

# New properties of a nonlinear conjugate gradient method

Yu-Hong Dai

State Key Laboratory of Scientific and Engineering Computing, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, P. O. Box 2719, Beijing 100080, China;  
e-mail: dyh@lsec.cc.ac.cn

Received March 12, 1999 / Revised version received April 25, 2000 /  
Published online February 5, 2001 – © Springer-Verlag 2001

**Summary.** This paper provides several new properties of the nonlinear conjugate gradient method in [5]. Firstly, the method is proved to have a certain self-adjusting property that is independent of the line search and the function convexity. Secondly, under mild assumptions on the objective function, the method is shown to be globally convergent with a variety of line searches. Thirdly, we find that instead of the negative gradient direction, the search direction defined by the nonlinear conjugate gradient method in [5] can be used to restart any optimization method while guaranteeing the global convergence of the method. Some numerical results are also presented.

*Mathematics Subject Classification (1991):* 49M37, 65K05, 90C30

## 1. Introduction

The object of this paper is to further analyze the properties of the nonlinear conjugate gradient method in [5]. We will focus our attention on the unconstrained optimization problem

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

where  $f$  is smooth and its gradient  $g$  is available. Conjugate gradient methods for solving (1.1) are iterative methods of the form

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

where  $\alpha_k$  is a steplength, and  $d_k$  is a search direction. The initial search direction  $d_1$  is always set to  $-g_1$ , and for  $k \geq 2$ ,  $d_k$  is recursively defined by

$$(1.3) \quad d_k = -g_k + \beta_k d_{k-1},$$

where  $g_k = \nabla f(x_k)$ , and  $\beta_k$  is a scalar.

Since R. Fletcher and C. Reeves [9] first proposed the nonlinear conjugate gradient method, there have been many formulae for the scalar  $\beta_k$ , for example see [7–10, 12, 15, 18–20]. Generally, in the analyses and implementations of these conjugate gradient methods, the steplength  $\alpha_k$  is required to satisfy the strong Wolfe conditions

$$(1.4) \quad f(x_k + \alpha_k d_k) - f(x_k) \leq \delta \alpha_k g_k^T d_k,$$

$$(1.5) \quad |g(x_k + \alpha_k d_k)^T d_k| \leq -\sigma g_k^T d_k,$$

where  $0 < \delta < \sigma < 1$ , or the sufficient descent condition

$$(1.6) \quad g_k^T d_k \leq -c \|g_k\|^2, \quad \text{for some constant } c > 0.$$

For example, under the conditions (1.4)–(1.5) or (1.6), reference [10] carefully analyzed the convergence properties of the methods related the Fletcher-Reeves method and those related to the Polak-Ribière-Polyak and Hestenes-Stiefel methods.

In [5], we proposed a new nonlinear conjugate gradient method, in which  $\beta_k$  has the form of

$$(1.7) \quad \beta_k = \frac{\|g_k\|^2}{d_{k-1}^T y_{k-1}},$$

where  $\|\cdot\|$  means the two norm, and  $y_{k-1} = g_k - g_{k-1}$ . This method is proved to produce a descent direction at each iteration and converge in the sense that

$$(1.8) \quad \liminf_{k \rightarrow \infty} \|g_k\| = 0,$$

provided that the steplength  $\alpha_k$  satisfies the standard Wolfe conditions, namely, (1.4) and

$$(1.9) \quad g(x_k + \alpha_k d_k)^T d_k \geq \sigma g_k^T d_k.$$

A recent reference [16] has listed the formula (1.7) as one of the four leading contenders for the choice of  $\beta_k$ . For convenience, we call the method (1.2)–(1.3) with  $\beta_k$  computed by (1.7) as the method (1.7).

If the objective function  $f$  is uniformly convex, reference [6] proved that the method (1.7) with several kinds of line searches also produces a descent

direction and converge globally. In this case, it was shown that the sufficient descent condition (1.6) holds for all  $k \geq 1$ .

In this paper, we will first study the method (1.7) for general objective functions and without doing any line searches. We prove that, if  $g_k^T d_k < 0$  for all  $k$  but  $\liminf_{k \rightarrow \infty} \|g_k\| \neq 0$ , then the sufficient descent condition (1.6) must hold for most of the iterates. More exactly, for any  $p \in (0, 1)$  there exists some constant  $c > 0$  such that, for any  $k$ , relation  $g_i^T d_i \leq -c\|g_i\|^2$  holds for at least  $[pk]$  values of  $i \in [1, k]$ . Secondly, under mild assumptions on  $f$ , the method is shown to be globally convergent with a variety of line searches, including several typical line searches such as the standard Wolfe line search, the Armijo line search and the one proposed in [13, 11]. Thus the result in [6] is extended to a great extent in this paper. Thirdly, we find that, instead of the negative gradient direction, the search direction defined by the method (1.7) can be used to restart any optimization method while guaranteeing the global convergence of the method. Some numerical results are also done, which shows that the new restart direction may be superior to the negative gradient direction. Conclusions are made in the last section.

## 2. Self-adjusting property

Throughout this paper, we assume that

$$(2.1) \quad g_k \neq 0, \quad \text{for all } k \geq 1,$$

for otherwise a stationary point has been found.

To begin with, let us define two important quantities that are

$$(2.2) \quad q_k = \frac{\|d_k\|^2}{(g_k^T d_k)^2}$$

and

$$(2.3) \quad r_k = -\frac{g_k^T d_k}{\|g_k\|^2}.$$

The quantity  $q_k$  shows the size of  $d_k$ , whereas  $r_k$  is a quantity showing the descent degree of  $d_k$ . In fact, if  $r_k > 0$ ,  $d_k$  is a descent direction. Furthermore, if  $r_k \geq c$  for some constant  $c > 0$ , then the sufficient descent condition (1.6) holds.

For the method (1.7), we get by multiplying (1.3) with  $g_k$  and using (1.7) that

$$(2.4) \quad g_k^T d_k = \frac{\|g_k\|^2}{d_{k-1}^T y_{k-1}} g_{k-1}^T d_{k-1}.$$

Since  $g_k^T d_k < 0$  follows  $g_{k-1}^T d_{k-1} < 0$  if  $d_{k-1}^T y_{k-1} > 0$ , and since  $d_1^T g_1 = -\|g_1\|^2 < 0$ , a direct consequence of (2.4) is that, the method (1.7) with the standard Wolfe line search produces a descent search direction at every iteration. Due to relation (2.4), the formula (1.7) can be also written as

$$(2.5) \quad \beta_k = \frac{g_k^T d_k}{g_{k-1}^T d_{k-1}}.$$

On the other hand, we have from (1.3) that  $d_k + g_k = \beta_k d_{k-1}$ . Hence

$$(2.6) \quad \|d_k\|^2 = \beta_k^2 \|d_{k-1}\|^2 - 2g_k^T d_k - \|g_k\|^2.$$

Substituting (2.5) into (2.6), we can then obtain

$$(2.7) \quad \frac{\|d_k\|^2}{(g_k^T d_k)^2} = \frac{\|d_{k-1}\|^2}{(g_{k-1}^T d_{k-1})^2} - \frac{2}{g_k^T d_k} - \frac{\|g_k\|^2}{(g_k^T d_k)^2}.$$

This with the definitions of  $q_k$  and  $r_k$  gives the relation

$$(2.8) \quad q_k = q_{k-1} + \frac{1}{\|g_k\|^2} \frac{2}{r_k} - \frac{1}{\|g_k\|^2} \frac{1}{r_k^2}.$$

We will see that relation (2.8) plays an important role in the coming analyses.

In fact, suppose that each  $d_k$  is a descent direction. Then the second term on the right side of (2.8) increases the value of  $q_{k-1}$ , whereas the third term decreases the value of  $q_{k-1}$ . Considering the two terms together, we see that  $q_{k-1}$  increases if and only if  $r_k \geq 1/2$ . If  $r_k$  is close to zero, then  $q_{k-1}$  will be significantly reduced, since the order of  $1/r_k$  in the second term is only one but its order in the third term is two. This and the fact that  $q_k \geq 0$  for all  $k$  imply that, in the case when  $q_{k-1}$  is very small,  $r_k$  must be relatively large. Such observations make us be able to give an estimation to the lower bound of the quantity  $r_k$ . To do so, we still need assume that there exist positive constants  $\gamma$  and  $\bar{\gamma}$  such that

$$(2.9) \quad 0 < \gamma \leq \|g_k\| \leq \bar{\gamma}, \quad \text{for all } k \geq 1.$$

**Theorem 2.1** *Consider the method (1.2), (1.3) and (1.7) where  $d_k$  is a descent direction. If (2.9) holds, there exist positive constants  $\delta_1, \delta_2$  and  $\delta_3$  such that relations*

$$(2.10) \quad -g_k^T d_k \geq \frac{\delta_1}{\sqrt{k}},$$

$$(2.11) \quad \|d_k\|^2 \geq \frac{\delta_2}{k},$$

and

$$(2.12) \quad r_k \geq \frac{\delta_3}{\sqrt{k}}$$

hold for all  $k \geq 1$ .

*Proof.* Summing (2.8) over the iterates and noting that  $d_1 = -g_1$ , we get that

$$(2.13) \quad q_k = \sum_{i=1}^k \frac{1}{\|g_i\|^2} \left( \frac{2}{r_i} - \frac{1}{r_i^2} \right).$$

Since  $q_k \geq 0$ , it follows from (2.13) that

$$(2.14) \quad \frac{1}{\|g_k\|^2} \left( -\frac{2}{r_k} + \frac{1}{r_k^2} \right) \leq \sum_{i=1}^{k-1} \frac{1}{\|g_i\|^2} \left( \frac{2}{r_i} - \frac{1}{r_i^2} \right),$$

which with (2.9) and the fact that

$$(2.15) \quad \frac{2}{r_i} - \frac{1}{r_i^2} \leq 1$$

yields the relation

$$(2.16) \quad \frac{1}{r_k^2} - \frac{2}{r_k} - \frac{\bar{\gamma}^2}{\gamma^2} (k-1) \leq 0.$$

This with the assumption that  $r_k > 0$  shows that

$$(2.17) \quad \frac{1}{r_k} \leq 1 + \sqrt{1 + \frac{\bar{\gamma}^2}{\gamma^2} (k-1)} \leq 1 + \frac{\bar{\gamma}}{\gamma} \sqrt{k} \leq \frac{2\bar{\gamma}}{\gamma} \sqrt{k}.$$

Thus (2.12) holds with  $\delta_3 = \gamma/(2\bar{\gamma})$ . Noting that

$$(2.18) \quad -g_k^T d_k = \|g_k\|^2 r_k$$

and that

$$(2.19) \quad \|d_k\| \geq \|g_k\| r_k,$$

we know from (2.12) and (2.9) that relations (2.10) and (2.11) hold with  $\delta_1 = \delta_3 \gamma^2$  and  $\delta_2 = \delta_3^2 \gamma^2$ , respectively. This completes our proof.  $\square$

Relation (2.12) does not imply that the sufficient descent condition holds. Under the same assumption, however, we can show that the sufficient descent condition must hold for most of the iterates.

**Theorem 2.2** Consider the method (1.2), (1.3) and (1.7) where  $d_k$  is a descent direction. If (2.9) holds, then for any  $p \in (0, 1)$  there exist constants  $\delta_4, \delta_5, \delta_6 > 0$  such that, for any  $k$ , the relations

$$(2.20) \quad -g_i^T d_i \geq \delta_4,$$

$$(2.21) \quad \|d_i\|^2 \geq \delta_5,$$

and

$$(2.22) \quad r_i \geq \delta_6$$

hold for at least  $[pk]$  values of  $i \in [1, k]$ .

*Proof.* For any  $p \in (0, 1)$ , we choose  $\delta_6 > 0$  to be so small that

$$(2.23) \quad \delta' \triangleq \frac{1}{\delta_6^2} - \frac{2}{\delta_6 \gamma} \geq \frac{\bar{\gamma}^2 p}{\gamma^2 (1 - p)}.$$

For this  $\delta_6$  and any  $k$ , we define

$$(2.24) \quad I_k = \{i \in [1, k] : r_i \geq \delta_6\}$$

and denote  $|I_k|$  to be the number of elements in  $I_k$ . By (2.8), (2.9) and the fact that  $q_k \geq 0$ , we can get that

$$(2.25) \quad \sum_{i \in [1, k] \setminus I_k} \left( -\frac{2}{r_i} + \frac{1}{r_i^2} \right) \leq \frac{\bar{\gamma}^2}{\gamma^2} \sum_{i \in I_k} \left( \frac{2}{r_i} - \frac{1}{r_i^2} \right).$$

It follows from this, (2.15) and the definition of  $I_k$  that

$$(2.26) \quad \delta' (k - |I_k|) \leq \frac{\bar{\gamma}^2}{\gamma^2} |I_k|,$$

where  $\delta'$  is given in (2.23). The above relation and (2.23) imply that

$$(2.27) \quad |I_k| \geq \frac{\delta' \gamma^2}{\delta' \gamma^2 + \bar{\gamma}^2} k \geq pk \geq [pk].$$

Therefore, for any  $p \in (0, 1)$ , if we choose  $\delta_6 > 0$  satisfying (2.23),  $\delta_4 = \delta_6 \gamma^2$  and  $\delta_5 = \delta_6^2 \gamma^2$ , we know from (2.27), (2.18), (2.19) and (2.9) that this theorem is true.  $\square$

Thus by Theorem 2.1 and 2.2, we expose the self-adjusting property of the method (1.7), that is independent of the line search and the function convexity. It is also interesting to note that Theorem 2.2 is very similar to one property of the BFGS variable metric method. Assuming that  $f$  is uniformly

convex, [3] proved that for any  $p \in (0, 1)$  there exists some positive constant  $c$  such that, for any  $k \geq 1$ , the relation

$$(2.28) \quad \cos \theta_i = \frac{-g_i^T d_i}{\|g_i\| \|d_i\|} \geq c$$

holds for at least  $[pk]$  values of  $i \in [1, k]$ . The differences between the two results are in that Theorem 2.2 needs not assume the uniform convexity of the function, and that Theorem 2.2 considers about the sufficient descent condition not the angle between  $-g_k$  and  $d_k$ .

### 3. Global convergence properties

In the above section, we have proved that the method (1.7) has certain self-adjusting property that is independent of the line search and the function convexity. In this section, we will make use of this property to establish the global convergence for the method (1.7) using a variety of line searches.

Suppose that the objective function satisfies the following assumption.

**Assumption 3.1** (i) The level set  $\mathcal{L} = \{x \in \mathbb{R}^n : f(x) \leq f(x_1)\}$  is bounded; (ii) In some neighborhood  $\mathcal{N}$  of  $\mathcal{L}$ ,  $f$  is differentiable and its gradient  $g$  is Lipschitz continuous, namely, there exists a constant  $L > 0$  such that

$$(3.1) \quad \|g(x) - g(\tilde{x})\| \leq L\|x - \tilde{x}\|, \quad \text{for all } x, \tilde{x} \in \mathcal{N}.$$

Suppose also that the line search is such that the following relation holds:

$$(3.2) \quad f_k - f_{k+1} \geq c \min \{-g_k^T d_k, \|d_k\|^2, q_k^{-1}\},$$

where  $c > 0$  is constant, and  $q_k$  is given in (2.2). Then we can show the following general convergence result for the method (1.7). The proof given below is by way of Theorem 2.1.

**Theorem 3.2** Suppose that  $x_1$  is a starting point for which Assumption 3.1 holds. Consider the method (1.2), (1.3) and (1.7) where  $d_k$  is a descent direction. If the line search is such that relation (3.2) holds for all  $k$ , we have that

$$(3.3) \quad \liminf_{k \rightarrow \infty} \|g_k\| = 0.$$

*Proof.* Assumption 3.1 implies that  $f$  is bounded below. Thus we can get by summing (3.2) over the iterates that

$$(3.4) \quad \sum_{k \geq 1} \min \{-g_k^T d_k, \|d_k\|^2, q_k^{-1}\} < +\infty.$$

Now we proceed by contradiction and assume that there exists some constant  $\gamma > 0$  such that

$$(3.5) \quad \|g_k\| \geq \gamma, \quad \text{for all } k \geq 1.$$

Because Assumption 3.1 implies that there exists some constant  $\bar{\gamma} > 0$  such that

$$(3.6) \quad \|g_k\| \leq \bar{\gamma}, \quad \text{for all } k \geq 1,$$

we know that (2.9) holds. Thus by Theorem 2.1, relations (2.10) and (2.11) hold for some positive constants  $\delta_1$  and  $\delta_2$ . Using (2.15) and (3.5) in (2.13), we can get that

$$(3.7) \quad q_k \leq q_{k-1} + \frac{1}{\gamma^2},$$

which with  $q_1 = 1$  implies that

$$(3.8) \quad q_k^{-1} \geq \frac{\gamma^2}{k}.$$

It follows from (2.10), (2.11) and (3.8) that

$$(3.9) \quad \sum_{k \geq 1} \min \{ -g_k^T d_k, \|d_k\|^2, q_k^{-1} \} = \infty.$$

The above relation contradicts (3.4). Therefore this theorem is true.  $\square$

The above theorem clearly extends [5]'s convergence result for the method (1.7) using the standard Wolfe line search, since under Assumption 3.1 on  $f$ , [21] proved that any descent method (1.2) with the standard Wolfe line search give the relation

$$(3.10) \quad f_k - f_{k+1} \geq c q_k^{-1},$$

where  $c > 0$  is some constant.

Besides the standard Wolfe line search, relation (3.2) can also be achieved by many other kinds of line searches. For example, the Armijo line search ([1]) is to choose the smallest nonnegative integer  $m$  such that the relation (1.4) holds with  $\alpha_k = \lambda^m$ , where  $\lambda \in (0, 1)$  is constant. For this line search, under Assumption 3.1 on  $f$ , one can show that

$$(3.11) \quad f_k - f_{k+1} \geq c \min \{ -g_k^T d_k, q_k^{-1} \}.$$

Another example is the line search proposed in [13] and [11]. This line search is to choose the smallest nonnegative integer  $m$  such that the steplength  $\alpha_k = \lambda^m$  satisfies

$$(3.12) \quad f(x_k + \alpha_k d_k) - f_k \leq -\delta \alpha_k^2 \|d_k\|^2,$$



where  $\lambda, \delta \in (0, 1)$  is constant. For this line search, one can establish the relation

$$(3.13) \quad f_k - f_{k+1} \geq c \min \{ \|d_k\|^2, q_k^{-1} \}.$$

Thus the line search condition (3.2) also holds.

If the objective function  $f$  is strictly convex, we can show that each search direction generated by the method (1.7) must be downhill. Thus by Theorem 3.2, we immediately have the following result.

**Corollary 3.3** *Suppose that  $x_1$  is a starting point for which Assumption 3.1 holds. Consider the method (1.2), (1.3) and (1.7). If  $f$  is strictly convex on the level set  $\mathcal{L}$ , namely,*

$$(3.14) \quad (g(x) - g(\tilde{x}))^T (x - \tilde{x}) > 0, \quad \text{for any } x, \tilde{x} \in \mathcal{L}, x \neq \tilde{x},$$

*then each  $d_k$  is a descent direction. Further, if the line search is such that relation (3.2) holds, the method converges in the sense that (3.3) holds.*

*Proof.* By Theorem 3.2, it suffices to show that

$$(3.15) \quad g_k^T d_k < 0, \quad \text{for all } k \geq 1.$$

In fact, since  $d_1 = -g_1$ , (3.15) clearly holds for  $k = 1$ . Suppose that  $g_{k-1}^T d_{k-1} < 0$ . It follows from (3.14) that

$$(3.16) \quad d_{k-1}^T y_{k-1} > 0.$$

Thus we know from (2.4) that  $g_k^T d_k < 0$ . Therefore by induction, (3.15) is true.  $\square$

In contrast with Theorems 2.3 and 2.4 in [6], the above Corollary needs not assume the uniform convexity of the objective function. For general objective functions, the search direction generated by the method (1.7) needs not be downhill if the line search is only such that (3.2) holds. For this, instead of the method (1.7), we may consider the corresponding method of

$$(3.17) \quad \beta_k = \frac{\|g_k\|^2}{\max\{d_{k-1}^T y_{k-1}, -g_{k-1}^T d_{k-1}\}}.$$

As is known, the conjugate descent method ([8]) computes  $\beta_k$  as follows:

$$(3.18) \quad \beta_k = \frac{\|g_k\|^2}{-g_{k-1}^T d_{k-1}}.$$

Therefore the method (3.17) can be regarded as a hybrid method of the method (1.7) and the conjugate descent method. In the case that the line

search satisfies (3.2), we can show the descent property and the global convergence of this hybrid method for general objective functions. The proof of this result can be achieved similarly to the results for the method (1.7), and hence is omitted here.

**Theorem 3.4** *Suppose that  $x_1$  is a starting point for which Assumption 3.1 holds. Consider the method (1.2), (1.3) and (3.17). Then each  $d_k$  is a descent direction. Further, if the line search is such that relation (3.2) holds, the method converges in the sense that (3.3) holds.*

Interestingly enough, we may also consider the following variant of the method (1.7):

$$(3.19) \quad d_k = -\frac{d_{k-1}^T y_{k-1}}{\|g_k\|^2} g_k + d_{k-1}.$$

If  $d_1 = -g_1$ , we have by direct calculations that

$$(3.20) \quad g_k^T d_k = -\|g_1\|^2, \quad \text{for all } k \geq 1,$$

which implies that each search direction is downhill if  $g_1 \neq 0$ . Similarly, we can prove that such a variant of the method (1.7) is globally convergent for general objective functions provided that the line search satisfies (3.2).

#### 4. Property as a restart direction of optimization methods

In the implementations of many optimization methods, one may often meet the difficulty that the search direction at some iteration is very poor. For example, the Newton's direction is not well defined if the Hessian of the objective function is singular. Even if the Hessian is nonsingular but not positive, the Newton's direction is not necessarily a descent direction. Another example is the Polak-Ribière-Polyak conjugate gradient method ([18, 19]). This method is now generally believed to be one of the most efficient conjugate gradient methods. Even for strictly convex quadratic functions, however, the Polak-Ribière-Polyak method with the strong Wolfe line search (1.4)–(1.5) may produce an uphill search direction ([4]).

In the case when the search direction  $d_k$  is poor, a simple remedy is to restart the method with the negative gradient direction  $-g_k$ . Such a remedy can easily guarantee the global convergence of the method, but has a major defect, that is, the second derivative information obtained along the previous direction  $d_{k-1}$  is discarded. A detailed description for this can be seen in [17]. In the following, we will show that the search direction defined by the method (1.7) can also be used to restart an optimization method while ensuring the global convergence of the method. Since the direction defined

by the method (1.7) includes the second derivative information that is found along  $d_{k-1}$ , it is reasonable to expect that the new restart direction would be more efficient than the negative gradient direction.

Now we denote  $d_k^{(1)}$  to be the direction defined by some optimization method, and  $d_k^{(2)}$  to be the direction defined by the method (1.7). Thus we can write

$$(4.1) \quad d_k^{(2)} = -g_k + \frac{\|g_k\|^2}{d_{k-1}^T y_{k-1}} d_{k-1}.$$

Given an initial direction  $d_1$  satisfying  $g_1^T d_1 < 0$ , for  $k \geq 2$  we consider the following direction:

$$(4.2) \quad d_k = \begin{cases} d_k^{(1)}, & \text{if } \cos \theta_k^{(1)} \geq \cos \theta_k^{(2)}; \\ d_k^{(2)}, & \text{otherwise.} \end{cases}$$

In the above relation,  $\theta_k^{(1)}$  and  $\theta_k^{(2)}$  stand for the angles between  $-g_k$  and  $d_k^{(1)}$ , and between  $-g_k$  and  $d_k^{(2)}$ , respectively. For the method (1.2) where  $d_k$  is given in (4.2), we can prove the following general result.

**Theorem 4.1** *Suppose that  $x_1$  is a starting point for which Assumption 3.1 holds. Consider the method (1.2) and (4.2), where  $d_k^{(1)}$  is generated by any optimization method and  $d_k^{(2)}$  is defined in (4.1), and where the line search satisfies the standard Wolfe conditions (1.4) and (1.9). Then if  $g_1^T d_1 < 0$ , we have that*

$$(4.3) \quad g_k^T d_k < 0, \quad \text{for all } k \geq 1.$$

Further, we have that  $\liminf_{k \rightarrow \infty} \|g_k\| = 0$ .

*Proof.* First, we show (4.3) by induction. The assumption that  $g_1^T d_1 < 0$  implies that (4.3) is true for  $k = 1$ . Suppose that

$$(4.4) \quad g_{k-1}^T d_{k-1} < 0.$$

Then we have from the line search condition (1.5) that

$$(4.5) \quad d_{k-1}^T y_{k-1} > 0.$$

Then it follows from (4.4), (4.5) and (2.4) with  $d_k$  replaced by  $d_k^{(2)}$  that

$$(4.6) \quad g_k^T d_k^{(2)} < 0.$$

This together with the definition of  $d_k$  shows that  $g_k^T d_k < 0$ . Thus by induction, (4.3) holds for all  $k \geq 1$ .

By the definition of  $d_k$ , we also have that

$$(4.7) \quad \frac{(g_k^T d_k)^2}{\|d_k\|^2} \geq \frac{(g_k^T d_k^{(2)})^2}{\|d_k^{(2)}\|^2}.$$

For the direction  $d_k^{(2)}$  given in (4.1), we have similarly to (2.7) that

$$(4.8) \quad \frac{\|d_k^{(2)}\|^2}{(g_k^T d_k^{(2)})^2} = \frac{\|d_{k-1}\|^2}{(g_{k-1}^T d_{k-1})} - \frac{2}{g_k^T d_k^{(2)}} - \frac{\|g_k\|^2}{(g_k^T d_k^{(2)})^2},$$

which implies that

$$(4.9) \quad \begin{aligned} \frac{\|d_k^{(2)}\|^2}{(g_k^T d_k^{(2)})^2} &= \frac{\|d_{k-1}\|^2}{(g_{k-1}^T d_{k-1})} + \frac{1}{\|g_k\|^2} \left[ 1 - \left( 1 + \frac{\|g_k\|^2}{g_k^T d_k^{(2)}} \right)^2 \right] \\ &\leq \frac{\|d_{k-1}\|^2}{(g_{k-1}^T d_{k-1})} + \frac{1}{\|g_k\|^2}. \end{aligned}$$

Combining (4.7) and (4.9), we get that

$$(4.10) \quad \frac{\|d_k\|^2}{(g_k^T d_k)^2} \leq \frac{\|d_{k-1}\|^2}{(g_{k-1}^T d_{k-1})} + \frac{1}{\|g_k\|^2}.$$

Suppose that the theorem is not true and there exists a constant  $\gamma > 0$  such that

$$(4.11) \quad \|g_k\| \geq \gamma, \quad \text{for all } k \geq 1.$$

Then by summing (4.10), we can obtain

$$(4.12) \quad \frac{\|d_k\|^2}{(g_k^T d_k)^2} \leq \frac{\|d_1\|^2}{(g_1^T d_1)^2} + \frac{k-1}{\gamma^2},$$

which shows that

$$(4.13) \quad \sum_{k \geq 1} \frac{(g_k^T d_k)^2}{\|d_k\|^2} = \infty.$$

This contradicts the Zoutendijk condition, since [21] showed that, under Assumption 3.1 on  $f$ , any descent method (1.2) with the standard Wolfe line search gives

$$(4.14) \quad \sum_{k \geq 1} \frac{(g_k^T d_k)^2}{\|d_k\|^2} < \infty.$$

The contradiction shows that  $\liminf_{k \rightarrow \infty} \|g_k\| = 0$ . □

The above theorem shows that the direction defined by the method (1.7) can be used to restart any optimization method while ensuring the descent property and the global convergence of the method.

To compare the efficiency of using the new restart direction with that of using the negative gradient direction, some numerical experiments have been done with double precisions on an SGI Indigo workstation. Our tests use the memoryless BFGS method ([2]) to generate the direction  $d_k^{(1)}$ , namely,

$$(4.15) \quad d_k^{(1)} = -P_k g_k,$$

where

$$(4.16) \quad P_k = I - \frac{s_{k-1}^T y_{k-1} + y_{k-1} s_{k-1}^T}{s_{k-1}^T y_{k-1}} + \left( 1 + \frac{\|y_{k-1}\|^2}{s_{k-1}^T y_{k-1}} \right) \frac{s_{k-1} s_{k-1}^T}{s_{k-1}^T y_{k-1}},$$

and where  $s_{k-1} = x_k - x_{k-1}$ ,  $y_{k-1} = g_k - g_{k-1}$ . Besides the original memoryless BFGS method, we also tested the method with two different restart strategies. The first is to restart the method with the negative gradient direction  $-g_k$  if the angle  $\theta_k^{(1)}$  between  $-g_k$  and  $d_k^{(1)}$  is close to  $90^\circ$ ; more exactly, we tested the method

$$(4.17) \quad \bar{d}_k = \begin{cases} d_k^{(1)}, & \text{if } \cos \theta_k^{(1)} \geq 0.1; \\ -g_k, & \text{otherwise.} \end{cases}$$

The second is to restart the method with the direction  $d_k^{(2)}$  given in (4.1); more exactly, we tested the method

$$(4.18) \quad \hat{d}_k = \begin{cases} d_k^{(1)}, & \text{if } \cos \theta_k^{(1)} \geq \min\{\cos \theta_k^{(2)}, 0.1\}; \\ d_k^{(2)}, & \text{otherwise,} \end{cases}$$

where  $\theta_k^{(2)}$  still denotes the angle between  $-g_k$  and  $d_k^{(2)}$ . For convenience, we call the above three versions of the memoryless BFGS method as Algorithm 1, Algorithm 2, and Algorithm 3, respectively.

The test problems are taken from Moré, Garbow and Hillstom [14]. Our line search subroutine computes the steplength  $\alpha_k$  such that the strong Wolfe conditions (1.4)–(1.5) hold with  $\delta = 0.01$  and  $\sigma = 0.1$ . The stopping condition is

$$(4.19) \quad \|g_k\|_2 \leq 10^{-6}.$$

The numerical results are listed in Table 1. They are written in the form of NI/NF/NG, where NI, NF, NG are numbers of iterations, function evaluations, and gradient evaluations, respectively. In addition, column “P” denotes the number of the test problem, and “n” the number of variables.

**Table 1.** Testing different versions of memoryless BFGS method

P	n	Alg. 1	Alg. 2	Alg. 3
1	3	103/353/136	59/206/85	56/203/83
2	6	130/273/212	184/377/284	130/273/212
3	3	3/7/5	3/7/5	3/7/5
5	3	8/28/21	8/24/17	8/28/21
6	6	4/14/10	4/14/10	4/14/10
8	8	85/268/232	32/86/73	30/89/76
9	3	7/18/11	7/18/11	7/18/11
13	20	55/111/109	55/111/109	55/111/109
14	14	24/120/67	35/132/77	22/90/51
15	16	194/558/248	97/293/124	114/331/149
16	2	10/31/20	10/31/20	10/31/20
17	4	135/566/255	164/552/217	86/306/122
18	8	34/99/47	34/99/47	34/99/47

**Table 2.** Numerical comparisons

Alg.	Alg. 1	Alg. 2	Alg. 3
NI	792	692	559
NF	2446	1950	1600
NG	1373	1079	916

From Table 1, we can see that the performances of the three algorithms for seven of the test problems are the same. We can also see that Algorithm 2 sometimes performs worse than Algorithm 1, for example for Problem 2 and 14, whereas Algorithm 3 performs not worse than Algorithm 1 for all the test problems. Nevertheless, Algorithm 3 does not perform uniformly better than Algorithm 2, for example for Problem 15.

For further comparisons, we sum all the numerical results for each of the three algorithms. See Table 2, in which *NI*, *NF*, *NG* denote the total numbers of iterations, function evaluations, and gradient evaluations, respectively. From Table 2, we can clearly find that Algorithm 3 is better than Algorithm 2. Therefore our numerical results show that instead of the negative gradient direction, the direction defined by the method (1.7) can be used to restart an optimization method and the numerical performances may be better.

## 5. Conclusions

We have further analyzed the nonlinear conjugate gradient method in [5] and exposed several new properties of the method. Specifically, we have found that the method has a certain self-adjusting property. Such a property is independent of the line search and the function convexity, and is very useful

in the convergence analyses of the method. Under mild assumptions on the objective function, the method is shown to be globally convergent with a variety of line searches. We have also found that instead of the negative gradient direction, the direction defined by the method in [5] can be used to restart any optimization method while guaranteeing the global convergence of the method. Some numerical results are made, which showed that the new restart direction may be better than the negative gradient direction.

*Acknowledgements.* The author is much indebted to Professor Ya-xiang Yuan for his meticulous supervisions during the last five years and many valuable suggestions on the draft of this paper. Thanks are also due to the two anonymous referees, whose comments much improved this paper.

## References

1. L. Armijo, Minimization of functions having Lipschitz continuous first partial derivatives. *Paci. J. Math.* **16**(1), 1–3 (1966)
2. A. Buckley, Conjugate gradient methods, in: M. J. D. Powell, ed., *Nonlinear Optimization* 1981, pp. 17–22, London: Academic Press (1982)
3. R. H. Byrd, J. Nocedal, A Tool for the Analysis of Quasi-Newton Methods with Application To Unconstrained Minimization. *SIAM J. Numer. Anal.* **26**(3), 727–739 (1989)
4. Y. H. Dai, Analyses of Nonlinear Conjugate Gradient Methods, Ph.D. thesis, Institute of Computational Mathematics and Scientific/Engineering Computing, Chinese Academy of Sciences, 1997
5. Y. H. Dai, Y. Yuan, A Nonlinear Conjugate Gradient Method with A Strong Global Convergence Property. *SIAM Journal on Optimization* **10**(1), 177–182 (1999)
6. Y. H. Dai, Y. Yuan, Some properties of a new conjugate gradient method, in: Y. Yuan ed., *Advances in Nonlinear Programming* pp. 251–262, Boston: Kluwer 1998
7. J. W. Daniel, The conjugate gradient method for linear and nonlinear operator equations. *SIAM J. Numer. Anal.* **4**, 10–26 (1967)
8. R. Fletcher, *Practical Methods of Optimization* vol. 1: Unconstrained Optimization. New York: John Wiley & Sons (1987)
9. R. Fletcher, C. Reeves, Function minimization by conjugate gradients, *Comput. J.* **7**, 149–154 (1964)
10. J. C. Gilbert, J. Nocedal, Global convergence properties of conjugate gradient methods for optimization. *SIAM J. Opt.* **2**(1), 21–42 (1992)
11. L. Grippo, F. Lampariello, S. Lucidi, Global convergence and stabilization of unconstrained minimization methods without derivatives. *J. Opt. Theory Appl.* **56**, 385–406 (1988)
12. M. R. Hestenes, E. L. Stiefel, Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Standards Sect.* **5**(49), 409–436 (1952)
13. R. De Leone, M. Gaudioso, L. Grippo, Stopping criteria for linesearch methods without derivatives. *Mathematical Programming* **30**, 285–300 (1984)
14. J. J. Moré, B. S. Garbow, K. E. Hillstom, Testing unconstrained optimization software. *ACM Transact. on Math. Software* **7**, 17–41 (1981)
15. Y. Liu, C. Storey, Efficient Generalized Conjugate Gradient Algorithms, Part 1: Theory. *J. Opt. Theory Appl.* **69**, 129–137 (1991)

16. J. L. Nazareth, Conjugate-gradient Methods, to appear in: C. Floudas, P. Pardalos, eds., *Encyclopedia of Optimization* (Kluwer Academic Publishers, Boston, USA and Dordrecht, The Netherlands 1999)
17. M. J. D. Powell, Restart procedures of the conjugate gradient method. *Math. Program.* **2**, 241–254 (1977)
18. E. Polak, G. Ribière, Note sur la convergence de directions conjuguées. *Rev. Francaise Informat Recherche Opertionelle*, 3e Année **16**, 35–43 (1969)
19. B. T. Polyak, The conjugate gradient method in extremem problems, *USSR Comp. Math. Math. Phys.* **9**, 94–112 (1969)
20. D. F. Shanno, Conjugate gradient methods with inexact searches, *Math. Oper. Res.* **3**, 244–256 (1978)
21. G. Zoutendijk, Nonlinear Programming, Computational Methods, in: *Integer and Non-linear Programming* (J. Abadie, ed.), North-Holland (Amsterdam), pp. 37–86, 1970