Annals of Operations Research 103, 161–173, 2001

Global Convergence of Conjugate Gradient Methods without Line Search

JIE SUN * jsun@nus.edu.sg
Department of Decision Sciences, National University of Singapore, Republic of Singapore

JIAPU ZHANG **
Department of Mathematics, University of Melbourne, Melbourne, Australia

Abstract. Global convergence results are derived for well-known conjugate gradient methods in which the line search step is replaced by a step whose length is determined by a formula. The results include the following cases: (1) The Fletcher–Reeves method, the Hestenes–Stiefel method, and the Dai–Yuan method applied to a strongly convex LC^1 objective function; (2) The Polak–Ribière method and the Conjugate Descent method applied to a general, not necessarily convex, LC^1 objective function.

Keywords: conjugate gradient methods, convergence of algorithms, line search

AMS subject classification: 90C25, 90C30

1. Introduction

The conjugate gradient method is quite useful in finding an unconstrained minimum of a high-dimensional function $f(x)$. In general, the method has the following form:

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

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

where g_k denotes the gradient $\nabla f(x_k)$, α_k is a steplength obtained by a line search, d_k is the search direction, and β_k is chosen so that d_k becomes the kth conjugate direction when the function is quadratic and the line search is exact. Varieties of this method differ in the way of selecting β_k. Some well-known formulae for β_k are given by

$$\beta_k^{FR} = \frac{\|g_k\|^2}{\|g_{k-1}\|^2} \quad \text{(Fletcher–Reeves [7])}, \tag{3}$$

$$\beta_k^{PR} = \frac{g_k^{\mathrm{T}}(g_k - g_{k-1})}{\|g_{k-1}\|^2} \quad \text{(Polak–Ribière [13])}, \tag{4}$$

* Corresponding author. This author is a fellow of the Singapore-MIT Alliance. His research was partially supported by grant R314-000-009-112 and R314-000-026-112 of National University of Singapore.
** This author's work was partially supported by a scholarship from National University of Singapore.

$$\beta_k^{HS} = \frac{g_k^{\mathrm{T}}(g_k - g_{k-1})}{d_{k-1}^{\mathrm{T}}(g_k - g_{k-1})} \quad \text{(Hestenes–Stiefel [9])}, \tag{5}$$

and

$$\beta_k^{CD} = \frac{\|g_k\|^2}{-d_{k-1}^{\mathrm{T}} g_{k-1}} \quad \text{(The Conjugate Descent Method [6])}, \tag{6}$$

where $\|\cdot\|$ is the Euclidean norm and "T" stands for the transpose. Recently Dai and Yuan [5] also introduced a formula for β_k:

$$\beta_k^{DY} = \frac{\|g_k\|^2}{d_{k-1}^{\mathrm{T}}(g_k - g_{k-1})}. \tag{7}$$

For ease of presentation we call the methods corresponding to (3)–(7) the FR method, the PR method, the HS method, the CD method, and the DY method, respectively. The global convergence of methods (3)–(7) has been studied by many authors, including Al-Baali [1], Gilbert and Nocedal [8], Hestenes and Stiefel [9], Hu and Storey [10], Liu, Han and Yin [11], Powell [14,15], Touati-Ahmed and Storey [16], and Zoutendijk [18] among others. A key factor of global convergence is how to select the steplength α_k. The most natural choice of α_k is to do the exact line search, i.e., to let $\alpha_k = \operatorname{argmin}_{\alpha \geqslant 0} f(x_k + \alpha d_k)$. However, somewhat surprisingly, this natural choice could result in a non-convergent sequence in the case of PR and HS methods, as shown by Powell [15]. Inspired by Powell's work, Gilbert and Nocedal [8] conducted an elegant analysis and showed that the PR method is globally convergent if β_k^{PR} is restricted to be non-negative and α_k is determined by a line search step satisfying the sufficient descent condition $g_k^{\mathrm{T}} d_k \leqslant -c\|g_k\|^2$ in addition to the standard Wolfe conditions [17]. Recently, Dai and Yuan [3,4] showed that both the CD method and the FR method are globally convergent if the following line search conditions for α_k are satisfied:

$$\begin{cases} f(x_k) - f(x_k + \alpha_k d_k) \geqslant -\delta \alpha_k g_k^{\mathrm{T}} d_k, \\ \sigma_1 g_k^{\mathrm{T}} d_k \leqslant g(x_k + \alpha_k d_k)^{\mathrm{T}} d_k \leqslant -\sigma_2 g_k^{\mathrm{T}} d_k, \end{cases} \tag{8}$$

where $\sigma_1 \geqslant 0$, $0 < \delta < \sigma_1 < 1$, $0 < \sigma_2 < 1$ for the CD method and in addition $\sigma_1 + \sigma_2 \leqslant 1$ for the FR method.

In this paper we consider the problem from a different angle. We show that a unified formula for α_k (see (10) below) can ensure global convergence for many cases, which include:

(1) The FR method, the HS method, and the DY method applied to a strongly convex first order Lipschitz continuous (LC^1 for short) objective function.

(2) The PR method and the CD method applied to a general, not necessarily convex, LC^1 objective function. For convenience we will refer to these methods as *conjugate gradient methods without line search* since the steplength is determined by a formula rather than a line search process.

2. The proof for convergence

We adopt the following assumption on function f which is commonly used in the literature.

Assumption 1. The function f is LC^1 in a neighborhood N of the level set $L := \{x \in \mathbb{R}^n \mid f(x) \leqslant f(x_1)\}$ and L is bounded. Here, by LC^1 we mean that the gradient $\nabla f(x)$ is Lipschitz continuous with modulus μ, i.e., there exists $\mu > 0$ such that $\|\nabla f(x_{k+1}) - \nabla f(x_k)\| \leqslant \mu \|x_{k+1} - x_k\|$ for any $x_{k+1}, x_k \in N$.

Assumption 1 is sufficient for the global convergence of the PR method and the CD method, but seems to be not enough for the convergence of the FR method, the HS method, and the DY method without line search. Thus, for these three methods we impose the following stronger assumption.

Assumption 2. The function f is LC^1 and strongly convex on N. In other words, there exists $\lambda > 0$ such that $[\nabla f(x_{k+1}) - \nabla f(x_k)]^{\mathrm{T}}(x_{k+1} - x_k) \geqslant \lambda \|x_{k+1} - x_k\|^2$ for any $x_{k+1}, x_k \in N$.

Note that assumption 2 implies assumption 1 since a strongly convex function has bounded level sets.

Let $\{Q_k\}$ be a sequence of positive definite matrices. Assume that there exist $\nu_{\min} > 0$ and $\nu_{\max} > 0$ such that $\forall d \in \mathbb{R}^n$

$$\nu_{\min} d^{\mathrm{T}} d \leqslant d^{\mathrm{T}} Q_k d \leqslant \nu_{\max} d^{\mathrm{T}} d. \tag{9}$$

This condition would be satisfied, for instance, if $Q_k = Q$ and Q is positive definite. We analyze the conjugate gradient methods that use the following steplength formula:

$$\alpha_k = -\frac{\delta g_k^{\mathrm{T}} d_k}{\|d_k\|_{Q_k}^2}, \tag{10}$$

where $\|d_k\|_{Q_k} := \sqrt{d_k^{\mathrm{T}} Q_k d_k}$ and $\delta \in (0, \nu_{\min}/\mu)$. Note that the specification of δ ensures $\delta\mu/\nu_{\min} < 1$.

Lemma 1. Suppose that x_k is given by (1), (2), and (10). Then

$$g_{k+1}^{\mathrm{T}} d_k = \rho_k g_k^{\mathrm{T}} d_k \tag{11}$$

holds for all k, where

$$\rho_k = 1 - \frac{\delta \phi_k \|d_k\|^2}{\|d_k\|_{Q_k}^2} \tag{12}$$

and

$$\phi_k = \begin{cases} 0 & \text{for } \alpha_k = 0, \\ \dfrac{(g_{k+1} - g_k)^{\mathrm{T}}(x_{k+1} - x_k)}{\|x_{k+1} - x_k\|^2} & \text{for } \alpha_k \neq 0. \end{cases} \tag{13}$$

Proof. The case of $\alpha_k = 0$ implies that $\rho_k = 1$ and $g_{k+1} = g_k$. Hence (11) is valid. We now prove for the case of $\alpha_k \neq 0$. From (2) and (10) we have

$$\begin{aligned} g_{k+1}^{\mathrm{T}} d_k &= g_k^{\mathrm{T}} d_k + (g_{k+1} - g_k)^{\mathrm{T}} d_k = g_k^{\mathrm{T}} d_k + \alpha_k^{-1} (g_{k+1} - g_k)^{\mathrm{T}} (x_{k+1} - x_k) \\ &= g_k^{\mathrm{T}} d_k + \alpha_k^{-1} \phi_k \|x_{k+1} - x_k\|^2 = g_k^{\mathrm{T}} d_k + \alpha_k \phi_k \|d_k\|^2 \\ &= g_k^{\mathrm{T}} d_k - \frac{\delta g_k^{\mathrm{T}} d_k}{\|d_k\|_{Q_k}^2} \phi_k \|d_k\|^2 = \left(1 - \frac{\delta \phi_k \|d_k\|^2}{\|d_k\|_{Q_k}^2}\right) g_k^{\mathrm{T}} d_k = \rho_k g_k^{\mathrm{T}} d_k. \end{aligned} \tag{14}$$

The proof is complete. □

Corollary 2. In formulae (6) and (7) we have $\beta_k^{DY} = \beta_k^{CD}/(1 - \rho_{k-1})$.

Proof. This is straightforward from (6), (7), and (11). □

Corollary 3. There holds

$$1 - \frac{\delta \mu}{\nu_{\min}} \leqslant \rho_k \leqslant 1 + \frac{\delta \mu}{\nu_{\min}} \tag{15}$$

for all k if assumption 1 is valid; and this estimate can be sharpened to

$$0 < \rho_k \leqslant 1 - \frac{\delta \lambda}{\nu_{\max}} \tag{16}$$

if assumption 2 is valid.

Proof. By (12) and (13) we have

$$\rho_k = 1 - \delta \phi_k \frac{\|d_k\|^2}{\|d_k\|_{Q_k}^2} = 1 - \frac{\delta (g_k - g_{k-1})^{\mathrm{T}} (x_k - x_{k-1})}{\|x_k - x_{k-1}\|^2} \frac{\|d_k\|^2}{\|d_k\|_{Q_k}^2}. \tag{17}$$

Then from assumption 1 or 2 we have either

$$\|g_{k+1} - g_k\| \leqslant \mu \|x_{k+1} - x_k\| \tag{18}$$

or

$$(g_k - g_{k-1})^{\mathrm{T}} (x_k - x_{k-1}) \geqslant \lambda \|x_k - x_{k-1}\|^2, \tag{19}$$

which, together with (9) and (17), leads to the corresponding bounds for ρ_k. □

Lemma 4. Suppose that assumptions 1 holds and that x_k is given by (1), (2), and (10). Then

$$\sum_{d_k\neq 0} \frac{(g_k^{\mathrm{T}} d_k)^2}{\|d_k\|^2} < \infty. \tag{20}$$

Proof. By the mean-value theorem we have

$$f(x_{k+1}) - f(x_k) = \bar{g}^{\mathrm{T}}(x_{k+1} - x_k), \tag{21}$$

where $\bar{g} = \nabla f(\bar{x})$ for some $\bar{x} \in [x_k, x_{k+1}]$. Now by the Cauchy–Schwartz inequality, (2), (10), and assumption 1 we obtain

$$\begin{aligned}
\bar{g}^{\mathrm{T}}(x_{k+1} - x_k) &= g_k^{\mathrm{T}}(x_{k+1} - x_k) + (\bar{g} - g_k)^{\mathrm{T}}(x_{k+1} - x_k) \\
&\leqslant g_k^{\mathrm{T}}(x_{k+1} - x_k) + \|\bar{g} - g_k\| \, \|x_{k+1} - x_k\| \\
&\leqslant g_k^{\mathrm{T}}(x_{k+1} - x_k) + \mu\|x_{k+1} - x_k\|^2 = \alpha_k g_k^{\mathrm{T}} d_k + \mu\alpha_k^2\|d_k\|^2 \\
&= \alpha_k g_k^{\mathrm{T}} d_k - \frac{\mu\alpha_k\delta g_k^{\mathrm{T}} d_k\|d_k\|^2}{\|d_k\|_{Q_k}^2} = \alpha_k g_k^{\mathrm{T}} d_k\left(1 - \frac{\mu\delta\|d_k\|^2}{\|d_k\|_{Q_k}^2}\right) \\
&\leqslant -\delta\left(1 - \frac{\mu\delta}{\nu_{\min}}\right)\frac{(g_k^{\mathrm{T}} d_k)^2}{\|d_k\|_{Q_k}^2};
\end{aligned} \tag{22}$$

i.e.,

$$f(x_{k+1}) - f(x_k) \leqslant -\delta\left(1 - \frac{\mu\delta}{\nu_{\min}}\right)\frac{(g_k^{\mathrm{T}} d_k)^2}{\|d_k\|_{Q_k}^2}, \tag{23}$$

which implies $f(x_{k+1}) \leqslant f(x_k)$. It follows by assumption 1 that $\lim_{k\to\infty} f(x_k)$ exists. Thus, from (9) and (23) we obtain

$$\frac{(g_k^{\mathrm{T}} d_k)^2}{\|d_k\|^2} \leqslant \nu_{\max}\frac{(g_k^{\mathrm{T}} d_k)^2}{\|d_k\|_{Q_k}^2} \leqslant \frac{\nu_{\max}}{\delta(1 - \mu\delta/\nu_{\min})}\big[f(x_k) - f(x_{k+1})\big]. \tag{24}$$

This finishes our proof. □

Lemma 5. Suppose that assumption 1 holds and that x_k is given by (1), (2), and (10). Then

$$\liminf_{k\to\infty} \|g_k\| \neq 0 \quad \text{implies} \quad \sum_{d_k\neq 0} \frac{\|g_k\|^4}{\|d_k\|^2} < \infty. \tag{25}$$

Proof. If $\liminf_{k\to\infty} \|g_k\| \neq 0$, then there exists $\gamma > 0$ such that $\|g_k\| \geqslant \gamma$ for all k. Let

$$\lambda_k = \frac{|g_k^{\mathrm{T}} d_k|}{\|d_k\|}. \tag{26}$$

Then by lemma 4 there holds

$$\lambda_k \leqslant \frac{\gamma}{4} \tag{27}$$

for all large k. By lemma 1 and corollary 3 we have

$$\left|g_k^{\mathrm{T}} d_{k-1}\right| = \left|\rho_k g_{k-1}^{\mathrm{T}} d_{k-1}\right| \leqslant \left(1 + \frac{\delta\mu}{\nu_{\min}}\right)\left|g_{k-1}^{\mathrm{T}} d_{k-1}\right| < 2\left|g_{k-1}^{\mathrm{T}} d_{k-1}\right|. \tag{28}$$

Considering (1), we have

$$g_k = \beta_k d_{k-1} - d_k. \tag{29}$$

By multiplying g_k on both sides of (29), we obtain

$$\|g_k\|^2 = \beta_k g_k^{\mathrm{T}} d_{k-1} - g_k^{\mathrm{T}} d_k. \tag{30}$$

From (30) and (28) it follows that

$$\begin{aligned}
\frac{\|g_k\|^2}{\|d_k\|} &= \frac{\beta_k g_k^{\mathrm{T}} d_{k-1} - g_k^{\mathrm{T}} d_k}{\|d_k\|} \leqslant \frac{|\beta_k|\,|g_k^{\mathrm{T}} d_{k-1}| + |g_k^{\mathrm{T}} d_k|}{\|d_k\|} \\
&\leqslant 2\lambda_{k-1}\frac{\|\beta_k d_{k-1}\|}{\|d_k\|} + \lambda_k = \lambda_k + 2\lambda_{k-1}\frac{\|d_k + g_k\|}{\|d_k\|} \\
&\leqslant \lambda_k + 2\lambda_{k-1} + 2\lambda_{k-1}\frac{\|g_k\|^2}{\|d_k\|\|g_k\|} \leqslant \lambda_k + 2\lambda_{k-1} + 2\lambda_{k-1}\frac{\|g_k\|^2}{\gamma\|d_k\|} \\
&\leqslant \lambda_k + 2\lambda_{k-1} + 2\left(\frac{\gamma}{4}\right)\frac{\|g_k\|^2}{\gamma\|d_k\|} = \lambda_k + 2\lambda_{k-1} + \frac{\|g_k\|^2}{2\|d_k\|};
\end{aligned}$$

i.e.,

$$\frac{\|g_k\|^2}{\|d_k\|} \leqslant \lambda_k + 2\lambda_{k-1} + \frac{\|g_k\|^2}{2\|d_k\|}. \tag{31}$$

The above relation can be re-written as

$$\frac{\|g_k\|^2}{\|d_k\|} \leqslant 2\lambda_k + 4\lambda_{k-1} \leqslant 4(\lambda_k + \lambda_{k-1}). \tag{32}$$

Hence we have

$$\frac{\|g_k\|^4}{\|d_k\|^2} \leqslant 16(\lambda_k + \lambda_{k-1})^2 \leqslant 32\left(\lambda_k^2 + \lambda_{k-1}^2\right). \tag{33}$$

By lemma 4 we know that $\sum \lambda_k^2 < \infty$. Hence we obtain from (33)

$$\sum_{d_k \neq 0} \frac{\|g_k\|^4}{\|d_k\|^2} < \infty.$$

□

Remarks. Note that the two lemmae above are also true if assumption 2 holds because assumption 2 implies assumption 1. The proof of lemma 5 is similar to the one of theorem 2.3 and corollary 2.4 in [2]. The difference is that here we do not need to assume the strong Wolfe conditions.

Lemma 6. The CD method satisfies

$$\frac{\|d_k\|^2}{\|g_k\|^4} \leqslant \frac{\|d_{k-1}\|^2}{\|g_{k-1}\|^4} + \frac{\Omega}{\|g_k\|^2} \tag{34}$$

under assumption 1, where Ω is a positive constant.

Proof. For the CD method, by lemma 1 and corollary 3 we have

$$\begin{aligned} g_k^{\mathrm{T}} d_k &= g_k^{\mathrm{T}}\left(-g_k + \frac{\|g_k\|^2}{-g_{k-1}^{\mathrm{T}} d_{k-1}} d_{k-1}\right) = -\|g_k\|^2 - \|g_k\|^2 \frac{g_k^{\mathrm{T}} d_{k-1}}{g_{k-1}^{\mathrm{T}} d_{k-1}} \\ &= -\|g_k\|^2 - \|g_k\|^2 \frac{\rho_{k-1} g_{k-1}^{\mathrm{T}} d_{k-1}}{g_{k-1}^{\mathrm{T}} d_{k-1}} = -(1+\rho_{k-1})\|g_k\|^2 < 0. \end{aligned} \tag{35}$$

Note that from the last equality above and corollary 3 we have

$$\left(g_k^{\mathrm{T}} d_k\right)^2 \geqslant \|g_k\|^4, \tag{36}$$

which is due to $\rho_{k-1} \geqslant 0$ that is deduced by corollary 3 and $\delta \in (0, \nu_{\min}/\mu)$. Applying (1), (6), lemma 1 and (36), we obtain

$$\begin{aligned} \|d_k\|^2 &= \left\|-g_k + \beta_k^{CD} d_{k-1}\right\|^2 = \left\|-g_k + \frac{\|g_k\|^2}{-g_{k-1}^{\mathrm{T}} d_{k-1}} d_{k-1}\right\|^2 \\ &= \|g_k\|^2 + \frac{\|g_k\|^4}{(g_{k-1}^{\mathrm{T}} d_{k-1})^2}\|d_{k-1}\|^2 + \frac{2\|g_k\|^2}{g_{k-1}^{\mathrm{T}} d_{k-1}} g_k^{\mathrm{T}} d_{k-1} \\ &= \|g_k\|^2 + \frac{\|g_k\|^4}{(g_{k-1}^{\mathrm{T}} d_{k-1})^2}\|d_{k-1}\|^2 + \frac{2\rho_{k-1} g_{k-1}^{\mathrm{T}} d_{k-1}}{g_{k-1}^{\mathrm{T}} d_{k-1}} \|g_k\|^2 \\ &= (1+2\rho_{k-1})\|g_k\|^2 + \frac{\|g_k\|^4}{(g_{k-1}^{\mathrm{T}} d_{k-1})^2}\|d_{k-1}\|^2 \\ &\leqslant \left[1 + 2\left(1 + \frac{\delta\mu}{\nu_{\min}}\right)\right]\|g_k\|^2 + \frac{\|g_k\|^4}{\|g_{k-1}\|^4}\|d_{k-1}\|^2, \end{aligned} \tag{37}$$

where

$$\frac{\|g_k\|^4}{(g_{k-1}^{\mathrm{T}} d_{k-1})^2}\|d_{k-1}\|^2 \leqslant \frac{\|g_k\|^4}{\|g_{k-1}\|^4}\|d_{k-1}\|^2$$

was gotten by using (36). By defining $\Omega := 1 + 2(1 + \delta\mu/\nu_{\min})$ and dividing both sides by $\|g_k\|^4$ we get

$$\frac{\|d_k\|^2}{\|g_k\|^4} \leqslant \frac{\|d_{k-1}\|^2}{\|g_{k-1}\|^4} + \frac{\Omega}{\|g_k\|^2}. \tag{38}$$

The proof for the CD method is complete. □

Lemma 7. Suppose assumption 2 holds. Then the FR method satisfies

$$\frac{\|d_k\|^2}{\|g_k\|^4} \leqslant \frac{\|d_{k-1}\|^2}{\|g_{k-1}\|^4} + \frac{\Omega}{\|g_k\|^2}, \tag{39}$$

where Ω is a positive constant.

Proof.

$$\begin{aligned} -g_k^{\mathrm{T}} d_{k-1} &= -\rho_{k-1} g_{k-1}^{\mathrm{T}} d_{k-1} = -\rho_{k-1} g_{k-1}^{\mathrm{T}} \big(-g_{k-1} + \beta_{k-1}^{FR} d_{k-2}\big) \\ &= -\rho_{k-1} g_{k-1}^{\mathrm{T}} \left(-g_{k-1} + \frac{\|g_{k-1}\|^2}{\|g_{k-2}\|^2} d_{k-2}\right) \\ &= \rho_{k-1}\|g_{k-1}\|^2 - \frac{\rho_{k-1}\|g_{k-1}\|^2}{\|g_{k-2}\|^2} g_{k-1}^{\mathrm{T}} d_{k-2}. \end{aligned} \tag{40}$$

Thus, we have a recursive equation which leads to

$$\begin{aligned} -g_k^{\mathrm{T}} d_{k-1} &= \rho_{k-1}\|g_{k-1}\|^2 - \frac{\rho_{k-1}\|g_{k-1}\|^2}{\|g_{k-2}\|^2} g_{k-1}^{\mathrm{T}} d_{k-2} \\ &= \rho_{k-1}\|g_{k-1}\|^2 + \rho_{k-1}\rho_{k-2}\|g_{k-1}\|^2 - \rho_{k-1}\rho_{k-2}\frac{\|g_{k-1}\|^2}{\|g_{k-3}\|^2} g_{k-2}^{\mathrm{T}} d_{k-3} \\ &= \cdots = \|g_{k-1}\|^2(\rho_{k-1} + \rho_{k-1}\rho_{k-2} + \cdots + \rho_{k-1}\rho_{k-2}\cdots\rho_2) \\ &\leqslant \|g_{k-1}\|^2 \sum_{k=1}^{\infty}\left(1 - \frac{\delta\lambda}{\nu_{\max}}\right)^k =: \Omega_1\|g_{k-1}\|^2. \end{aligned} \tag{41}$$

Hence we obtain

$$\begin{aligned} \frac{\|d_k\|^2}{\|g_k\|^4} &= \frac{\|g_k\|^2 - 2\beta_k^{FR} g_k^{\mathrm{T}} d_{k-1} + (\beta_k^{FR})^2\|d_{k-1}\|^2}{\|g_k\|^4} \\ &\leqslant \frac{\|g_k\|^2 + 2\Omega_1\beta_k^{FR}\|g_{k-1}\|^2 + (\beta_k^{FR})^2\|d_{k-1}\|^2}{\|g_k\|^4} \\ &= \frac{1 + 2\Omega_1}{\|g_k\|^2} + \frac{\|d_{k-1}\|^2}{\|g_{k-1}\|^4}. \end{aligned} \tag{42}$$

By defining $\Omega := 1 + 2\Omega_1$, the proof is complete. □

Lemma 8. The DY method satisfies

$$\frac{\|d_k\|^2}{\|g_k\|^4} = \left(\frac{1 - \rho_{k-2}}{1 - \rho_{k-1}}\right)^2 \frac{\|d_{k-1}\|^2}{\|g_{k-1}\|^4} + \left(\frac{1 + \rho_{k-1}}{1 - \rho_{k-1}}\right)\frac{1}{\|g_k\|^2} \tag{43}$$

under assumption 2.

Proof. For the DY method, by (1), (7), lemma 1 and corollary 3 we have

$$g_k^{\mathrm{T}} d_k = g_k^{\mathrm{T}}\left(-g_k + \frac{\|g_k\|^2}{d_{k-1}^{\mathrm{T}}(g_k - g_{k-1})} d_{k-1}\right) = -\|g_k\|^2 + \|g_k\|^2 \frac{g_k^{\mathrm{T}} d_{k-1}}{g_k^{\mathrm{T}} d_{k-1} - g_{k-1}^{\mathrm{T}} d_{k-1}}$$
$$= -\|g_k\|^2 + \|g_k\|^2 \frac{\rho_{k-1} g_{k-1}^{\mathrm{T}} d_{k-1}}{(\rho_{k-1} - 1) g_{k-1}^{\mathrm{T}} d_{k-1}} = -\frac{1}{1 - \rho_{k-1}} \|g_k\|^2 < 0, \tag{44}$$

where $0 < \rho_{k-1} < 1$ because of corollary 3. Applying (1), (7), lemma 1 and (44), we obtain

$$\|d_k\|^2 = \left\|-g_k + \beta_k^{DY} d_{k-1}\right\|^2 = \left\|-g_k + \frac{\beta_k^{CD} d_{k-1}}{1 - \rho_{k-1}}\right\|^2$$
$$= \left\|-g_k + \frac{\|g_k\|^2}{-g_{k-1}^{\mathrm{T}} d_{k-1}} \frac{d_{k-1}}{1 - \rho_{k-1}}\right\|^2$$
$$= \|g_k\|^2 + \frac{\|g_k\|^4}{(1 - \rho_{k-1})^2 (g_{k-1}^{\mathrm{T}} d_{k-1})^2} \|d_{k-1}\|^2 + \frac{2\|g_k\|^2}{(1 - \rho_{k-1}) g_{k-1}^{\mathrm{T}} d_{k-1}} g_k^{\mathrm{T}} d_{k-1}$$
$$= \|g_k\|^2 + \frac{\|g_k\|^4}{(1 - \rho_{k-1})^2 (g_{k-1}^{\mathrm{T}} d_{k-1})^2} \|d_{k-1}\|^2 + \frac{2\rho_{k-1} g_{k-1}^{\mathrm{T}} d_{k-1}}{(1 - \rho_{k-1}) g_{k-1}^{\mathrm{T}} d_{k-1}} \|g_k\|^2$$
$$= \left(\frac{1 - \rho_{k-2}}{1 - \rho_{k-1}}\right)^2 \frac{\|d_{k-1}\|^2}{\|g_{k-1}\|^4} \|g_k\|^4 + \left(\frac{1 + \rho_{k-1}}{1 - \rho_{k-1}}\right) \|g_k\|^2. \tag{45}$$

Note in the last step above we used equation (44). Dividing both sides of the above formula by $\|g_k\|^4$, the proof is complete. □

Theorem 9. Suppose that assumption 1 holds. Then the CD method will generate a sequence $\{x_k\}$ such that $\liminf_{k\to\infty} \|g_k\| = 0$. The same conclusion holds for the FR and DY methods under assumption 2.

Proof. We first show this for the CD and FR methods. If $\liminf_{k\to\infty} \|g_k\| \neq 0$, then there exists $\gamma > 0$ such that $\|g_k\| \geqslant \gamma$ for all k, and by lemma 4 we have

$$\sum_{d_k \neq 0} \frac{\|g_k\|^4}{\|d_k\|^2} < \infty. \tag{46}$$

From lemmas 6 and 7 the two methods satisfy

$$\frac{\|d_k\|^2}{\|g_k\|^4} \leqslant \frac{\|d_{k-1}\|^2}{\|g_{k-1}\|^4} + \frac{\Omega}{\|g_k\|^2} \tag{47}$$

under either assumption 1 or assumption 2. Hence we get

$$\frac{\|d_k\|^2}{\|g_k\|^4} \leqslant \frac{\|d_{k-1}\|^2}{\|g_{k-1}\|^4} + \frac{\Omega}{\gamma^2} \leqslant \frac{\|d_{k-2}\|^2}{\|g_{k-2}\|^4} + \frac{2\Omega}{\gamma^2} \leqslant \cdots \leqslant \frac{\|d_1\|^2}{\|g_1\|^4} + \frac{(k-1)\Omega}{\gamma^2}. \tag{48}$$

Thus, (48) means that

$$\frac{\|g_k\|^4}{\|d_k\|^2} \geqslant \frac{1}{ka - a + b}, \tag{49}$$

where $a = \Omega/\gamma^2$ and $b = \|d_1\|^2/\|g_1\|^4 = 1/\|g_1\|^2$ – both are constants. From (49) we have

$$\sum_{d_k \neq 0} \frac{\|g_k\|^4}{\|d_k\|^2} = +\infty, \tag{50}$$

which is contradictory to (46). Hence the theorem is valid for the CD and FR methods.

We next consider the DY method. If $\liminf_{k\to\infty} \|g_k\| \neq 0$, then there exists $\gamma > 0$ such that $\|g_k\| \geqslant \gamma$ for all k, and by lemma 4 we have

$$\sum_{d_k \neq 0} \frac{\|g_k\|^4}{\|d_k\|^2} < \infty. \tag{51}$$

From lemma 8 the method satisfies

$$\frac{\|d_k\|^2}{\|g_k\|^4} = \left(\frac{1-\rho_{k-2}}{1-\rho_{k-1}}\right)^2 \frac{\|d_{k-1}\|^2}{\|g_{k-1}\|^4} + \left(\frac{1+\rho_{k-1}}{1-\rho_{k-1}}\right)\frac{1}{\|g_k\|^2}, \tag{52}$$

under assumption 2. Hence we get

$$\begin{aligned}
\frac{\|d_k\|^2}{\|g_k\|^4} &\leqslant \left(\frac{1-\rho_{k-2}}{1-\rho_{k-1}}\right)^2 \frac{\|d_{k-1}\|^2}{\|g_{k-1}\|^4} + \left(\frac{1+\rho_{k-1}}{1-\rho_{k-1}}\right)\frac{1}{\gamma^2} \\
&\leqslant \left(\frac{1-\rho_{k-3}}{1-\rho_{k-1}}\right)^2 \frac{\|d_{k-2}\|^2}{\|g_{k-2}\|^4} + \left[\frac{1-\rho_{k-1}^2}{(1-\rho_{k-1})^2} + \frac{1-\rho_{k-2}^2}{(1-\rho_{k-1})^2}\right]\frac{1}{\gamma^2} \leqslant \cdots \\
&\leqslant \left(\frac{1-\rho_1}{1-\rho_{k-1}}\right)^2 \frac{\|d_2\|^2}{\|g_2\|^4} + \left[\frac{1-\rho_{k-1}^2}{(1-\rho_{k-1})^2} + \cdots + \frac{1-\rho_2^2}{(1-\rho_{k-1})^2}\right]\frac{1}{\gamma^2}.
\end{aligned} \tag{53}$$

The quantities

$$\left(\frac{1-\rho_1}{1-\rho_{k-1}}\right)^2, \frac{1-\rho_{k-1}^2}{(1-\rho_{k-1})^2}, \ldots, \frac{1-\rho_2^2}{(1-\rho_{k-1})^2} \tag{54}$$

have a common upper bound due to corollary 3. Denoting this common bound by Ω, from (53) we obtain

$$\frac{\|g_k\|^4}{\|d_k\|^2} \geqslant \frac{1}{ka - 2a + b}, \tag{55}$$

where $a = \Omega/\gamma^2$, $b = \Omega\|d_2\|^2/\|g_2\|^4$ are constants. From (55) we have

$$\sum_{d_k \neq 0} \frac{\|g_k\|^4}{\|d_k\|^2} = +\infty, \tag{56}$$

which is contradictory to (51). Hence the proof for the DY method is complete and the theorem is valid. □

Now we turn to the PR and the HS methods.

Theorem 10. Suppose that assumption 1 holds. Then the PR method will generate a sequence $\{x_k\}$ such that $\liminf_{k\to\infty} \|g_k\| = 0$. If assumption 2 holds, then the same conclusion holds for the HS method as well.

Proof. We consider the PR method first. Assume in contrary that $\|g_k\| \geqslant \gamma$ for all k. From lemma 4 we have

$$\sum \|x_{k+1} - x_k\|^2 = \sum \|\alpha_k d_k\|^2 = \sum_{d_k\neq 0} \frac{(g_k^{\mathrm{T}} d_k)^2}{\|d_k\|_{Q_k}^2} \leqslant \frac{1}{\nu_{\min}} \sum_{d_k\neq 0} \frac{(g_k^{\mathrm{T}} d_k)^2}{\|d_k\|^2} < \infty. \tag{57}$$

Hence $\|x_{k+1} - x_k\|^2 \to 0$ and

$$\left|\beta_k^{PR}\right| = \left|\frac{g_{k+1}^{\mathrm{T}}(g_{k+1} - g_k)}{\|g_k\|^2}\right| \to 0 \tag{58}$$

as $k \to \infty$. Since L is bounded, both $\{x_k\}$ and $\{g_k\}$ are bounded. By using

$$\|d_k\| \leqslant \|g_k\| + |\beta_k|\|d_{k-1}\|, \tag{59}$$

one can show that $\|d_k\|$ is uniformly bounded. Thus we have

$$\begin{aligned} \left|g_k^{\mathrm{T}} d_k\right| &= \left|g_k^{\mathrm{T}}\left(-g_k + \beta_k^{PR} d_{k-1}\right)\right| \geqslant \|g_k\|^2 - \left|\beta_k^{PR}\right|\|g_k\|\|d_{k-1}\| \geqslant \frac{\|g_k\|^2}{2} \\ &\left(\text{since } \left\|\beta_k^{PR} d_{k-1}\right\| \leqslant \frac{\|g_k\|}{2}\right) \end{aligned} \tag{60}$$

for large k. Thus we have

$$\frac{(g_k^{\mathrm{T}} d_k)^2}{\|d_k\|^2\|g_k\|^2} \geqslant \frac{1}{2}\frac{\|g_k\|^2}{\|d_k\|^2}. \tag{61}$$

Since $\|g_k\| \geqslant \gamma$ and $\|d_k\|$ is bounded above, we conclude that there is $\varepsilon > 0$ such that

$$\frac{(g_k^{\mathrm{T}} d_k)^2}{\|d_k\|^2\|g_k\|^2} \geqslant \varepsilon, \tag{62}$$

which implies

$$\sum_{d_k\neq 0} \|g_k\|^2 \frac{(g_k^{\mathrm{T}} d_k)^2}{\|d_k\|^2\|g_k\|^2} = \infty. \tag{63}$$

This is a contradiction to lemma 4. The proof for the PR method is complete. We now consider the HS method. Again, we assume $\|g_k\| \geqslant \gamma$ for all k and note that $g_k - g_{k-1} \to 0$. Since

$$g_k^{\mathrm{T}} d_k = g_k^{\mathrm{T}}\big(-g_k + \beta_k^{HS} d_{k-1}\big) = g_k^{\mathrm{T}}\left(-g_k - \frac{g_k^{\mathrm{T}}(g_k - g_{k-1})}{(1-\rho_{k-1})d_{k-1}^{\mathrm{T}} g_{k-1}} d_{k-1}\right)$$
$$= -\|g_k\|^2 - \frac{\rho_{k-1}}{1-\rho_{k-1}} g_k^{\mathrm{T}}(g_k - g_{k-1}), \tag{64}$$

we have

$$\left(\frac{g_k^{\mathrm{T}} d_k}{\|g_k\|^2}\right)^2 = \left[1 + \frac{\rho_{k-1}}{1-\rho_{k-1}} \frac{g_k^{\mathrm{T}}(g_k - g_{k-1})}{\|g_k\|^2}\right]^2 \geqslant \frac{1}{2} \tag{65}$$

for large k. This is the same as (61). Moreover, from

$$\big|\beta_k^{HS}\big| = \left|\frac{g_k^{\mathrm{T}}(g_k - g_{k-1})}{(1-\rho_{k-1})d_{k-1}^{\mathrm{T}} g_{k-1}}\right| \leqslant \frac{\sqrt{2}}{\delta\lambda/\nu_{\max}} \left|\frac{g_k^{\mathrm{T}}(g_k - g_{k-1})}{\|g_{k-1}\|^2}\right| \to 0 \tag{66}$$

we conclude that $\|d_k\|$ is bounded through (59). The rest of the proof is the same as the proof for the PR method. □

3. Final remarks

We show that by taking a "fixed" steplength α_k defined by formula (10), the conjugate gradient method is globally convergent for several popular choices of β_k. The result discloses an interesting property of the conjugate gradient method – its global convergence can be guaranteed by taking a pre-determined steplength rather than following a set of line search rules. This steplength might be practical in cases that the line search is expensive or hard. Our proofs require that the function is at least LC^1 (sometimes strongly convex in addition) and the level set L is bounded. We point out that the latter condition can be relaxed to that the function is lower bounded on L in the case of the CD method because an upper bound of $\|g_k\|$ is not assumed in the corresponding proof. We allow certain flexibility in selecting the sequence $\{Q_k\}$ in practical computation. At the simplest case we may just let all Q_k be the identity matrix. It would be more interesting to define Q_k as a matrix of certain simple structure that carries some second order information of the objective function. This might be a topic of further research.

Acknowledgments

The authors are grateful to the referees for their comments, which improved the paper. The second author is much indebted to Professor C. Wang for his cultivation and to Professor L. Qi for his support.

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, J.Y. Han, G.H. Liu, D.F. Sun, H.X. Yin and Y.X. Yuan, Convergence properties of nonlinear conjugate gradient methods, SIAM J. Optim. 10 (1999) 345–358.

[3] Y.H. Dai and Y. Yuan, Convergence properties of the conjugate descent method, Advances in Mathematics 25 (1996) 552–562.

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

[5] Y.H. Dai and Y. Yuan, A nonlinear conjugate gradient method with a strong global convergence property, SIAM J. Optimization 10 (1999) 177–182.

[6] R. Fletcher, *Practical Method of Optimization, Vol I: Unconstrained Optimization*, 2nd edn. (Wiley, New York, 1987).

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

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

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

[10] Y. Hu and C. Storey, Global convergence result for conjugate gradient methods, Journal of Optimization Theory and Applications 71 (1991) 399–405.

[11] G. Liu, J. Han and H. Yin, Global convergence of the Fletcher–Reeves algorithm with inexact line search, Appl. Math. J. Chinese Univ. Ser. B 10 (1995) 75–82.

[12] B. Polak, The conjugate gradient method in extreme problems, Comput. Math. Math. Phys. 9 (1969) 94–112.

[13] B. Polak and G. Ribière, Note sur la convergence des méthodes de directions conjuguées, Rev. Fran. Informat. Rech. Opér. 16 (1969) 35–43.

[14] M.J.D. Powell, Nonconvex minimization calculations and the conjugate gradient method, in: *Lecture Notes in Mathematics* 1066 (1984) pp. 121–141.

[15] M.J.D. Powell, Convergence properties of algorithms for nonlinear optimization, SIAM Review 28 (1986) 487–500.

[16] D. Touati-Ahmed and C. Storey, Efficient hybrid conjugate gradient techniques, Journal of Optimization Theory and Applications 64 (1990) 379–397.

[17] P. Wolfe, Convergence conditions for ascent methods, SIAM Review 11 (1969) 226–235.

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