A three-term conjugate gradient method with sufficient descent property for unconstrained optimization

Yasushi Narushima, Hiroshi Yabe and John A. Ford

(December 10, 2008; Revised October 31, 2009)

Abstract

Conjugate gradient methods are widely used for solving large-scale unconstrained optimization problems, because they do not need the storage of matrices. In this paper, we propose a general form of three-term conjugate gradient methods which always generate a sufficient descent direction. We give a sufficient condition for the global convergence of the proposed general method. Moreover, we present a specific three-term conjugate gradient method based on the multi-step quasi-Newton method. Finally, some numerical results of the proposed method are given.

keyword; Unconstrained optimization, three-term conjugate gradient method, sufficient descent condition, global convergence

1 Introduction

In this paper, we deal with conjugate gradient methods for solving the following unconstrained optimization problem:

$$\text{minimize } f(x),$$

where f is a continuously differentiable function. We denote its gradient ∇f by g. Usually, iterative methods are used for solving unconstrained optimization problems, and they are of the form

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

where $x_k \in \boldsymbol{R}^n$ is the k-th approximation to a solution, α_k is a positive step size and $d_k \in \boldsymbol{R}^n$ is a search direction.

In 1952, Hestenes and Stiefel [15] first proposed a conjugate gradient method for solving a linear system of equations with a symmetric positive definite coefficient matrix,

or equivalently for minimizing a strictly convex quadratic function. Later on, in 1964, Fletcher and Reeves [6] applied the conjugate gradient method to general unconstrained optimization problems. Recently, conjugate gradient methods are paid attention to as iterative methods for solving large-scale unconstrained optimization problems, because they do not need the storage of matrices. The search direction of conjugate gradient methods is defined by the following:

$$d_k = \begin{cases} -g_k, & \text{for} \quad k = 0, \\ -g_k + \beta_k d_{k-1}, & \text{for} \quad k \geq 1, \end{cases} \tag{1.1}$$

where g_k denotes $\nabla f(x_k)$ and $\beta_k \in \boldsymbol{R}$ is a parameter that characterizes the method. It is known that choices of β_k affect numerical performance of the method, and hence many researchers studied choices of β_k. Well-known formulas for β_k are the Hestenes-Stiefel (HS) [15, 16], Fletcher-Reeves (FR) [6], Polak-Ribière (PR) [16], Polak-Ribière Plus (PR+) [10], and Dai-Yuan (DY) [4] formulas, which are respectively given by

$$\begin{gathered} \beta_k^{HS} = \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}}, \qquad \beta_k^{FR} = \frac{\|g_k\|^2}{\|g_{k-1}\|^2}, \\ \beta_k^{PR} = \frac{g_k^T y_{k-1}}{\|g_{k-1}\|^2}, \qquad \beta_k^{PR+} = \max\left\{ \frac{g_k^T y_{k-1}}{\|g_{k-1}\|^2},\ 0 \right\}, \qquad \beta_k^{DY} = \frac{\|g_k\|^2}{d_{k-1}^T y_{k-1}}, \end{gathered} \tag{1.2}$$

where y_{k-1} is defined by

$$y_{k-1} = g_k - g_{k-1}$$

and $\|\cdot\|$ denotes the ℓ_2 norm. Furthermore, we define

$$s_{k-1} = x_k - x_{k-1},$$

which is used in the subsequent sections. Note that these formulas for β_k are equivalent each other if the objective function is a strictly convex quadratic function and α_k is the one dimensional minimizer. There are many researches on convergence properties of these methods (see [13, 16], for example).

For this decade, many other conjugate gradient methods are proposed and these are classified by two classes. The first approach makes use of the second-order information of the objective function to accelerate conjugate gradient methods. Dai and Liao [3] proposed a conjugate gradient method based on the secant condition and proved its global convergence property. Later some researchers proposed its variants based on other secant conditions, and they proved global convergence properties of their proposed methods [9, 18, 22]. Although these methods are effective for solving large-scale unconstrained optimization problems in our numerical experiments, they do not necessarily satisfy the descent condition (i.e. $g_k^T d_k < 0$ for all k). The second approach aims to generate a descent search direction. Dai and Yuan [4] proposed a conjugate gradient method which generates descent search directions under the Wolfe conditions. Later Yabe and Sakaiwa [17] gave

its variant which also generates descent search directions. Independently of Dai-Yuan's research, Hager and Zhang [12] proposed a conjugate gradient method which generates the descent search direction under the Wolfe conditions. However, these methods depend on line searches to satisfy the descent condition. Conjugate gradient methods which have the both characteristics of the two approaches above have not been proposed.

More recently, Zhang, Zhou and Li. [19–21] proposed three-term conjugate gradient methods which always satisfy the sufficient descent condition:

$$g_k^T d_k \leq -\bar{c}\|g_k\|^2 \quad \text{for all } k, \tag{1.3}$$

for a positive constant $\bar{c}$, independently of line searches. They proposed the modified FR method [20] defined by

$$d_k = -\bar{\theta}_k g_k + \beta^{FR} d_{k-1},$$

where $\bar{\theta}_k = d_{k-1}^T y_{k-1}/\|g_{k-1}\|^2$. Since this search direction satisfies $g_k^T d_k = -\|g_k\|^2$ for all k, it can be rewritten by the three-term form:

$$d_k = -g_k + \beta^{FR} d_{k-1} - \theta_k^{(1)} g_k, \tag{1.4}$$

where $\theta_k^{(1)} = g_k^T d_{k-1}/\|g_{k-1}\|^2$. They also proposed the modified PR method [19] and the modified HS method [21], which are respectively given by

$$d_k = -g_k + \beta^{PR} d_{k-1} - \theta_k^{(2)} y_{k-1}, \tag{1.5}$$

$$d_k = -g_k + \beta^{HS} d_{k-1} - \theta_k^{(3)} y_{k-1}, \tag{1.6}$$

where $\theta_k^{(2)} = g_k^T d_{k-1}/\|g_{k-1}\|^2$ and $\theta_k^{(3)} = g_k^T d_{k-1}/d_{k-1}^T y_{k-1}$. Cheng [2] gave another modified PR method:

$$d_k = -g_k + \beta_k^{PR}\left(I - \frac{g_k g_k^T}{g_k^T g_k}\right) d_{k-1} = -g_k + \beta_k^{PR} d_{k-1} - \beta_k^{PR}\frac{g_k^T d_{k-1}}{g_k^T g_k} g_k. \tag{1.7}$$

They showed their global convergence properties under appropriate line searches. We note that these methods always satisfy $g_k^T d_k = -\|g_k\|^2 < 0$ for all k, which implies the sufficient descent condition with $\bar{c} = 1$.

In this paper, by modifying (1.1), we propose a general form of three-term conjugate gradient methods which always satisfy (1.3), independently of choices of β_k and line searches. Moreover, we establish its global convergence property. The present paper is organized as follows. In Section 2, we construct a general form of three-term conjugate gradient methods which satisfy (1.3), and give a sufficient condition for its global convergence. In Section 3, we propose a specific three-term conjugate gradient method based on the multi-step quasi-Newton method, and prove its global convergence by using the result of Section 2. Finally, in Section 4, some numerical experiments are presented.

2 Three-term conjugate gradient method and its convergence property

In this section, we consider a three-term conjugate gradient method to obtain a descent search direction. Section 2.1 presents a general form of three-term conjugate gradient methods and Section 2.2 shows its global convergence property.

2.1 Three-term conjugate gradient method

We propose a new three-term conjugate gradient method of the form:

$$x_{k+1} = x_k + \alpha_k d_k, \tag{2.1}$$

$$d_k = \begin{cases} -g_k & k = 0, \\ -g_k + \beta_k (g_k^T p_k)^\dagger \{(g_k^T p_k) d_{k-1} - (g_k^T d_{k-1}) p_k\} & k \geq 1, \end{cases} \tag{2.2}$$

where $\beta_k \in \boldsymbol{R}$ is a parameter, $p_k \in \boldsymbol{R}^n$ is any vector and

$$a^\dagger = \begin{cases} \dfrac{1}{a} & a \neq 0, \\ 0 & a = 0. \end{cases}$$

We emphasize that the method (2.1)–(2.2) always satisfies

$$g_k^T d_k = -\|g_k\|^2, \tag{2.3}$$

independently of choices of p_k and line searches. It means that the sufficient descent condition (1.3) holds with $\bar{c} = 1$.

Note that (2.2) can be rewritten by

$$d_k = \begin{cases} -g_k & \text{if } k = 0 \text{ or } g_k^T p_k = 0, \\ -g_k + \beta_k d_{k-1} - \beta_k \dfrac{g_k^T d_{k-1}}{g_k^T p_k} p_k & \text{otherwise.} \end{cases} \tag{2.4}$$

Accordingly, if $g_k^T p_k \neq 0$ is satisfied, the form (2.2) becomes

$$d_k = -g_k + \beta_k \left(I - \frac{p_k g_k^T}{g_k^T p_k} \right) d_{k-1}. \tag{2.5}$$

The matrix $(I - p_k g_k^T / g_k^T p_k)$ is a projection matrix into the orthogonal complement of $\mathrm{Span}\{g_k\}$ along $\mathrm{Span}\{p_k\}$. Especially, if we choose $p_k = g_k$, then $(I - g_k g_k^T / \|g_k\|^2)$ is an orthogonal projection matrix.

If we use the exact line search and p_k such that $g_k^T p_k \neq 0$, then our method (2.4) becomes the nonlinear conjugate gradient method (1.1). The most simple choices are

$p_k = g_k$ and $p_k = y_{k-1}$. On the other hand, if we choose $p_k = d_{k-1}$, then (2.2) implies $d_k = -g_k$ for all k.

We should note that the present method includes the three-term conjugate gradient methods proposed by Zhang et al. [19–21]. The method (2.1)–(2.2) with $\beta_k = \beta_k^{FR}$ and $p_k = g_k$ becomes the method by [20] (see (1.4)), and, if $g_k^T y_{k-1} \neq 0$, the method (2.1)–(2.2) with $\beta_k = \beta_k^{PR}$ and $p_k = y_{k-1}$ becomes the method by [19] (see (1.5)). If $g_k^T y_{k-1} \neq 0$, the method (2.1)–(2.2) with $\beta_k = \beta_k^{HS}$ and $p_k = y_{k-1}$ becomes the method by [21] (see (1.6)). In addition, the method (2.1)–(2.2) with $\beta_k = \beta_k^{PR}$ and $p_k = g_k$ becomes the method by [2] (see (1.7)).

2.2 Convergence analysis

In order to establish the global convergence property, we make the following standard assumptions for the objective function.

Assumption 2.1.

1. *The level set $\mathcal{L} = \{x | f(x) \leq f(x_0)\}$ at x_0 is bounded, namely, there exists a constant $\widehat{a} > 0$ such that*

$$\|x\| \leq \widehat{a} \qquad \textit{for all} \quad x \in \mathcal{L}. \tag{2.6}$$

2. *In some neighborhood $\mathcal{N}$ of $\mathcal{L}$, f is continuously differentiable, and its gradient is Lipschitz continuous with Lipschitz constant $L > 0$, i.e.*

$$\|g(u) - g(v)\| \leq L\|u - v\| \qquad \textit{for all } u, v \in \mathcal{N}.$$

Assumption 2.1 implies that there exists a positive constant $\widehat{\gamma}$ such that

$$\|g(x)\| \leq \widehat{\gamma}, \qquad \text{for all} \quad x \in \mathcal{L}. \tag{2.7}$$

In the line search, we require α_k to satisfy the Wolfe conditions:

$$f(x_k) - f(x_k + \alpha_k d_k) \geq -\delta \alpha_k g_k^T d_k, \tag{2.8}$$

$$g(x_k + \alpha_k d_k)^T d_k \geq \sigma g_k^T d_k \tag{2.9}$$

where $0 < \delta < \sigma < 1$, or the strong Wolfe conditions: (2.8) and

$$|g(x_k + \alpha_k d_k)^T d_k| \leq \sigma |g_k^T d_k| \tag{2.10}$$

where $0 < \delta < \sigma < 1$.

In the rest of this section, we assume $g_k \neq 0$ for all k, otherwise a stationary point has been found.

Under Assumption 2.1, we have the following well-known lemma which was proved by Zoutendijk (see [16]). The following lemma is the result for general iterative methods with the Wolfe condition (2.8) and (2.9).

Lemma 2.1. *Suppose that Assumption* 2.1 *is satisfied. Consider any method in the form* (2.1)*, where d_k is a descent search direction and α_k satisfies the Wolfe conditions* (2.8) *and* (2.9)*. Then*

$$\sum_{k=0}^{\infty} \frac{(g_k^T d_k)^2}{\|d_k\|^2} < \infty.$$

Using Lemma 2.1, we have the following lemma, which is useful in showing the global convergence of our method.

Lemma 2.2. *Suppose that Assumption* 2.1 *is satisfied. Consider the method* (2.1)–(2.2)*, where α_k satisfies the Wolfe conditions* (2.8) *and* (2.9)*. If*

$$\sum_{k=0}^{\infty} \frac{1}{\|d_k\|^2} = \infty \tag{2.11}$$

holds, then the following holds:

$$\liminf_{k\to\infty} \|g_k\| = 0. \tag{2.12}$$

Proof. If (2.12) is not true, there exists a constant $\varepsilon > 0$ such that

$$\|g_k\| \geq \varepsilon$$

for all k. Therefore from (2.3) and (2.11), we have

$$\sum_{k=0}^{\infty} \frac{\varepsilon^4}{\|d_k\|^2} \leq \sum_{k=0}^{\infty} \frac{\|g_k\|^4}{\|d_k\|^2} = \sum_{k=0}^{\infty} \frac{(g_k^T d_k)^2}{\|d_k\|^2} = \infty.$$

Since this contradicts Lemma 2.1, the proof is complete. □

Now we consider a sufficient condition to establish the global convergence property of the method (2.1)–(2.2). First, we estimate the norm of the search direction of the proposed method. If $g_k^T p_k = 0$, the following relation

$$\|d_k\| = \|g_k\| \tag{2.13}$$

holds. Otherwise, by squaring both sides of (2.5), we have from the orthogonality of g_k and $(I - p_k g_k^T / g_k^T p_k) d_{k-1}$

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

and hence it follows from $\left\| I - \dfrac{p_k g_k^T}{g_k^T p_k} \right\| = \dfrac{\|g_k\|\|p_k\|}{|g_k^T p_k|}$ that

$$\|d_k\|^2 \leq \beta_k^2 \left(\frac{\|g_k\|\|p_k\|}{|g_k^T p_k|} \right)^2 \|d_{k-1}\|^2 + \|g_k\|^2. \tag{2.14}$$

Therefore, by defining

$$\psi_k = \beta_k \|g_k\|\|p_k\|(g_k^T p_k)^\dagger, \tag{2.15}$$

relations (2.13) and (2.14) yield

$$\|d_k\|^2 \leq \psi_k^2 \|d_{k-1}\|^2 + \|g_k\|^2 \tag{2.16}$$

for all k.

For standard conjugate gradient methods, Gilbert and Nocedal [10] derived *Property* $(*)$, which shows that β_k will be small when the step s_{k-1} is small (see also Dai and Liao [3]). The following property corresponds with *Property* $(*)$ except for using ψ_k instead of β_k.

Property A. *Consider the method* (2.1)–(2.2). *Assume that there exists a positive constant* ε *such that* $\varepsilon \leq \|g_k\|$ *holds for all* k. *Then we say that the method has Property A if there exist constants* $b > 1$ *and* $\xi > 0$ *such that for all* k*:*

$$|\psi_k| \leq b, \tag{2.17}$$

and

$$\|s_{k-1}\| \leq \xi \implies |\psi_k| \leq \frac{1}{b}. \tag{2.18}$$

We note that (2.17) implies that if there exists a positive constant ε such that $\varepsilon \leq \|g_k\|$ for all k, then

$$|\beta_k| \ \|p_k\| \ |g_k^T p_k|^\dagger \leq c \tag{2.19}$$

holds with $c = b/\varepsilon$.

The next lemma corresponds to Lemma 3.4 in Dai and Liao [3].

Lemma 2.3. *Suppose that Assumption* 2.1 *is satisfied. Consider the method* (2.1)–(2.2), *where* α_k *satisfies the strong Wolfe conditions* (2.8) *and* (2.10). *Assume that there exists a positive constant* ε *such that the following relation holds* $\varepsilon \leq \|g_k\|$ *holds for all* k. *If the method has Property A and* $\beta_k \geq 0$ *holds, then* $d_k \neq 0$ *and the following relation holds*

$$\sum_{k=0}^{\infty} \|u_k - u_{k-1}\|^2 < \infty,$$

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

Proof. Since $d_k \neq 0$ follows from (2.3) and $\varepsilon \leq \|g_k\|$, the vector u_k is well-defined. Using Lemma 2.2 and $\varepsilon \leq \|g_k\|$, we have

$$\sum_{k=0}^{\infty} \frac{1}{\|d_k\|^2} < \infty. \tag{2.20}$$

By defining

$$v_k = -\left(g_k + \beta_k (g_k^T p_k)^\dagger (g_k^T d_{k-1}) p_k\right) \frac{1}{\|d_k\|} \quad \text{and} \quad \eta_k = \beta_k (g_k^T p_k)^\dagger (g_k^T p_k) \frac{\|d_{k-1}\|}{\|d_k\|},$$

equation (2.2) is written as

$$u_k = v_k + \eta_k u_{k-1}.$$

Then we have from the fact that $\|u_k\| = \|u_{k-1}\| = 1$,

$$\|v_k\| = \|u_k - \eta_k u_{k-1}\| = \|\eta_k u_k - u_{k-1}\|. \tag{2.21}$$

It follows from $\beta_k \geq 0$ and (2.21) that

$$\begin{aligned} \|u_k - u_{k-1}\| &\leq (1+\eta_k)\|u_k - u_{k-1}\| \\ &= \|u_k - \eta_k u_{k-1} + \eta_k u_k - u_{k-1}\| \\ &\leq \|u_k - \eta_k u_{k-1}\| + \|\eta_k u_k - u_{k-1}\| \\ &= 2\|v_k\|. \end{aligned} \tag{2.22}$$

From (2.19), we have

$$\beta_k |g_k^T p_k|^\dagger \|p_k\| \leq c$$

for all k. Therefore by (2.10), (2.3), (2.7) and (2.19), we have

$$\begin{aligned} \beta_k |g_k^T d_{k-1}| |g_k^T p_k|^\dagger \|p_k\| &\leq \sigma \beta_k |g_{k-1}^T d_{k-1}| |g_k^T p_k|^\dagger \|p_k\| \\ &= \sigma \beta_k |g_k^T p_k|^\dagger \|p_k\| \|g_{k-1}\|^2 \\ &\leq \sigma c \widehat{\gamma}^2. \end{aligned}$$

Thus (2.22), (2.7) and (2.20) yield

$$\begin{aligned} \sum_{k=0}^{\infty} \|u_k - u_{k-1}\|^2 &\leq 4 \sum_{k=0}^{\infty} \|v_k\|^2 \\ &\leq 4 \sum_{k=0}^{\infty} (\|g_k\| + \beta_k |g_k^T d_{k-1}| |g_k^T p_k|^\dagger \|p_k\|)^2 \cdot \frac{1}{\|d_k\|^2} \\ &\leq 4(\widehat{\gamma} + \sigma \widehat{\gamma}^2 c)^2 \sum_{k=0}^{\infty} \frac{1}{\|d_k\|^2} \\ &< \infty. \end{aligned}$$

Therefore the lemma is proved. □

Let $\boldsymbol{N}$ denote the set of all positive integers. For $\lambda > 0$ and a positive integer Δ, we define the set of indices:

$$\mathcal{K}^{\lambda}_{k,\Delta} := \{i \in \boldsymbol{N} \mid k \leq i \leq k + \Delta - 1,\ \|s_{i-1}\| > \lambda\}.$$

Let $|\mathcal{K}^{\lambda}_{k,\Delta}|$ denote the number of elements in $\mathcal{K}^{\lambda}_{k,\Delta}$. The following lemma shows that if the gradients are bounded away from zero and (2.17)–(2.18) hold, then a certain fraction of the steps cannot be too small. This lemma corresponds to [3, Lemma 3.5] and [10, Lemma 4.2].

Lemma 2.4. *Suppose that all assumptions of Lemma* 2.3 *hold. If the method has Property A, then there exists $\lambda > 0$ such that, for any $\Delta \in \boldsymbol{N}$ and any index k_0, there is an index $\widehat{k} \geq k_0$ such that*

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

Proof. We prove this lemma by contradiction. Assume that for any $\lambda > 0$, there exist $\Delta \in \boldsymbol{N}$ and k_0 such that

$$|\mathcal{K}^{\lambda}_{k,\Delta}| \leq \frac{\Delta}{2} \tag{2.23}$$

for all $k \geq k_0$. Let $b > 1$ and $\xi > 0$ be given in Property A. For $\lambda = \xi$, we choose Δ and k_0 such that (2.23) holds. Then from (2.17), (2.18) and (2.23), we have

$$\prod_{k=k_0+i\Delta+1}^{k_0+(i+1)\Delta} |\psi_k| = \prod_{k \in \mathcal{K}^{\lambda}_{k',\Delta}} |\psi_k| \prod_{k \notin \mathcal{K}^{\lambda}_{k',\Delta}} |\psi_k| \leq b^{\Delta/2} \left(\frac{1}{b}\right)^{\Delta/2} = 1 \quad \text{for any } i \geq 0, \tag{2.24}$$

where $k' = k_0 + i\Delta + 1$. If $\psi_k = 0$ holds, then the search direction becomes $d_k = -g_k$. Therefore, if ψ_k equals zero infinitely many times, the search direction becomes the steepest descent direction infinitely many times, which implies that $\liminf_{k\to\infty} \|g_k\| = 0$. Otherwise, we have $\psi_k \neq 0$ for k sufficiently large. Therefore we assume without loss of generality that

$$\psi_k \neq 0 \tag{2.25}$$

for all $k \geq 1$. It follows from (2.24) that

$$\prod_{j=2}^{k_0+i\Delta} |\psi_j| = \left(\prod_{j=2}^{k_0} |\psi_j|\right) \cdot \left(\prod_{j=k_0+1}^{k_0+\Delta} |\psi_j|\right) \cdots \left(\prod_{j=k_0+(i-1)\Delta+1}^{k_0+i\Delta} |\psi_j|\right) \leq \prod_{j=2}^{k_0} |\psi_j| \qquad \text{for any } i \geq 0,$$

which implies by (2.25)

$$\prod_{j=2}^{k_0+i\Delta} \psi_j^{-2} \geq \prod_{j=2}^{k_0} \psi_j^{-2} \qquad \text{for any } i \geq 0. \tag{2.26}$$

By summing (2.26), we have

$$\sum_{k=2}^{\infty}\prod_{j=2}^{k}\psi_j^{-2} \geq \sum_{i=0}^{\infty}\prod_{j=2}^{k_0+i\Delta}\psi_j^{-2} \geq \sum_{i=0}^{\infty}\prod_{j=2}^{k_0}\psi_j^{-2} = \infty. \tag{2.27}$$

From Lemma 2.1 and the assumption $0 < \varepsilon \leq \|g_k\|$, we have

$$\sum_{k=0}^{\infty}\frac{(g_k^T d_k)^2}{\|d_k\|^2\|g_k\|^2} \leq \sum_{k=0}^{\infty}\frac{(g_k^T d_k)^2}{\varepsilon^2\|d_k\|^2} < \infty.$$

Thus there exist a integer j_0 and a constant $c_2 > 0$ such that

$$\prod_{j=j_0}^{k}\left(1 - \frac{(g_j^T d_j)^2}{\|g_j\|^2\|d_j\|^2}\right) \geq c_2 \tag{2.28}$$

holds for any $k \geq j_0$. On the other hand, (2.16) and (2.3) yield

$$\|d_k\|^2 \leq \psi_k^2\|d_{k-1}\|^2 + \|g_k\|^2 = \psi_k^2\|d_{k-1}\|^2 + \frac{(g_k^T d_k)^2}{\|g_k\|^2},$$

and hence it follows from (2.28) that

$$\begin{aligned}
\|d_k\|^2 &\leq \left(1 - \frac{(g_k^T d_k)^2}{\|g_k\|^2\|d_k\|^2}\right)^{-1}\psi_k^2\|d_{k-1}\|^2 \\
&\leq \cdots \\
&\leq \prod_{j=j_0}^{k}\left(1 - \frac{(g_j^T d_j)^2}{\|g_j\|^2\|d_j\|^2}\right)^{-1}\left(\prod_{j=j_0}^{k}\psi_j^2\right)\|d_{j_0-1}\|^2 \\
&\leq \frac{\|d_{j_0-1}\|^2}{c_2}\left(\prod_{j=2}^{j_0-1}\psi_j^{-2}\right)\left(\prod_{j=2}^{k}\psi_j^2\right) \\
&\leq c_3\prod_{j=2}^{k}\psi_j^2
\end{aligned}$$

for all $k \geq j_0$, where $c_3 = \dfrac{\|d_{j_0-1}\|^2}{c_2}\displaystyle\prod_{j=2}^{j_0-1}\psi_j^{-2}$. Note that c_3 is a positive constant, because j_0 is a fixed integer in (2.28). Therefore, we get by (2.27)

$$\sum_{k=j_0}^{\infty}\frac{1}{\|d_k\|^2} \geq \frac{1}{c_3}\sum_{k=j_0}^{\infty}\prod_{j=2}^{k}\psi_j^{-2} = \infty.$$

It follows from Lemma 2.2 that $\liminf_{k\to\infty}\|g_k\| = 0$ holds. Since this contradicts the assumption $0 < \varepsilon \leq \|g_k\|$, we obtain the desired result. □

Now we can give a sufficient condition for the global convergence of the method (2.1)–(2.2) by using Lemmas 2.3 and 2.4 and *Property* A. This theorem corresponds to Theorem 3.6 in [3] and the proof is exactly same as that of Theorem 3.6, but we write it for the readability.

Theorem 2.1. *Consider the method* (2.1)–(2.2) *that satisfies the following conditions:*

(C1) $\beta_k \geq 0$ *for all* k,

(C2) *Property A holds.*

Assume that α_k *satisfies the strong Wolfe conditions* (2.8) *and* (2.10). *If Assumption* 2.1 *holds, then the method converges in the sense that* $\liminf_{k\to\infty} \|g_k\| = 0$.

Proof. Since we prove this theorem by contradiction, we assume that there exists ε such that $0 < \varepsilon \leq \|g_k\|$ holds for all k. Then Lemmas 2.3 and 2.4 hold. From the definition of u_k, we have for any l and k with $l \geq k$,

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

It follows from this relation, the fact $\|u_{k-1}\| = 1$ and (2.6) that

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

which implies that

$$2\widehat{a} \geq \sum_{i=k}^{l} \|s_{i-1}\| (1 - \|u_{i-1} - u_{k-1}\|). \tag{2.29}$$

Let $\lambda > 0$ be given by Lemma 2.4 and define $\Delta = \lceil 8\widehat{a}/\lambda \rceil$ to be the smallest integer not less than $8\widehat{a}/\lambda$. By Lemma 2.3, we can find an index k_0 such that

$$\sum_{i=k_0}^{\infty} \|u_i - u_{i-1}\|^2 \leq \frac{1}{4\Delta}. \tag{2.30}$$

For Δ and k_0 defined above, Lemma 2.4 gives an index $k \geq k_0$ such that

$$|\mathcal{K}^{\lambda}_{k,\Delta}| > \frac{\Delta}{2}. \tag{2.31}$$

By (2.30) and the fact that $\|v\|_1 \leq \sqrt{n}\|v\|$ for any vector $v \in \boldsymbol{R}^n$, we have

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

for any i ($k \leq i \leq k+\Delta-1$). Therefore it follows from (2.29) with $l = k+\Delta-1$, the definition of $\mathcal{K}^{\lambda}_{k,\Delta}$ and (2.31) that

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

Thus we get $\Delta < 8\widehat{a}/\lambda$, which contradicts the definition of Δ. Therefore, the theorem is true. □

Theorem 2.1 plays an important role to establish global convergence properties of various kinds of three-term conjugate gradient methods. For instance, we obtain the following convergence results as a corollary of Theorem 2.1.

Corollary 2.1. *Suppose that Assumption* 2.1 *is satisfied. Consider the method* (2.1)–(2.2)*, where α_k satisfies the strong Wolfe conditions* (2.8) *and* (2.10)*. Then the following hold* :

(i) *The method with $\beta_k = \beta_k^{PR+}$ and $p_k = y_{k-1}$ (or $p_k = g_k$) converges in the sense that* $\liminf_{k\to\infty} \|g_k\| = 0$.

(ii) *The method with $\beta_k = \beta_k^{HS+} \equiv \max\{\beta_k^{HS},\ 0\}$ and $p_k = y_{k-1}$ (or $p_k = g_k$) converges in the sense that* $\liminf_{k\to\infty} \|g_k\| = 0$.

Proof. In each case, since $\beta_k \geq 0$ holds, condition (C1) of Theorem 2.1 is satisfied. It suffices to prove that (C2) holds in each case. Accordingly, we assume that there exists ε such that $0 < \varepsilon \leq \|g_k\|$ holds for all k.

(i) It follows from $\beta_k = \beta_k^{PR+}$ and $p_k = y_{k-1}$ that

$$\begin{aligned} |\psi_k| &= \left|\max\left\{\frac{g_k^T y_{k-1}}{\|g_{k-1}\|^2},\ 0\right\}\|g_k\|\|y_{k-1}\|(g_k^T y_{k-1})^{\dagger}\right| \\ &\leq \frac{\|g_k\|\|y_{k-1}\|}{\|g_{k-1}\|^2} \\ &\leq \frac{2L\widehat{\gamma}\widehat{a}}{\varepsilon^2} = \bar{b}. \end{aligned}$$

If $\bar{b}$ is not greater than 1, define $b = 1+\bar{b}$, so that $b > 1$ and $b \geq \bar{b}$, else define $b = \bar{b}$. Now, we define $\xi = \varepsilon^2/(L\widehat{\gamma}b)$. If $\|s_{k-1}\| \leq \xi$, we have

$$|\psi_k| \leq \frac{L\widehat{\gamma}\|s_{k-1}\|}{\varepsilon^2} \leq \frac{1}{b},$$

which implies that *Property* A holds.

Next we consider the case of $\beta_k = \beta_k^{PR+}$ and $p_k = g_k$. Then we have

$$|\psi_k| = \left|\max\left\{\frac{g_k^T y_{k-1}}{\|g_{k-1}\|^2},\ 0\right\}\right| \leq \frac{\|g_k\|\|y_{k-1}\|}{\|g_{k-1}\|^2},$$

and hence we can prove that Property A holds for the case $p_k = g_k$ in the same way as for the case $p_k = y_{k-1}$. Therefore the proof of (i) is complete.

(ii) It follows from $\beta_k = \beta_k^{HS+}$, $p_k = y_{k-1}$ and (2.10) that

$$\begin{aligned} |\psi_k| &= \left| \max\left\{ \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}},\ 0 \right\} \|g_k\| \|y_{k-1}\| (g_k^T y_{k-1})^\dagger \right| \\ &\leq \frac{\|g_k\| \|y_{k-1}\|}{(1-\sigma)\|g_{k-1}\|^2} \\ &\leq \frac{2L\widehat{\gamma}\widehat{a}}{(1-\sigma)\varepsilon^2} = \bar{b}. \end{aligned}$$

If $\bar{b}$ is not greater than 1, define $b = 1 + \bar{b}$, so that $b > 1$ and $b \geq \bar{b}$, else define $b = \bar{b}$. Now, we define $\xi = (1-\sigma)\varepsilon^2/(L\widehat{\gamma}b)$. If $\|s_{k-1}\| \leq \xi$, we have

$$|\psi_k| \leq \frac{L\widehat{\gamma}\|s_{k-1}\|}{(1-\sigma)\varepsilon^2} \leq \frac{1}{b},$$

which implies that *Property* A holds.

Next we consider the case of $\beta_k = \beta_k^{HS+}$ and $p_k = g_k$. Then we have

$$|\psi_k| = \left| \max\left\{ \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}},\ 0 \right\} \right| \leq \frac{\|g_k\| \|y_{k-1}\|}{(1-\sigma)\|g_{k-1}\|^2},$$

and hence we can prove that Property A holds for the case $p_k = g_k$ in the same way as for the case $p_k = y_{k-1}$. Therefore the proof of (ii) is complete. □

3 Three-term conjugate gradient method based on multi-step quasi-Newton method

In this section, we propose a three-term conjugate gradient method based on the multi-step quasi-Newton method. In order to introduce a new choice of β_k and p_k, let us briefly refer to the multi-step quasi-Newton method by Ford and Moghrabi [7, 8]. The search direction d_k of their method is given by $d_k = -H_k g_k$, where H_k approximates the inverse Hessian of the objective function and it is updated by the multi-step BFGS formula:

$$H_k = \left(I - \frac{\widehat{w}_{k-1}\widehat{r}_{k-1}^T}{\widehat{r}_{k-1}^T \widehat{w}_{k-1}} \right)^T H_{k-1} \left(I - \frac{\widehat{w}_{k-1}\widehat{r}_{k-1}^T}{\widehat{r}_{k-1}^T \widehat{w}_{k-1}} \right) + \frac{\widehat{r}_{k-1}\widehat{r}_{k-1}^T}{\widehat{r}_{k-1}^T \widehat{w}_{k-1}}$$

and

$$\widehat{r}_{k-1} = s_{k-1} - \widehat{\phi}_k s_{k-2}, \quad \widehat{w}_{k-1} = y_{k-1} - \widehat{\phi}_k y_{k-2} \quad \text{and} \quad \widehat{\phi}_k = \frac{g_k^T s_{k-1}}{g_k^T s_{k-2}}.$$

Incorporating a parameter $t_k \geq 0$ into $\widehat{w}_k$, we redefine

$$\widehat{w}_{k-1} = y_{k-1} - t_k \widehat{\phi}_k y_{k-2}.$$

If $H_{k-1} \equiv I$, then the above multi-step BFGS method becomes the multi-step limited-memory BFGS method, where the memory equals 1. Since $g_k^T \widehat{r}_{k-1} = 0$, the search direction d_k is given by

$$\begin{aligned} d_k &= -\left(I - \frac{\widehat{w}_{k-1}\widehat{r}_{k-1}^T}{\widehat{r}_{k-1}^T \widehat{w}_{k-1}}\right)^T \left(I - \frac{\widehat{w}_{k-1}\widehat{r}_{k-1}^T}{\widehat{r}_{k-1}^T \widehat{w}_{k-1}}\right) g_k - \frac{\widehat{r}_{k-1}\widehat{r}_{k-1}^T}{\widehat{r}_{k-1}^T \widehat{w}_{k-1}} g_k \\ &= -g_k + \frac{g_k^T \widehat{w}_{k-1}}{\widehat{r}_{k-1}^T \widehat{w}_{k-1}} \widehat{r}_{k-1}. \end{aligned}$$

This search direction can be rewritten as the form:

$$d_k = -g_k + \beta_k^{MS} d_{k-1} - \beta_k^{MS} \phi_k d_{k-2}, \tag{3.1}$$

where

$$\phi_k = \frac{g_k^T d_{k-1}}{g_k^T d_{k-2}}, \tag{3.2}$$

$$r_{k-1} = d_{k-1} - \phi_k d_{k-2}, \tag{3.3}$$

$$w_{k-1} = y_{k-1} - t_k \frac{\alpha_{k-1}}{\alpha_{k-2}} \phi_k y_{k-2}, \tag{3.4}$$

and

$$\beta_k^{MS} = \frac{g_k^T w_{k-1}}{r_{k-1}^T w_{k-1}}. \tag{3.5}$$

Since (3.2) cannot be defined for the case $g_k^T d_{k-2} = 0$, we replace (3.2) with

$$\phi_k = g_k^T d_{k-1} (g_k^T d_{k-2})^\dagger \tag{3.6}$$

as a safeguard, and by considering (2.2), the direction (3.1) can be rewritten by

$$d_k = -g_k + \beta_k^{MS} (g_k^T d_{k-2})^\dagger \{(g_k^T d_{k-2}) d_{k-1} - (g_k^T d_{k-1}) d_{k-2}\}. \tag{3.7}$$

We note that this corresponds to the three-term conjugate gradient method (2.2) with $p_k = d_{k-2}$ and $\beta_k = \beta_k^{MS}$. In addition, in order to establish the global convergence of our method, we modify (3.5) as follows:

$$\beta_k^{MS+} = \max\left\{\frac{g_k^T w_{k-1}}{r_{k-1}^T w_{k-1}},\ 0\right\}. \tag{3.8}$$

If we use the exact line search, then $\phi_k = 0$ and $\beta_k^{MS+} = \max\{g_k^T y_{k-1}/d_{k-1}^T y_{k-1},\ 0\}$, and hence our method reduces to a modified HS (HS+) method.

Now we consider the global convergence of the proposed method. For this purpose, we make the following additional assumptions.

Assumption 3.1.

1. *Assume that there exists a positive constant* τ_1 *such that, for all* k,

$$\|g_k\|\|d_{k-2}\||g_k^T d_{k-2}|^\dagger \leq \tau_1. \tag{3.9}$$

2. *Assume that there exists a positive constant* τ_2 *such that, for all* k,

$$|g_{k-1}^T r_{k-1}| \geq \tau_2 |g_{k-1}^T d_{k-1}|. \tag{3.10}$$

3. *Assume that there exists a constant* τ_3 *that satisfies* $0 \leq \tau_3 < 1$ *and*

$$t_k \frac{\alpha_{k-1}}{\alpha_{k-2}} |\phi_k| \leq \tau_3 \min\left\{|g_k^T y_{k-1}||g_k^T y_{k-2}|^\dagger,\ |r_{k-1}^T y_{k-1}||r_{k-1}^T y_{k-2}|^\dagger\right\} \quad \textit{for all } k. \tag{3.11}$$

Using Theorem 2.1, we obtain the following global convergence property.

Theorem 3.1. *Suppose that Assumptions* 2.1 *and* 3.1 *are satisfied. Consider the method* (2.1)–(2.2) *with* (3.8) *and* $p_k = d_{k-2}$. *Assume that* α_k *satisfies the strong Wolfe conditions* (2.8) *and* (2.10). *Then the method converges in the sense that* $\liminf_{k\to\infty} \|g_k\| = 0$.

Proof. By (3.8), $\beta_k \geq 0$ clearly holds. So we only prove that the proposed method satisfies condition (C2) of Theorem 2.1. To this end, we assume that there exists a constant $\varepsilon > 0$ such that

$$\|g_k\| \geq \varepsilon \qquad \text{for all } k.$$

It follows from (3.4) and (3.11) that

$$\begin{aligned} |g_k^T w_{k-1}| &\leq |g_k^T y_{k-1}| + t_k \frac{\alpha_{k-1}}{\alpha_{k-2}} |\phi_k g_k^T y_{k-2}| \\ &\leq (1+\tau_3)|g_k^T y_{k-1}| \\ &\leq (1+\tau_3) L \|g_k\|\|s_{k-1}\|. \end{aligned} \tag{3.12}$$

By (3.4), (3.11) and the fact $g_k^T r_{k-1} = 0$, we have

$$\begin{aligned} |r_{k-1}^T w_{k-1}| &\geq |r_{k-1}^T y_{k-1}| - t_k \frac{\alpha_{k-1}}{\alpha_{k-2}} |\phi_k r_{k-1}^T y_{k-2}| \\ &\geq (1-\tau_3)|r_{k-1}^T y_{k-1}| \\ &= (1-\tau_3)|g_{k-1}^T r_{k-1}|. \end{aligned} \tag{3.13}$$

It follows from (3.10) and (2.3) that

$$|g_{k-1}^T r_{k-1}| \geq \tau_2 |g_{k-1}^T d_{k-1}| = \tau_2 \|g_{k-1}\|^2.$$

Therefore (3.13) yields

$$|r_{k-1}^T w_{k-1}| \geq \tau_2 (1-\tau_3) \|g_{k-1}\|^2. \tag{3.14}$$

By (3.8), (3.12) and (3.14), we have

$$\begin{aligned}\beta_k^{MS+} &\leq \frac{|g_k^T w_{k-1}|}{|r_{k-1}^T w_{k-1}|} \\ &\leq \frac{(1+\tau_3)L\|g_k\|\|s_{k-1}\|}{\tau_2(1-\tau_3)\|g_{k-1}\|^2} \\ &\leq \frac{(1+\tau_3)L\widehat{\gamma}\|s_{k-1}\|}{\tau_2(1-\tau_3)\varepsilon^2}. \end{aligned} \tag{3.15}$$

Since the choice $p_k = d_{k-2}$ in (2.2) and (2.15) yield

$$\psi_k = \beta_k^{MS+}\|g_k\|\|p_k\|(g_k^T p_k)^\dagger = \beta_k^{MS+}\|g_k\|\|d_{k-2}\|(g_k^T d_{k-2})^\dagger,$$

(3.15) and (3.9) give

$$\begin{aligned}|\psi_k| &\leq \frac{\tau_1(1+\tau_3)L\widehat{\gamma}\|s_{k-1}\|}{\tau_2(1-\tau_3)\varepsilon^2} \\ &\leq \frac{2\tau_1(1+\tau_3)L\widehat{a\gamma}}{\tau_2(1-\tau_3)\varepsilon^2} = \bar{b}. \end{aligned}$$

We define $b = 1 + \bar{b}$ and

$$\xi = \frac{\tau_2(1-\tau_3)\varepsilon^2}{\tau_1(1+\tau_3)L\widehat{\gamma}b}.$$

Then, if $\|s_{k-1}\| \leq \xi$, we have

$$|\psi_k| \leq \frac{\tau_1(1+\tau_3)L\widehat{\gamma}\xi}{\tau_2(1-\tau_3)\varepsilon^2} \leq \frac{1}{b}.$$

Therefore, Property A holds. Thus from Theorem 2.1, the theorem is true. □

If $g_k^T d_{k-2}$ equals zero infinitely many times, the search direction becomes the steepest descent direction infinitely many times, which implies that $\liminf_{k\to\infty}\|g_k\| = 0$. So it is sufficient to consider the case $g_k^T d_{k-2} \neq 0$ for all k sufficiently large. We note that assumption (3.9) yields

$$|g_{k-1}^T r_{k-1}| \geq |g_{k-1}^T d_{k-1}| - |\phi_k||g_{k-1}^T d_{k-2}| \geq \left(1 - \frac{\tau_1\sigma^2\|g_{k-2}\|^2}{\|g_k\|\|d_{k-2}\|}\right)|g_{k-1}^T d_{k-1}|.$$

If σ is chosen to be sufficiently small and $\dfrac{\|g_{k-2}\|^2}{\|g_k\|\|d_{k-2}\|}$ is bounded, then (3.10) holds. If $\dfrac{\|g_{k-2}\|^2}{\|g_k\|\|d_{k-2}\|}$ is unbounded, then $\liminf_{k\to\infty}\|g_k\|\|d_{k-2}\| = 0$ holds from (2.7), and it implies $\liminf_{k\to\infty}\|g_k\| = 0$ or $\liminf_{k\to\infty}\|d_k\| = 0$. By Lemma 2.2, $\liminf_{k\to\infty}\|d_k\| = 0$ leads $\liminf_{k\to\infty}\|g_k\| = 0$, which is the desired result. Thus if (3.9) holds, then assumption (3.10) is not unreasonable. In our numerical experiments of Section 4, if (3.9) with $\tau_1 = 10^{15}$ does not hold, then we use the steepest descent direction. However, such a case did not occur in our numerical results.

4 Numerical results

In this section, we report some numerical results. We investigated numerical performance of the proposed algorithms on 79 problems in the CUTEr [1,11] library. Except for 8 problems, we used the default value of parameter included in each problem. Dimensions of the test problems lay on the range from 2 to 10000. We examined the following methods, where we denote CG and 3TCG by conjugate gradient methods and three-term conjugate gradient methods, respectively:

1. CG-DESCENT : CG by Hager and Zhang [12, 14]
2. HS : CG with $\beta_k = \beta^{HS}$
3. PR+ : CG with $\beta_k = \beta^{PR+}$
4. FR : CG with $\beta_k = \beta^{FR}$
5. DY : CG with $\beta_k = \beta^{DY}$
6. 3HS+ : 3TCG with $\beta_k = \beta^{HS+}$ and $p_k = y_{k-1}$
7. 3PR+ : 3TCG with $\beta_k = \beta^{PR+}$ and $p_k = y_{k-1}$
8. 3MS+ : 3TCG with $\beta_k = \beta^{MS+}$, $p_k = d_{k-2}$ and $t_k = 1$.

In order to compare three-term conjugate gradient methods with conjugate gradient methods, we coded HS, PR+, FR, DY, 3HS+, 3PR+ and 3MS+ by using the software package CG-DESCENT developed by Hager and Zhang [12, 14], in which the line search and parameters were set as default. Since CG methods except for CG-DESCENT do not generally generate a descent search direction, we restart as the direction of steepest descent when a descent search direction is not produced. As stated in Section 3, for 3MS+, if $\|g_k\|\|d_{k-2}\||g_k^T d_{k-2}|^\dagger \leq 10^{15}$, then we use the restart technique. However, such a case did not occur in our numerical experiments. We recognize that these numerical experiments are against 3HS+, 3PR+ and 3MS+, because the code CG-DESCENT is suitably tuned to the CG method by Hager and Zhang. Computational costs of 3HS+, 3PR+ and 3MS+ may be reduced by effectively tuning the code, but it is beyond the scope of this paper. In the line search, we used the Wolfe conditions (2.8) and (2.9). Although we also tested 3HS+, 3PR+ and 3MS+ with the strong Wolfe conditions (2.8) and (2.10) for some problems, the results are not so different from results of the methods using the Wolfe conditions.

As stated in Section 2, if $g_k^T y_{k-1} \neq 0$, the search directions of 3HS+ and 3PR+ become those given by Zhang et al. [19, 21]. However their line search is not same as ours, and hence 3HS+ and 3PR+ are different from the algorithms by Zhang et al.

The stopping condition was

$$\|g_k\|_1 \leq 10^{-6}.$$

We stopped the algorithm if CPU time exceeds 500(sec) or if a numerical overflow occurs while the method tries to compute $f(x_k + \alpha_k d_k)$. However the second case did not occur.

We adopt the performance profiles by Dolan and Moré [5] to compare the performance among the tested methods. Figure 1–4 are the performance profile measured by CPU time,

the number of iterations, the number of function evaluations and the number of gradient evaluations, respectively. In Figure 1, CG-DESCENT performed well from the viewpoint of CPU time. Since the code was not tuned for our methods, there was a case where our methods needed more CPU time. For example, for small-scale problems, there are the cases that CPU time of CG-DESCENT is 0.01(sec) and CPU time of 3MS+ is 0.02(sec), and hence the line of 3MS+ in Figure 1 much goes up at $\tau = 2$. Accordingly, the numerical performance should be compared by measures different from CPU time. This is a reason why we give Figures 2–4. In Figures 2–4, we see that CG-DESCENT also performed well, and 3PR+, 3HS+ and PR+ are comparable with CG-DESCENT. On the other hand, 3MS+ is slightly outperformed by CG-DESCENT and is comparable with HS.

From our numerical experiments, we see that 3TCG (especially 3PR+ and 3HS+) performed as well as CG-DESCENT did. However, there is room to improve 3TCG. Especially, since the line search in CG-DESCENT is also tuned for CG by Hager and Zhang, we need to develop a suitable line search for 3TCG. It is our further work.

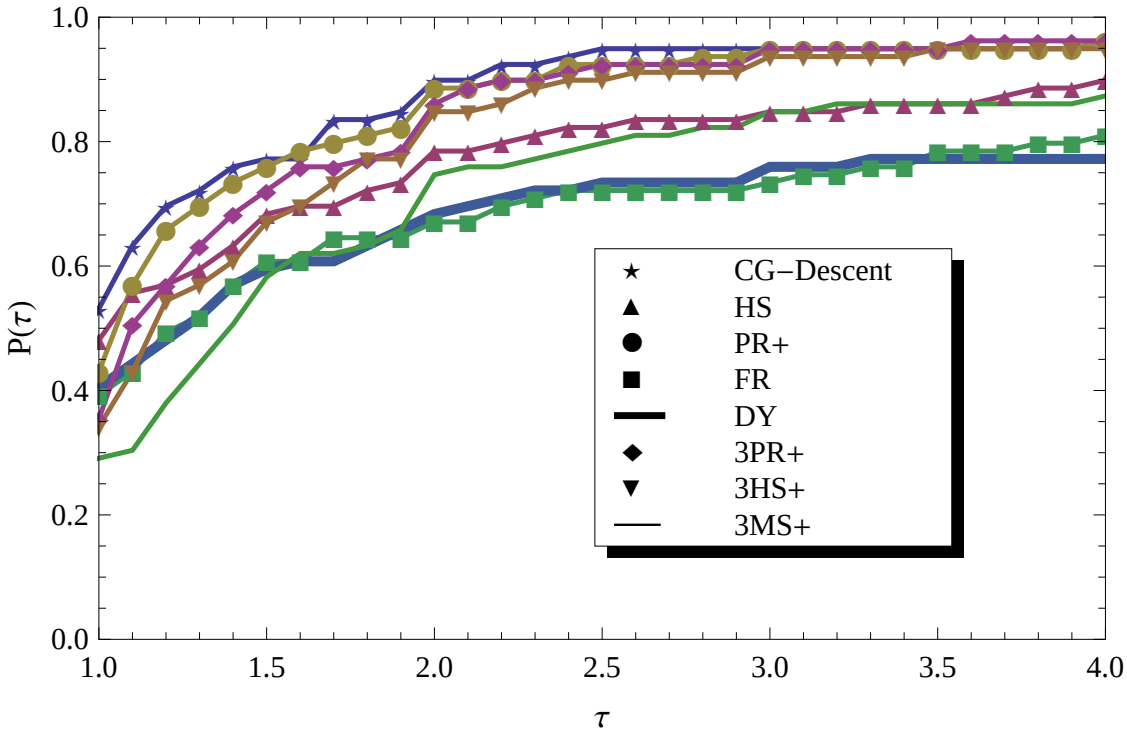

Figure 1: Performance profile by CPU time

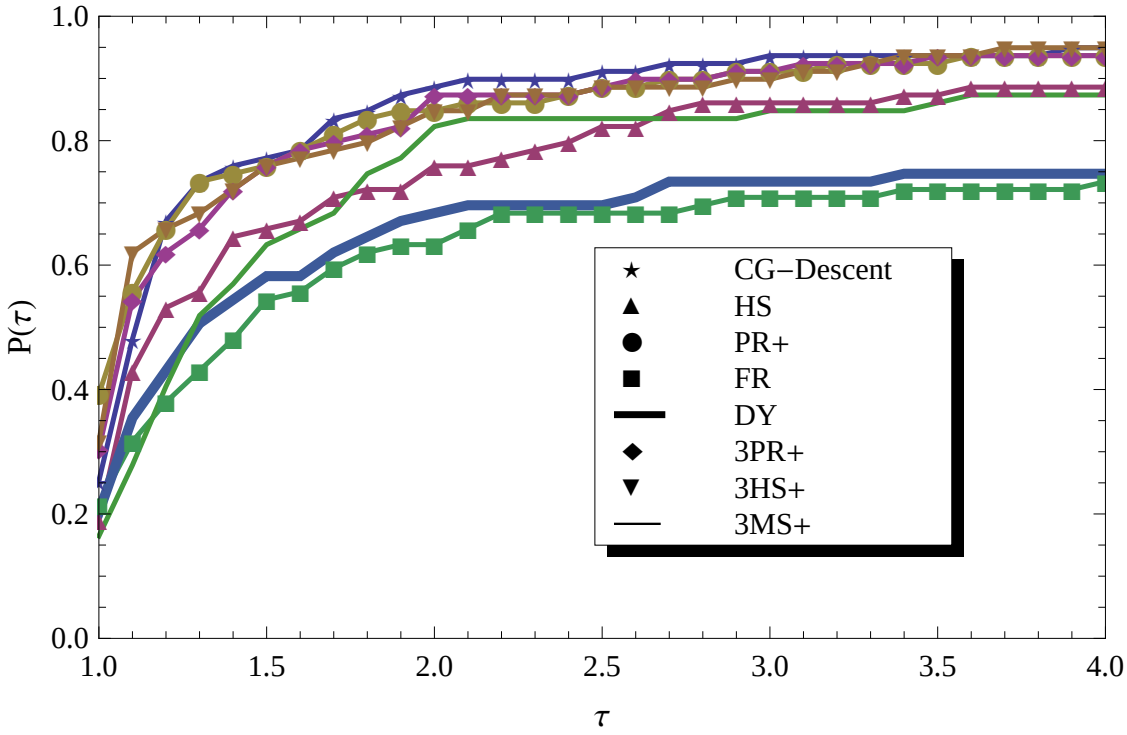

Figure 2: Performance profile by iterations

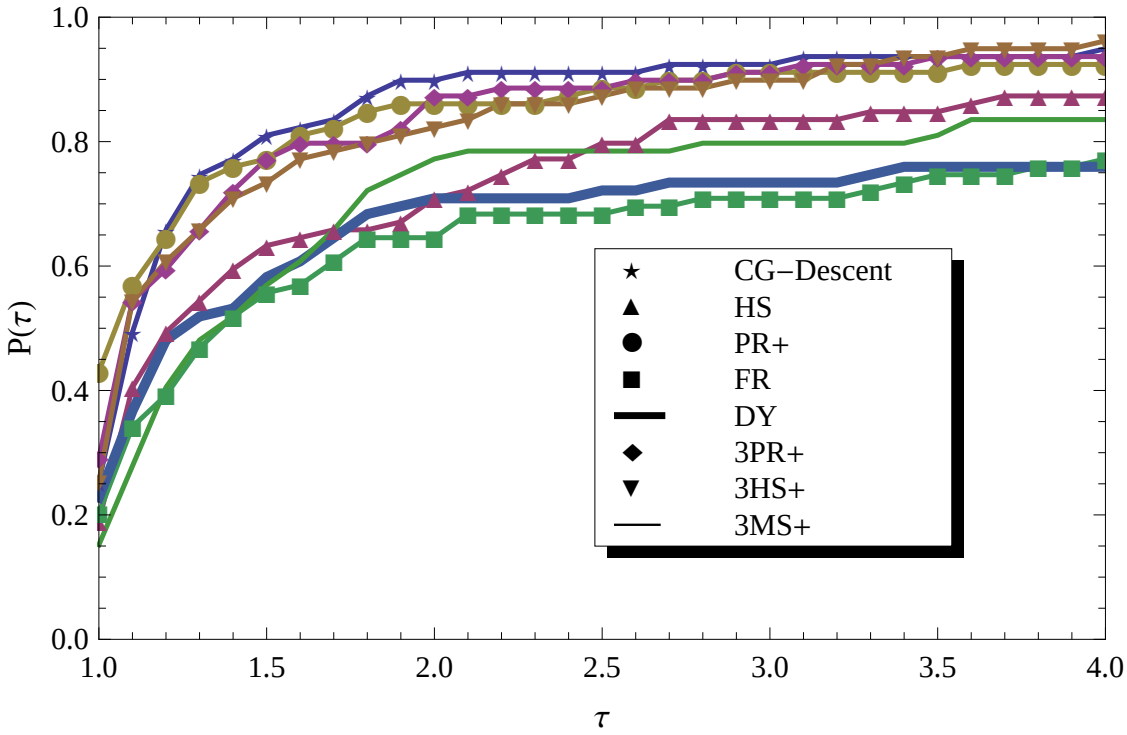

Figure 3: Performance profile by function evaluations

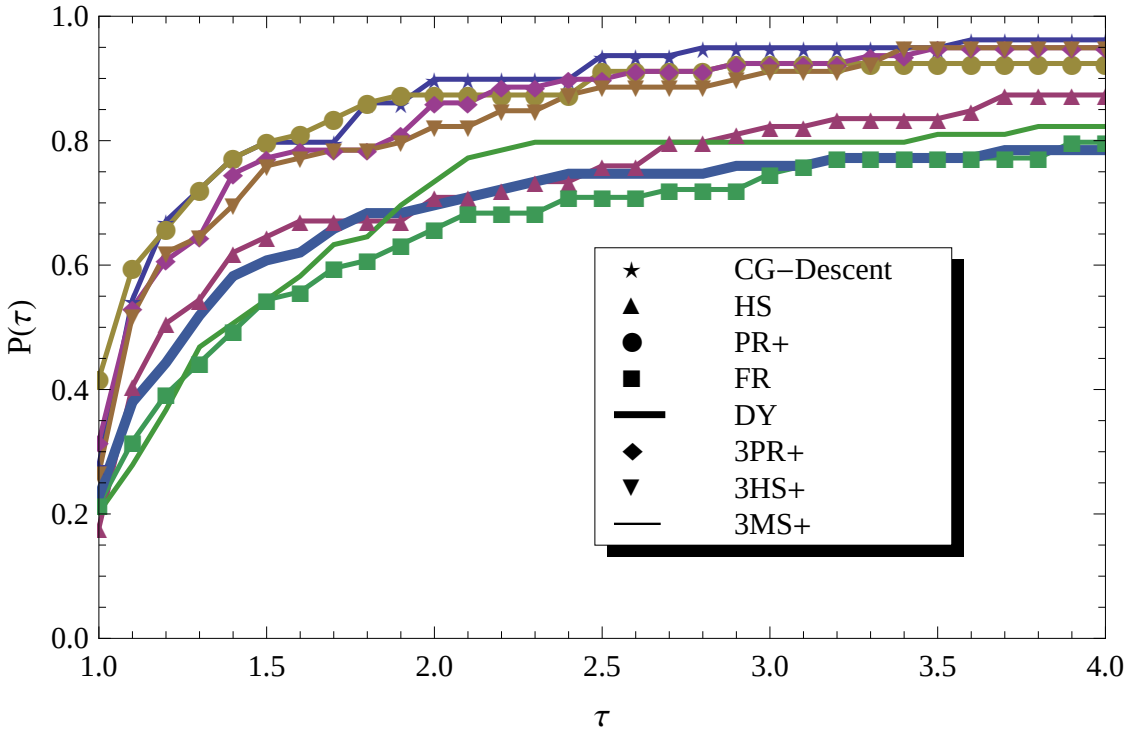

Figure 4: Performance profile by gradient evaluations

5 Conclusion

In this paper, we have proposed a general form of three-term conjugate gradient methods which always satisfy the sufficient descent condition independently of line searches and a choice of β_k. Moreover, we have given a sufficient condition for the global convergence of the proposed method. We have also proposed a new three-term conjugate gradient method based on the multi-step quasi-Newton method as a specific method. We have given the numerical results of our method by using commonly used benchmark problems, and have shown that our method perform effectively. Our further works are to find a suitable choice of p_k and to develop an efficient line search for three-term conjugate gradient methods.

6 Acknowledgements

The authors would like to thank the referees for valuable comments. The first and second authors are supported in part by the Grant-in-Aid for Scientific Research (C) 21510164 of Japan Society for the Promotion of Science.

References

[1] I. Bongartz, A.R. Conn, N.I.M. Gould and P.L. Toint, CUTE: constrained and unconstrained testing environments, *ACM Transactions on Mathematical Software*, **21** 1995, 123–160.

[2] W. Cheng, A two-term PRP-based descent method, *Numerical Functional Analysis and Optimization*, **28** (2007), 1217–1230.

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

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

[5] E.D. Dolan and J.J. Moré, Benchmarking optimization software with performance profiles, *Mathematical Programming*, **91** (2002), 201–213.

[6] R. Fletcher and C.M. Reeves, Function minimization by conjugate gradients, *Computer Journal*, **7** (1964), 149–154.

[7] J.A. Ford and I.A. Moghrabi, Alternative parameter choices for multi-step quasi-Newton methods, *Optimization Methods and Software*, **2** (1993), 357–370.

[8] J.A. Ford and I.A. Moghrabi, Multi-step quasi-Newton methods for optimization, *Journal of Computational and Applied Mathematics*, **50** (1994), 305–323.

[9] J.A. Ford, Y. Narushima and H. Yabe, Multi-step nonlinear conjugate gradient methods for unconstrained minimization, *Computational Optimization and Applications*, **40** (2008), 191–216.

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

[11] N.I.M. Gould, D. Orban and P.L. Toint, CUTEr web site, *http://cuter.rl.ac.uk/cuter-www/index.html*.

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

[13] W.W. Hager and H. Zhang, A survey of nonlinear conjugate gradient methods, *Pacific Journal of Optimization*, **2** (2006), 35–58.

[14] W.W. Hager and H. Zhang, CG_DESCENT Version 1.4 User' Guide, University of Florida, November 2005, *http://www.math.ufl.edu/~hager/*.

[15] M.R. Hestenes and E. Stiefel, Methods of conjugate gradients for solving linear systems, *Journal of Research of the National Bureau of Standards*, **49** (1952), 409–436.

[16] J. Nocedal and S.J. Wright, Numerical Optimization (Second Edition), Springer Series in Operations Research, Springer Verlag, New York, 2006.

[17] H. Yabe and N. Sakaiwa, A new nonlinear conjugate gradient method for unconstrained optimization, *Journal of the Operations Research Society of Japan*, **48** (2005), 284–296.

[18] H. Yabe and M. Takano, Global convergence properties of nonlinear conjugate gradient methods with modified secant condition, *Computational Optimization and Applications*, **28** (2004), 203–225.

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

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

[21] L. Zhang, W. Zhou and D.H. Li, Some descent three-term conjugate gradient methods and their global convergence, *Optimization Methods and Software*, **22** (2007), 697–711 .

[22] W. Zhou and L. Zhang, A nonlinear conjugate gradient method based on the MBFGS secant condition, *Optimization Methods and Software*, 21 (2006), 707-714.