Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 10, No. 1, pp. 85–104, January 2014

ON THE USE OF POWELL'S RESTART STRATEGY TO CONJUGATE GRADIENT METHODS*

Cai-Xia Kou, Wen-Hui Zhang, Wen-Bao Ai and Ya-Feng Liu

Abstract: Restart strategy is often used in conjugate gradient (CG) methods to improve their computational efficiency. In this paper, we propose to apply Powell's restart strategy [20] to four classical CG methods including the Fletcher-Reeves (FR) method [10], the Polak-Ribière-Polyak (PRP) method [18, 19], the Hestenes-Stiefel (HS) method [15], and the Dai-Yuan (DY) method [6]. We show that the corresponding variants of all aforementioned CG methods enjoy the sufficient descent property and are globally convergent if a generalized improved Wolfe line search is performed. Numerical results show that the use of Powell's restart strategy to the four classical CG methods can significantly improve their computational efficiency. In particular, the numerical performance of the corresponding variant of the HS method compares favorably with the one of the Dai-Kou's CGOPT in [3].

Key words: *CG method, unconstrained optimization, Powell's restart strategy, global convergence, generalized improved Wolfe line search*

Mathematics Subject Classification: *90C30, 49M37*

1 Introduction

Consider the unconstrained optimization problem

$$\min_{x \in \mathcal{R}^n} f(x), \tag{1.1}$$

where f is smooth and its gradient ∇f is available. Moreover, we assume that f satisfies

Assumption 1.1. (i) f is bounded below; namely, $f(x) > -\infty$ for all $x \in \mathcal{R}^n$;

(ii) f is differentiable and its gradient ∇f is Lipschitz continuous; i.e., there exists a constant $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \text{for all } x,\, y \in \mathcal{R}^n, \tag{1.2}$$

where $\|\cdot\|$ stands for the Euclidean norm.

*This work was partially supported by the NSFC grant 10831006, 10971017, 11331012 and 11301516, the Chinese NSF grant 81173633 and the CAS Program for Cross & Cooperative Team of the Science & Technology Innovation.

CG methods are very useful for solving problem (1.1), especially when the dimension n of the problem is large. They are iterative methods, of the form

$$x_{k+1} = x_k + \alpha_k d_k, \; k = 1, 2, \cdots , \tag{1.3}$$

where d_k is the search direction, and the stepsize $\alpha_k > 0$ is obtained by some line search (This shall be discussed later). The next search direction d_{k+1} is generated as follows:

$$d_{k+1} = -g_{k+1} + \beta_k \, d_k, \tag{1.4}$$

where $g_{k+1} := \nabla f(x_{k+1})$ and $d_1 = -g_1$. The scalar $\beta_k \in \mathcal{R}$ in (1.4) should be chosen such that (1.3)–(1.4) reduces to the linear CG method, which is originally developed for solving problem (1.1) with the objective being a strictly convex quadratic function. For general nonlinear objectives in (1.1), different choices of β_k shall lead to different CG methods. Well-known formulae for β_k include the FR, PRP, HS, and DY formulae; (see [10]; [18], [19]; [15]; [6], respectively):

$$\beta_k^{FR} = \frac{\|g_{k+1}\|^2}{\|g_k\|^2}, \tag{1.5}$$

$$\beta_k^{PRP} = \frac{g_{k+1}^T \left(g_{k+1} - g_k\right)}{\|g_k\|^2}, \tag{1.6}$$

$$\beta_k^{HS} = \frac{g_{k+1}^T \left(g_{k+1} - g_k\right)}{d_k^T \left(g_{k+1} - g_k\right)}, \tag{1.7}$$

$$\beta_k^{DY} = \frac{\|g_{k+1}\|^2}{d_k^T \left(g_{k+1} - g_k\right)}. \tag{1.8}$$

The convergence rate of the CG methods is linear unless restart is invoked from time to time. We could restart the CG methods every n iterations. However, it is known that the better restart strategy should be related to the objective function. For instance, Powell [20] suggested a restart strategy, which checks the quantity $|g_k^T g_{k+1}|/\|g_{k+1}\|^2$ at each iteration. If this quantity is greater than 0.2, then the Beale's three-term CG method is restarted. In such a way, satisfactory numerical results have been observed in [20]. The same restart strategy was also used by Shanno and his collaborators [22, 23] in building the CONMIN software.

There are other variants of CG methods which check the quantity $g_k^T g_{k+1}$ in an implicit fashion at each iteration and changes the computations of the formulae β_k based on the ratio $g_k^T g_{k+1}/\|g_{k+1}\|^2$. One of them is the DYHS method [8] with

$$\beta_k = \max\left\{0, \, \min\left\{\beta_k^{HS}, \beta_k^{DY}\right\}\right\}. \tag{1.9}$$

If $g_k^T g_{k+1} > \|g_{k+1}\|^2$, then β_k in (1.9) is equal to zero, and the DYHS method is actually restarted. Another variant is the FRPRP method [11] with the choice of β_k being

$$\beta_k = \begin{cases} -\beta_k^{FR}, & \text{if} \quad \beta_k^{PRP} < -\beta_k^{FR}; \\ \beta_k^{PRP}, & \text{if} \quad \beta_k^{PRP} \le \beta_k^{FR}; \\ \beta_k^{FR}, & \text{if} \quad \beta_k^{PRP} > \beta_k^{FR}. \end{cases} \tag{1.10}$$

As can be seen from (1.10), the FRPRP method chooses β_k to be β_k^{PRP} if the quantity $g_k^T g_{k+1}/\|g_{k+1}\|^2$ lies in the interval $[0, 2]$; the method chooses β_k to be $-\beta_k^{FR}$ if the quantity is greater than 2; and the method chooses β_k to be β_k^{FR} if the quantity is less than 0.

In this paper, we propose to apply Powell's restart strategy to the four classical CG methods corresponding to (1.5)–(1.8). The motivations for the use of Powell's restart strategy to these classical CG methods are as follows. First, as is known, the FR and DY methods are globally convergent when they are used for solving problem (1.1). However, these two methods may generate many short steps without making productive progress to the minimum (See Section 4 for detailed discussions). In this case, restarting these two methods is an effective way of overcoming this numerical problem. Second, when applying the CG methods to solve problem (1.1) with quadratic objective functions, it holds

$$g_k^T g_{k+1} = 0; \tag{1.11}$$

while, for general nonlinear objective functions, the value $g_k^T g_{k+1}$ may not be zero. Therefore, it is natural to use the quantity $\left|g_k^T g_{k+1}\right| / \|g_{k+1}\|^2$ to measure the loss of orthogonality between g_k and g_{k+1}, and restart the CG methods if $\left|g_k^T g_{k+1}\right| / \|g_{k+1}\|^2$ is far away from zero.

As can be seen from (1.3), a stepsize α_k needs to be determined along the direction d_k at each iteration of CG methods. For example, it is often required to satisfy the Wolfe conditions

$$f(x_k + \alpha d_k) \leq f(x_k) + \delta \alpha d_k^T g_k, \tag{1.12}$$

$$d_k^T g(x_k + \alpha d_k) \geq \sigma d_k^T g_k, \tag{1.13}$$

or some of its variants such as the strong Wolfe conditions that is (1.12) and

$$|d_k^T g(x_k + \alpha d_k)| \leq -\sigma d_k^T g_k, \tag{1.14}$$

and the generalized strong Wolfe conditions introduced in [3]. Theoretically, there must exist some positive α_k satisfying the Wolfe conditions (and its variants) if d_k is a descent direction. While in practical implementations, the first Wolfe condition (1.12) may never be satisfied due to the existence of numerical errors. This numerical drawback of the Wolfe condition was analyzed by using a one-dimension quadratic function in [13], and was also observed on a problem called JENSMP from the CUTEr collection in [3].

The main contributions of this paper are as follows. First, we propose to apply Powell's restart strategy to improve the computational efficiency of the four classical CG methods. We show that the directions generated by the corresponding variants of the classical CG methods satisfy the sufficient descent property at each iteration; namely,

$$d_k^T g_k \leq -C\|g_k\|^2, \quad \forall \;\; k \geq 1, \tag{1.15}$$

where $C > 0$ is some constant. The above property (1.15) plays a key role in the global convergence of the CG methods [11]. In addition, we propose generalized improved Wolfe conditions to overcome the numerical drawback of the Wolfe conditions, which shall be introduced in Section 2. By combining the sufficient descent property (1.15) and the proposed generalized improved Wolfe conditions, we establish the global convergence of the variants of all four classical CG methods. Second, numerical results are very promising, showing that all four CG methods after equipping with Powell's restart strategy significantly outperform the classical ones. In particular, the numerical performance of the HS method with Powell's restart strategy compares favorably with the one of the Dai-Kou's CGOPT in [3]. To the best of our knowledge, the Dai-Kou's CGOPT is one of the best CG methods in terms of the numerical performance.

The rest of the paper is organized as follows. The generalized improved Wolfe conditions are first introduced in Section 2. Then, the application of Powell's restart strategy to the HS and PRP methods is presented in Section 3, and the application to the FR and DY methods is presented in Section 4. Also, we shall show, in Sections 3 and 4, that the directions generated by the variants of the four classical CG methods satisfy the sufficient descent property (1.15) and the iterates generated by these variants are global convergent if the proposed line search conditions in Section 2 are satisfied. Numerical results are reported in Section 5. Finally, conclusion is drawn in Section 6.

2 The Generalized Improved Wolfe Conditions

As is known to us, the line search strategy plays an important role in CG methods. In this paper, we propose the following line search strategy:

$$f(x_k + \alpha d_k) \le f(x_k) + \min\left\{\epsilon|f(x_k)|,\, \delta\alpha d_k^T g_k + \eta_k\right\}, \tag{2.1}$$

$$\sigma_1 d_k^T g_k \le d_k^T g_{k+1} \le -\sigma_2 d_k^T g_k, \tag{2.2}$$

where $\epsilon > 0$, $0 < \delta < \sigma_1 < 1$, $\sigma_2 > 0$, and the positive sequence $\{\eta_k\}$ satisfies $\sum \eta_k < +\infty$. The condition (2.2) is firstly used in [5]. In particular, if the parameter σ_2 in (2.2) is set to be $+\infty$, then the conditions (2.1)–(2.2) reduce to the Dai-Kou's improved Wolfe conditions [3]. We thus call the conditions (2.1)–(2.2) the *generalized improved Wolfe conditions*. It is worthwhile remarking that, if $d_k^T g_k < 0$, then there must exist a suitable positive stepsize α_k satisfying (2.1) and (2.2), since these two conditions are weaker than the strong Wolfe conditions (1.14).

Throughout this paper, we shall use the line search strategy (2.1)–(2.2). The motivation for the use of the line search conditions (2.1)–(2.2) lies in its numerical efficiency. Specifically, at the beginning of the iterations, the value of $|f(x_k)|$ is often relatively large, condition (2.1) allows the stepsize α_k satisfying the first Wolfe condition (1.12), which can give a sufficient descent to the objective function. While, when the iterate point is close to the minimum, it often occurred that the trial point is close to the current point x_k, in which case,

$$f(x_k + \alpha d_k) \le f(x_k) + \epsilon|f(x_k)|, \tag{2.3}$$

then the condition (2.1) becomes

$$f(x_k + \alpha d_k) \le f(x_k) + \delta\alpha d_k^T g_k + \eta_k. \tag{2.4}$$

Comparing the above condition with the first Wolfe line search condition (1.12), it can be seen that the extra positive term η_k allows a slight increase in the function value and thus is helpful in avoiding the numerical drawback of the line search condition (1.12). At the same time, the fact that the sequence $\{\eta_k\}$ is summable can guarantee the global convergence of the method.

On the other hand, in typical implementations of the Wolfe conditions (1.12)-(1.13), it is often desirable to choose σ to be close to 1. However, Dai and Yuan [7] showed that when $\sigma > 1/2$, the FR method may not yield a descent direction even for the function $f(x) = \lambda\|x\|^2$. As can be seen later, the FR method is globally convergent under the proposed generalized improved Wolfe conditions (2.1)–(2.2) with the parameters satisfying $\sigma_1 + \sigma_2 \le 1$. Therefore, using condition (2.2) instead of (1.13) makes it possible to take σ_1 arbitrarily close to 1, by taking σ_2 close to 0.

Using the similar argument as in [3], we can obtain the following lemma.

Lemma 2.1. *Assume that f satisfies Assumption 1.1. Consider the iterative method of the form (1.3) where the direction d_k satisfies $g_k^T d_k < 0$ and the step size α_k satisfies the generalized improved Wolfe conditions* (2.1) *and* (2.2)*. Then we have*

$$\sum_{k\geq 1} \frac{(g_k^T d_k)^2}{\|d_k\|^2} < \infty. \tag{2.5}$$

From the lemma above, it follows that the generalized improved Wolfe line search also satisfies the Zoutendijk condition (see [24]), which is the basis in showing the global convergence of CG methods.

In the end of this section, we present the following lemma, which is often used to prove the global convergence of CG methods.

Lemma 2.2 ([11, 14])**.** *Suppose the CG method* (1.3) *and* (1.4) *satisfies the following conditions:*

(a) $\beta_k \geq 0$.

(b) *The search directions satisfy the sufficient descent condition* (1.15).

(c) *The Zoutendijk condition* (2.5) *holds.*

(d) *Property (*) holds. That is, assume that $0 < \gamma \leq \|g_k\| \leq \bar{\gamma}$ for all $k \geq 1$, then there exist constants $b > 1$ and $\lambda > 0$ such that for all k:*

$$|\beta_k| \leq b,$$

and

$$\|s_k\| \leq \lambda \Rightarrow |\beta_k| \leq \frac{1}{b},$$

where $s_k := x_{k+1} - x_k$. If, additionally, Assumption 1.1 holds, then the corresponding CG method is globally convergent.

3 On HS and PRP methods

In this section, Powell's restart strategy is applied to the HS and PRP methods. The sufficient descent property and the global convergence are also established for the HS and PRP methods equipped with Powell's restart strategy.

The PRP and HS methods (corresponding to (1.6) and (1.7) respectively) share the common numerator $g_{k+1}^T(g_{k+1} - g_k)$. If $x_{k+1} - x_k$ is small, the factor $y_k := g_{k+1} - g_k$ in the numerator of β_k is close to zero, and the new search direction d_{k+1} is essentially the steepest descent direction $-g_{k+1}$. Thus, the PRP and HS methods possess a built-in restart feature that addresses the jamming problem. That is, they take a large number of small steps without making productive progress to the minimum (Detailed discussions can be found in Section 4). This actually explains why the performance of the PRP and HS methods is generally better than the performance of the FR and DY methods. However, the PRP and HS methods are not globally convergent. In this paper, we consider modifying the PRP and HS methods to guarantee their global convergence.

We modify the classical HS method as follows:

$$\beta_k^{HS*} = \begin{cases} \beta_k^{HS}, & \text{if } |g_k^T g_{k+1}| \leq c_{hs}\|g_{k+1}\|^2; \\ 0, & \text{otherwise,} \end{cases} \tag{3.1}$$

where c_{hs} is a constant satisfying

$$0 < c_{hs} < \min\left\{1, \frac{1}{\sigma_2}\right\}. \tag{3.2}$$

This formula for β_k basically says that if $|g_k^T g_{k+1}|$ is greater than $c_{hs}\|g_{k+1}\|^2$, then the HS method should be restarted.

Next, we establish the sufficient descent property for the search direction d_{k+1} in (1.4) with β_k chosen according to (3.1).

Lemma 3.1. *Consider the CG method* (1.3)–(1.4)*, where β_k is given by* (3.1)*, and the stepsize α_k satisfies* (2.2)*. If the parameter c_{hs} in* (3.1) *satisfies* (3.2)*, then the sufficient descent condition* (1.15) *holds with* $C = 1 - \dfrac{\sigma_2(1+c_{hs})}{1+\sigma_2}$.

Proof. It suffices to consider the case where $\beta_k \neq 0$. In the sequential, we establish the lemma by induction. First of all, the lemma holds true for $k = 1$ due to $d_1 = -g_1$. Suppose that $d_k^T g_k \leq -C\|g_k\|^2$ holds true, we show that $d_{k+1}^T g_{k+1} \leq -C\|g_{k+1}\|^2$ is also true. By (3.1), the direction (1.4) always satisfies

$$|g_k^T g_{k+1}| \leq c_{hs}\|g_{k+1}\|^2 \quad \text{with} \quad 0 < c_{hs} < 1, \tag{3.3}$$

which, together with the choice of c_{hs}, further implies

$$0 \leq (1 - c_{hs})\|g_{k+1}\|^2 \leq g_{k+1}^T y_k \leq (1 + c_{hs})\|g_{k+1}\|^2. \tag{3.4}$$

From the condition (2.2) and the assumption $d_k^T g_k \leq -C\|g_k\|^2$, it follows that

$$-\sigma_2 \leq \frac{d_k^T g_{k+1}}{d_k^T g_k} \leq \sigma_1. \tag{3.5}$$

Consequently, there holds

$$\frac{d_k^T g_{k+1}}{d_k^T y_k} = \frac{\frac{d_k^T g_{k+1}}{d_k^T g_k}}{\frac{d_k^T g_{k+1}}{d_k^T g_k} - 1} \leq \frac{\sigma_2}{1+\sigma_2}. \tag{3.6}$$

By combining (3.4) and (3.6), we immediately obtain

$$\begin{aligned} d_{k+1}^T g_{k+1} &= -\|g_{k+1}\|^2 + \frac{d_k^T g_{k+1}}{d_k^T y_k} g_{k+1}^T y_k \\ &\leq \left(-1 + \frac{\sigma_2(1+c_{hs})}{1+\sigma_2}\right)\|g_{k+1}\|^2, \end{aligned} \tag{3.7}$$

which completes the proof. Notice that the positivity of C is due to the fact $0 < c_{hs} < \frac{1}{\sigma_2}$. □

A detailed description of the proposed modified HS method is given as follows.

Algorithm 3.2 (The Sufficient Descent HS* Method)**.**

Step 0. Given $x_1 \in \mathcal{R}^n$, $\varepsilon > 0$, $0 < \delta \leq \sigma_1 < 1$, $\sigma_2 > 0$, and c_{hs} satisfies (3.2).

Step 1. Set $k = 1$. If $\|g_1\| \leq \varepsilon$, terminate the algorithm; else set $d_1 = -g_1$.

Step 2. Compute the stepsize $\alpha_k > 0$ such that conditions (2.1) and (2.2) are satisfied.

Step 3. Let $x_{k+1} = x_k + \alpha_k d_k$. If $\|g_{k+1}\| \leq \varepsilon$, terminate the algorithm.

Step 4. If $|g_k^T g_{k+1}| > c_{hs}\|g_{k+1}\|^2$, set $d_{k+1} = -g_{k+1}$, $k = k + 1$, and go to Step 2.

Step 5. Compute β_k by (1.7) and d_{k+1} by (1.4). Set $k = k + 1$, and go to Step 2.

Applying the sufficient descent HS* method in Algorithm 3.2 to solve problem (1.1) with uniformly convex functions, i.e., there exists a constant $\mu > 0$ such that

$$(\nabla f(x) - \nabla f(\bar{x}))^T (x - \bar{x}) \geq \mu \|x - \bar{x}\|^2, \quad \forall\, x, \bar{x} \in \mathcal{R}^n, \tag{3.8}$$

we have the following global convergence result.

Theorem 3.3. *Assume that f satisfies Assumption 1.1. Apply the sufficient descent HS* method in Algorithm 3.2 to solve problem* (1.1) *with uniformly convex functions. Then, we have*

$$\lim_{k \to \infty} \|g_k\| = 0. \tag{3.9}$$

Proof. It follows from (1.2) and (3.8) that

$$\|y_k\| \leq L\,\|s_k\|, \tag{3.10}$$

$$d_k^T y_k \geq \mu\,\|d_k\|\,\|s_k\|. \tag{3.11}$$

By using (3.10) and (3.11), we have

$$\|d_{k+1}\| \quad \leq \quad \|g_{k+1}\| + \left|\frac{y_k^T g_{k+1}}{d_k^T y_k}\right| \|d_k\| \leq \left(1 + \frac{L}{\mu}\right) \|g_{k+1}\|. \tag{3.12}$$

Since the sufficient descent property (1.15) holds true, it follows from Lemma 2.1 that

$$\sum_{k \geq 1} \frac{\|g_k\|^4}{\|d_k\|^2} < \infty. \tag{3.13}$$

Combining (3.12) and (3.13) yields $\sum_{k\geq 1} \|g_k\|^2 < \infty$, which implies (3.9). □

Theorem 3.4. *Assume f satisfies Assumption 1.1. If the generated sequence $\{x_k\}$ is bounded, then the sufficient descent HS* method given in Algorithm 3.2 is globally convergent.*

Proof. To prove the global convergence of the HS* method for general nonlinear functions, we only need to check that all conditions in Lemma 2.2 are satisfied.

First of all, we show that $\beta_k^{HS*} \geq 0$. Note that, to show this, we only need to consider the case where $\beta_k^{HS*} = \beta_k^{HS}$. In this case, it follows from (2.2) and (3.7) that

$$d_k^T y_k \geq (\sigma_1 - 1) d_k^T g_k \geq (1 - \sigma_1)\left(1 - \frac{\sigma_2(1 + c_{hs})}{1 + \sigma_2}\right) \|g_k\|^2 \geq 0. \tag{3.14}$$

This, together with (3.4), implies β_k^{HS} in (1.7) is nonnegative.

Second, it follows from Lemma 3.1 and Lemma 2.1 that the search directions satisfy the sufficient descent properties and the Zoutendijk condition (2.5) holds true.

Finally, we show that Property (*) is also true. Assume that there exists $\gamma > 0$ such that

$$\|g_k\| \geq \gamma, \quad \forall\, k \geq 1. \tag{3.15}$$

By the continuity of ∇f and the boundedness of $\{x_k\}$, there must exist some positive constant $\bar{\gamma}$ such that

$$\|x_k\| \leq \bar{\gamma},\ \|g_k\| \leq \bar{\gamma}, \quad \forall\, k \geq 1. \tag{3.16}$$

Moreover, by combining (2.2), (1.15), and (3.15), we obtain

$$d_k^T y_k \geq -(1-\sigma_1) d_k^T g_k \geq C(1-\sigma_1)\gamma^2. \tag{3.17}$$

This, together with (1.2) and (3.16), further implies

$$\beta_k = \frac{g_{k+1}^T y_k}{d_k^T y_k} \leq \frac{\bar{\gamma}\|y_k\|}{C(1-\sigma_1)\gamma^2} \leq \frac{L\bar{\gamma}\|s_k\|}{C(1-\sigma_1)\gamma^2} := c_\beta \|s_k\|. \tag{3.18}$$

Define $b = 2c_\beta\bar{\gamma}$ and $\lambda = \frac{1}{2c_\beta^2\bar{\gamma}}$. It follows from (3.18) and (3.16) that

$$|\beta_k| \leq b, \tag{3.19}$$

and

$$\|s_k\| \leq \lambda \Rightarrow |\beta_k| \leq \frac{1}{b}. \tag{3.20}$$

The relations (3.19) and (3.20) show that β_k in (1.7) has Property(*) (See Lemma 2.2). □

We modify the classical PRP method in a similar fashion as follows:

$$\beta_k^{PRP*} = \begin{cases} \beta_k^{PRP} & \text{if } |g_k^T g_{k+1}| \leq c_{prp}\|g_{k+1}\|^2; \\ 0, & \text{otherwise,} \end{cases} \tag{3.21}$$

where c_{prp} is a positive constant in $(0,1)$ satisfying

$$\sigma_1(1-c_{prp}) + \sigma_2(1+c_{prp}) < 1. \tag{3.22}$$

Here we choose the parameters σ_1 and σ_2 in (2.2) satisfying $0 < \sigma_1, \sigma_2 < 1$.

The following lemma gives the sufficient descent property of the search direction (1.4) with β_k chosen according to (3.21).

Lemma 3.5. *Consider the CG method* (1.3)–(1.4)*, where β_k is given by* (3.21)*, and the stepsize α_k satisfies* (2.2)*. If the parameter c_{prp} in* (3.21) *satisfies* (3.22)*, then the sufficient descent condition* (1.15) *holds with* $C = 1 - \dfrac{\sigma_2(1+c_{prp})}{1-\sigma_1(1-c_{prp})} > 0$.

Proof. To show the Lemma, it is sufficient to consider the case where $\beta \neq 0$. Similar to the proof of Lemma 3.1, we can show that (3.4) remains true if c_{hs} there is replaced with c_{prp}, which, together with (1.6), further implies that

$$-1 + (1-c_{prp})\frac{d_k^T g_{k+1}}{\|g_k\|^2} \leq \frac{d_{k+1}^T g_{k+1}}{\|g_{k+1}\|^2} \leq -1 + (1+c_{prp})\frac{d_k^T g_{k+1}}{\|g_k\|^2}. \tag{3.23}$$

By combining the above relation and the second inequality of (2.2), we can show

$$\frac{d_{k+1}^T g_{k+1}}{\|g_{k+1}\|^2} \le -1 - \sigma_2(1+c_{prp})\frac{d_k^T g_k}{\|g_k\|^2}.$$

Inductively using the above inequality and the first inequality of (2.2) yields

$$\frac{d_{k+1}^T g_{k+1}}{\|g_{k+1}\|^2} \le -1 + \sigma_2(1+c_{prp})\sum_{j=0}^{k-1}\sigma_1^j(1-c_{prp})^j \le -1 + \frac{\sigma_2(1+c_{prp})}{1-\sigma_1(1-c_{prp})}. \tag{3.24}$$

Finally, it follows from (3.22) that $C := 1 - \dfrac{\sigma_2(1+c_{prp})}{1-\sigma_1(1-c_{prp})} > 0.$ □

For convenience, we present the modified PRP method as follows:

Algorithm 3.6 (The Sufficient Descent PRP* Method)**.** Replace Step 5 of Algorithm 3.2 with

Step 5. Compute β_k by (3.21) and d_{k+1} by (1.4). Set $k = k+1$, and go to Step 2.

By the definition (3.21) of β_k^{PRP*}, we see that either $\beta_k^{PRP*} = 0$ or it satisfies

$$(1-c_{prp})\frac{\|g_{k+1}\|^2}{\|g_k\|^2} \le \beta_k^{PRP*} \le (1+c_{prp})\frac{\|g_{k+1}\|^2}{\|g_k\|^2}. \tag{3.25}$$

Since $c_{prp} \in (0,1)$, it follows that

$$\beta_k^{PRP*} \ge 0. \tag{3.26}$$

Furthermore, we know from [11] that the PRP* method has Property(*). Specifically, using the constants γ and $\bar{\gamma}$ in (3.15) and choosing $b = 2\bar{\gamma}^2/\gamma^2$ and $\lambda = \gamma^2/(2L\bar{\gamma}b)$, we have

$$|\beta_k^{PRP}| \le \frac{(\|g_k\| + \|g_{k+1}\|)\|g_{k+1}\|}{\|g_k\|^2} \le \frac{2\bar{\gamma}^2}{\gamma^2} = b, \tag{3.27}$$

and

$$\|s_k\| \le \lambda \Rightarrow |\beta_k^{PRP}| \le \frac{\|y_k\|\|g_{k+1}\|}{\|g_k\|^2} \le \frac{L\lambda\bar{\gamma}}{\gamma^2} = \frac{1}{2b}. \tag{3.28}$$

The above analysis shows that β_k^{PRP*} is positive, satisfies Property(*), and guarantees the sufficient descent condition (Lemma 3.5). By Lemma 2.2, we immediately obtain the following convergence theorem.

Theorem 3.7. *Assume f satisfies Assumption 1.1. The sufficient descent PRP* method given in Algorithm 3.6 is globally convergent.*

4 On FR and DY methods

In this section, we apply Powell's restart strategy to the FR and DY methods. We shall also establish the sufficient descent property and the global convergence of the variants of the corresponding methods for solving problem (1.1) with general nonlinear objectives.

As can be seen from (1.5) and (1.8), the FR and DY methods have a common term $\|g_{k+1}\|^2$ in the numerator. Different from the PRP and HS methods, the global convergence of these two methods can be guaranteed without any modification on β_k. Zoutendijk [24] firstly proved the FR method is globally convergent with exact line search. Powell [20] pointed out that the FR method with exact line search will produce many short steps without making significant progress to the minimum in practical computations. Actually, this phenomenon can also be observed when the inexact line search is performed. Denoting the angle between $-g_k$ and d_k by θ_k, then

$$\cos\theta_k = -\frac{d_k^T g_k}{\|d_k\|\|g_k\|}. \tag{4.1}$$

If the inexact line search is performed, then there exist positive constants c_1, c_2 such that

$$c_1 \frac{\|g_k\|}{\|d_k\|} \le \cos\theta_k \le c_2 \frac{\|g_k\|}{\|d_k\|}. \tag{4.2}$$

Suppose that an unfortunate search direction d_k is generated in terms that $\cos\theta_k \approx 0$ at some iteration k. In this case, it is likely that $x_{k+1} \approx x_k$. If this happens, then we have $\|g_{k+1}\| \approx \|g_k\|$, and

$$\beta_{k+1}^{FR} \approx 1. \tag{4.3}$$

Moreover, by (4.2), we obtain

$$\|g_{k+1}\| \approx \|g_k\| \ll \|d_k\|. \tag{4.4}$$

Combining the above relation and (4.3) shows $\|d_{k+1}\| \approx \|d_k\| \gg \|g_{k+1}\|$, which further implies $\cos\theta_{k+1} \approx 0$ according to (4.2). Therefore, this behavior can start all over again and a large number of subsequent steps are small, unless the method is restarted. Gilbert and Nocedal in [11] gave a numerical example illustrating this behavior. The poor performance of the FR method in applications are often due to this jamming phenomenon.

To overcome the numerical drawback of the FR method, we consider applying Powell's restart idea to it. We modify the FR method as follows:

$$\beta_k^{FR*} = \begin{cases} \beta_k^{FR} & \text{when } |g_k^T g_{k+1}| \le c_{fr}\|g_{k+1}\|^2; \\ 0, & \text{otherwise,} \end{cases} \tag{4.5}$$

where the parameter c_{fr} is set to be in the interval $(0, 1)$.

In the FR* method, we also use the generalized improved Wolfe conditions (2.1) and (2.2), where the parameters σ_1 and σ_2 are chosen such that

$$\sigma_1 + \sigma_2 \le 1.$$

Under the generalized improved Wolfe line search, we can establish the sufficient descent property for the FR* method. In fact, in a similar fashion as showing (3.24), we can show, by combining (1.4) and (2.2), that

$$\frac{d_{k+1}^T g_{k+1}}{\|g_{k+1}\|^2} \le -1 + \sigma_2 \sum_{j=0}^{k-1} \sigma_1^j \le -1 + \frac{\sigma_2}{1-\sigma_1}. \tag{4.6}$$

Lemma 4.1. *Suppose that the generalized improved Wolfe line search conditions are satisfied at every iteration. Then, the direction generated by the FR^* method satisfies the sufficient descent property* (1.15) *with* $C = 1 - \dfrac{\sigma_2}{1-\sigma_1} > 0$.

A specification of the FR* method is given as follows.

Algorithm 4.2 (The Sufficient Descent FR* Method)**.** Replace Step 5 of Algorithm 3.2 with

Step 5. Compute β_k by (1.5) and d_{k+1} by (1.4). Set $k = k + 1$, and go to Step 2.

Lemma 4.1 shows that the FR* method always generates the sufficient descent directions. The similar argument as in [7] shows the following convergence result of the FR* method.

Theorem 4.3. *Assume that f satisfies Assumption 1.1. The sufficient descent FR* method given in Algorithm 4.2 is globally convergent.*

Now, we apply Powell's restart strategy to the DY method developed in [6]. With a standard Wolfe line search, the DY method always generates the descent directions. In addition, global convergence of the DY method is guaranteed if Assumption 1.1 holds true. The motivation here for applying Powell's restart strategy is to improve its numerical performance. The β_k in the DY method (c.f. (1.8)) is modified as follows:

$$\beta_k^{DY*} = \begin{cases} \beta_k^{DY}, & \text{when } |g_k^T g_{k+1}| \le c_{dy}\|g_{k+1}\|^2; \\ 0, & \text{otherwise,} \end{cases} \tag{4.7}$$

where the parameter c_{dy} lies in the interval $(0, 1)$. Under the generalized improved Wolfe line search, the DY* method can obtain the search directions satisfying the sufficient descent property. In fact, it follows from the generalized improved Wolfe line search conditions that

$$\frac{d_k^T g_{k+1}}{d_k^T g_k} \ge -\sigma_2. \tag{4.8}$$

Thus, we have

$$\frac{d_{k+1}^T g_{k+1}}{\|g_{k+1}\|^2} = -1 + \frac{\beta_k^{DY} d_k^T g_{k+1}}{\|g_{k+1}\|^2} = \frac{d_k^T g_k}{d_k^T g_{k+1} - d_k^T g_k} \le \frac{-1}{1+\sigma_2}. \tag{4.9}$$

Lemma 4.4. *Suppose that the generalized improved Wolfe line search conditions are satisfied at each iteration. Then the directions generated by the DY* method satisfy the sufficient descent property with $C = 1/(1+\sigma_2)$.*

Algorithm 4.5 (The Sufficient Descent DY* Method)**.** Replace Step 5 of Algorithm 3.2 with

Step 5. Compute β_k by (1.8) and d_{k+1} by (1.4). Set $k = k + 1$ and go to Step 2.

Similar to the DY method, the DY* method is also globally convergent for problem (1.1) with general nonlinear objectives.

Theorem 4.6. *Assume that f satisfies Assumption 1.1. Then the sufficient descent DY* method given in Algorithm 4.5 is globally convergent.*

5 Numerical Results

In this section, we present some simulation results to show the efficiency of the use of Powell's restart strategy to the FR, PRP, HS, and DY methods. For convenience, we compare the performance of the following methods:

- FR*: Algorithm 4.2 with the parameters $\delta = 0.1, \sigma_1 = 0.8, \sigma_2 = 0.1$, and $c_{fr} = 0.8$.
- FR: The FR method using the same line search as in the FR* method.
- PRP*: Algorithm 3.6 with the parameters $\delta = 0.1, \sigma_1 = 0.8, \sigma_2 = 0.1$, and $c_{prp} = 0.8$.
- PRP$^+$: The PRP method allows only positive β_k^{prp} and uses strong Wolfe line search with the parameters $\delta = 10^{-4}$ and $\sigma = 0.1$. This variant of the PRP method is proposed by Gilbert and Nocedal [11].
- HS*: Algorithm 3.2 with the parameters $\delta = 0.1, \sigma_1 = \sigma_2 = 0.9$, and $c_{hs} = 0.8$.
- HS$^+$: A similar variant of the HS method as the PRP$^+$ method.
- DY*: Algorithm 4.5 with the parameters $\delta = 0.1, \sigma_1 = 0.9, \sigma_2 = +\infty$, and $c_{dy} = 0.8$.
- DY: The Dai-Yuan method in [6] using the same line search as in the DY* method.
- DYHS: The method in [8] using the same line search as in the HS* method.
- FRPRP: The method in [11] using the same line search as in the PRP* method.

All of the above algorithms were implemented in C language and tested in Fedora 12 Linux environment. The computer used is a Lenovo X200 laptop with 2G RAM memory and Centrino2 processor. The termination criterion of all algorithms is set to be $\|\nabla f(x_k)\|_\infty \leq 10^{-6}$.

The test problems are 135 unconstrained optimization problems drawn from the CUTEr [12] collection. For each comparison, we excluded those problems for which different solvers converge to different solutions.

The performance profile [9] is used to display the performance of the algorithms. Define $\mathcal{P}$ as the whole set of n_p test problems and $\mathcal{S}$ the set of the interested solvers. Let

$$l_{p,s} = n_f + 3n_g$$

be the cost required by solver s for solving problem p, where n_f and n_g denote the number of objective function and gradient evaluations, respectively. Define the performance ratio as

$$r_{p,s} = \frac{l_{p,s}}{l_p^*},$$

where $l_p^* = \min\{l_{p,s} : s \in \mathcal{S}\}$. It is obvious that $r_{p,s} \geq 1$ for all p and s. If the solver s fails to solve problem p, then $r_{p,s}$ is set to be a large number M. The performance profile for each solver s is defined as the following cumulative distribution function based on the performance ratio $r_{p,s}$:

$$\rho_s(\tau) = \frac{size\{p \in \mathcal{P} : r_{p,s} \leq \tau\}}{n_p}.$$

Obviously, $\rho_s(1)$ represents the percentage of problems for which the solver s is the best. See [9] for more details about the performance profile.

Firstly, we test the efficiency of using Powell's restart strategy to the HS method. Figure 1 plots the perfomance profile of the HS* and HS^+ methods on the number of function and gradient evaluations. After eliminating the problems for which the two algorithms converge to different solutions, there are 129 problems left. It can be seen from Figure 1 that the HS* method is faster for about 80% of test problems, while the HS^+ method is faster for about 35% of test problems, which indicates that Powell's restart strategy gives a substantial improvement on the HS method.

Figures 2-4 show the comparison of the PRP* method with the PRP^+ method, the FR* method with the FR method, and the DY* method with the DY method, respectively. After eliminating the problems for which the two algorithms converge to different solutions, 130, 135 and 133 problems are left, respectively. It can be observed from these figures that the use of Powell's restart strategy can significantly improve the computational efficiency of the corresponding methods.

In Figure 5, we compare the Algorithm 3.2, which ranks the first among the four variants of the CG methods, with CGOPT on function and gradient evaluations. After eliminating the problems for which the two algorithms converge to different solutions, 132 problems are left. Observe that the HS* method is faster for about 65% of the test problems, while CGOPT is faster for less than 45% of the test problems. This shows that the HS method with Powell's restart strategy is competitive with the CGOPT method.

In Figure 6, we compare the HS* method with Dai-Kou's CGOPT where the same generalized improved Wolfe line search is used. We abbreviated CGOPT with Generalized Improved Wolfe line search to *CGOPT-GIW*. Figure 6 shows that the HS* method still has a better performance.

Figure 7 and 8 plot comparisons of the proposed HS* with the DYHS method [8] and the proposed PRP* method with the FRPRP method [11], respectively. These two figures show that the proposed methods outperform the DYHS and FRPRP methods.

6 Conclusion

In this paper, we apply Powell's restart strategy to the FR, PRP, HS, and DY methods, and obtain the corresponding variants of the four classical CG methods. Also, we propose a generalized improved Wolfe line search condition, which can be regarded as an extension of the Dai-Kou's improved Wolfe line search condition in [3]. Based on the proposed generalized improved Wolfe line search condition, we show that the variants of all aforementioned CG methods enjoy the sufficient descent property and are globally convergent. Numerical simulation results show that the use of Powell's restart strategy can significantly improve the numerical performance of the four classical CG methods. In particular, the HS* method performs as well as the Dai-Kou's CGOPT method [3], which is, by far, one of the best CG methods.

As we know, the performance improvement of the variants over the classical CG methods is a result of the use of Powell's restart strategy, i.e., using the quantity $|g_k^T g_{k+1}|$ measures nonlinear extent of the objective function, and restarting the corresponding CG methods if this quantity is sufficiently large. In the CGOPT method, there is another dynamical restart strategy based on r_{k-1} defined by (3.4) in [3], which reflects the similarity of an objective with some quadratic function. Such restart strategy accelerates the CGOPT method to a great extent. Therefore, to extend this work, we may consider how to use these two restart strategies to CGOPT. This will be one of our future works.

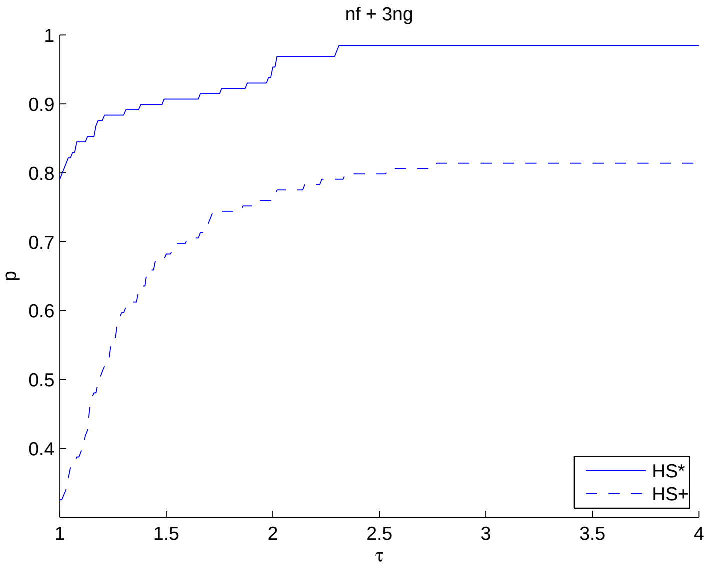

Figure 1: Performance profile of Algorithm 3.2 (the HS* method) and the HS$^+$ method for the test problems from CUTEr set.

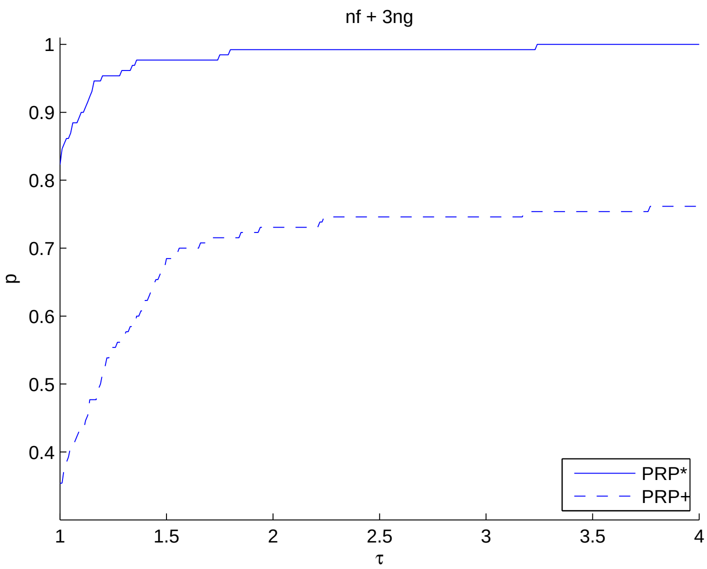

Figure 2: Performance profile of Algorithm 3.6 (the PRP* method) and the PRP$^+$ method for the test problems from CUTEr set.

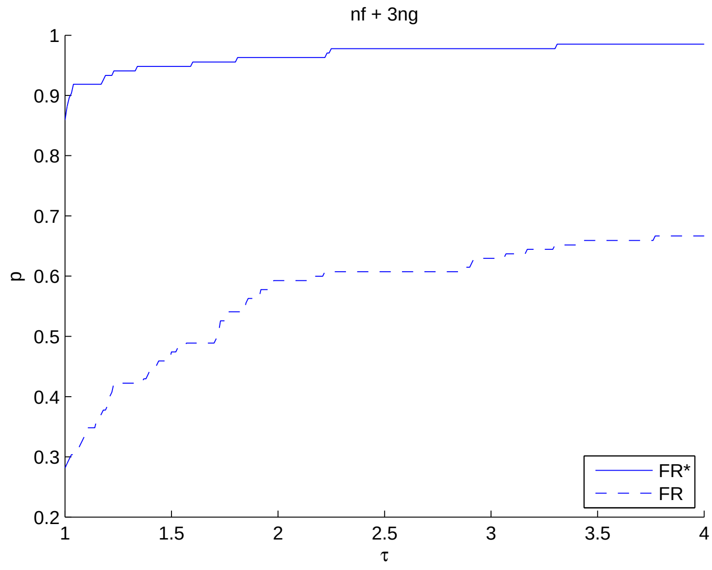

Figure 3: Performance profile of Algorithm 4.2 (the FR* method) and the FR method for the test problems from CUTEr set.

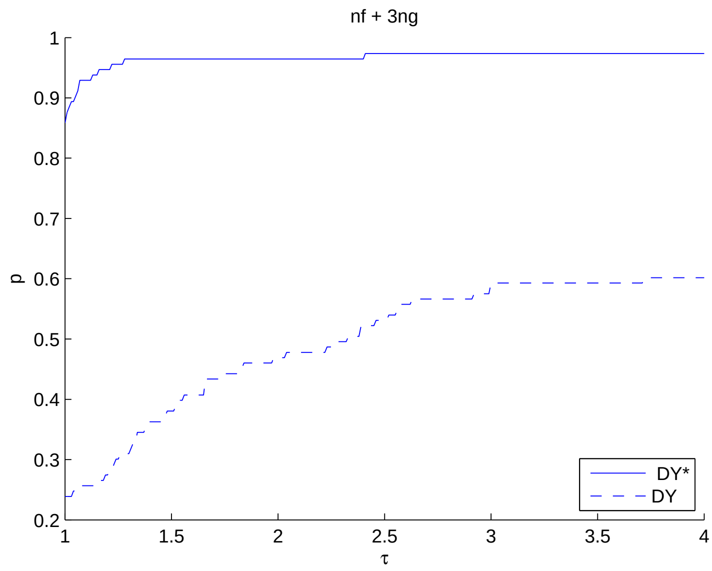

Figure 4: Performance profile of Algorithm 4.5 (the DY* method) and the DY method for the test problems from CUTEr set.

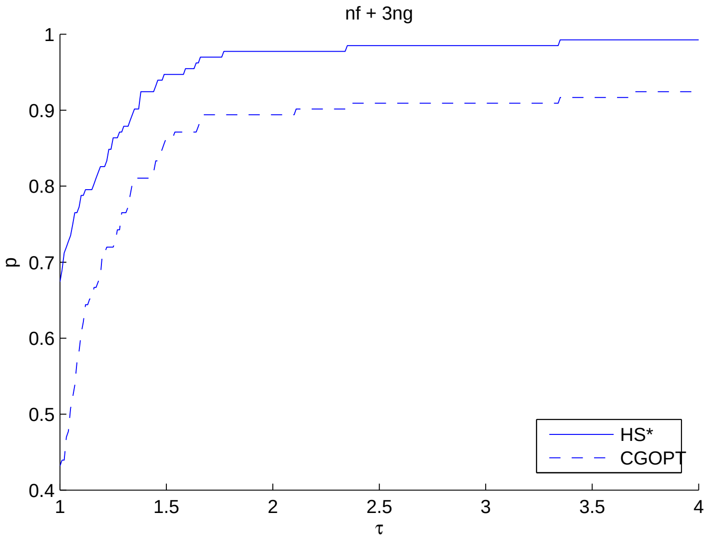

Figure 5: Performance profile of Algorithm 3.2 (the HS* method) and the CGOPT method for the test problems from CUTEr set.

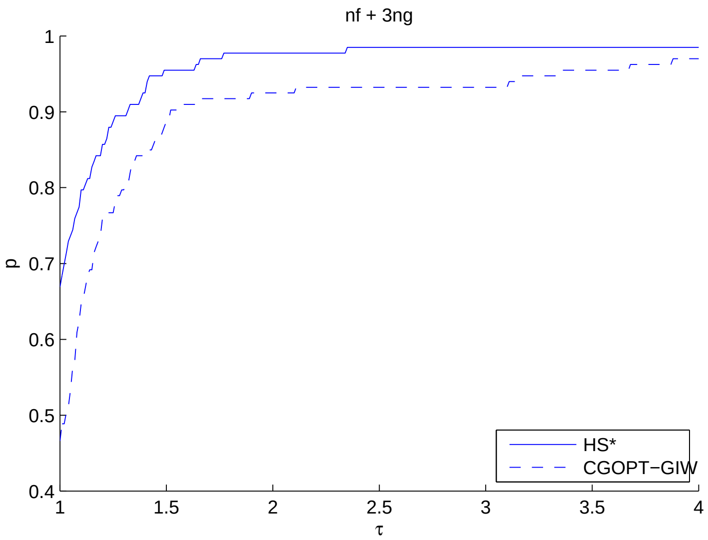

Figure 6: Performance profile of Algorithm 3.2 (the HS* method) and the CGOPT-GIW method for the test problems from CUTEr set.

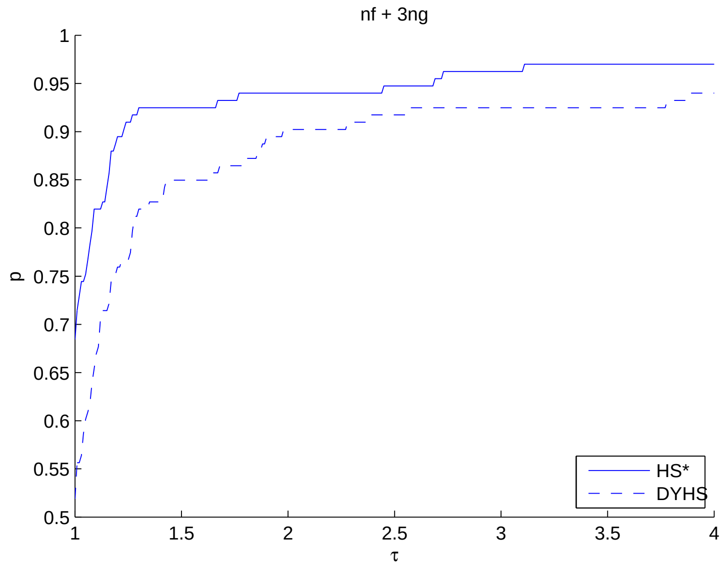

Figure 7: Performance profile of Algorithm 3.2 (the HS* method) and the DYHS method for the test problems from CUTEr set.

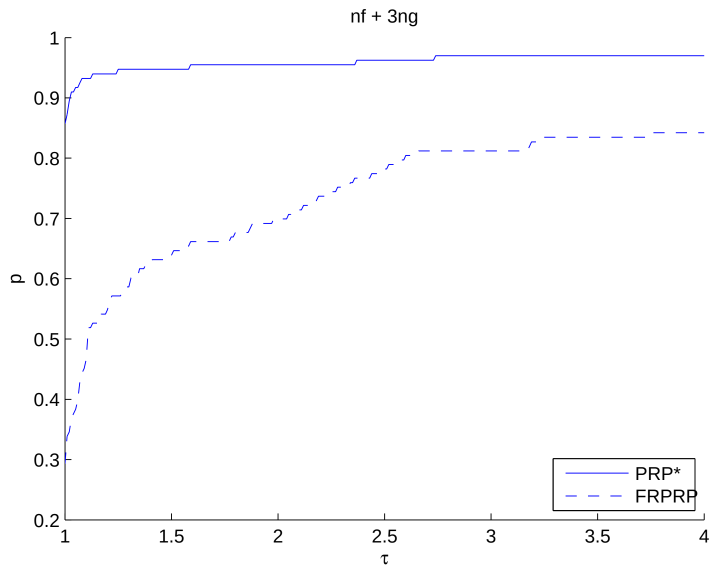

Figure 8: Performance profile of Algorithm 3.6 (the PRP* method) and the FRPRP method for the test problems from CUTEr set.

Acknowledgements

The authors would like to thank Professor Yu-Hong Dai from State Key Laboratory of Scientific and Engineering Computing, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, for his very 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] E.M.L. Beale, A derivation of conjugate gradients, in *Numerical Methods for Nonlinear Optimization*, F.A. Lootsman (ed.), Academic Press, London, 1972, pp. 39–43.

[3] Y.H. Dai and C.X. Kou, A nonlinear conjugate gradient algorithm with an optimal property and an improved Wolfe line search, Research report, Academy of Mathematics and Systems Science, *SIAM J. Optim.* 23 (2013) 296–320.

[4] Y.H. Dai and L.Z. Liao, New conjugacy conditions and related nonlinear conjugate gradient methods, *Appl. Math. Optim.* 43 (2001) 87–101.

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

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

[7] Y.H. Dai and Y. Yuan, *Nonlinear Conjugate Gradient Methods*, Shanghai Science and Technology Publisher, Shanghai, 2000.

[8] Y.H. Dai and Y. Yuan, An efficient hybrid conjugate gradient method for unconstrained optimization, *Annals of Operations Research* 103 (2001) 33–47.

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

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

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

[12] N.I.M. Gould, D. Orban and Ph. L. Toint, CUTEr (and SifDec), a constrained and unconstrained testing environment, revisited, Technical Report TR/PA/01/04, CERFACS, Toulouse, France, 2001.

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

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

[15] M.R. Hestenes and E.L. Stiefel, Methods of conjugate gradients for solving linear systems, *J. Research Nat. Bur. Standards* 49 (1952) 409–436.

[16] 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.

[17] J.M. Perry, A class of conjugate gradient algorithms with a two-step variable-metric memory, Discussion Paper 269, Center for Mathematical Studies in Economics and Management Sciences, Northwestern University, Evanston, Illinois, 1977.

[18] E. Polak and G. Ribière, Note sur la convergence de méthodes de directions conjugées, *Rev. Francaise Informat. Recherche Opértionelle* 3 (1969) 35–43.

[19] B.T. Polyak, The conjugate gradient method in extreme problems, *USSR Comp. Math. and Math. Phys.* 9 (1969) 94–112.

[20] M.J.D. Powell, Restart Procedures for the Conjugate Gradient Method, *Math. Programming* 12 (1977) 241–254.

[21] M.J.D. Powell, Nonconvex minimization calculations and the conjugate gradient method, in *Lecture Notes in Mathematics 1066*, D.F. Griffiths (ed.), Springer, Berlin, 1984, pp. 122–141.

[22] D.F. Shanno, On the convergence of a new conjugate gradient algorithm, *SIAM J. Numer. Anal.* 15 (1978) 1247–1257.

[23] D.F. Shanno and K.H. Phua, Matrix conditioning and nonlinear optimization, *Math. Programming* 14 (1978) 149–160.

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

Manuscript received 20 March 2013
revised 14 June 2013
accepted for publication 17 June 2013

Cai-Xia Kou
School of Science, Beijing University of Posts and Telecommunications
Beijing 100876, P. R. China
E-mail address: koucx@bupt.edu.cn

Wen-Hui Zhang
State Key Laboratory of Scientific and Engineering Computing
Institute of Computational Mathematics and Scientific/Engineering Computing
Academy of Mathematics and Systems Science
Chinese Academy of Sciences, Beijing 100190, P. R. China
E-mail address: wenhui@lsec.cc.ac.cn

Wen-Bao Ai
School of Science, Beijing University of Posts and Telecommunications
Beijing 100876, P. R. China
E-mail address: wenbaoaibupt@vip.126.com

Ya-Feng Liu
State Key Laboratory of Scientific and Engineering Computing
Institute of Computational Mathematics and Scientific/Engineering Computing
Academy of Mathematics and Systems Science
Chinese Academy of Sciences, Beijing 100190, P. R. China
E-mail address: yafliu@lsec.cc.ac.cn