Annals of Operations Research 103, 33–47, 2001


# An Efficient Hybrid Conjugate Gradient Method for Unconstrained Optimization *

Y.H. DAI ** and Y. YUAN {dyh, yyx}@lsec.cc.ac.cn
*State Key Laboratory of Scientific and Engineering Computing, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, P.O. Box 2719, Beijing 100080, P.R. China*

**Abstract.** Recently, we propose a nonlinear conjugate gradient method, which produces a descent search direction at every iteration and converges globally provided that the line search satisfies the weak Wolfe conditions. In this paper, we will study methods related to the new nonlinear conjugate gradient method. Specifically, if the size of the scalar $\beta_k$ with respect to the one in the new method belongs to some interval, then the corresponding methods are proved to be globally convergent; otherwise, we are able to construct a convex quadratic example showing that the methods need not converge. Numerical experiments are made for two combinations of the new method and the Hestenes–Stiefel conjugate gradient method. The initial results show that, one of the hybrid methods is especially efficient for the given test problems.

**Keywords:** unconstrained optimization, conjugate gradient method, line search, descent property, global convergence

**AMS subject classification:** 49M37, 65K05, 90C30

## 1. Introduction

In [6], we propose a nonlinear conjugate gradient method. An important property of the method is that, it produces a descent search direction at every iteration and converges globally provided that the line search satisfies the weak Wolfe conditions. The purpose of this paper is to study some methods related to the new nonlinear conjugate gradient method and find efficient algorithms among them.

Consider the following unconstrained optimization problem,

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

where $f$ is smooth and its gradient $g$ is available. Conjugate gradient methods are very efficient for solving (1.1) especially when the dimension $n$ is large, and have the following form

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

* Research partly supported by Chinese National Natural Sciences Foundation Grants 19731010 and 19801033.
** Corresponding author.

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

where $g_k = -\nabla f(x_k)$, $\alpha_k > 0$ is a steplength obtained by a line search, and $\beta_k$ is a scalar. The formula for $\beta_k$ should be so chosen that the method reduces to the linear conjugate gradient method in the case when $f$ is a strictly convex quadratic and the line search is exact. Well-known formulae for $\beta_k$ are called the Fletcher–Reeves (FR), conjugate descent (CD), Polak–Ribière–Polyak (PRP), and Hestenes–Stiefel (HS) formulae (see [7,8,10,14,15]), and are given by

$$\beta_k^{FR} = \frac{\|g_k\|^2}{\|g_{k-1}\|^2}, \tag{1.4}$$

$$\beta_k^{PRP} = \frac{g_k^{\mathrm{T}} y_{k-1}}{\|g_{k-1}\|^2}, \tag{1.5}$$

$$\beta_k^{CD} = -\frac{\|g_k\|^2}{d_{k-1}^{\mathrm{T}} g_{k-1}}, \tag{1.6}$$

$$\beta_k^{HS} = \frac{g_k^{\mathrm{T}} y_{k-1}}{d_{k-1}^{\mathrm{T}} y_{k-1}}, \tag{1.7}$$

respectively, where $y_{k-1} = g_k - g_{k-1}$ and $\|\cdot\|$ means the Euclidean norm.

In the convergence analyses and implementations of conjugate gradient methods, one often requires the line search to be exact or satisfy the strong Wolfe conditions, namely,

$$f(x_k) - f(x_k + \alpha_k d_k) \geqslant -\delta \alpha_k g_k^{\mathrm{T}} d_k, \tag{1.8}$$

$$\left| g(x_k + \alpha_k d_k)^{\mathrm{T}} d_k \right| \leqslant -\sigma g_k^{\mathrm{T}} d_k, \tag{1.9}$$

where $0 < \delta < \sigma < 1$ (for the latter, we call the line search as the strong Wolfe line search). For example, the FR method is shown to be globally convergent under strong Wolfe line searches with $\sigma \leqslant 1/2$ [1,3,12]. If $\sigma > 1/2$, the FR method may fail due to producing an ascent search direction [3]. The PRP method with exact line searches may cycle without approaching any stationary point, see Powell's counter-example [17].

Recently, in [6], we propose a nonlinear conjugate gradient method, in which

$$\beta_k^{DY} = \frac{\|g_k\|^2}{d_{k-1}^{\mathrm{T}} y_{k-1}}. \tag{1.10}$$

The method is proved to produce a descent search direction at every iteration and converge globally provided that the line search satisfies the weak Wolfe conditions, namely, (1.8) and

$$g(x_k + \alpha_k d_k)^{\mathrm{T}} d_k \geqslant \sigma g_k^{\mathrm{T}} d_k, \tag{1.11}$$

where also $0 < \delta < \sigma < 1$ (in this case, we call the line search as the weak Wolfe line search). Other nice properties of the method can be found in [2,5]. In this paper, we call the method defined by (1.2), (1.3) where $\beta_k$ is computed by (1.10) as the method (1.10).

Although one would be satisfied with its global convergence properties, the FR method sometimes performs much worse than the PRP method in real computations. Powell [16] observed one major evidence for the inefficient behaviors of the FR method with exact line searches; that is, if a very small step is generated away from the solution, then the subsequent steps will be likely to be also very short. In contrast, the PRP method with exact line searches could recover from this situation. Gilbert and Nocedal [9] extended Powell's analyses to the case of inexact line searches.

For the method (1.10) with strong Wolfe line searches, we can deduce the same drawback as the FR method. In fact, by multiplying (1.3) with $g_k$ and using (1.10), we have that

$$g_k^{\mathrm{T}} d_k = \frac{g_{k-1}^{\mathrm{T}} d_{k-1}}{d_{k-1}^{\mathrm{T}} y_{k-1}} \|g_k\|^2, \tag{1.12}$$

which with (1.10) gives an equivalent formula to (1.10):

$$\beta_k^{DY} = \frac{g_k^{\mathrm{T}} d_k}{g_{k-1}^{\mathrm{T}} d_{k-1}}. \tag{1.13}$$

On the other hand, writing (1.3) as $d_k + g_k = \beta_k d_{k-1}$ and squaring it, we get

$$\|d_k\|^2 = -\|g_k\|^2 - 2g_k^{\mathrm{T}} d_k + \beta_k^2 \|d_{k-1}\|^2. \tag{1.14}$$

Dividing (1.14) by $(g_k^{\mathrm{T}} d_k)^2$ and substituting (1.12) and (1.13), we can obtain

$$\frac{\|d_k\|^2}{(g_k^{\mathrm{T}} d_k)^2} = \frac{\|d_{k-1}\|^2}{(g_{k-1}^{\mathrm{T}} d_{k-1})^2} + \frac{1 - l_{k-1}^2}{\|g_k\|^2}, \tag{1.15}$$

where $l_{k-1}$ is given by

$$l_{k-1} = \frac{g_k^{\mathrm{T}} d_{k-1}}{g_{k-1}^{\mathrm{T}} d_{k-1}}. \tag{1.16}$$

Denote $\theta_k$ to be the angle between $-g_k$ and $d_k$, namely,

$$\cos\theta_k = \frac{-g_k^{\mathrm{T}} d_k}{\|g_k\| \|d_k\|}. \tag{1.17}$$

Then it follows from (1.15) that

$$\cos^{-2}\theta_k = \frac{\|g_k\|^2}{\|g_{k-1}\|^2} \cos^{-2}\theta_{k-1} + \left(1 - l_{k-1}^2\right). \tag{1.18}$$

In case of strong Wolfe line searches, we have by (1.9) that $|l_{k-1}| \leqslant \sigma$. Suppose that at $(k-1)$th iteration an unfortunate search direction is generated, such that $\cos\theta_{k-1} \approx 0$, and that $x_k \approx x_{k-1}$. Then $\|g_k\| \approx \|g_{k-1}\|$. It follows from this, (1.18) and $|l_{k-1}| \leqslant \sigma$ that $\cos\theta_k \approx 0$. The argument can therefore start all over again.

To combine the good numerical performances of the PRP method and the nice global convergence properties of the FR method, Touati-Ahmed and Storey [18] extended Al-Baali's result [1] to any method (1.2), (1.3) with $\beta_k$ satisfying

$$\beta_k \in \left[0, \beta_k^{FR}\right]. \tag{1.19}$$

Gilbert and Nocedal [9] extended this result to the case that

$$\beta_k \in \left[-\beta_k^{FR}, \beta_k^{FR}\right]. \tag{1.20}$$

Their studies suggested, for example, the following hybrid conjugate gradient method

$$\beta_k = \max\left\{-\beta_k^{FR}, \min\left\{\beta_k^{PRP}, \beta_k^{FR}\right\}\right\}. \tag{1.21}$$

The hybrid method (1.21) has the same advantage of avoiding the propensity of short steps as the PRP method. In real computations, however, the method (1.21) does not perform better than the PRP method (see, for example, [9]). Therefore it is doubtful whether the global convergence study will yield a better conjugate gradient algorithm.

In this paper, methods related to the method (1.10) are carefully studied and some encouraging numerical results are presented. Denote $r_k$ to be the size of $\beta_k$ with respect to $\beta_k^{DY}$, namely,

$$r_k = \frac{\beta_k}{\beta_k^{DY}}. \tag{1.22}$$

We prove that any method (1.2), (1.3) with the weak Wolfe line search produces a descent search direction at every iteration and converges globally if the scalar $\beta_k$ is such that

$$-c \leqslant r_k \leqslant 1, \tag{1.23}$$

where $c = (1-\sigma)/(1+\sigma) > 0$. This result will be established in section 2. A convex quadratic example is given in section 3, showing that the bounds of $r_k$ in (1.23) can not be relaxed in some sense. Preliminary numerical results of two combinations of the method (1.10) and the HS method are reported in section 4, where $\beta_k$ is given by

$$\beta_k = \max\left\{-c\beta_k^{DY}, \min\left\{\beta_k^{HS}, \beta_k^{DY}\right\}\right\} \tag{1.24}$$

and

$$\beta_k = \max\left\{0, \min\left\{\beta_k^{HS}, \beta_k^{DY}\right\}\right\}, \tag{1.25}$$

respectively. The results show that the two hybrid conjugate gradient methods, even the hybrid method (1.25), perform better than the PRP method. Conclusions and discussions are made in the last section.

## 2. Methods related to the method (1.10)

We give the following basic assumption on the objective function.

**Assumption 2.1.**

(1) $f$ is bounded below in the level set $\mathcal{L} = \{x \in \mathbb{R}^n\colon\ f(x) \leqslant f(x_1)\}$;

(2) In a neighborhood $\mathcal{N}$ of $\mathcal{L}$, $f$ is differentiable and its gradient $g$ is Lipschitz continuous, namely, there exists a constant $L > 0$ such that

$$\left\|\nabla f(x) - \nabla f(y)\right\| \leqslant L\|x - y\| \quad \text{for any } x, y \in \mathcal{N}. \tag{2.1}$$

Under assumption 2.1 on $f$, we give a useful lemma, which was obtained by Zoutendijk [21] and Wolfe [19,20].

**Lemma 2.2.** Suppose that $x_1$ is a starting point for which assumption 2.1 holds. Consider any method in the form (1.2), where $d_k$ is a descent direction and $\alpha_k$ satisfies the weak Wolfe conditions (1.8) and (1.11). Then we have that

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

*Proof.* From (1.11) we have that

$$(g_{k+1} - g_k)^{\mathrm{T}} d_k \geqslant (\sigma - 1) g_k^{\mathrm{T}} d_k. \tag{2.3}$$

Besides it, the Lipschitz condition (2.1) gives

$$(g_{k+1} - g_k)^{\mathrm{T}} d_k \leqslant \alpha_k L \|d_k\|^2. \tag{2.4}$$

Combing these two relations, we obtain

$$\alpha_k \geqslant \frac{\sigma - 1}{L} \cdot \frac{g_k^{\mathrm{T}} d_k}{\|d_k\|^2}, \tag{2.5}$$

which with (1.8) implies that

$$f_k - f_{k+1} \geqslant c_1 \frac{(g_k^{\mathrm{T}} d_k)^2}{\|d_k\|^2}, \tag{2.6}$$

where $c_1 = \delta(1 - \sigma)/L$. Thus

$$f_1 - f_{k+1} \geqslant c_1 \sum_{i=1}^{k} \frac{(g_i^{\mathrm{T}} d_i)^2}{\|d_i\|^2}. \tag{2.7}$$

Noting that $f$ is bounded below, (2.2) holds. □

For methods related to the method (1.10). We have the following result, where $r_k$ is given in (1.22) and $c$ is a positive constant given by

$$c = \frac{1 - \sigma}{1 + \sigma}. \tag{2.8}$$

**Theorem 2.3.** Suppose that $x_1$ is a starting point for which assumption 2.1 holds. Consider the method (1.2), (1.3), where $\alpha_k$ is computed by the weak Wolfe line search, and $\beta_k$ is such that

$$r_k \in [-c, 1]. \tag{2.9}$$

Then if $g_k \neq 0$ for all $k \geqslant 1$, we have that

$$g_k^{\mathrm{T}} d_k < 0 \quad \text{for all } k \geqslant 1. \tag{2.10}$$

Further, the method converges in the sense that

$$\liminf_{k\to\infty} \|g_k\| = 0. \tag{2.11}$$

*Proof.* Multiplying (1.3) with $g_k$ and noting that $\beta_k = r_k \beta_k^{DY}$, we have that

$$g_k^{\mathrm{T}} d_k = \frac{g_{k-1}^{\mathrm{T}} d_{k-1} + (r_k - 1) g_k^{\mathrm{T}} d_{k-1}}{d_{k-1}^{\mathrm{T}} y_{k-1}} \|g_k\|^2. \tag{2.12}$$

From this, (2.9) and the formula for $\beta_k^{DY}$, we get

$$\beta_k = r_k \beta_k^{DY} = \frac{r_k g_k^{\mathrm{T}} d_k}{g_{k-1}^{\mathrm{T}} d_{k-1} + (r_k - 1) g_k^{\mathrm{T}} d_{k-1}} = \xi_k \frac{g_k^{\mathrm{T}} d_k}{g_{k-1}^{\mathrm{T}} d_{k-1}}, \tag{2.13}$$

where

$$\xi_k = \frac{r_k}{1 + (r_k - 1) l_{k-1}}, \tag{2.14}$$

and $l_{k-1}$ is given in (1.16). At the same time, if we define

$$\zeta_k = \frac{1 + (r_k - 1) l_{k-1}}{l_{k-1} - 1}, \tag{2.15}$$

it follows from (2.12) and (1.16) that

$$g_k^{\mathrm{T}} d_k = \zeta_k \|g_k\|^2. \tag{2.16}$$

Since $d_1 = -g_1$, it is obvious that $g_1^{\mathrm{T}} d_1 < 0$. Assume that $g_{k-1}^{\mathrm{T}} d_{k-1} < 0$. Then we have by (1.11) with $k$ replaced by $k - 1$ that

$$l_{k-1} \leqslant \sigma. \tag{2.17}$$

It follows from this and (2.9) that

$$1 + (r_k - 1) l_{k-1} \geqslant 1 + \left( -\frac{1-\sigma}{1+\sigma} - 1 \right) \sigma = \frac{1-\sigma}{1+\sigma}. \tag{2.18}$$

The above relation, (2.17), (2.16) and the fact that $\sigma < 1$ imply that $g_k^{\mathrm{T}} d_k < 0$. Thus by induction, (2.10) holds.

We now prove (2.11) by contradiction and assume that there exists some constant $\gamma > 0$ such that

$$\|g_k\| \geqslant \gamma \quad \text{for all } k \geqslant 1. \tag{2.19}$$

Since $d_k + g_k = \beta_k d_{k-1}$, we have that

$$\|d_k\|^2 = \beta_k^2 \|d_{k-1}\|^2 - 2g_k^{\mathrm{T}} d_k - \|g_k\|^2. \tag{2.20}$$

Dividing both sides of (2.20) by $(g_k^{\mathrm{T}} d_k)^2$ and using (2.13) and (2.16), we obtain

$$\begin{aligned}\frac{\|d_k\|^2}{(g_k^{\mathrm{T}} d_k)^2} &= \xi_k^2 \frac{\|d_{k-1}\|^2}{(g_{k-1}^{\mathrm{T}} d_{k-1})^2} - \frac{1}{\|g_k\|^2}\left(\frac{2}{\zeta_k} + \frac{1}{\zeta_k^2}\right) \\ &= \xi_k^2 \frac{\|d_{k-1}\|^2}{(g_{k-1}^{\mathrm{T}} d_{k-1})^2} + \frac{1}{\|g_k\|^2}\left[1 - \left(1 + \frac{1}{\zeta_k}\right)^2\right].\end{aligned} \tag{2.21}$$

(2.9) and (2.18) clearly imply that

$$1 + (r_k - 1)l_{k-1} \geqslant -r_k. \tag{2.22}$$

In addition, since $l_{k-1} < 1$ and $r_k \leqslant 1$, we have that $(1-r_k)(1-l_{k-1}) \geqslant 0$, or equivalently

$$1 + (r_k - 1)l_{k-1} \geqslant r_k. \tag{2.23}$$

Thus we have that

$$\left|1 + (r_k - 1)l_{k-1}\right| \geqslant |r_k|, \tag{2.24}$$

which with (2.14) yields

$$|\xi_k| \leqslant 1. \tag{2.25}$$

By (2.25) and (2.21), we obtain

$$\frac{\|d_k\|^2}{(g_k^{\mathrm{T}} d_k)^2} \leqslant \frac{\|d_{k-1}\|^2}{(g_{k-1}^{\mathrm{T}} d_{k-1})^2} + \frac{1}{\|g_k\|^2}. \tag{2.26}$$

Using (2.26) recursively and noting that $\|d_1\|^2 = -g_1^{\mathrm{T}} d_1 = \|g_1\|^2$,

$$\frac{\|d_k\|^2}{(g_k^{\mathrm{T}} d_k)^2} \leqslant \sum_{i=1}^{k} \frac{1}{\|g_i\|^2}. \tag{2.27}$$

Then we get from this and (2.19) that

$$\frac{(g_k^{\mathrm{T}} d_k)^2}{\|d_k\|^2} \geqslant \frac{\gamma^2}{k}, \tag{2.28}$$

which indicates

$$\sum_{k \geqslant 1} \frac{(g_k^{\mathrm{T}} d_k)^2}{\|d_k\|^2} = +\infty. \tag{2.29}$$

This contradicts the Zoutendijk condition (2.2). Therefore (2.11) holds. □

By theorem 2.3, we can immediately give the following convergence result for the CD method, which was first obtained in [4].

**Corollary 2.4.** Suppose that $x_1$ is a starting point for which assumption 2.1 holds. Consider the CD method (1.2), (1.3) and (1.6), where $\alpha_k$ satisfies the line search conditions (1.8) and

$$\sigma g_k^{\mathrm{T}} d_k \leqslant g(x_k + \alpha_k d_k)^{\mathrm{T}} d_k \leqslant 0. \tag{2.30}$$

Then we have either $g_k = 0$ for some finite $k$ or $\liminf_{k\to\infty} \|g_k\| = 0$.

*Proof.* It follows from (2.30) and the definitions of $\beta_k^{CD}$ and $\beta_k^{DY}$ that

$$0 \leqslant \beta_k^{CD} \leqslant \beta_k^{DY}. \tag{2.31}$$

Therefore the statement follows theorem 2.3. □

## 3. Optimality of the bounds in (2.9)

In this section, we will consider whether the bounds in (2.9) of $r_k$ can be relaxed. For any constant $c > 1$, Dai and Yuan [4] constructed an example showing that the method (1.2), (1.3) where

$$\beta_k = c\beta_k^{FR} \tag{3.1}$$

needs not converge even if the line search is exact. Since $\beta_k^{DY} = \beta_k^{FR}$ in case of exact line searches, we know that the example in [4] also applies to the method (1.10). Hence the upper bound 1 of $r_k$ in (2.9) cannot be relaxed. In the following, we will show by a convex quadratic example that the lower bound $(\sigma - 1)/(1 + \sigma)$ can not be relaxed, either.

Consider the following quadratic function with the unit Hessian:

$$f(x) = \frac{1}{2}x^{\mathrm{T}}x, \quad x \in \mathbb{R}^n. \tag{3.2}$$

We will prove that for any constant $r$ satisfying

$$r < -\frac{1-\sigma}{1+\sigma}, \tag{3.3}$$

the method (1.2), (1.3) with strong Wolfe line searches and with

$$\beta_k = r\beta_k^{DY} \tag{3.4}$$

may fail to reach the unique minimizer $x^* = 0$ of the function in (3.2).

In fact, for any $r$ satisfying (3.3), let $\widehat{\sigma}$ be the largest number in $(0, \sigma]$ such that

$$1 + r\frac{\widehat{\sigma}}{1-\widehat{\sigma}} \geqslant \frac{1}{2}. \tag{3.5}$$

Our definition of $\widehat{\sigma}$ implies that

$$-\frac{1-\widehat{\sigma}}{2\widehat{\sigma}} \leqslant r < -\frac{1-\widehat{\sigma}}{1+\widehat{\sigma}}. \tag{3.6}$$

To satisfy

$$\nabla f(x_k + \alpha_k d_k)^{\mathrm{T}} d_k = \widehat{\sigma} g_k^{\mathrm{T}} d_k, \tag{3.7}$$

we choose the steplength $\alpha_k$ as follows:

$$\alpha_k = (\widehat{\sigma} - 1) \frac{g_k^{\mathrm{T}} d_k}{\|d_k\|^2}. \tag{3.8}$$

In this case, it is easy to show that

$$f(x_k) - f(x_k + \alpha_k d_k) = \frac{1-\widehat{\sigma}^2}{2} \frac{(g_k^{\mathrm{T}} d_k)^2}{\|d_k\|^2}. \tag{3.9}$$

Relations (3.8) and (3.9) imply that

$$f(x_k) - f(x_k + \alpha_k d_k) = -\frac{1+\widehat{\sigma}}{2} \alpha_k g_k^{\mathrm{T}} d_k. \tag{3.10}$$

Thus if $\delta < 1/2$, the steplength $\alpha_k$ in (3.8) satisfies the strong Wolfe conditions (1.8) and (1.9). In addition, we have from (3.2) that

$$f(x_k) = \frac{1}{2} \|g_k\|^2, \tag{3.11}$$

which with (3.9) gives

$$\|g_{k+1}\|^2 = \|g_k\|^2 - (1-\widehat{\sigma}^2) \frac{(g_k^{\mathrm{T}} d_k)^2}{\|d_k\|^2}. \tag{3.12}$$

Summing this expression, we obtain

$$\|g_{k+1}\|^2 = \|g_1\|^2 - (1-\widehat{\sigma}^2) \sum_{i=1}^{k} \frac{(g_i^{\mathrm{T}} d_i)^2}{\|d_i\|^2}. \tag{3.13}$$

Again, we define $l_k$, $\xi_k$, $\zeta_k$ by (1.16), (2.14) and (2.15), respectively. It follows from (3.7) that for all $k \geqslant 1$,

$$l_k = \widehat{\sigma}, \tag{3.14}$$

$$\xi_k = \frac{r}{1 + (r-1)\widehat{\sigma}}, \tag{3.15}$$

$$\zeta_k = \frac{1 + (r-1)\widehat{\sigma}}{\widehat{\sigma} - 1}. \tag{3.16}$$

Since the values of $l_k$, $\xi_k$ and $\zeta_k$ are independent of $k$, we now remove their subscripts and only use $l$, $\xi$ and $\zeta$ to denote them. Using (3.6), it is easy to show that

$$-\frac{1}{\widehat{\sigma}} \leqslant \xi < -1 \tag{3.17}$$

and

$$\left(1+\frac{1}{\zeta}\right)^2 \leqslant 1. \tag{3.18}$$

Applying (3.18) in (2.21), we get that

$$\frac{\|d_k\|^2}{(g_k^{\mathrm{T}} d_k)^2} \geqslant \xi^2 \frac{\|d_{k-1}\|^2}{(g_{k-1}^{\mathrm{T}} d_{k-1})^2}, \tag{3.19}$$

from which we can obtain

$$\frac{(g_k^{\mathrm{T}} d_k)^2}{\|d_k\|^2} \leqslant \xi^{-2(k-1)} \frac{(g_1^{\mathrm{T}} d_1)^2}{\|d_1\|^2}. \tag{3.20}$$

This and (3.13) imply that

$$\begin{aligned}
\|g_{k+1}\|^2 &\geqslant \|g_1\|^2 - \left(1-\widehat{\sigma}^2\right) \frac{(g_1^{\mathrm{T}} d_1)^2}{\|d_1\|^2} \sum_{i=1}^{k} \xi^{-2(i-1)} \\
&= \|g_1\|^2 - \left(1-\widehat{\sigma}^2\right) \frac{(g_1^{\mathrm{T}} d_1)^2}{\|d_1\|^2} \frac{1-\xi^{-2k}}{1-\xi^{-2}} \\
&\geqslant \|g_1\|^2 - \frac{1-\widehat{\sigma}^2}{1-\xi^{-2}} \frac{(g_1^{\mathrm{T}} d_1)^2}{\|d_1\|^2}.
\end{aligned} \tag{3.21}$$

Therefore for any $x_1 \neq 0$, if $d_1$ is so chosen that

$$\frac{-g_1^{\mathrm{T}} d_1}{\|g_1\|\|d_1\|} \leqslant \frac{1}{2} \sqrt{\frac{1-\xi^{-2}}{1-\widehat{\sigma}^2}}, \tag{3.22}$$

and if $\alpha_k$ is computed by (3.8), we have from (3.21) that

$$\|g_{k+1}\| \geqslant \frac{\sqrt{2}}{2} \sqrt{\frac{1-\widehat{\sigma}^2}{1-\xi^{-2}}} \|g_1\|. \tag{3.23}$$

The above relation implies that the method (3.4) with $r$ satisfying (3.3) may fail to minimize (3.2) under strong Wolfe line searches.

Thus neither the upper bound nor the lower bound of $r_k$ in (2.9) can be relaxed in some sense even if the line search satisfies the strong Wolfe conditions. We write this result as the following theorem.

**Theorem 3.1.** Consider the method (1.2), (1.3) with $\beta_k = r\beta_k^{DY}$. Assume that the line search conditions are (1.8), (1.9) with the parameters satisfying $0 < \delta < 1/2$ and

$\delta < \sigma < 1$. Then for any constant $r \notin [(\sigma - 1)/(1 + \sigma), 1]$, there exists a twice continuously differentiable objective function and a starting point such that the sequence of gradient norms $\{\|g_k\|\}$ is bounded away from zero.

We see that the lower bound $(\sigma - 1)/(1 + \sigma)$ of $r_k$ in (2.9) depends on the scalar $\sigma$ in the line search condition. If $\sigma$ is close to 1, the lower bound tends to 0, whereas if $\sigma$ is close to 0, the lower bound tends to $-1$. In addition, it is obvious that

$$\left[\frac{\sigma - 1}{1 + \sigma}, 1\right] \subset [-1, 1], \tag{3.24}$$

which indicates that the convergent interval of the size of $\beta_k$ with respect to $\beta_k^{DY}$ is narrower than that of the size of $\beta_k$ with respect to $\beta_k^{FR}$ provided that $\sigma \neq 0$.

## 4. An efficient hybrid conjugate gradient method

Since formula (1.7) has the same denominator as formula (1.10), we consider the following hybrid method:

$$\beta_k = \max\{-c\beta_k^{DY}, \min\{\beta_k^{HS}, \beta_k^{DY}\}\}, \tag{4.1}$$

where $c$ is the constant in (2.8). The above method is still a conjugate gradient method, since (4.1) reduces the FR formula for $\beta_k$ if $f$ is a strictly convex quadratic and the line search is exact. By theorem 2.3, we know that the hybrid method (4.1) with weak Wolfe line searches produces a descent direction at every iteration and converges globally. Since it is easier to compute a steplength satisfying the weak Wolfe conditions than to compute a steplength satisfying the strong Wolfe conditions, we will test the hybrid method (4.1) with weak Wolfe line searches. It turns out that this algorithm performs slightly better than the PRP metod with strong Wolfe line searches.

In addition to (4.1), we are also interesting in the following hybrid conjugate gradient method:

$$\beta_k = \max\{0, \min\{\beta_k^{HS}, \beta_k^{DY}\}\}. \tag{4.2}$$

We suggest the hybrid method (4.2) for two reasons. The first is related to the restart strategy proposed in [17]. While dealing with the Beale three-term method, Powell [17] introduced a restart if the following condition holds:

$$\left|g_k^{\mathrm{T}} g_{k-1}\right| > 0.2\|g_k\|^2, \tag{4.3}$$

and obtained satisfactory numerical results. If $\beta_k^{HS} \leqslant 0$, we have that $g_{k-1}^{\mathrm{T}} g_k > \|g_k\|^2$ and hence (4.3) holds. Thus in this case, it is suitable to set $\beta_k = 0$, which implies that a restart along $-g_k$ will be done. Another reason is that, we know from (1.3) that $d_k$ may tend to almost opposite to $d_{k-1}$ if $\beta_k < 0$ and $\|d_k\| \gg \|g_k\|$. Thus the restriction that $\beta_k \geqslant 0$ will prevent two consecutive search directions from tending to be almost opposite. Our numerical results showed that the hybrid method (4.2) really performs better than the hybrid method (4.1).

Both the hybrid methods (4.1) and (4.2) can avoid the propensity of short steps. For example, for the hybrid method (4.2), we define

$$\xi_k = \max\left\{0, \min\left\{\frac{g_k^{\mathrm{T}} y_{k-1}}{\|g_k\|^2}, 1\right\}\right\}. \tag{4.4}$$

Then similarly to the second relation in (2.21), we can establish

$$\frac{\|d_k\|^2}{(g_k^{\mathrm{T}} d_k)^2} \leqslant \xi_k^2 \frac{\|d_{k-1}\|^2}{(g_{k-1}^{\mathrm{T}} d_{k-1})^2} + \frac{1}{\|g_k\|^2}. \tag{4.5}$$

Recalling the definition of $\theta_k$, we have by (4.5) that

$$\cos^{-2}\theta_k \leqslant \xi_k^2 \frac{\|g_k\|^2}{\|g_{k-1}\|^2} \cos^{-2}\theta_{k-1} + 1. \tag{4.6}$$

Suppose that at $(k-1)$th iteration an unfortunate search direction is generated, such that $\cos\theta_{k-1} \approx 0$, and that $g_k \approx g_{k-1}$. Thus $\xi_k \approx 0$. Therefore by (4.6), we have that $\cos\theta_k \gg \cos\theta_{k-1}$, which indicates that the hybrid method (4.2) would avoid the propensity of short steps.

We tested the hybrid methods (4.1) and (4.2) on an SGI Indigo workstation. The used line search conditions are (1.8) and (1.11) with $\delta = 0.01$ and $\sigma = 0.1$. The initial value of $\alpha_k$ is always set to 1. By theorem 2.3, we know that the line search conditions ensure the descent property and global convergence of the two hybrid methods. Since the PRP method is generally believed to be one of the most efficient conjugate gradient algorithms, we compared the hybrid methods with the PRP method. For the PRP method, our line search subroutine computes $\alpha_k$ such that the strong Wolfe conditions (1.8), (1.9) hold with $\delta = 0.01$ and $\sigma = 0.1$. Although the strong Wolfe conditions can not ensure the descent property of $d_k$ for the PRP method, uphill search directions seldom occur in our numerical experiments. In the case when an uphill search direction is produced, we restart the algorithm with $d_k = -g_k$.

The test problems are drawn from Moré et al. [13]. The first column "P" in table 1 denotes the problem number in [13], whereas the second gives the name of the problem. We tested each problem with two different values of $n$ ranging from $n = 20$ to $n = 10000$. The numerical results are given in the form of I/F/G, where I, F, G denote numbers of iterations, function evaluations, and gradient evaluations, respectively. The stopping condition is

$$\|g_k\| \leqslant 10^{-6}. \tag{4.7}$$

From table 1, we see that the hybrid method (4.1) requires fewer function evaluations and gradient evaluations than the PRP method for 9 problems, whereas the PRP method outperforms the hybrid method (4.1) only for 6 problems. For the other 3 test problems, the PRP method requires fewer function evaluations but the hybrid method (4.1) does require fewer gradient evaluations. In addition, for some problems such as Penalty 2 and Extended Powell, the advantage of the hybrid method (4.1) over the PRP

Table 1
Comparing different conjugate gradient methods.

| P | Name | $n$ | PRP | (4.1) | (4.2) |
|---|---|---|---|---|---|
| 24 | Penalty 2 | 20 | 530/1641/912 | 290/821/370 | 135/419/228 |
| | | 40 | 1312/3650/1590 | 487/1492/539 | 122/366/177 |
| 25 | Variably dimensioned | 20 | 6/33/12 | 5/30/10 | 5/30/10 |
| | | 50 | 5/25/11 | 9/53/18 | 9/51/17 |
| 35 | Chebyquad | 20 | 104/340/132 | 145/453/162 | 100/321/119 |
| | | 50 | 365/1203/432 | 359/1205/426 | 350/1156/406 |
| 30 | Broyden tridiagonal | 50 | 32/102/37 | 50/158/58 | 50/158/58 |
| | | 500 | 32/103/39 | 58/183/67 | 58/183/67 |
| 31 | Broyden banded | 50 | 37/142/64 | 31/115/49 | 30/113/49 |
| | | 500 | 34/128/58 | 23/74/27 | 23/74/27 |
| 22 | Extended Powell | 100 | 118/358/163 | 110/317/117 | 66/203/87 |
| | | 1000 | 396/1176/545 | 128/365/135 | 66/203/87 |
| 26 | Trigonome-tric | 100 | 55/98/97 | 58/97/95 | 58/97/95 |
| | | 1000 | 54/97/97 | 52/87/87 | 52/87/87 |
| 21 | Extended Rosenbrock | 1000 | 23/107/60 | 34/125/57 | 28/87/39 |
| | | 10000 | 23/107/60 | 37/133/60 | 28/87/39 |
| 23 | Penalty 1 | 1000 | 21/66/49 | 51/130/92 | 54/154/110 |
| | | 10000 | 30/113/82 | 37/118/72 | 35/111/66 |

method is impressive. On average, the hybrid method (4.1) performs slightly better than the PRP method for the given test problems.

From table 1, we also see that the hybrid method (4.2) clearly dominates the PRP method and the hybrid method (4.1). The numerical performances of the three methods can also be reflected by their CPU time. To solve all the 18 problems, the CPU time (in seconds) required by the PRP method, the hybrid method (4.1), and the hybrid method (4.2) are 18.80, 16.56 and 14.04, respectively. To sum up, our numerical results suggest two promising hybrid conjugate gradient methods, even the hybrid method (4.2).

## 5. Conclusions and discussions

In this paper, we have carefully studied methods related to a new nonlinear conjugate gradient method proposed by the authors – the method (1.10). Denote $r_k$ to be the size of $\beta_k$ with respect to $\beta_k^{DY}$. If $r_k$ belongs to some interval, the corresponding methods are shown to produce a descent search direction at every iteration and converge globally provided that the line search satisfies the weak Wolfe conditions. Otherwise, a convex quadratic counter-example can be constructed, showing that the corresponding methods need not converge.

Although it was doubtful whether the global convergence study would yield a better conjugate gradient algorithm, we tested two variants of the method (1.10), namely, the hybrid methods (4.1) and (4.2). The two hybrid methods are combinations of the method (1.10) and the HS method, and do not show any propensity for short steps. Initial numerical experiments were done for the hybrid methods with weak Wolfe line searches. These experiments show that both hybrid methods are competitive with respect to the PRP conjugate gradient method. Morever, the hybrid method (4.2) appears to outperform the two others, even though it only uses the weak Wolfe conditions in its line search. This shows that efficient conjugate gradient algorithms can be designed that use these weak conditions. More numerical experiments are of course needed to assess the true potential of the methods discussed here, but the preliminary results are encouraging.

## Acknowledgments

The authors are very grateful to the referees for their valuable comments and suggestions on the early version of the paper.

## References

[1] M. Al-Baali, Descent property and global convergence of the Fletcher–Reeves method with inexact line search, IMA J. Numer. Anal. 5 (1985) 121–124.
[2] Y.H. Dai, New properties of a nonlinear conjugate gradient method, Numer. Math. 89 (2001) 83–98.
[3] Y.H. Dai and Y. Yuan, Convergence properties of the Fletcher–Reeves method, IMA J. Numer. Anal. 16(2) (1996) 155–164.
[4] Y.H. Dai and Y. Yuan, Convergence properties of the conjugate descent method, Advances in Mathematics 6 (1996) 552–562.
[5] Y.H. Dai and Y. Yuan, Some properties of a new conjugate gradient method, in: *Advances in Nonlinear Programming*, ed. Y. Yuan (Kluwer Academic, Boston, 1998) pp. 251–262.
[6] Y.H. Dai and Y. Yuan, A nonlinear conjugate gradient method with a strong global convergence property, SIAM J. Optimization 10(1) (1999) 177–182.
[7] R. Fletcher, *Practical Methods of Optimization, Vol. 1, Unconstrained Optimization* (Wiley, New York, 1987).
[8] R. Fletcher and C. Reeves, Function minimization by conjugate gradients, Comput. J. 7 (1964) 149–154.
[9] J.C. Gilbert and J. Nocedal, Global convergence properties of conjugate gradient methods for optimization, SIAM J. Optimization 2(1) (1992) 21–42.
[10] M.R. Hestenes and E.L. Stiefel, Methods of conjugate gradients for solving linear systems, J. Res. Nat. Bur. Standards Sect. 5(49) (1952) 409–436.
[11] Y.F. Hu and C. Storey, Global convergence result for conjugate gradient methods, JOTA 71(2) (1991) 399–405.
[12] 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) 75–82.
[13] J.J. Moré, B.S. Garbow and K.E. Hillstrom, Testing unconstrained optimization software, ACM Transactions on Mathematical Software 7 (1981) 17–41.
[14] E. Polak and G. Ribière, Note sur la convergence de directions conjugées, Rev. Française Informat. Recherche Opertionelle, 3e année 16 (1969) 35–43.

[15] B.T. Polyak, The conjugate gradient method in extremem problems, Comput. Math. Math. Phys. 9 (1969) 94–112.
[16] M.J.D. Powell, Restart procedures of the conjugate gradient method, Math. Programming 2 (1977) 241–254.
[17] M.J.D. Powell, Nonconvex minimization calculations and the conjugate gradient method, in: *Lecture Notes in Mathematics* 1066 (Springer, Berlin, 1984) pp. 122–141.
[18] D. Touati-Ahmed and C. Storey, Efficient hybrid conjugate gradient techniques, JOTA 64 (1990) 379–397.
[19] P. Wolfe, Convergence conditions for ascent methods, SIAM Review 11 (1969) 226–235.
[20] P. Wolfe, Convergence conditions for ascent methods. II: Some corrections, SIAM Review 13 (1971) 185–188.
[21] G. Zoutendijk, Nonlinear programming, computational methods, in: *Integer and Nonlinear Programming*, ed. J. Abadie (North-Holland, Amsterdam, 1970) pp. 37–86.