

New Conjugacy Conditions and Related Nonlinear Conjugate Gradient Methods*

Y.-H. Dai¹ and L.-Z. Liao²

¹ State Key Laboratory of Scientific and Engineering Computing,
Institute of Computational Mathematics and Scientific/Engineering Computing,
Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences,
Box 2719, Beijing 100080, The People's Republic of China
dyh@lsec.cc.ac.cn

² Department of Mathematics, Hong Kong Baptist University,
Kowloon Tong, Kowloon, Hong Kong
liliao@hkbu.edu.hk

Abstract. Conjugate gradient methods are a class of important methods for unconstrained optimization, especially when the dimension is large. This paper proposes a new conjugacy condition, which considers an inexact line search scheme but reduces to the old one if the line search is exact. Based on the new conjugacy condition, two nonlinear conjugate gradient methods are constructed. Convergence analysis for the two methods is provided. Our numerical results show that one of the methods is very efficient for the given test problems.

Key Words. Unconstrained optimization, Conjugate gradient, Line search, Global convergence.

AMS Classification. 65K, 90C.

1. Introduction

Our problem is to minimize a function of n variables,

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

* This research was supported in part by the Chinese NSF Grant 19801033 and Grant FRG/97-98/II-42 of Hong Kong Baptist University.

where f is smooth and its gradient ∇f is available. The conjugate gradient method is very useful for solving (1.1) especially when n is large, and has the following form:

$$x_{k+1} = x_k + \alpha_k d_k, \quad (1.2)$$

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

where $\alpha_k > 0$ is a step-length, β_k is a scalar, and g_k denotes $\nabla f(x_k)$. In the case when f is a convex quadratic function,

$$f(x) = g^T x + \frac{1}{2} x^T H x, \quad (1.4)$$

and when α_k is the one-dimensional minimizer along d_k , i.e.,

$$\alpha_k = \arg \min_{\alpha > 0} f(x_k + \alpha d_k), \quad (1.5)$$

the conjugate gradient method is such that the conjugacy condition holds, namely,

$$d_i^T H d_j = 0, \quad \forall i \neq j. \quad (1.6)$$

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

$$y_{k-1} = g_k - g_{k-1}. \quad (1.7)$$

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

$$\alpha_{k-1}^{-1} d_k^T y_{k-1} = d_k^T \nabla^2 f(x_{k-1} + t \alpha_{k-1} d_{k-1}) d_{k-1}, \quad (1.8)$$

therefore it is reasonable to replace (1.6) with the following conjugacy condition:

$$d_k^T y_{k-1} = 0. \quad (1.9)$$

Multiplying y_{k-1} in (1.3) and using (1.9), we can deduce a formula for the scalar β_k :

$$\beta_k^{\text{HS}} = \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}}. \quad (1.10)$$

This is the so-called HS formula, which was given by Hestenes and Stiefel [4]. In practical computation, the HS method resembles the PRP method (see [8] and [9] for the PRP method); both methods are generally believed to be two of the most efficient conjugate gradient methods.

However, both the conjugacy conditions (1.6) and (1.9) depend on exact line searches. In practical computation, one normally carries out inexact line searches instead of exact line searches. In the case when $g_{k+1}^T d_k \neq 0$, the conjugacy conditions (1.6) and (1.9) may have some disadvantages (for instance, see [11]). Suppose we minimize the convex quadratic function (1.4) on a subspace spanned by a set of mutually conjugate directions $\{d_1, \dots, d_k\}$. Suppose that the line search along d_1 is not exact, that is, $\alpha_1 \neq \alpha_1^*$ where α_1^* is the step-length that solves (1.5). Then no matter what line searches are used in the subsequent iterations, it is always true that

$$(x_{k+1} - x^*)^T H (x_{k+1} - x^*) \geq (\alpha_1 - \alpha_1^*)^2 d_1^T H d_1, \quad (1.11)$$

where $x^* = -H^{-1}g$ is the minimum of the objective function (1.4). Hence we see that the error left in the current iteration cannot be eliminated in the subsequent iterations as long as the subsequent search directions are conjugate to the current search condition.

In [7], Nazareth develops a three-term-recurrence (TTR) algorithm, in which the search direction d_k is of the form

$$d_{k+1} = -y_k + \frac{y_k^T y_k}{d_k^T y_k} d_k + \frac{y_{k-1}^T y_k}{d_{k-1}^T y_{k-1}} d_{k-1}. \quad (1.12)$$

For convex quadratic functions, the search directions generated by the TTR algorithm are mutually conjugate even when line searches are inexact or the initial direction is not along the negative gradient. After n iterations, the algorithm implements a line search along the vector

$$-\sum_{k=1}^n \frac{g_{k+1}^T d_k}{y_k^T d_k} \alpha_k d_k \quad (1.13)$$

with the initial step being unity, and hence finite quadratic termination is retained. However, despite theoretical advantages on quadratics, the TTR algorithm has not been proved to be significantly superior to the PRP method. One possible reason is that if f is very nonlinear or the dimension n is large, then the coefficients $(g_{k+1}^T d_k / y_k^T d_k)$ in (1.13), which are computed in the previous n iterations and attempt to approximate the second-order information, may not provide accurate information.

In [11], Yuan and Stoer consider the search direction of the form

$$d_k = \mu_k g_k + \nu_k d_{k-1}, \quad (1.14)$$

and compute the scalars μ_k and ν_k by minimizing an approximate quadratic model in the two-dimensional subspace spanned by the current gradient and the previous search direction:

$$\min_{d \in \Omega_k} \varphi_k(d) = g_k^T d + \frac{1}{2} d^T H_k d, \quad (1.15)$$

where $\Omega_k = \text{span}\{g_k, d_{k-1}\}$. Then by approximating H_k through the memoryless BFGS update matrix or estimating the quantity $g_k^T H_k g_k$ suitably, they obtain satisfactory numerical results.

The main object of this paper is to find some new and efficient conjugate gradient methods with the search direction d_k having the simple form (1.3). For this purpose, we propose a new conjugacy condition, which considers an inexact line search scheme but reduces to (1.9) if the line search is exact. Based on the new conjugacy condition, we propose two new nonlinear conjugate gradient methods (see the next section). Convergence analysis for the two methods is presented in Section 3, and numerical results are reported in the last section.

2. New Conjugacy Condition and Its Resulting Formula for β_k

Our idea originated mainly from the following observation: For many unconstrained optimization methods, including quasi-Newton methods, the memoryless BFGS method,

and the limited memory BFGS method, the search direction d_k can be written in the form

$$d_k = -B_k g_k, \quad (2.1)$$

where B_k is some $n \times n$ symmetric and positive definite matrix satisfying the quasi-Newton equation:

$$B_k y_{k-1} = s_{k-1}, \quad (2.2)$$

where $s_{k-1} = \alpha_{k-1} d_{k-1}$ is the step. By (2.1) and (2.2), we have that

$$d_k^T y_{k-1} = -(B_k g_k)^T y_{k-1} = -g_k^T (B_k y_{k-1}) = -g_k^T s_{k-1}. \quad (2.3)$$

The above relation implies that (1.9) holds if the line search is exact since in this case $g_k^T s_{k-1} = 0$. However, practical numerical algorithms normally adopt inexact line searches instead of exact line searches. For this reason, it seems more reasonable to replace the conjugacy condition (1.9) with the condition

$$d_k^T y_{k-1} = -t g_k^T s_{k-1}, \quad (2.4)$$

where $t \geq 0$ is a scalar.

To ensure the search direction d_k in (1.3) satisfies the conjugacy condition (2.4), we only need to multiply (1.3) with y_{k-1} and use (2.4), yielding

$$\beta_k = \frac{g_k^T (y_{k-1} - t s_{k-1})}{d_{k-1}^T y_{k-1}}. \quad (2.5)$$

It is obvious that

$$\beta_k = \beta_k^{\text{HS}} - t \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}, \quad (2.6)$$

from which we see that formula (2.5) with $t \in [0, \infty)$ really defines a class of nonlinear conjugate gradient methods. For simplicity, we call the method defined by (1.2)–(1.3) with β_k from (2.5), method (2.5). Notice that if $d_{k-1}^T y_{k-1} > 0$, which is required by the (strong) Wolfe line search, we have that $\beta_k g_k^T d_{k-1} \leq \beta_k^{\text{HS}} g_k^T d_{k-1}$.

From (2.3), one reasonable value of t in (2.4) is

$$t = 1. \quad (2.7)$$

In this case, it follows from (2.5) that

$$\beta_k = \frac{g_k^T (y_{k-1} - s_{k-1})}{d_{k-1}^T y_{k-1}}. \quad (2.8)$$

Similarly, we call the method defined by (1.2)–(1.3) with β_k from (2.8), method (2.8). A remarkable property of formula (2.8) is that it is the solution of the following one-parameter quadratic model on β :

$$\min_{\beta} g_k^T d(\beta) + \frac{1}{2} d(\beta)^T H_k d(\beta), \quad (2.9)$$

where

$$d(\beta) := -g_k + \beta d_{k-1} \quad (2.10)$$

and the matrix $H_k = B_k^{-1}$ is such that $H_k s_{k-1} = y_{k-1}$. For any $t \geq 0$, denote d_k and \bar{d}_k to be the search directions given by method (2.5) and the HS method, respectively, namely,

$$d_k = -g_k + \beta_k d_{k-1} \quad (2.11)$$

and

$$\bar{d}_k = -g_k + \beta_k^{HS} d_{k-1}. \quad (2.12)$$

Assume that $g_k^T \bar{d}_k < 0$. Then from (2.11), (2.12), (2.6), and $d_{k-1}^T y_{k-1} > 0$, we also have that $g_k^T d_k < 0$. Thus if the direction generated by the HS method is descent, and if the line search provides the relation $d_{k-1}^T y_{k-1} > 0$, then the direction given by method (2.5) must also be a descent direction. Denote also α_k^* and $\bar{\alpha}_k$ to be the one-dimensional minimizers of f along the directions d_k and \bar{d}_k , respectively. We have the following lemma for quadratic functions.

Lemma 2.1. *Suppose that f is given in (1.4). Then we have that*

$$\begin{aligned} & f(x_k + \alpha_k^* d_k) - f(x_k + \bar{\alpha}_k \bar{d}_k) \\ &= \frac{(g_k^T d_{k-1})^2 t^2}{2(d_{k-1}^T H d_{k-1})(d_k^T H d_k)} \left[\left(\frac{2}{t} - \bar{\alpha}_k \right) g_k^T \bar{d}_k - \frac{(g_k^T s_{k-1})^2}{s_{k-1}^T y_{k-1}} \right]. \end{aligned} \quad (2.13)$$

Proof. By the definitions of α_k^* and $\bar{\alpha}_k$, it is easy to know that

$$\alpha_k^* = -\frac{g_k^T d_k}{d_k^T H d_k} \quad \text{and} \quad \bar{\alpha}_k = -\frac{g_k^T \bar{d}_k}{\bar{d}_k^T H \bar{d}_k}. \quad (2.14)$$

Hence we have that

$$\begin{aligned} & f(x_k + \alpha_k^* d_k) - f(x_k + \bar{\alpha}_k \bar{d}_k) \\ &= [f(x_k + \alpha_k^* d_k) - f(x_k)] - [f(x_k + \bar{\alpha}_k \bar{d}_k) - f(x_k)] \\ &= \frac{(g_k^T \bar{d}_k)^2}{2\bar{d}_k^T H \bar{d}_k} - \frac{(g_k^T d_k)^2}{2d_k^T H d_k} = \frac{\Gamma_k}{2(d_k^T H d_k)(\bar{d}_k^T H \bar{d}_k)}, \end{aligned} \quad (2.15)$$

where

$$\Gamma_k = (d_k^T H d_k)(g_k^T \bar{d}_k)^2 - (\bar{d}_k^T H \bar{d}_k)(g_k^T d_k)^2. \quad (2.16)$$

Define

$$\lambda_k := \frac{-g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}} = \frac{-g_k^T d_{k-1}}{d_{k-1}^T H d_{k-1}}. \quad (2.17)$$

It follows by (2.11), (2.12), (2.6), and (2.17) that

$$d_k = \bar{d}_k + t \lambda_k d_{k-1}. \quad (2.18)$$

In addition, since $\bar{d}_k^T y_{k-1} = 0$, we clearly have that

$$\bar{d}_k^T H d_{k-1} = 0. \quad (2.19)$$

Substituting (2.18) and (2.17) into (2.16) and using (2.19), we obtain

$$\begin{aligned} \Gamma_k &= t^2 \lambda_k^2 (d_{k-1}^T H d_{k-1}) (g_k^T \bar{d}_k)^2 - 2t \lambda_k (\bar{d}_k^T H \bar{d}_k) (g_k^T \bar{d}_k) (g_k^T d_{k-1}) \\ &\quad - t^2 \lambda_k^2 (\bar{d}_k^T H \bar{d}_k) (g_k^T d_{k-1})^2 \\ &= \frac{(\bar{d}_k^T H \bar{d}_k) (g_k^T d_{k-1})^2 t^2}{d_{k-1}^T H d_{k-1}} \left[\left(\frac{2}{t} + \frac{g_k^T \bar{d}_k}{\bar{d}_k^T H \bar{d}_k} \right) (g_k^T \bar{d}_k) - \frac{(g_k^T d_{k-1})^2}{d_{k-1}^T H d_{k-1}} \right]. \end{aligned}$$

Therefore (2.13) follows from the above relation, (2.14), and (2.15). \square

The above lemma indicates that if $\bar{\alpha}_k$ has been well estimated, and if

$$\tau_k := \bar{\alpha}_k g_k^T \bar{d}_k + \frac{(g_k^T s_{k-1})^2}{s_{k-1}^T y_{k-1}} < 0, \quad (2.20)$$

then a good choice of t is that

$$t = \frac{g_k^T \bar{d}_k}{\tau_k}, \quad (2.21)$$

which minimizes the value in (2.13). However, we should see that the information already achieved along \bar{d}_k is not enough for us to have a good estimate on $\bar{\alpha}_k$. In fact, how to estimate the initial step-lengths in conjugate gradient methods still remains under study. Thus not dealing with how to estimate $\bar{\alpha}_k$ in a good manner, we only simply regard that there exists some constant M such that

$$\bar{\alpha}_k \leq M, \quad \text{for all } k \geq 1. \quad (2.22)$$

In this case we can choose t to be some constant such that

$$t \leq \frac{2}{M}. \quad (2.23)$$

Then by this and (2.22), we have that

$$\frac{2}{t} - \bar{\alpha}_k \geq 0. \quad (2.24)$$

Equations (2.24), (2.13), and the condition $g_k^T \bar{d}_k < 0$ indicate that

$$f(x_k + \alpha_k^* d_k) \leq f(x_k + \bar{\alpha}_k \bar{d}_k). \quad (2.25)$$

In numerical experiments we obtained good results by setting $t = 0.1$ (see Section 4).

In the next section we prove the global convergence of method (2.5) for uniformly convex functions. For general functions, Powell [10] constructed an example showing that the PRP method may cycle without approaching any solution point if the step-length α_k is chosen to be the first local minimizer along d_k . Since method (2.5) reduces to the

PRP method in the case that $g_k^T d_{k-1} = 0$ holds, the example also shows that method (2.5) need not converge for general functions. Therefore, like Gilbert and Nocedal [3], who have proved the global convergence of the PRP method with the restriction that $\beta_k^{\text{PRP}} \geq 0$, we replace (2.5) by

$$\beta_k = \max \left\{ \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}}, 0 \right\} - t \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}, \quad (2.26)$$

and prove that such a modification of (2.5) is globally convergent for general functions. We call the method defined by (1.2)–(1.3) with β_k from (2.26), as method (2.26).

3. Convergence Analysis

Throughout this section we assume that

$$g_k \neq 0, \quad \text{for all } k \geq 1, \quad (3.1)$$

otherwise a stationary point is found. We also assume that there exists a constant $c \geq 0$ such that

$$g_k^T d_k < -c \|g_k\|^2, \quad \text{for all } k \geq 1. \quad (3.2)$$

The relation (3.2) implies that each search direction d_k is a descent direction. If the constant c is strictly greater than zero, which is needed in Theorem 3.6, then from (3.2) we know that the so-called sufficient descent condition holds.

We make the following basic assumptions on the objective function.

Assumption 3.1.

- (i) *The level set $\mathcal{L} = \{x \mid f(x) \leq f(x_1)\}$ is bounded, namely, there exists a constant $B > 0$ such that*

$$\|x\| \leq B, \quad \text{for all } x \in \mathcal{L}. \quad (3.3)$$

- (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*

$$\|\nabla f(x) - \nabla f(\bar{x})\| \leq L \|x - \bar{x}\|, \quad \text{for all } x, \bar{x} \in \mathcal{N}. \quad (3.4)$$

Under the above assumptions on f , there exists a constant $\bar{\gamma} \geq 0$ such that

$$\|\nabla f(x)\| \leq \bar{\gamma}, \quad \text{for all } x \in \mathcal{L}. \quad (3.5)$$

The step-length α_k in (1.2) is obtained by some line search scheme. In conjugate gradient methods, the strong Wolfe conditions, namely,

$$f(x_k + \alpha_k d_k) - f_k \leq \delta \alpha_k g_k^T d_k, \quad (3.6)$$

$$|g(x_k + \alpha_k d_k)^T d_k| \leq -\sigma g_k^T d_k, \quad (3.7)$$

where $0 < \delta < \sigma < 1$, are often imposed on the line search (in this case we call the line search the strong Wolfe line search).

For any conjugate gradient method with the strong Wolfe line search, we have the following general result, which is obtained in [2].

Lemma 3.2. *Suppose that Assumption 3.1 holds. Consider any conjugate gradient method in the form (1.2)–(1.3), where d_k is a descent direction and α_k is obtained by the strong Wolfe line search. If*

$$\sum_{k \geq 1} \frac{1}{\|d_k\|^2} = \infty, \quad (3.8)$$

we have that

$$\liminf_{k \rightarrow \infty} \|g_k\| = 0. \quad (3.9)$$

For uniformly convex functions, we can prove that the norm of d_k generated by method (2.5) is bounded above. Thus by Lemma 3.2 we immediately have the following result.

Theorem 3.3. *Suppose that Assumption 3.1 holds. Consider method (2.5), where d_k is a descent direction and α_k is obtained by the strong Wolfe line search. If there exists a constant $\mu > 0$ such that*

$$(\nabla f(x) - \nabla f(\bar{x}))^T (x - \bar{x}) \geq \mu \|x - \bar{x}\|^2, \quad \text{for all } x, \bar{x} \in \mathcal{L}, \quad (3.10)$$

we have that

$$\lim_{k \rightarrow \infty} g_k = 0. \quad (3.11)$$

Proof. It follows from (3.10) that f is a uniformly convex function in \mathcal{L} and

$$d_{k-1}^T y_{k-1} \geq \mu \alpha_{k-1} \|d_{k-1}\|^2. \quad (3.12)$$

By (1.3), (2.5), (3.4), (3.5), and (3.12), we have that

$$\|d_k\| \leq \|g_k\| + \frac{(L+t)\|g_k\| \|s_{k-1}\|}{\mu \alpha_{k-1} \|d_{k-1}\|^2} \|d_{k-1}\| \leq \mu^{-1}(L+t+\mu)\bar{\gamma}, \quad (3.13)$$

which implies the truth of (3.8). Therefore by Lemma 3.2 we have (3.9), which is equivalent to (3.11) for uniformly convex functions. \square

For the general function, because method (2.5) is the same as the PRP method in the case of exact line searches, it is known from the counterexample in [10] that method (2.5) may also cycle without approaching any solution point. Nevertheless, we will show that its modification, method (2.26), is globally convergent for general functions. The proof of this result follows the same one as in [3] for their method with $\beta_k = \max\{\beta_k^{\text{PRP}}, 0\}$. However, our result allows negative values for β_k since, by (2.26), we only have

$$\beta_k \geq -t \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}. \quad (3.14)$$

In addition, we should note that if the following relation holds,

$$\frac{g_k^T g_{k-1}}{\|g_k\|^2} \geq 1, \quad (3.15)$$

then method (2.26) is restarted with the direction

$$d_k = -g_k - t \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}} d_{k-1}. \quad (3.16)$$

Since this direction might include some second-order information, it is reasonable to expect that this direction would be better than the negative gradient $-g_k$ as a restart direction. With $t = 1$ in (3.16), a neural network model for nonconvex optimization has been constructed in [5]. The numerical results in [5] show that the new model is much better than the gradient model.

Lemma 3.4. *Suppose that Assumption 3.1 holds. Consider method (2.26), where d_k is a descent direction and α_k is obtained by the strong Wolfe line search. If there exists a constant $\gamma > 0$ such that*

$$\|g_k\| \geq \gamma, \quad \text{for all } k \geq 1, \quad (3.17)$$

then $d_k \neq 0$ and

$$\sum_{k \geq 2} \|u_k - u_{k-1}\|^2 < \infty, \quad (3.18)$$

where $u_k = d_k / \|d_k\|$.

Proof. First, note that $d_k \neq 0$, otherwise (3.2) is false. Therefore u_k is well defined. In addition, by relation (3.17) and Lemma 3.2, we have that

$$\sum_{k \geq 1} \frac{1}{\|d_k\|^2} < \infty, \quad (3.19)$$

for otherwise we have that (3.9) holds, contradicting (3.17). Now, we divide formula (2.26) for β_k into two parts as follows:

$$\beta_k^{(1)} = \max \left\{ \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}}, 0 \right\} \quad \text{and} \quad \beta_k^{(2)} = -t \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}, \quad (3.20)$$

and define

$$r_k := \frac{v_k}{\|d_k\|} \quad \text{and} \quad \delta_k := \frac{\beta_k^{(1)} \|d_{k-1}\|}{\|d_k\|}, \quad (3.21)$$

where

$$v_k = -g_k + \beta_k^{(2)} d_{k-1}. \quad (3.22)$$

Then by (1.3) we have for $k \geq 2$,

$$u_k = r_k + \delta_k u_{k-1}. \quad (3.23)$$

Using the identity $\|u_k\| = \|u_{k-1}\| = 1$ and (3.23), we can show

$$\|r_k\| = \|u_k - \delta_k u_{k-1}\| = \|\delta_k u_k - u_{k-1}\|. \quad (3.24)$$

Using the condition $\delta_k \geq 0$, the triangle inequality, and (3.24), we obtain

$$\begin{aligned} \|u_k - u_{k-1}\| &\leq \|(1 + \delta_k)u_k - (1 + \delta_k)u_{k-1}\| \\ &\leq \|u_k - \delta_k u_{k-1}\| + \|\delta_k u_k - u_{k-1}\| \\ &= 2\|r_k\|. \end{aligned} \quad (3.25)$$

On the other hand, the line search condition (3.7) gives

$$d_{k-1}^T y_{k-1} \geq (\sigma - 1)g_{k-1}^T d_{k-1}. \quad (3.26)$$

Equations (3.26), (3.7), and the assumption $d_{k-1}^T g_{k-1} < 0$ imply that

$$\left| \frac{g_k^T d_{k-1}}{d_{k-1}^T y_{k-1}} \right| \leq \frac{\sigma}{1 - \sigma}. \quad (3.27)$$

It follows from the definition of v_k , (3.27), (3.3), and (3.5) that

$$\|v_k\| \leq \|g_k\| + t \left| \frac{g_k^T d_{k-1}}{d_{k-1}^T y_{k-1}} \right| \|s_{k-1}\| \leq \bar{\gamma} + 2t\sigma(1 - \sigma)^{-1}B. \quad (3.28)$$

Therefore by (3.25), (3.21), (3.28), and (3.19), we know that (3.18) holds, which completes our proof. \square

Now we state a property of formula (2.26) for β_k , which is similar to but slightly different from Property (*) in [3]. Suppose that Assumption 3.1 and relations (3.7) and (3.17) hold. Then if (3.2) holds with some constant $c > 0$, we claim that there exist constants $b > 1$ and $\lambda > 0$ such that for all k ,

$$|\beta_k| \leq b, \quad (3.29)$$

and

$$\|s_{k-1}\| \leq \lambda \implies |\beta_k| \leq \frac{1}{b}. \quad (3.30)$$

In fact, by (3.26), (3.2), and (3.17), we have that

$$d_{k-1}^T y_{k-1} \geq (\sigma - 1)g_{k-1}^T d_{k-1} \geq (1 - \sigma)c\|g_{k-1}\|^2 \geq (1 - \sigma)c\gamma^2. \quad (3.31)$$

Using this, (3.3), (3.4), and (3.5), we obtain

$$|\beta_k| \leq \frac{(L + t)\|g_k\|\|s_{k-1}\|}{(1 - \sigma)c\gamma^2} \leq \frac{2(L + t)\bar{\gamma}B}{(1 - \sigma)c\gamma^2} =: b. \quad (3.32)$$

Note that b can be defined such that $b > 1$. Therefore we can say $b > 1$. As a result, we define

$$\lambda := \frac{(1 - \sigma)c\gamma^2}{b(L + t)\bar{\gamma}}. \quad (3.33)$$

We get from the first inequality in (3.32) that if $\|s_{k-1}\| \leq \lambda$, then

$$|\beta_k| \leq \frac{(L + t)\bar{\gamma}\lambda}{(1 - \sigma)c\gamma^2} = \frac{1}{b}. \quad (3.34)$$

Thus for the b and λ in (3.32) and (3.33), relations (3.29) and (3.30) hold.

Let N^* denote the set of positive integers. For $\lambda > 0$ and a positive integer Δ , denote

$$\mathcal{K}_{k,\Delta}^\lambda := \{i \in N^* : k \leq i \leq k + \Delta - 1, \|s_{i-1}\| > \lambda\}. \quad (3.35)$$

Let $|\mathcal{K}_{k,\Delta}^\lambda|$ denote the number of elements in $\mathcal{K}_{k,\Delta}^\lambda$. From the above property of formula (2.26), we can prove the following lemma.

Lemma 3.5. *Suppose that Assumption 3.1 holds. Consider method (2.26), where d_k satisfies condition (3.2) with $c > 0$, and α_k is obtained by the strong Wolfe line search. Then if (3.17) holds, there exists $\lambda > 0$ such that, for any $\Delta \in N^*$ and any index k_0 , there is an index $k > k_0$ such that*

$$|\mathcal{K}_{k,\Delta}^\lambda| > \frac{\Delta}{2}. \quad (3.36)$$

Proof. We proceed by contradiction. Suppose that for any $\lambda > 0$, there exist $\Delta \in N^*$ and k_0 such that

$$|\mathcal{K}_{k,\Delta}^\lambda| \leq \frac{\Delta}{2}, \quad \text{for all } k \geq k_0. \quad (3.37)$$

Let $b > 1$ and $\lambda > 0$ be given in (3.32) and (3.33). For $\lambda > 0$, we choose Δ and k_0 such that (3.37) holds. Then it follows from (3.29), (3.30), and (3.37) that

$$\prod_{k_0+i\Delta+1}^{k_0+(i+1)\Delta} |\beta_k| \leq b^{\Delta/2} \left(\frac{1}{b}\right)^{\Delta/2} = 1, \quad \text{for any } i \geq 0. \quad (3.38)$$

If $\beta_k = 0$, the direction in (1.3) reduces to $-g_k$. Then either the method gives the convergence relation (3.9) or we can take some x_k as a new initial point. Thus we assume without loss of generality that

$$\beta_k \neq 0, \quad \text{for all } k \geq 1. \quad (3.39)$$

It follows from (3.38) and (3.39) that

$$\prod_{j=2}^{k_0+i\Delta} \beta_j^{-2} \geq \prod_{j=2}^{k_0} \beta_j^{-2}, \quad \text{for any } i \geq 0, \quad (3.40)$$

which indicates

$$\sum_{k \geq 2} \prod_{j=2}^k \beta_j^{-2} = \infty. \quad (3.41)$$

By (3.41), it can be shown [1] that any conjugate gradient method with the strong Wolfe line search gives the convergence relation (3.9). In fact, it follows from (1.3) that for all $k \geq 2$,

$$d_k + g_k = \beta_k d_{k-1}. \quad (3.42)$$

Squaring both sides of (3.42), we can get

$$\begin{aligned} \|d_k\|^2 &= -2g_k^T d_k - \|g_k\|^2 + \beta_k^2 \|d_{k-1}\|^2 \\ &\leq \frac{(g_k^T d_k)^2}{\|g_k\|^2} + \beta_k^2 \|d_{k-1}\|^2. \end{aligned} \quad (3.43)$$

Hence we have

$$\begin{aligned} \|d_k\|^2 &\leq \left(1 - \frac{(g_k^T d_k)^2}{\|g_k\|^2 \|d_k\|^2}\right)^{-1} \beta_k^2 \|d_{k-1}\|^2 \\ &\leq \dots \\ &\leq \prod_{j=j_0}^k \left(1 - \frac{(g_j^T d_j)^2}{\|g_j\|^2 \|d_j\|^2}\right)^{-1} \left(\prod_{j=j_0}^k \beta_j^2\right) \|d_{j_0-1}\|^2, \end{aligned} \quad (3.44)$$

where $j_0 \geq 2$ is any integer. It is well known [12] that any descent method (1.2) with the Wolfe line search gives the relation

$$\sum_{k \geq 1} \frac{(g_k^T d_k)^2}{\|d_k\|^2} < \infty. \quad (3.45)$$

This together with (3.17) indicates that there exists some integer j_0 such that

$$\prod_{j \geq j_0} \left(1 - \frac{(g_j^T d_j)^2}{\|g_j\|^2 \|d_j\|^2}\right) \geq c_1, \quad \text{for some constant } c_1 > 0. \quad (3.46)$$

From (3.41), (3.44), and (3.46), we know that (3.8) holds. Thus by Lemma 3.2 we have (3.9). This gives a contradiction to (3.17). So (3.36) must be true. \square

Here we note that (3.41) is also a sufficient condition for the global convergence of any conjugate gradient method with the Wolfe line search, as is shown in [1]. We are now ready to prove the following convergence theorem for method (2.26).

Theorem 3.6. *Suppose that Assumption 3.1 holds. Consider method (2.26), where d_k satisfies condition (3.2) with $c > 0$, and α_k is obtained by the strong Wolfe line search. Then we have $\liminf_{k \rightarrow \infty} \|g_k\| = 0$.*

Proof. We proceed by contradiction. Since $\liminf_{k \rightarrow \infty} \|g_k\| > 0$, (3.17) must hold. Then the conditions of Lemmas 3.4 and 3.5 hold. Defining $u_i = d_i / \|d_i\|$, we have for any two indices l, k , with $l \geq k$,

$$\begin{aligned} x_l - x_{k-1} &= \sum_{i=k}^l \|s_{i-1}\| u_{i-1} \\ &= \sum_{i=k}^l \|s_{i-1}\| u_{k-1} + \sum_{i=k}^l \|s_{i-1}\| (u_{i-1} - u_{k-1}). \end{aligned} \quad (3.47)$$

This relation, the fact that $\|u_{k-1}\| = 1$, and (3.3) give that

$$\begin{aligned} \sum_{i=k}^l \|s_{i-1}\| &\leq \|x_l - x_{k-1}\| + \sum_{i=k}^l \|s_{i-1}\| \|u_{i-1} - u_{k-1}\| \\ &\leq 2B + \sum_{i=k}^l \|s_{i-1}\| \|u_{i-1} - u_{k-1}\|. \end{aligned} \quad (3.48)$$

Let $\lambda > 0$ be given by Lemma 3.5 and define $\Delta := \lceil 8B/\lambda \rceil$ to be the smallest integer not less than $8B/\lambda$. By Lemma 3.4, we can find an index $k_0 \geq 1$ such that

$$\sum_{i \geq k_0} \|u_i - u_{i-1}\|^2 \leq \frac{1}{4\Delta}. \quad (3.49)$$

With this Δ and k_0 , Lemma 3.5 gives an index $k \geq k_0$ such that

$$|\mathcal{K}_{k,\Delta}^\lambda| > \frac{\Delta}{2}. \quad (3.50)$$

Next, for any index $i \in [k, k + \Delta - 1]$, by the Cauchy–Schwartz inequality and (3.49),

$$\begin{aligned} \|u_i - u_{k-1}\| &\leq \sum_{j=k}^i \|u_j - u_{j-1}\| \\ &\leq (i - k + 1)^{1/2} \left(\sum_{j=k}^i \|u_j - u_{j-1}\|^2 \right)^{1/2} \\ &\leq \Delta^{1/2} \left(\frac{1}{4\Delta} \right)^{1/2} = \frac{1}{2}. \end{aligned} \quad (3.51)$$

From this relation, (3.50), and taking $l = k + \Delta - 1$ in (3.48), we get that

$$2B \geq \frac{1}{2} \sum_{i=k}^{k+\Delta-1} \|s_{i-1}\| > \frac{\lambda}{2} |\mathcal{K}_{k,\Delta}^\lambda| > \frac{\lambda \Delta}{4}. \quad (3.52)$$

Thus $\Delta < 8B/\lambda$, which contradicts the definition of Δ . Therefore, the theorem is true. \square

Table 4.1. Numerical comparisons.

P	Name	n	HS method	Method (2.8)	Method (2.26)
24	Penalty 2	20	2092/6528/3158	1483/4521/2208	476/1486/796
		40	798/2398/1104	1326/4085/1838	342/1081/520
25	Variably dimensioned	20	<i>Failed</i>	9/42/17	5/28/11
		50	11/67/28*	14/72/27	12/55/24
35	Chebyquad	20	142/450/163	116/380/147	158/504/185
		50	348/1162/418	395/1312/466	349/1156/420
30	Broyden tridiagonal	50	32/103/38	32/103/38	32/103/38
		500	34/109/41	34/109/41	34/109/41
31	Broyden banded	50	39/152/70	38/146/66	30/120/56
		500	36/131/56	37/138/60	23/79/32
22	Extended Powell	100	126/373/170	208/629/277	98/292/134
		1,000	151/448/210	265/793/349	149/433/199
26	Trigonometric	100	54/101/100	58/106/104	54/105/105
		1,000	55/98/98	60/112/111	55/100/100
21	Extended Rosenbrock	1,000	24/114/60	23/119/68	23/96/55
		10,000	25/117/61	23/119/68	23/96/55
23	Penalty 1	1,000	28/98/74	29/93/69	24/77/57
		10,000	72/278/176	36/138/81	38/136/98

4. Numerical Results

We tested the HS method, method (2.8), and method (2.26) on an SGI Indigo workstation. Our line search subroutine computes α_k such that the strong Wolfe conditions (3.6)–(3.7) hold with $\delta = 0.01$ and $\sigma = 0.1$. The initial value of α_k is always set to 1. Although our line search cannot always ensure the descent property of d_k for all three methods, uphill search directions seldom occur in our numerical experiments. In the case when an uphill search direction does occur, we restart the algorithm by setting $d_k = -g_k$. For method (2.26), $t = 0.1$ is selected.

The test problems are drawn from [6]. The numerical results of our tests are reported in Table 4.1.

The first column “P” and the second column represent the problem number and problem name in [6], respectively. Each problem was tested with two different values of n ranging from $n = 20$ to $n = 10,000$. The numerical results are given in the form of I/F/G, where I, F, G denote the numbers of iterations, function evaluations, and gradient evaluations, respectively. The stopping condition is

$$\|g_k\| \leq 10^{-6}. \quad (4.1)$$

The iteration is also terminated if the number of function evaluations exceed 9999, but we find that this never occurs. We also terminate the iteration if the function value improvement is too small. More exactly, iterations are terminated whenever

$$\frac{f(x_k) - f(x_{k+1})}{1 + |f(x_k)|} \leq 10^{-16}. \quad (4.2)$$

In this case, we use a superscript “*” to show that the iteration is terminated due to (4.2) but (4.1) is not satisfied. In addition, we write “Failed” if d_k is so large that a numerical overflow occurs while the method tries to compute $f(x_k + d_k)$.

From Table 4.1, we see that for some problems method (2.8) really performs much better than the HS method, for example Problem 25 with $n = 20$ and Problem 23 with $n = 10,000$; whereas for some other problems, method (2.8) performs worse than the HS method, for example Problem 24 with $n = 40$ and Problem 22 with $n = 1000$. On the whole, the method (2.8) and the HS method perform quite similarly for the given test problems.

Comparing method (2.26) with the HS method, we find that there are quite a number of test problems for which method (2.26) outperforms the HS method prominently; whereas for the rest of these problems, method (2.26) and the HS method perform very similarly. Therefore we could say that method (2.26) is much better than the HS method.

Acknowledgment

The authors thank an anonymous referee for his useful comments and suggestions on an early version of this paper.

References

1. Dai YH (1999) Convergence properties of nonlinear conjugate gradient methods (II). Research Report AMSS-1999-082, Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences (submitted to SIAM J Optim)
2. Dai YH, Han JY, Liu GH, Sun DF, Yin HX, Yuan Y (1999) Convergence properties of nonlinear conjugate gradient methods. SIAM J Optim 10(2):348–358
3. Gilbert JC, Nocedal J (1992) Global convergence properties of conjugate gradient methods for optimization. SIAM J Optim 2(1):21–42
4. Hestenes MR, Stiefel EL (1952) Methods of conjugate gradients for solving linear systems. J Res Nat Bur Standards 49(5):409–436
5. Liao LZ, Dai YH (1999) A time delay neural network model for unconstrained nonconvex optimization (submitted)
6. Moré JJ, Garbow BS, Hillstom KE (1981) Testing unconstrained optimization software. ACM Trans Math Software 7:17–41
7. Nazareth JL (1977) A conjugate direction algorithm without line searches. J Optim Theory Appl 23(3):373–387
8. Polak E, Ribière G (1969) Note sur la convergence de méthodes directions conjuguées. Rev Francaise Inform Rech Opér 16:35–43
9. Polyak BT (1969) The conjugate gradient method in extreme problems. USSR Comp Math Math Phys 9:94–112
10. Powell MJD (1984) Nonconvex minimization calculations and the conjugate gradient method. In: Lecture Notes in Mathematics 1066. Springer-Verlag, Berlin, pp 122–141
11. Yuan Y, Stoer J (1995) A subspace study on conjugate gradient algorithms. Z Angew Math Mech 75(11):69–77
12. Zoutendijk G (1970) Nonlinear programming, computational methods. In: Integer and Nonlinear Programming (Abadie J ed). North-Holland, Amsterdam, pp 37–86

Accepted 15 September 2000. Online publication 27 November 2000.