, in: *Optimization and Optimal Control*, (Auslander, Oettli and Stoer, eds.), Lecture Notes in Control and Information 30, Springer Verlag, 1981, pp. 59-78.

[50] G. Y. Li, C. M. Tang and Z. X. Wei, *New conjugacy condition and related new conjugate gradient methods for unconstrained optimization*, Journal of Computational and Applied Mathematics 202 (2007), pp. 523-539.

[51] G. H. Liu, J. Y. Han and H. X. Yin, *Global convergence of the Fletcher-Reeves algorithm with an inexact line search*, Appl. Math. J. Chinese Univ. Ser. B, 10 (1995), pp. 75-82.

[52] D. C. Liu and J. Nocedal, *On the limited memory BFGS method for large scale optimization*, Mathematical Programming 45 (1989), pp. 503-528.

[53] Y. Liu and C. Storey, *Efficient generalized conjugate gradient algorithms, part 1: theory*, Journal of Optimization Theory and Applications 69 (1991), pp. 129-137.

[54] P. McCormick and K. Ritter, *Alternative proofs of the convergenc properties of the conjugate-gradient method*, JOTA 13:5 (1975) pp. 497-518.

[55] J. L. Morales, J. Nocedal, "Automatic Preconditioning by Limited Memory Quasi-Newton Updating", SIAM J. Optimization 10:4 (2000), pp. 1079-1096.

[56] J. J. Moré, B. S. Garbow and K. E. Hillstrom, *Testing unconstrained optimization software*, ACM Transactions on Mathematical Software 7 (1981), pp. 17-41.

[57] J. Moré and D. J. Thuente, *On line search algorithms with guaranteed sufficient decrease*, ACM Trans. Math. Software 20 (1994), pp. 286-307.

[58] J. L. Nazareth, *A conjugate direction algorithm without line searches*, J. Optim. Theory App. 23:3 (1977), pp. 373-387.

[59] J. L. Nazareth, *The method of successive affine reduction for nonlinear minimization*, Mathematical Programming 35 (1986) pp. 97–109.

[60] L. Nazareth, *Conjugate-gradient methods*, Encyclopedia of Optimization (C. Floudas and P. Pardalos, eds.), Kluwer Academic Publishers, Boston, USA and Dordrecht, The Netherlands, 2001, pp. 319-323.

[61] J. L. Nazareth, *Conjugate gradient method*, Wiley Interdisciplinary Reviews: Computational Statistics 1:3 (2009), pp. 348–353.

[62] J. Nocedal, *Theory of algorithms for unconstrained optimization*, Acta Numerica (1991), pp. 199-242.

[63] J. Nocedal, *Conjugate gradient methods and nonlinear optimization*, in: Linear and Nonlinear Conjugate Gradient-Related Methods (L. Adams and J. L. Nazareth, eds.), SIAM, Philadelphia, 1996, pp. 9-23.

[64] J. Nocedal, *Large scale unconstrained optimization*, in: The State of the Art in Numerical Analysis (A. Watson and I. Duff, eds.), Oxford University Press, 1997, pp. 311-338.

[65] J. M. Perry, *A class of conjugate gradient algorithms with a two-step variable-metric memory*, Discussion Paper 269, Center for Mathematical Studies in Economics and Management Sciences, Northwestern University (Evanston, Illinois, 1977).

[66] E. Polak and G. Ribière, *Note sur la convergence de méthods de directions conjugées*, Revue Franccaise d'Informatique et de Recherche Opiérationnelle 16 (1969), pp. 35–43.

[67] B. T. Polyak. *The conjugate gradient method in extremem problems*, USSR Comp Math and Math. Phys. 9 (1969), pp. 94-112.

[68] M. J. D. Powell, *Restart procedures of the conjugate gradient method*, Mathematical Programming 12 (1977) pp. 241-254.

[69] M. J. D. Powell, *Nonconvex minimization calculations and the conjugate gradient method*, in: D.F. Griffiths, ed., Numerical Analysis, Lecture Notes in Mathematics 1066 (Springer-Verlag, Berlin, 1984), pp. 122–141.

[70] M. J. D. Powell, *Convergence properties of algorithms for nonlinear optimization*, SIAM Rev. 28 (1986), pp. 487–500.

[71] D. Pu and W. Yu, *On the convergence properties of the DFP algorithms*, Annals of Operations Research 24 (1990), pp. 175-184.

[72] R. Pytlak, *On the convergence of conjugate gradient algorithm*, IMA Journal of Numerical Analysis 14 (1989), pp. 443-460.

[73] R. Pytlak and T. Tarnawski, *On the method of shortest residuals for unconstrained optimization*, J. Optim. Theory App. 133:1 (2007), pp. 99-110.

[74] H. D. Qi, J. Y. Han, G. H. Liu, *A modified Hestenes-Stiefel conjugate gradient algorithm*, Chinese Annals of Mathematics, Series A, 17:3 (1996), pp. 177-184.

[75] M. Raydan, *The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem*, SIAM J. Optim. 7: 1 (1997), pp. 26-33.

[76] D. F. Shanno, *Conjugate gradient methods with inexact searches*, Math. Oper. Res. 3 (1978), 244-256.

[77] J. R. Shewchuk, *An introduction to the conjugate gradient method without the agonizing pain*, Technical Report CS-94-125, Carnegie Mellon University, Pittsburgh, PA, USA, 1994.

[78] Z. J. Shi and J. Guo, *A new family of conjugate gradient methods*, Journal of Computational and Applied Mathematics 224 (2009), pp. 444-457.

[79] L. P. Sun, *Updating the self-scaling symmetric rank one algorithm with limited memory for large-scale unconstrained optimization*, Computational Optimization and Applications 27 (2004), pp. 23-29.

[80] D. Touati-Ahmed and C. Storey, *Global convergent hybrid conjugate gradient methods*, J. Optim. Theory Appl. 64 (1990), pp. 379–397.

[81] C. Y. Wang and Y. Z. Zhang, *Global convergence properties of s-related conjugate gradient methods*, Chinese Science Bulletin 43:23 (1998), pp. 1959-1965.

[82] Z. Wei, G. Yu, G. Yuan and Z. Lian, *The superlinear convergence of a modified BFGS-type method for unconstrained optimization*, Comput. Optim. Appl. 29:3 (2004), pp. 315-332.

[83] P. Wolfe, *Convergence conditions for ascent methods*, SIAM Rev. 11 (1969), pp. 226–235.

[84] P. Wolfe, *Convergence conditions for ascent methods* II*: some corrections*, SIAM Rev. 13 (1971), pp. 185–188.

[85] H. Yabe and M. Takano, *Global convergence properties of nonlinear conjugate gradient methods with modified secant condition*, Comput. Optim. Appl. 28 (2004), pp. 203-225.

[86] G. H. Yu, L. T. Guan, *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.

[87] G. H. Yu, L. T. Guan and W. F. Chen, *Spectral conjugate gradient methods with sufficient descent property for large-scale unconstrained optimization*, Optimization Methods and Software 23:2 (2008), pp. 275-293.

[88] Y. Yuan, *Numerical Methods for Nonlinear Programming*, Shanghai Scientific & Technical Publishers, 1993 (in Chinese).

[89] Y. Yuan and J. Stoer, *A subspace study on conjugate gradient algorithms*, Z. Angew. Math. Mech. 75:1 (1995), pp. 69-77.

[90] J. Z. Zhang, N. Y. Deng and L. H. Chen, *New quasi-Newton equation and related methods for unconstrained optimization*, J. Optim. Theory Appl. 102 (1999), pp. 147-167.

[91] J. Z. Zhang and C. X. Xu, *Properties and numerical performance of quasi-Newton methods with modified quasi-Newton equations*, J. Comput. Appl. Math. 137 (2001), pp. 269-278.

[92] L. Zhang, W. Zhou and D. Li, *Global convergence of a modified Fletcher-Reeves conjugate gradient method with Armijo-type line search*, Numerische Mathematik 104:4 (2006), pp. 561 - 572.

[93] L. Zhang, W. Zhou and D. Li, *A descent modified Polak-Ribière-Polyak conjugate gradient method and its global convergence*, IMA Journal of Numerical Analysis 26:4 (2006) pp. 629-640.

[94] G. Zoutendijk, *Nonlinear programming, computational methods*, in: J. Abadie, ed., Integer and Nonlinear Programming (North-holland, Amsterdam, 1970), pp. 37–86.