Chapter 8
Convergence Analysis of Nonlinear Conjugate Gradient Methods

Yuhong Dai

Abstract. Conjugate gradient methods are a class of important methods for unconstrained optimization and vary only with a scalar β_k. In this chapter, we analyze general conjugate gradient method using the Wolfe line search and propose a condition on the scalar β_k, which is sufficient for the global convergence. An example is constructed, showing that the condition is also necessary in some sense for the global convergence of general conjugate gradient method. To make better use of the condition, we introduce a new property for conjugate gradient methods. It is shown that many conjugate gradient methods have such property, including the FR, PRP, HS, and DY methods and the FR-PRP, and DY-HS hybrid methods. Consequently, convergence results are gained for these methods under mild assumptions. In addition, an analysis is also given to a new conjugate gradient method, which further demonstrates the usefulness of the condition and the new property. Some discussions about the bound in the hybrid conjugate gradient methods are also given.

8.1 Introduction

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

$$\min f(x), \quad x \in R^n, \tag{8.1.1}$$

Yuhong 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,
P. O. Box 2719, Beijing 100190, China.
e-mail: dyh@lsec.cc.ac.cn

Y. Wang et al. (eds.), *Optimization and Regularization for Computational Inverse Problems and Applications*

especially if the dimension n is large. The methods without regular restarts are of the form

$$x_{k+1} = x_k + \lambda_k d_k, \tag{8.1.2}$$

$$d_k = -g_k + \beta_k d_{k-1}, \tag{8.1.3}$$

where $d_1 = -g_1 = -\nabla f(x_1)$, λ_k is a step-length obtained by a line search, and β_k is a scalar. The step-length λ_k is often required to satisfy the strong Wolfe conditions, namely,

$$f(x_k + \lambda_k d_k) - f(x_k) \leq \rho \lambda_k g_k^T d_k, \tag{8.1.4}$$

$$|g(x_k + \lambda_k d_k)^T d_k| \leq -\sigma g_k^T d_k, \tag{8.1.5}$$

where $0 < \rho < \sigma < 1$. The scalar β_k should be so chosen that the method (8.1.2)–(8.1.3) reduces to the linear conjugate gradient method in the case when f is a convex quadratic and the line search is exact. Some well-known formulae for β_k are called the FR [9], PRP [15, 16], HS [12] and DY [7] formulae, and are given by

$$\beta_k^{\mathrm{FR}} = ||g_k||^2 / ||g_{k-1}||^2, \tag{8.1.6}$$

$$\beta_k^{\mathrm{PRP}} = g_k^T y_{k-1} / ||g_{k-1}||^2, \tag{8.1.7}$$

$$\beta_k^{\mathrm{HS}} = g_k^T y_{k-1} / d_{k-1}^T y_{k-1}, \tag{8.1.8}$$

$$\beta_k^{DY} = \|g_k\|^2 / d_{k-1}^T y_{k-1}, \tag{8.1.9}$$

respectively, where $\|\cdot\|$ denotes the l_2-norm of R^n, and $y_{k-1} = g_k - g_{k-1}$.

Although all these methods have the quadratic termination property, their convergence properties and numerical performances may be very different for general objective functions. Basically, nonlinear conjugate gradient methods can be divided into the following three categories. The first category includes the FR, and DY methods, etc. In practical computations, this category of methods perform worse than the second, and third categories for they may produce small steps continuously ([17, 8]). However, their convergences can be achieved under mild assumptions. For example, the FR method with the strong Wolfe line search is shown to converge globally for general functions if the scalar σ in (8.1.5) is not greater than 0.5 (for example, see [1, 5]). The DY method is globally convergent provided that λ_k satisfies the Wolfe conditions, namely, (8.1.4) and

$$g(x_k + \lambda_k d_k)^T d_k \geq \sigma g_k^T d_k, \tag{8.1.10}$$

where $0 < \rho < \sigma < 1$ [7]. The second category includes the PRP, and HS methods, etc. If a small step occurs, the methods in this category can generate search directions close to the negative gradient direction and hence avoid the propensity of small steps [17, 10]. As a result, they perform often much better than the first category of methods. However, Powell [17] was able to show that the PRP method with exact line searches can cycle round eight nonstationary points.

The example also applies to the HS method since the two methods are the same in case of exact line searches. Till 1992, Gilbert and Nocedal [10] gave the global convergence of the PRP method with the restriction $\beta_k \geq 0$ for general functions. The third category includes the FR-PRP, and DY-HS hybrid methods, etc. This category was first proposed by Touati-Ahmed and Storey [20]. They suggested the following hybrid method of the FR, and PRP methods:

$$\beta_k = \max\{0, \min\{\beta_k^{PRP}, \beta_k^{FR}\}\}. \tag{8.1.11}$$

Gilbert and Nocedal [10] further considered the hybrid method

$$\beta_k = \max\{-\beta_k^{FR}, \min\{\beta_k^{PRP}, \beta_k^{FR}\}\}, \tag{8.1.12}$$

that allows negative values. The hybrid methods (8.1.11) and (8.1.12) have the following advantages: (i) their convergences can be achieved similar to FR; and (ii) they can avoid the propensity of small steps like PRP. However, their numerical performances are worse than PRP, though better than FR (for example, see [10]). Dai and Yuan [8]) extended the convergence result of the DY method in [7] to the following hybrid method:

$$\beta_k = \max\{0, \min\{\beta_k^{HS}, \beta_k^{DY}\}\}. \tag{8.1.13}$$

Promising numerical results were also obtained for this hybrid method [8, 4]. For a collection of medium and large-scale problems drawn from CUTE [2], it was shown that the use of the Wolfe line search in the hybrid method (8.1.13) is better than the use of the strong Wolfe line search, and that the method with the Wolfe line search performs better than the PRP method with the strong Wolfe line search for most of the test problems.

Although the already-existing results offer fascinating glimpses into the behavior of conjugate gradient methods, its theory still remains fragmentary [14]. A comprehensive theory of conjugate gradient methods, which is regarded in [14] as one of the outstanding challenges in theoretical optimization, is then to be developed. Reference [3] analyzed general conjugate gradient method with the strong Wolfe line search, and showed that the method converges globally if $\|d_k\|^2$ increases mostly linearly (see Lemma 8.2.3). Since it is possible to get convergence results and develop efficient algorithms in the conjugate gradient field via the Wolfe line search, as stated in the previous paragraph, we will analyze general conjugate gradient method with the Wolfe line search in this chapter. Specifically, since conjugate gradient methods vary only with the scalar β_k, we ask the following question: what condition on β_k can ensure the global convergence of general conjugate gradient method?

We will provide an answer to this question in Section 8.3 after giving some preliminaries in the next section. See (8.3.5) for the condition. An example is also constructed in Section 8.3, which shows that the condition (8.3.5) is necessary in some sense for the global convergence of general conjugate gradient method. To make better use of the condition, we will introduce a new property, namely,

Property (#) , for conjugate gradient methods in Section 8.4. It is shown that all the three categories of conjugate gradient methods may have such property. As a result, convergence results can be obtained conveniently for these methods under suitable assumptions. An analysis is also given to a new conjugate gradient method in Section 8.4, which further demonstrates the usefulness of the condition (8.3.5) and Property (#) . Some discussions about the bound in hybrid methods are made in the last section.

8.2 Some preliminaries

Throughout this chapter, we assume that $g_k \neq 0$ for all $k \geq 1$ for otherwise a stationary point has been found. We also assume that $\beta_k \neq 0$ for all $k \geq 1$. This is because if $\beta_k = 0$, the direction in (8.1.3) reduces to the negative gradient direction. Thus either the method converges globally if $\beta_k = 0$ for infinite number of k, or one can regard some x_k as the new initial point.

We give the following basic assumptions of the objective function.

Assumption 8.2.1. *(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*

$$\|g(x) - g(y)\| \leq L\|x - y\|, \quad for\ all\ x, y \in \mathcal{N}. \tag{8.2.1}$$

Under Assumption 8.2.1 on f, we state a very useful result, which was mainly obtained by Zoutendijk [23] and Wolfe [21, 22].

Lemma 8.2.2. *Suppose that Assumption 8.2.1 holds. Consider any iterative method of the form (8.1.2), where d_k satisfies $g_k^T d_k < 0$ and λ_k is obtained by the Wolfe line search. Then*

$$\sum_{k=1}^{\infty} \frac{(g_k^T d_k)^2}{||d_k||^2} < +\infty. \tag{8.2.2}$$

Relation (8.2.2) is normally called the Zoutendijk condition. In the case that the sufficient descent condition holds,

$$-g_k^T d_k \geq c\|g_k\|^2, \quad \text{for some } c > 0 \text{ and all } k \geq 1, \tag{8.2.3}$$

we can conclude from (8.2.2) that if $\|d_k\|^2$ increases at most linearly,

$$\sum_{k \geq 1} \frac{1}{\|d_k\|^2} = +\infty, \tag{8.2.4}$$

the iterative method converges in the sense that

$$\liminf_{k\to\infty} \|g_k\| = 0. \tag{8.2.5}$$

In fact, the sufficient descent condition (8.2.3) is often implied or required in many convergence analyses of conjugate gradient methods, for example see [1, 10, 11, 13, 20]. However, for general method (8.1.2)–(8.1.3) with strong Wolfe line searches, Dai et al. [3] showed that this result still holds even if the sufficient descent condition (8.2.3) is replaced with the descent condition $g_k^T d_k < 0$.

Lemma 8.2.3. *Suppose that Assumption 8.2.1 holds. Consider any iterative method of the form (8.1.2)–(8.1.3), where d_k satisfies $g_k^T d_k < 0$ and λ_k is obtained by the strong Wolfe line search. Then if the condition (8.2.4) holds, then the method gives the convergence relation (8.2.5).*

In the above lemma, the condition (8.2.4) is also necessary in some sense for the global convergence, as will be briefly discussed in Section 8.5.

8.3 A sufficient and necessary condition on β_k

The purpose of this section is to provide a condition on β_k, which is sufficient for the global convergence of general conjugate gradient method with the Wolfe line search. To do so, we will first give some analyses for general conjugate gradient method with the strong Wolfe line search with the help of Lemma 8.2.3 (see §8.3.1). In §8.3.2, we will give a basic lemma for any method in the form of (8.1.2)–(8.1.3) and prove that the condition (8.3.5) can really ensure the global convergence of general conjugate gradient method with the Wolfe line search. An example is constructed in §8.3.3, which shows that the condition (8.3.5) is also necessary in some sense for the global convergence.

8.3.1 Proposition of the condition

In this subsection, we assume that the step-length λ_k satisfies the strong Wolfe conditions (8.1.4)–(8.1.5). By Lemma 8.2.3, we know that if the relation (8.2.4) holds, then the method gives the convergence relation (8.2.5). Otherwise, we have that

$$\sum_{k\geq 1} \frac{1}{\|d_k\|^2} < +\infty, \tag{8.3.1}$$

which gives

$$\lim_{k\to\infty} \|d_k\| = +\infty. \tag{8.3.2}$$

It follows from Assumption 8.2.1 that

$$\|g_k\| \leq \bar{\gamma}, \quad \text{for some } \bar{\gamma} > 0 \text{ and all } k \geq 1. \tag{8.3.3}$$

Then by (8.1.3), (8.3.2) and (8.3.3), we have that

$$\|d_k\| \approx |\beta_k| \, \|d_{k-1}\|. \tag{8.3.4}$$

Thus if the scalar β_k is such that

$$\sum_{k\geq 1} \prod_{j=2}^{k} \beta_j^{-2} = +\infty, \tag{8.3.5}$$

it is possible for us to establish (8.2.4) and then by Lemma 8.2.3 obtain a contradiction to (8.3.1). We formally describe this result as follows and give a strict proof, since the proof here is easy to understand and is quite different from the one for the Wolfe line search.

Theorem 8.3.1. *Suppose that Assumption 8.2.1 holds. Consider any method of the form (8.1.2)–(8.1.3) with d_k satisfying $g_k^T d_k < 0$ and with the strong Wolfe line searches (8.1.4) and (8.1.5). If β_k satisfies (8.3.5), we have that* $\liminf_{k\to\infty} \|g_k\| = 0$.

Proof. We write (8.1.3) as follows:

$$d_k + g_k = \beta_k d_{k-1}. \tag{8.3.6}$$

Squaring both sides of (8.3.6), we get that

$$\|d_k\|^2 = -2g_k^T d_k - \|g_k\|^2 + \beta_k^2 \|d_{k-1}\|^2. \tag{8.3.7}$$

Noting that

$$-2g_k^T d_k - \|g_k\|^2 \leq \frac{(g_k^T d_k)^2}{\|g_k\|^2}, \tag{8.3.8}$$

it follows from this and (8.3.7) that

$$\|d_k\|^2 \leq \frac{(g_k^T d_k)^2}{\|g_k\|^2} + \beta_k^2 \|d_{k-1}\|^2 \tag{8.3.9}$$

Letting θ_k be the angle between $-g_k$ and d_k, namely,

$$\cos\theta_k = \frac{-g_k^T d_k}{\|g_k\| \, \|d_k\|}, \tag{8.3.10}$$

we get from (8.3.9) and (8.3.10) that

$$\begin{aligned}\|d_k\|^2 &\le (1-\cos^2\theta_k)^{-1}\beta_k^2\|d_{k-1}\|^2\\ &\le \cdots\cdots\\ &\le \prod_{j=j_0}^{k}(1-\cos^2\theta_j)^{-1}\,(\prod_{j=j_0}^{k}\beta_j^2)\|d_{j_0-1}\|^2,\end{aligned} \tag{8.3.11}$$

where $j_0 \ge 2$ is any integer. We now assume that (8.2.5) does not hold and hence there exists some constant $\gamma > 0$ such that

$$\|g_k\| \ge \gamma, \quad \text{for all } k \ge 1. \tag{8.3.12}$$

Then it follows from (8.2.2), (8.3.10) and (8.3.12) that

$$\sum_{k\ge 1}\cos^2\theta_k < +\infty. \tag{8.3.13}$$

The above relation clearly implies that

$$\prod_{j\ge j_0}(1-\cos^2\theta_j) \ge c, \quad \text{for some } c>0 \text{ and integer } j_0 \ge 2. \tag{8.3.14}$$

By (8.3.11), (8.3.14) and (8.3.5), we know that (8.2.4) holds. Thus by Lemma 8.2.3, (8.2.5) holds. This with (8.3.12) gives a contradiction, which ends the proof. □

8.3.2 Sufficiency of (8.3.5)

In the above subsection, we propose a condition on (8.3.5) for the global convergence of general conjugate gradient method with the strong Wolfe line search. Its proof is based on Lemma 8.2.3. In this subsection, we prove that Theorem 8.3.1 still holds if the strong Wolfe conditions are replaced with the Wolfe conditions.

At first, we present a basic lemma for general method (8.1.2)–(8.1.3) without line searches (see [6] for a similar lemma).

Lemma 8.3.2. *Consider any method of the form (8.1.2)–(8.1.3). Define ϕ_k and t_k as follows:*

$$\phi_k := \begin{cases}\|g_k\|^2, & \text{for } k=1;\\ \prod_{j=2}^{k}\beta_j^2, & \text{for } k\ge 2\end{cases} \tag{8.3.15}$$

and

$$t_k := \frac{\|d_k\|^2}{\phi_k^2}. \tag{8.3.16}$$

Then we have for all $k \ge 1$,

$$t_k = -2\sum_{i=1}^{k} \frac{g_i^T d_i}{\phi_i^2} - \sum_{i=1}^{k} \frac{\|g_i\|^2}{\phi_i^2}. \tag{8.3.17}$$

Proof. Since $d_1 = -g_1$ and $\phi_1 = \|g_1\|^2$, (8.3.17) holds for $k = 1$. For $k \geq 2$, dividing (8.3.7) by ϕ_k^2 and using the definitions of t_k and ϕ_k, we get that

$$t_k = t_{k-1} - \frac{2g_k^T d_k}{\phi_k^2} - \frac{\|g_k\|^2}{\phi_k^2}. \tag{8.3.18}$$

Summing this expression for $k = 2, \ldots, k$, we obtain

$$t_k = t_1 - 2\sum_{i=2}^{k} \frac{g_i^T d_i}{\phi_i^2} - \sum_{i=2}^{k} \frac{\|g_i\|^2}{\phi_i^2}. \tag{8.3.19}$$

Since $d_1 = -g_1$ and $t_1 = \|g_1\|^2/\phi_1^2$, the above relation is equivalent to (8.3.17). So (8.3.17) holds for all $k \geq 1$. □

To show our main result, we still require the following lemma. See, for example, Pu and Yu [19] for its proof.

Lemma 8.3.3. *Suppose that $\{a_i\}$ and $\{b_i\}$ are positive number sequences, satisfying*

$$b_k \leq c_1 + c_2 \sum_{i=1}^{k} a_i, \quad \textit{for all } k, \tag{8.3.20}$$

where c_1 and c_2 are positive constants. If the sum $\sum_{k\geq 1} a_k$ is divergent, then $\sum_{k\geq 1} a_k/b_k$ is also divergent.

Now we prove that the condition (8.3.5) on β_k is sufficient for the global convergence of any method of the form (8.1.2)–(8.1.3).

Theorem 8.3.4. *Suppose that Assumption 8.2.1 holds. Consider any method of the form (8.1.2)–(8.1.3) with d_k satisfying $g_k^T d_k < 0$ and with the Wolfe line searches (8.1.4) and (8.1.10). If β_k satisfies (8.3.5), we have that $\liminf_{k\to\infty} \|g_k\| = 0$.*

Proof. Define ϕ_k as in (8.3.15). It follows from (8.3.5) that

$$\sum_{k\geq 1} \frac{1}{\phi_k^2} = +\infty. \tag{8.3.21}$$

Using (8.3.8) in (8.3.17), we can get

$$t_k \leq \sum_{i=1}^{k} \frac{(g_i^T d_i)^2}{\|g_i\|^2 \phi_i^2}. \tag{8.3.22}$$

Since $t_k \geq 0$, we also have by (8.3.17) that

$$-2\sum_{i=1}^{k}\frac{g_i^T d_i}{\phi_i^2} \geq \sum_{i=1}^{k}\frac{\|g_i\|^2}{\phi_i^2}. \tag{8.3.23}$$

Noting that

$$-4g_k^T d_k - \|g_k\|^2 \leq 4\frac{(g_k^T d_k)^2}{\|g_k\|^2} \tag{8.3.24}$$

for any k, we obtain from this and (8.3.23) that

$$4\sum_{i=1}^{k}\frac{(g_i^T d_i)^2}{\|g_i\|^2\phi_i^2} \geq -4\sum_{i=1}^{k}\frac{g_i^T d_i}{\phi_i^2} - \sum_{i=1}^{k}\frac{\|g_i\|^2}{\phi_i^2} \geq \sum_{i=1}^{k}\frac{\|g_i\|^2}{\phi_i^2}. \tag{8.3.25}$$

Now we proceed by contradiction and assume that (8.3.12) holds. Then by (8.3.25), (8.3.21) and (8.3.12), we have that

$$\sum_{k\geq 1}\frac{(g_k^T d_k)^2}{\|g_k\|^2\phi_k^2} = +\infty. \tag{8.3.26}$$

Using relations (8.3.22), (8.3.26) and Lemma (8.3.3), we then obtain

$$\sum_{k\geq 1}\frac{(g_k^T d_k)^2}{\|g_k\|^2\phi_k^2}\,\frac{1}{t_k} = \sum_{k\geq 1}\frac{(g_k^T d_k)^2}{\|g_k\|^2\|d_k\|^2} = \sum_{k\geq 1}\cos^2\theta_k = +\infty, \tag{8.3.27}$$

which contradicts (8.3.13). The contradiction shows the truth of (8.2.5). □

Thus we have proved that the condition on β_k is sufficient for the global convergence of general conjugate gradient method with the Wolfe line search. Instead of the sufficient descent condition (8.2.3), only the descent condition $d_k^T g_k < 0$ is used in the Theorem. Since different nonlinear conjugate gradient methods vary with the scalar β_k and condition (8.3.5) only concerns β_k, we believe that Theorem 8.3.4 is very powerful in the convergence analyses of conjugate gradient methods, as will partly be shown in the coming section.

From the proof to Theorem 8.3.4, we can see that the relation (8.3.22) gives an upper bound for the quantity $t_k = \|d_k\|^2/\phi_k^2$, whereas (8.3.25) estimates the lower bound of a quantity related to $g_k^T d_k$. Both the relations (8.3.22) and (8.3.25) are derived from (8.3.17). Then by the two relations and Lemma 8.3.3, we are able to prove the sufficiency of (8.3.5) for the global convergence of general conjugate gradient method with the Wolfe line search.

For some conjugate gradient method, we know that if the sequence $\{\|d_k\|^2\}$ increases mostly linearly,

$$\|d_k\|^2 \leq c_1 + c_2 k, \tag{8.3.28}$$

where c_1 and c_2 are positive constants, and if the sufficient descent condition (8.2.3) holds for all k, then we can conclude the global convergence by the Zou-

tendijk condition (8.2.2) and the contradiction principle. Such approach is often used in the convergence analyses of many conjugate gradient methods, for example, the analyses of the FR method in Al-Baali [1] and the ones of the PRP method in Gilbert and Nocedal [10]. From the proof to Theorem 8.3.4, we see that the sufficient descent condition (8.2.3) is not necessary, but the roles of (8.3.22) and (8.3.25) are similar to those of (8.3.28) and (8.2.3).

For convenience in use, we give the following corollary of Theorem 8.3.4 at the end of this subsection.

Corollary 8.3.5. *Suppose that Assumption 8.2.1 holds. Consider any method of the form (8.1.2)–(8.1.3) with d_k satisfying $g_k^T d_k < 0$ and with the Wolfe line searches (8.1.4) and (8.1.10). If there exist nonnegative constants c_1 and c_2 such that*

$$\prod_{j=2}^{k} \beta_j^2 \leq c_1 + c_2 k, \tag{8.3.29}$$

we have that $\liminf_{k\to\infty} \|g_k\| = 0$.

Proof. Since (8.3.29) implies that (8.3.5) holds, the statement follows Theorem 8.3.4. □

8.3.3 Necessity of (8.3.5)

In this subsection, we consider the necessity of the condition (8.3.5). To make our analyses more general, we assume that the line search is exact, namely,

$$g_{k+1}^T d_k = 0, \quad \text{for all } k. \tag{8.3.30}$$

We also assume that the iterations $\{x_k; k = 2, 3, \ldots\}$ fall into a region Ω where f is a quadratic function with the unit Hessian,

$$f(x) = \frac{1}{2} x^T x, \quad x \in \Omega \subset R^n. \tag{8.3.31}$$

Then by (8.1.3) and (8.3.30), we have that

$$d_k^T g_k = -\|g_k\|^2, \tag{8.3.32}$$

$$\|d_k\|^2 = \|g_k\|^2 + \beta_k^2 \|d_{k-1}\|^2. \tag{8.3.33}$$

It follows from (8.1.2), (8.1.3) and (8.3.31) that

$$g_{k+1} = g_k + \alpha_k d_k, \tag{8.3.34}$$

which with (8.3.30) and (8.3.32) gives

$$\alpha_k = -\frac{g_k^T d_k}{\|d_k\|^2} = \frac{\|g_k\|^2}{\|d_k\|^2}. \tag{8.3.35}$$

By squaring both sides of (8.3.34) and using (8.3.32), (8.3.35) and (8.3.33), we get that

$$\|g_{k+1}\|^2 = \|g_k\|^2 + 2\alpha_k g_k^T d_k + \alpha_k^2 \|d_k\|^2 = \|g_k\|^2 \left[1 - \frac{\|g_k\|^2}{\|d_k\|^2}\right] = \beta_k^2 \frac{\|d_{k-1}\|^2}{\|d_k\|^2} \|g_k\|^2. \tag{8.3.36}$$

The recursion of the above relation yields

$$\|g_{k+1}\|^2 = \left(\prod_{j=2}^{k} \beta_j^2\right) \cdot \frac{\|d_1\|^2}{\|d_k\|^2} \cdot \|g_2\|^2. \tag{8.3.37}$$

Still define ϕ_k and t_k as in Lemma 8.3.2. Then we see that (8.3.37) is equivalent to

$$\|g_{k+1}\|^2 = t_k^{-1} \|d_1\|^2 \|g_2\|^2. \tag{8.3.38}$$

On the other hand, we have from (8.3.17), (8.3.32) and $d_1 = -g_1$ that

$$t_k = \sum_{i=1}^{k} \frac{\|g_i\|^2}{\phi_i^2}. \tag{8.3.39}$$

Note from the second equality in (8.3.36) that the sequence $\{\|g_k\|^2; k = 2, 3, \ldots\}$ is monotonically decreasing. Thus have that

$$\|g_k\| \leq \|g_2\|, \quad \text{for all } k \geq 2. \tag{8.3.40}$$

Therefore if (8.3.5) is false, namely,

$$\sum_{k \geq 2} \prod_{j=2}^{k} \beta_j^{-2} < +\infty, \tag{8.3.41}$$

we have from the definition of ϕ_k, (8.3.40) and (8.3.39) that

$$t_k \leq M, \quad \text{for some positive constant } M. \tag{8.3.42}$$

Relations (8.3.42) and (8.3.38) indicate that

$$\|g_k\| \geq M^{-1} \|d_1\|^2 \|g_2\|^2, \quad \text{for all } k, \tag{8.3.43}$$

which implies that the iterations can not approach the unique minimizer $x^* = 0$ of the function in (8.3.31). By contrast, if (8.3.5) is true, we have by the definition of ϕ_k, (8.3.39) and (8.3.38) that $t_k \to +\infty$ and $\lim_{k \to \infty} \|g_k\| = 0$. Therefore, in this

example, any method of the form (8.1.2)–(8.1.3) converges globally if and only if the condition (8.3.5) holds.

8.4 Applications of the condition (8.3.5)

In the above section, we have presented a sufficient condition, namely, (8.3.5), on β_k for the global convergence of general conjugate gradient method using the Wolfe line search. As a matter of fact, the previous analyses do not depend on the choice of β_k and hence apply to any method of the form (8.1.2)–(8.1.3). To make better use of the condition in the conjugate gradient field, we will introduce a new property, namely, Property (#), in §4.1. Such property may apply to all the three categories of conjugate gradient methods. Consequently, by Theorem 8.3.4 and Property (#), convergence results can conveniently be gained for some known conjugate gradient methods (see §4.2). An analysis is also given to a new conjugate gradient method, which further shows the usefulness of Theorem 8.3.4 and Property (#) (see §4.3).

8.4.1 Property (#)

In [10], Gilbert and Nocedal proposed the so-called Property (∗) for the second category of conjugate gradient methods and brought about the convergence results for the PRP, and HS methods with the restriction $\beta_k \geq 0$. The purpose of this subsection is to define a new property, that may apply to all the three categories of conjugate gradient methods.

Denoting $s_{k-1} = x_k - x_{k-1}$, we define Property (#) as follows:

Property (#). *Consider a method of the form (8.1.2)–(8.1.3), and suppose that*

$$0 < \gamma \leq \|g_k\| \leq \bar{\gamma}, \quad \text{for all } k. \tag{8.4.1}$$

Under this assumption we say that the method has Property (#) if there exist a positive and uniformly bounded sequence $\{\psi_k\}$, and constants $b \geq 1$ and $\lambda > 0$ such that for all k:

$$\text{(1) } |\beta_k| \leq b\,\frac{\psi_k}{\psi_{k-1}}; \tag{8.4.2}$$

$$\text{(2) if } \|s_{k-1}\| \leq \lambda, \text{ then } |\beta_k| \leq \frac{1}{b}\,\frac{\psi_k}{\psi_{k-1}}. \tag{8.4.3}$$

The above Property (#) clearly has Property (∗) in [10] as its special case. Under the same assumption (8.4.1), Property (∗) requires that there exist constants $b > 1$ and $\lambda > 0$ such that $|\beta_k| \leq b$ and if $\|s_{k-1}\| \leq \lambda$, then $|\beta_k| \leq \frac{1}{2b}$. So if Property (∗) holds, Property (#) must be true with $\psi_k \equiv 1$.

Similar to [10], we now present an analysis of the PRP method. Let

$$b = \frac{2\bar{\gamma}^2}{\gamma^2}, \quad \lambda = \frac{\gamma^2}{L\bar{\gamma}b}, \quad \psi_k \equiv 1, \tag{8.4.4}$$

where L is the Lipschitz constant in (8.2.1). Then by (8.1.7), (8.4.1) and (8.2.1), we can get that

$$|\beta_k^{PRP}| \le \frac{(\|g_k\| + \|g_{k-1}\|)\|g_k\|}{\|g_{k-1}\|} \le \frac{2\bar{\gamma}^2}{\gamma^2} = b\frac{\psi_k}{\psi_{k-1}}, \tag{8.4.5}$$

and if $\|s_{k-1}\| \le \lambda$,

$$|\beta_k^{PRP}| \le \frac{\|y_{k-1}\|\,\|g_k\|}{\|g_{k-1}\|} \le \frac{L\bar{\gamma}\|s_{k-1}\|}{\gamma^2} \le \frac{L\lambda\bar{\gamma}}{\gamma^2} = \frac{1}{b}\frac{\psi_k}{\psi_{k-1}}. \tag{8.4.6}$$

So Property (#) holds with the b, λ and ψ_k in (8.4.4). If we reduce the above λ by half, then Property $(*)$ in [10] also holds.

Since the ψ_k in Property (#) can be any bounded sequence, and since the factor $\frac{1}{2}$ of Property $(*)$ is missing in (8.4.3), Property (#) may apply to not only the second category of methods but the first and third categories. In fact, for the FR method, which belongs to the first category, we can choose

$$b = 1, \quad \psi_k = \|g_k\|^2, \quad \text{and } \lambda \text{ is any positive number.} \tag{8.4.7}$$

By the definition (8.1.6) of β_k^{FR}, (8.4.2)–(8.4.3) clearly holds. In addition, (8.4.1) implies that $\psi_k = \|g_k\|^2$ is uniformly bounded. Thus the FR method has Property (#). For the DY method with the Wolfe line search, we can get by multiplying (8.1.3) with g_k and using (8.1.9) that

$$g_k^T d_k = \frac{\|g_k\|^2}{d_{k-1}^T y_{k-1}} g_{k-1}^T d_{k-1}. \tag{8.4.8}$$

The above relation and (8.1.9) give an equivalent formula of β_k^{DY} (see also [7]):

$$\beta_k^{DY} = \frac{g_k^T d_k}{g_{k-1}^T d_{k-1}}. \tag{8.4.9}$$

Further, we have from this, (8.4.8), (8.4.1) and (8.1.10) that

$$-g_k^T d_k \le (1-\sigma)^{-1}\|g_k\|^2 \le (1-\sigma)^{-1}\bar{\gamma}^2. \tag{8.4.10}$$

Thus for the DY method, Property (#) holds with

$$b = 1, \quad \psi_k = -g_k^T d_k, \quad \text{and any } \lambda > 0. \tag{8.4.11}$$

For the FR-PRP, and DY-HS hybrid methods, that belong to the third category, we have by (8.1.11), (8.1.12) and (8.1.13) that $|\beta_k| \le \beta_k^{FR}$ or $\beta_k \in [0, \beta_k^{DY}]$.

Hence these hybrid methods have also Property (#), as will be seen in the proof of Corollaries 8.4.2 and 8.4.3.

To sum up, Property (#) includes Property (∗) in [10] as its special case, and may apply to all the three categories of conjugate gradient methods.

8.4.2 Applications to some known conjugate gradient methods

In this subsection, we will discuss how to use Theorem 8.3.4 and Property (#) to analyze the global convergence of some known conjugate gradient methods.

At first, we have the following theorem for those methods for which Property (#) holds with $b = 1$.

Theorem 8.4.1. *Suppose that Assumption 8.2.1 holds. Consider any method of the form (8.1.2)–(8.1.3), where the scalar β_k has Property (#) with $b = 1$. If the step-length λ_k satisfies the Wolfe conditions (8.1.4) and (8.1.10) and the descent condition $g_k^T d_k < 0$, then we have that $\liminf_{k\to\infty} \|g_k\| = 0$.*

Proof. We proceed by contradiction, assuming that (8.3.12) holds. Then we know by (8.3.12) and (8.3.3) that (8.4.1) is true. By Property (#) with $b = 1$, we then have that

$$\prod_{j=2}^{k} \beta_j^2 = \frac{\psi_k^2}{\psi_1^2}, \tag{8.4.12}$$

which with the boundedness of ψ_k implies that (8.3.29) holds. Thus by Corollary 8.3.5, we have $\liminf_{k\to\infty} \|g_k\| = 0$, contradicting (8.3.12). The contradiction shows the truth of this theorem. □

By the above theorem, we can analyze the global convergence of some conjugate gradient methods in the first and third categories. For example, we have the following result for the FR method and its related hybrid methods.

Corollary 8.4.2. *Suppose that Assumption 8.2.1 holds. Consider the method (8.1.2)–(8.1.3) with $|\beta_k| \le \beta_k^{FR}$. If the step-length λ_k satisfies the Wolfe conditions (8.1.4) and (8.1.10) and the descent condition $g_k^T d_k < 0$, then we have that $\liminf_{k\to\infty} \|g_k\| = 0$.*

Proof. Noting that the method is such that (8.4.2)–(8.4.3) holds with the parameters in (8.4.7), the statement follows Theorem 8.4.1. □

The above corollary clearly covers the FR method and the hybrid methods (8.1.11) and (8.1.12). If the strong Wolfe conditions (8.1.4)–(8.1.5) are used, and if $\sigma \le 0.5$, we can prove any method (8.1.2)–(8.1.3) with $|\beta_k| \le \beta_k^{FR}$ generates a descent direction at every iteration. Then by Corollary 8.4.2, we know that there is the global convergence, and hence obtain again those corresponding results in

[1, 5, 10] for the FR method. For the DY method and its related hybrid method, we also have the following corollary. See also [8] for the result.

Corollary 8.4.3. *Suppose that Assumption 8.2.1 holds. Consider the method (8.1.2)–(8.1.3) with* $\beta_k = r_k\beta_k^{DY}$. *If the step-length* λ_k *satisfies the Wolfe conditions (8.1.4) and (8.1.10), and if*

$$r_k \in \left[\frac{\sigma-1}{1+\sigma}, 1\right], \tag{8.4.13}$$

then we have that $\liminf_{k\to\infty} \|g_k\| = 0$.

Proof. First, we prove by induction that $g_k^T d_k < 0$ for all k. In fact, since $d_1 = -g_1$, it is obvious that $g_1^T d_1 < 0$. Suppose that $g_{k-1}^T d_{k-1} < 0$. Then we have by (8.1.10) that

$$l_k := \frac{g_k^T d_{k-1}}{g_{k-1}^T d_{k-1}} \le \sigma. \tag{8.4.14}$$

It follows from (8.1.3) and $\beta_k = r_k\beta_k^{DY}$ that

$$g_k^T d_k = -\|g_k\|^2 + r_k\beta_k^{DY} g_k^T d_{k-1}, \tag{8.4.15}$$

from (8.4.13) and (8.4.14), we obtain

$$g_k^T d_k = \frac{1+(r_k-1)l_k}{l_k-1}\|g_k\|^2 \in [-\frac{1}{1+\sigma}\|g_k\|^2, 0). \tag{8.4.16}$$

So $g_k^T d_k < 0$. By the induction principle, $g_k^T d_k < 0$ for all k.

Further, by $\beta_k = r_k\beta_k^{DY}$, (8.4.15) and the definition of l_k, we get that

$$\beta_k = \frac{r_k}{1+(r_k-1)l_k}\frac{g_k^T d_k}{g_{k-1}^T d_{k-1}}, \tag{8.4.17}$$

from (8.4.13) and (8.4.14), we can obtain

$$|\beta_k| \le \frac{g_k^T d_k}{g_{k-1}^T d_{k-1}}. \tag{8.4.18}$$

This relation and (8.4.16) indicate that Property (#) holds with the parameters in (8.4.11). Therefore the result follows Theorem 8.4.1. □

To use Property (#) to analyze the second category of conjugate gradient methods, we now provide the following general lemma. In the lemma, we denote N^* to be the set of positive integers and

$$\mathcal{K}_{k,\Delta}^{\lambda} := \{i \in N^* : k \le i \le k+\Delta-1, \|s_{i-1}\| \ge \lambda\}, \tag{8.4.19}$$

and let $|\mathcal{K}_{k,\Delta}^{\lambda}|$ be the number of elements of the set $\mathcal{K}_{k,\Delta}^{\lambda}$.

Lemma 8.4.4. *Suppose that Assumption 8.2.1 holds. Consider any method (8.1.2)–(8.1.3) having Property (#), where the step-length λ_k satisfies the Wolfe conditions (8.1.4) and (8.1.10) and the descent condition $g_k^T d_k < 0$. If there exist $\Delta \in N^*$ and integer k_0 such that*

$$|\mathcal{K}^{\lambda}_{k,\Delta}| \leq \frac{\Delta}{2}, \quad \textit{for any } k \geq k_0, \tag{8.4.20}$$

we have that $\liminf_{k\to\infty} \|g_k\| = 0$.

Proof. For any $i \geq 1$, denote

$$p_i := |\mathcal{K}^{\lambda}_{k_0,i\Delta}|. \tag{8.4.21}$$

It follows by (8.4.20) and the arbitrariness of $k \geq k_0$ in the relation that

$$p_i \leq \frac{i\Delta}{2}. \tag{8.4.22}$$

This means that in the range $[k_0, k_0 + i\Delta - 1)$ there are exactly p_i indices j such that $\|s_{j-1}\| > \lambda$, and thus there are $(i\Delta - p_i)$ indices with $\|s_{j-1}\| < \lambda$. Using this fact, (8.4.2), (8.4.3) and (8.4.22), we obtain for any $i \geq 1$

$$\begin{aligned}\prod_{j=k_0}^{k_0+i\Delta-1} \beta_j^2 &\leq b^{2p_i} \left(\frac{1}{b}\right)^{2(i\Delta-p_i)} \prod_{j=k_0}^{k_0+i\Delta-1} \frac{\psi_j^2}{\psi_{j-1}^2} \\ &\leq b^{2(2p_i-i\Delta)} \frac{\psi_{k_0+i\Delta-1}^2}{\psi_{k_0-1}^2} \leq \frac{\psi_{k_0+i\Delta-1}^2}{\psi_{k_0-1}^2}. \end{aligned} \tag{8.4.23}$$

Since $\{\psi_k\}$ is uniformly bounded, the above relation indicates that (8.3.5) holds. Thus the result follows Theorem 8.3.4. □

Now we prove a general result for any method with Property (#) and $\beta_k \geq 0$. The restriction that $\beta_k \geq 0$ was first suggested by Powell [18] for the PRP method and later used by Gilbert and Nocedal [10] in getting the convergence result for algorithms related to the PRP method.

Theorem 8.4.5. *Suppose that Assumption 8.2.1 holds. Consider any method (8.1.2)–(8.1.3) with Property (#) and $\beta_k \geq 0$. If the step-length λ_k satisfies the Wolfe conditions (8.1.4) and (8.1.10) and the descent condition $g_k^T d_k < 0$, then we have either*

$$\liminf_{k\to\infty} \|d_k\| < +\infty, \tag{8.4.24}$$

or the convergence relation (8.2.5) holds.

Proof. Suppose that (8.1.10) is false. Then we have that

$$\lim_{k\to\infty} \|d_k\| = +\infty. \tag{8.4.25}$$

Since $g_k^T d_k < 0$, we have that $d_k \neq 0$. Define $u_k := d_k/\|d_k\|$,

$$\rho_k := \frac{-g_k}{\|d_k\|} \quad \text{and} \quad \delta_k := \frac{\beta_k \|d_{k-1}\|}{\|d_k\|}. \tag{8.4.26}$$

From (8.1.3), we have for $k \geq 2$:

$$u_k = \rho_k + \delta_k u_{k-1}. \tag{8.4.27}$$

Note that $\|u_k\| = \|u_{k-1}\| = 1$ and by (8.3.3) and (8.4.25), $\lim_{k\to\infty} \|\rho_k\| = 0$. Hence, by (8.4.27),

$$\lim_{k\to\infty} |\delta_k| = 1, \tag{8.4.28}$$

which with (8.4.27) and the condition $\delta_k \geq 0$ implies that

$$\lim_{k\to\infty} \|u_k - u_{k-1}\| = \lim_{k\to\infty} \|\rho_k + (\delta_k - 1)u_{k-1}\| = 0. \tag{8.4.29}$$

In addition, using (8.1.2) and the definition of u_k, we can write for any indices l, k, with $l \geq k$:

$$x_l - x_{k-1} = \sum_{i=k}^{l} \|s_{i-1}\| u_{i-1} = \sum_{i=k}^{l} \|s_{i-1}\| u_{k-1} + \sum_{i=k}^{l} \|s_{i-1}\| (u_{i-1} - u_{k-1}). \tag{8.4.30}$$

So we have that

$$\begin{aligned} \sum_{i=k}^{l} \|s_{i-1}\| &\leq \|x_l - x_{k-1}\| + \sum_{i=k}^{l} \|s_{i-1}\| \|u_{i-1} - u_{k-1}\| \\ &\leq 2B + \sum_{i=k}^{l} \|s_{i-1}\| \, \|u_{i-1} - u_{k-1}\|, \end{aligned} \tag{8.4.31}$$

where B is a bound of the level set $\mathcal{L}$.

We now proceed by contradiction and assume that $\liminf_{k\to\infty} \|g_k\| \neq 0$. Then by Lemma 8.4.4, for any Δ and integer k_0, there exists an index $k \geq k_0$ such that

$$|\mathcal{K}_{k,\Delta}^{\lambda}| > \frac{\Delta}{2}. \tag{8.4.32}$$

Let $\Delta := \lceil 8B/\lambda \rceil$. For this Δ, by (8.4.29), we can choose k_0 such that

$$\|u_k - u_{k-1}\| \leq \frac{1}{2\Delta}, \quad \text{for all } k \geq k_0. \tag{8.4.33}$$

Then for any $i \in [k, k + \Delta - 1]$, we have by (8.4.29) and (8.4.33) that

$$\|u_{i-1} - u_{k-1}\| \leq \sum_{j=k}^{i-1} \|u_j - u_{j-1}\| \leq \Delta \cdot \left(\frac{1}{2\Delta}\right) = \frac{1}{2}. \tag{8.4.34}$$

Using (8.4.34) and (8.4.32) in (8.4.31), with $l = k + \Delta - 1$, we obtain

$$2B \geq \frac{1}{2} \sum_{i=k}^{k+\Delta-1} \|s_{i-1}\| > \frac{\lambda}{2} |\mathcal{K}^{\lambda}_{k,\Delta}| > \frac{\lambda\Delta}{4}. \tag{8.4.35}$$

Thus $\Delta < 8B/\lambda$, which contradicts the definition of Δ. Therefore (8.2.5) holds, which ends our proof. □

Note that if (8.4.24) holds, then (8.2.4) must be true. Thus by Lemma 8.2.3, we know that the convergence relation (8.2.5) holds if the Wolfe line search in Theorem 8.4.5 is replaced with the strong Wolfe line search. Consequently, we have the following corollary for the PRP method with $\beta_k \geq 0$. See also [3] for this result.

Corollary 8.4.6. *Suppose that Assumption 8.2.1 holds. Consider the method (8.1.2)–(8.1.3) with* $\beta_k = \max\{\beta_k^{PRP}, 0\}$. *If the step-length* λ_k *satisfies the strong Wolfe conditions (8.1.4)–(8.1.5) and the descent condition* $g_k^T d_k < 0$, *we have that* $\liminf_{k\to\infty} \|g_k\| = 0$.

Proof. Suppose that this corollary is false and (8.3.12) holds. Then, noting that β_k is nonnegative and that Property (#) holds with the parameters in (8.4.4), we have by Theorem 8.4.5 that relation (8.4.24) holds. It follows that (8.2.4) is true. Therefore by Lemma 8.2.3, we obtain $\liminf_{k\to\infty} \|g_k\| = 0$, contradicting (8.3.12). The contradiction shows the truth of this corollary. □

Noting that relations (8.2.2), (8.2.3) and (8.3.12) indicate the truth of (8.4.25), we know that there is also the global convergence if the descent condition $g_k^T d_k < 0$ in Theorem 8.4.5 is replaced with the sufficient descent condition (8.2.3). Thus we can prove the following result for the HS method with $\beta_k \geq 0$ ([10]). The proof here is different from the one in [10].

Corollary 8.4.7. *Suppose that Assumption 8.2.1 holds. Consider the method (8.1.2)–(8.1.3) with* $\beta_k = \max\{\beta_k^{HS}, 0\}$. *If the step-length* λ_k *satisfies the Wolfe conditions (8.1.4) and (8.1.10) and the sufficient descent condition (8.2.3), then we have that* $\liminf_{k\to\infty} \|g_k\| = 0$.

Proof. We proceed by contradiction and assume that (8.3.12) holds. Define

$$b = \frac{2\bar{\gamma}}{c\gamma(1-\sigma)}, \quad \psi_k = -g_k^T d_k, \quad \lambda = \frac{c\gamma(1-\sigma)}{Lb}. \tag{8.4.36}$$

Then it follows by (8.1.8), (8.3.12), (8.3.3), (8.1.10), (8.2.3) and (8.2.1) that

$$|\beta_k^{HS}| \leq \frac{2\bar{\gamma}\|g_k\|}{(\sigma-1)g_{k-1}^T d_{k-1}} \leq \frac{bg_k^T d_k}{g_{k-1}^T d_{k-1}} = b\frac{\psi_k}{\psi_{k-1}}, \tag{8.4.37}$$

and if $\|s_{k-1}\| \leq \lambda$,

$$|\beta_k^{HS}| \leq \frac{L\lambda\|g_k\|}{(\sigma-1)g_{k-1}^T d_{k-1}} \leq \frac{g_k^T d_k}{bg_{k-1}^T d_{k-1}} = \frac{1}{b}\frac{\psi_k}{\psi_{k-1}}. \tag{8.4.38}$$

In addition, the line search condition (8.1.10) and $g_{k-1}^T d_{k-1} < 0$ imply that

$$|g_k^T d_{k-1}| \leq |d_{k-1}^T y_{k-1}|. \tag{8.4.39}$$

By (8.1.3), (8.1.8) and (8.3.3), we have that

$$-g_k^T d_k = \|g_k\|^2 - \frac{g_k^T d_{k-1}}{d_{k-1}^T y_{k-1}} g_k^T y_{k-1} \leq \|g_k\|^2 + |g_k^T y_{k-1}| \leq 3\bar{\gamma}^2. \tag{8.4.40}$$

The above relation implies that the ψ_k is a bounded sequence. So Property (#) holds. Since $\beta_k \geq 0$, we have by Theorem 8.4.5 and (8.3.12) that (8.4.24) is true. However, we have from (8.2.2), (8.2.3) and (8.3.12) that $\lim\limits_{k\to\infty} \|d_k\| = +\infty$, contradicting (8.4.24). The contradiction shows the truth of this corollary. □

8.4.3 Application to a new conjugate gradient method

To further show the usefulness of Property (#) in the convergence analyses of conjugate gradient methods, we will consider a new conjugate gradient method in this subsection.

For any method of the form (8.1.2)–(8.1.3), noting that

$$d_k^T g_k = -\|g_k\|^2 + \beta_k g_k^T d_{k-1}, \tag{8.4.41}$$

we know that d_k is a descent direction if the $(k-1)$-th line search is enough exact. Since exact line searches are expensive, and since the line search is only to minimize the objective function in the one-dimensional subspace $\{x_{k-1}+\alpha d_{k-1}\}$, it is preferable to do some inexact line search in practical computations. Suppose that the Wolfe line search is used and d_{k-1} is a descent direction. In this case, to ensure the descent property of d_k, we know from (8.4.41) that the choice of β_k should satisfy

$$\beta_k g_k^T d_{k-1} < \|g_k\|^2. \tag{8.4.42}$$

Assuming that

$$\beta_k = \frac{\|g_k\|^2}{g_k^T d_{k-1} + \tau_k}, \tag{8.4.43}$$

where τ_k satisfies

$$g_k^T d_{k-1} + \tau_k > 0, \tag{8.4.44}$$

the condition (8.4.42) is equivalent to

$$\tau_k > 0. \tag{8.4.45}$$

If we choose $\tau_k = -g_{k-1}^T d_{k-1}$, then it follows from the descent property of d_{k-1} and the second Wolfe condition (8.1.10) that the relations (8.4.44) and (8.4.45) hold. This method is just the DY method, for which we have proved its descent property and global convergence (see [7] or Corollary 8.4.3). Another possible choice is that

$$\tau_k = \|g_{k-1}\|^2, \tag{8.4.46}$$

which with (8.4.43) gives

$$\beta_k = \frac{\|g_k\|^2}{g_k^T d_{k-1} + \|g_{k-1}\|^2}. \tag{8.4.47}$$

For such method, we can really prove that it can produce a descent direction at every iteration if the parameter σ in (8.1.10) is not greater than 0.25. However, due to the good numerical performances of the hybrid method (8.1.13), we are only interested in the following hybrid method of (8.4.47):

$$\beta_k = \frac{\max\{0, \min\{g_k^T y_{k-1}, \|g_k\|^2\}\}}{g_k^T d_{k-1} + \|g_{k-1}\|^2}. \tag{8.4.48}$$

Under mild assumptions, we can prove that the method (8.4.48) produces a descent direction at every iteration and converges globally. The proof is mainly based on Theorem 4.5.

Theorem 8.4.8. *Suppose that Assumption 8.2.1 holds. Consider the methods (8.1.2), (8.1.3), (8.4.48), where the step-length λ_k satisfies the Wolfe conditions (8.1.4) and (8.1.10). If the parameter σ is such that*

$$\sigma \leq 0.25, \tag{8.4.49}$$

we have that for all $k \geq 1$,

$$-2\|g_k\|^2 \leq g_k^T d_k < 0. \tag{8.4.50}$$

Further, we have that $\liminf_{k\to\infty} \|g_k\| = 0$.

Proof. Defining

$$\xi_k = \max\left\{0, \min\left\{\frac{g_k^T y_{k-1}}{\|g_k\|^2}, 1\right\}\right\}, \tag{8.4.51}$$

the formula (8.4.48) for β_k can be rewritten as

$$\beta_k = \frac{\xi_k \|g_k\|^2}{g_k^T d_{k-1} + \|g_{k-1}\|^2}. \tag{8.4.52}$$

Multiplying (8.1.3) by $-g_k$ and using (8.4.52), we can get

$$-g_k^T d_k = \frac{(1-\xi_k) g_k^T d_{k-1} + \|g_{k-1}\|^2}{g_k^T d_{k-1} + \|g_{k-1}\|^2} \|g_k\|^2. \tag{8.4.53}$$

We now prove by induction that (8.4.50) holds for all $k \geq 1$. In fact, since $d_1 = -g_1$, (8.4.50) holds for $k = 1$. Suppose that (8.4.50) holds for some $k-1$. Then by (8.1.10), (8.4.49) and the induction hypothesis, we get

$$g_k^T d_{k-1} \geq \sigma g_{k-1}^T d_{k-1} \geq -\frac{1}{2}\|g_{k-1}\|^2. \tag{8.4.54}$$

Then it follows from (8.4.53), (8.4.54) and $\xi_k \in [0, 1]$ that

$$\frac{-g_k^T d_k}{\|g_k\|^2} = \frac{(1-\xi_k)\frac{g_k^T d_{k-1}}{\|g_{k-1}\|^2} + 1}{\frac{g_k^T d_{k-1}}{\|g_{k-1}\|^2} + 1} \geq \min\left\{ \frac{\bar{\xi}_k a_k + 1}{a_k + 1} : a_k \geq -\frac{1}{2}, \bar{\xi}_k \in [0,1] \right\} = 2. \tag{8.4.55}$$

The above relation implies that (8.4.50) holds for k. By induction, (8.4.50) holds. So each d_k is a descent search direction.

Now we show that the method has Property (#). In fact, using (8.4.53), we can also write β_k as

$$\beta_k = \frac{\xi_k(-g_k^T d_k)}{(1-\xi_k) g_k^T d_{k-1} + \|g_{k-1}\|^2}. \tag{8.4.56}$$

By (8.4.50), (8.4.54) and the fact that $\xi_k \in [0, 1]$, we have

$$(1-\xi_k) g_k^T d_{k-1} + \|g_{k-1}\|^2 \geq \frac{1}{2}\|g_{k-1}\|^2 \geq \frac{1}{4}(-g_{k-1}^T d_{k-1}), \tag{8.4.57}$$

which with (8.4.56) implies that

$$|\beta_k| \leq 4\xi_k \frac{g_k^T d_k}{g_{k-1}^T d_{k-1}}. \tag{8.4.58}$$

Let $b = 4$, $\lambda = \gamma/(16L)$ and $\psi_k = -g_k^T d_k$. It follows by (8.4.58) and $\xi_k \in [0, 1]$ that (8.4.2) holds. If $\|s_{k-1}\| \leq \lambda$, we have from the definition (8.4.51) of ξ_k, (8.2.1) and (8.4.1) that

$$|\xi_k| \leq \frac{\|y_{k-1}\|}{\|g_k\|} \leq \frac{L\lambda}{\gamma} = \frac{1}{b}. \tag{8.4.59}$$

Thus (8.4.3) is also true. In addition, (8.4.50) and (8.3.3) imply that ψ_k is uniformly bounded. So Property (#) holds.

We now proceed by contradiction, assuming (8.3.12). In this case, by Theorem 8.4.5, we know that $\liminf_{k\to\infty} \|d_k\| < +\infty$ and hence there must exist constant $M > 0$ such that

$$\|d_{k_i}\| \le M \tag{8.4.60}$$

holds for some infinite subsequence $\{k_i\} \subset N^*$. It follows by this and (8.3.3) that

$$g_{k_i+1}^T d_{k_i} \le \|g_{k_i+1}\| \, \|d_{k_i}\| \le \bar{\gamma} M. \tag{8.4.61}$$

Applying (8.4.61) and (8.3.12) in (8.4.53) (with k replaced by $k_i + 1$), we obtain

$$-g_{k_i+1}^T d_{k_i+1} \ge \frac{\gamma^4}{\bar{\gamma} M + \gamma^2}, \tag{8.4.62}$$

which with the Zoutendijk condition means that

$$\lim_{i\to\infty} \|d_{k_i+1}\| = +\infty. \tag{8.4.63}$$

On the other hand, we have by (8.4.56), $\xi_k \in [0, 1]$, (8.4.54), (8.3.12) and (8.3.3) that

$$|\beta_k| \le \frac{2\|g_k\|^2}{\|g_{k-1}\|^2} \le \frac{2\bar{\gamma}^2}{\gamma^2}. \tag{8.4.64}$$

By (8.1.3), (8.4.64) and (8.3.3), we can prove

$$\|d_{k_i+1}\| \le \bar{\gamma} + \frac{2\bar{\gamma}^2}{\gamma^2}\|d_{k_i}\|. \tag{8.4.65}$$

Thus by (8.4.60) and (8.4.63), we obtain a contradiction by letting $i \to \infty$ in (8.4.65). The contradiction shows that $\liminf_{k\to\infty} \|g_k\| = 0$. □

According to our numerical experiences with the hybrid method (8.1.13) with the Wolfe line search [4, 8], the parameter σ in (8.1.10) can be chosen as 0.1. Thus, to some extent, we would be satisfied with the condition (8.4.49) in Theorem 8.4.8.

8.5 Discussion

In this chapter we have analyzed nonlinear conjugate gradient methods, where the step-length is computed by the Wolfe line search under the assumption that all the search directions are descent. A general condition on the scalar β_k, that is (8.3.5), was proposed which is sufficient for the global convergence. Since different conjugate gradient methods vary with the scalar β_k, we believe that the result is very powerful in the convergence analyses of conjugate gradient methods.

To use the result better, we have presented a new property, that is Property (#), for conjugate gradient methods. It was also shown that such property may apply to all the three categories of conjugate gradient methods, including the FR, PRP, HS and DY methods and the hybrid methods (8.1.11), (8.1.12) and (8.1.13). As a result, convergence analyses were provided for these methods under mild assumptions.

The result in Section 8.3 can also be used to analyze the bound in the hybrid conjugate gradient methods. Denote $r_k = \beta_k/\beta_k^{FR}$ and consider any method (8.1.2)–(8.1.3) related to the FR method. Assume that the line search satisfies the Wolfe conditions (8.1.4) and (8.1.10) and the descent condition $g_k^T d_k < 0$. If

$$\sum_{k\geq 2}\prod_{j=2}^{k} r_j^{-2} = +\infty, \tag{8.5.1}$$

we have by this, the definition of r_k, (8.1.6) and (8.3.3) that

$$\sum_{k\geq 2}\prod_{j=2}^{k} \beta_j^{-2} = \sum_{k\geq 2}\frac{\|g_1\|^4}{\|g_k\|^4}\prod_{j=2}^{k} r_j^{-2} \geq \frac{\|g_1\|^4}{\bar{\gamma}^4}\prod_{j=2}^{k} r_j^{-2} = +\infty. \tag{8.5.2}$$

Then we can conclude the global convergence by Theorem 8.3.4 and the contradiction principle. On the other hand, if

$$\sum_{k\geq 2}\prod_{j=2}^{k} r_j^{-2} < +\infty, \tag{8.5.3}$$

we can make use of the example in §8.3.3 that the method (8.1.2)–(8.1.3) with exact line searches need not converge. In fact, it follows from (8.3.37), the definition of r_k, (8.1.6), the monotonical decreasing of $\|g_k\|$ and $d_1 = -g_1$ that

$$\frac{\|g_k\|^2}{\|d_k\|^2} = \frac{\|g_1\|^2}{\|g_2\|^2}\frac{\|g_{k+1}\|^2}{\|g_k\|^2}\prod_{j=2}^{k} r_j^{-2} \leq \frac{\|g_1\|^2}{\|g_2\|^2}\prod_{j=2}^{k} r_j^{-2}. \tag{8.5.4}$$

Thus we know from the above relation and (8.5.3) that

$$\sum_{k\geq 1}\frac{\|g_k\|^2}{\|d_k\|^2} < +\infty, \tag{8.5.5}$$

which with the second equality of (8.3.36) implies that for all k,

$$\|g_{k+1}\|^2 = \|g_1\|^2\prod_{i=1}^{k}(1 - \frac{\|g_i\|^2}{\|d_i\|^2}) \geq c', \tag{8.5.6}$$

where c' is some positive constant. Therefore (8.5.1) is also necessary in some sense for the global convergence of general method (8.1.2)–(8.1.3) related to the FR method. A direct corollary to this result is that for any $c > 1$, any method (8.1.2)–(8.1.3) with the restriction $|r_k| \leq c$ need not converge. This result is stronger than Proposition 3.3 in [15], where Nocedal proved that there exists some constant $c > 1$ such that the method (8.1.2)–(8.1.3) with $|r_k| \leq c$ need not converge.

Here we also note that if

$$\sum_{k\geq 1} \frac{1}{\|d_k\|^2} < +\infty, \tag{8.5.7}$$

it follows by this and the boundness of $\|g_k\|^2$ that relation (8.5.5) holds. Then we also have (8.5.6). Since exact line searches are used in the example of §8.3.3, we know that the condition (8.2.4) is also necessary in some sense for the global convergence.

As an illustrative example, in this chapter we have also analyzed a hybrid conjugate gradient method, namely, the method (8.1.2)–(8.1.3) with β_k given by (8.4.48). With the help of Theorem 8.4.5, the descent property and global convergence of the method are proved under the Wolfe conditions with $\sigma \leq 0.25$. We wonder whether the method (8.4.48) is also efficient in practice or we can obtain a more efficient conjugate gradient algorithm by combining (8.4.48) and (8.1.13). This question still remains under studies.

Finally, we would like to mention that although most of the analyses in this chapter use the Wolfe line search, they are also efficient for the strong Wolfe line search. As is known, there is still lack of a similar theory for conjugate gradient methods using the strong Wolfe line search. We also expect that this chapter will arouse more attention to the use of the Wolfe line search in conjugate gradient methods, even from the aspect of numerical computation.

Acknowledgements The author thanks Professor Yaxiang Yuan very much for his useful discussion and suggestions. This research was partly supported by the Chinese NSF grants 19801033, 10571171 and 10831006 and the CAS grant kjcx-yw-s7-03.

References

1. M. Al-Baali, Descent property and global convergence of the Fletcher-Reeves method with inexact linesearch, *IMA J. Numer. Anal.*, **5**, 121–124, 1985.
2. I. Bongartz, A. R. Conn, N. Gould and Ph. L. Toint, CUTE: constrained and unconstrained testing environment, *ACM Transactions on Mathematical Software*, **21**, 123–160, 1995.
3. 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), 345–358, 1999.

4. Y. H. Dai and Q. Ni, Testing different conjugate gradient methods for large-scale unconstrained optimization, *Journal of Computational Mathematics*, **21**(3), 311–320, 2003.
5. Y. H. Dai and Y. Yuan, Convergence properties of the Fletcher-Reeves method, *IMA J. Numer. Anal.*, **16**, 155–164, 1996.
6. Y. H. Dai and Y. Yuan, A class of globally convergent conjugate gradient methods, *Sciences in China (Series A)*, **46**(2), 251–261, 2003.
7. Y. H. Dai and Y. Yuan, A nonlinear conjugate gradient method with a strong global convergence property, *SIAM Journal on Optimization*, **10**(1), 177–182, 1999.
8. Y. H. Dai and Y. Yuan, An efficient hybrid conjugate gradient method for unconstrained optimization, *Annals of Operations Research*, **103**, 33–47, 2001.
9. R. Fletcher and C. M. Reeves, Function minimization by conjugate gradients, *Comput. J.*, **7**, 149–154, 1964.
10. J. C. Gilbert and J. Nocedal, Global convergence properties of conjugate gradient methods for optimization, *SIAM J. Optimization*, **2**, 21–42, 1992.
11. L. Grippo and S. Lucidi, A globally convergent version of the Polak-Ribière conjugate gradient method, *Math. Prog.*, **78**, 375–391, 1997.
12. M. R. Hestenes and E. Stiefel, Method of conjugate gradient for solving linear system, *J. Res. Nat. Bur. Stand.*, **49**, 409–436, 1952.
13. Y. F. Hu and C. Storey, Global convergence result for conjugate gradient methods, *J. Optim. Theory Appl.*, **71**, 399–405, 1991.
14. J. Nocedal, Theory of algorithms for unconstrained optimization, *Acta Numerica*, 199–242, 1991.
15. E. Polak and G. Ribière, Note sur la convergence de méthods de directions conjugées, *Revue Française d'Informatique et de Recherche Opiérationnelle*, **16**, 35–43, 1969.
16. B. T. Polyak, Conjugate gradient method in extremal problems, *USSR Comp. Math. and Math. Phys.*, **9**, 94–112, 1969.
17. M. J. D. Powell, Nonconvex minimization calculations and the conjugate gradient method, in: D. F. Griffiths, ed., *Numerical Analysis, Lecture Notes in Mathematics*, **1066**, 122–141, Springer-Verlag, Berlin, 1984.
18. M. J. D. Powell, Convergence properties of algorithms for nonlinear optimization, *SIAM Rev.*, **28**, 487–500, 1986.
19. D. Pu and W. Yu, On the convergence properties of the DFP algorithms, *Annals of Operations Research*, **24**, 175–184, 1990.
20. D. Touati-Ahmed and C. Storey, Efficient hybrid conjugate gradient techniques, *J. Optim. Theory Appl.*, **64**, 379–397, 1990.
21. P. Wolfe, Convergence conditions for ascent methods, *SIAM Rev.*, **11**, 226–235, 1969.
22. P. Wolfe, Convergence conditions for ascent methods II: some corrections, *SIAM Rev.*, **13**, 185–188, 1971.
23. G. Zoutendijk, Nonlinear programming, computational methods, in: J. Abadie, ed., *Integer and Nonlinear Programming*, 37–86, North-holland, Amsterdam, 1970.