AN EXTENDED CLASS OF NONLINEAR CONJUGATE GRADIENT METHODS*

Y. H. Dai and Y. Yuan

State Key Laboratory of Scientific and Engineering Computing,
Institute of Computational Mathematics and Scientific/Engineering Computing,
Academy of Mathematics and System Sciences, Chinese Academy of Sciences,
P. O. Box 2719, Beijing 100080, P. R. China. Email: dyh,yyx@lsec.cc.ac.cn

Abstract

Conjugate gradient methods are very important methods for solving nonlinear optimization problems, especially for large-scale problems. However, unlike quasi-Newton methods, some of which can be expressed in a unified way and consequently can be analyzed uniformly, conjugate gradient methods are often analyzed individually. Recently, Dai and Yuan (1998) presented a one-parameter class of conjugate gradient methods and gave a uniform proof to their convergences under certain line search conditions. The class of methods can be viewed as some kind of CONVEX combinations of the FR method and the DY method. This paper extends the convergent interval $[0, 1]$ of the parameter in the class to the interval $(-\infty, +\infty)$. Therefore the extended class of methods are some kind of LINEAR combinations of the FR method and the DY method. Some explanatory numerical results are also provided.

Key words: conjugate gradient, line search, global convergence.
AMS classifications: 65k, 90c.

1. Introduction

Consider the unconstrained optimization problem

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

where f is smooth and its gradient g is available. Conjugate gradient methods are very important methods for solving (1.1), especially for large scale problems, which have the following form

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

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

where $g_k = \nabla f(x_k)$, α_k is a stepsize obtained by a one-dimensional line search, and β_k is a scalar. Many authors have proposed different formulae for the scalar β_k since the introduction

*Research partly supported by Chinese NSF grants 19525101, 19731010 and 19801033.

of nonlinear conjugate gradient method by Fletcher-Reeves [5]. However, unlike variable metric methods, nonlinear conjugate gradient methods are often analyzed individually.

In a recent paper [2], Dai and Yuan presented a one-parameter class of conjugate gradient methods, in which the scalar β_k is of the form

$$\beta_k = \frac{||g_k||^2}{\lambda||g_{k-1}||^2 + (1-\lambda)d_{k-1}^T y_{k-1}}, \tag{1.4}$$

where $||\cdot||$ means the Euclidean norm and $y_{k-1} = g_k - g_{k-1}$. The parameter λ takes any value in $[0,1]$. When λ equals 1 and 0, the method (1.2)–(1.4) reduces to the Fletcher-Reeves method and the conjugate gradient method in [1] respectively. Therefore the class of conjugate gradient methods in [2] can be regarded as some kind of convex combination of the Fletcher-Reeves method and the method in [1].

This paper will extend the interval $[0,1]$ of λ in the class to the interval $(-\infty,+\infty)$. Thus the extended class of methods is some kind of linear combination of the Fletcher-Reeves method and the method in [1]. Convergence properties of the extended class of methods are studied in this paper. Specifically, assuming the line search conditions are the generalized Wolfe conditions:

$$f(x_k + \alpha_k d_k) - f(x_k) \leq \delta \alpha_k g_k^T d_k, \tag{1.5}$$

$$\sigma_1 g_k^T d_k \leq g(x_k + \alpha_k d_k)^T d_k \leq -\sigma_2 g_k^T d_k, \tag{1.6}$$

where $0 < \delta < \sigma_1 < 1$ and $\sigma_2 > 0$, the method (1.2)–(1.4) with any $\lambda \in (-\infty,+\infty)$ is shown to produce a descent direction at every iteration and converges globally if the scalars σ_1 and σ_2 are such that

$$\sigma_1 - 1 \leq (\sigma_1 + \sigma_2)\lambda \leq 1. \tag{1.7}$$

This paper is organized as follows. Some preliminaries are provided in the next section. Section 3 analyzes the convergence properties of the extended class of conjugate gradient methods; the hybrid methods related to the extended class are also briefly considered in this section. Numerical results are presented in Section 4, which demonstrate our convergence results. Some discussions are made in the last section.

2. Preliminaries

Throughout this paper, we assume that $g_k \neq 0$ for all k, for otherwise a stationary point has been found. We also assume that the objective function always satisfies the following assumption.

Assumption 2.1 *(i) f is bounded below on the level set $\mathcal{L} = \{x \in \Re^n : f(x) \leq f(x_1)\}$; (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*

$$\|g(x) - g(\tilde{x})\| \leq L\|x - \tilde{x}\|, \qquad \textit{for all } x,\ \tilde{x} \in \mathcal{N}. \tag{2.1}$$

Since the Wolfe conditions are corresponding to (1.5)–(1.6) with $\sigma_2 = +\infty$, we can derive the following lemma from [8].

Lemma 2.2 *Suppose that x_1 is a starting point for which Assumption 2.1 holds. Consider any iterative method (1.2), where d_k is a descent direction and α_k satisfies the line search conditions (1.5)–(1.6), where $\sigma_1 \in (0,1)$ and $\sigma_2 \geq 0$. Then*

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

By direct calculations (or see [2]), we can without difficulty obtain the following equivalent formula of formula (1.4):

$$\beta_k = \frac{\lambda||g_k||^2 + (1-\lambda)(-g_k^T d_k)}{\lambda||g_{k-1}||^2 + (1-\lambda)(-g_{k-1}^T d_{k-1})}. \tag{2.3}$$

Denoting

$$\phi_k = \lambda||g_k||^2 + (1-\lambda)(-g_k^T d_k), \tag{2.4}$$

the formula (1.4) can be written as

$$\beta_k = \frac{\phi_k}{\phi_{k-1}}. \tag{2.5}$$

To analyze the class of the methods (1.4), we draw the following lemmas for the general method from [2]. In the lemmas,

$$r_k = -\frac{g_k^T d_k}{\phi_k}. \tag{2.6}$$

Lemma 2.3 *Suppose that x_1 is a starting point for which Assumption 2.1 holds. Consider the method (1.2), (1.3) and (2.5), if for all k, d_k is a descent direction and α_k satisfies the line search conditions (1.5)–(1.6), and if*

$$\sum_{k\geq 1} r_k^2 = +\infty, \tag{2.7}$$

we have that

$$\liminf_{k\to+\infty} \|g_k\| = 0. \tag{2.8}$$

Lemma 2.4 *Suppose that x_1 is a starting point for which Assumption 2.1 holds. Consider the method (1.2), (1.3) and (2.5), if for all k, d_k is a descent direction and α_k satisfies the line search conditions (1.5)–(1.6), and if*

$$\sum_{k\geq 1} \frac{\|g_k\|^2}{\phi_k^2} = +\infty, \tag{2.9}$$

we have that (2.8) holds.

In addition, we also need the following lemma. Its proof is self-evident.

Lemma 2.5 *Consider the following 1-dimensional function,*

$$\rho(t) = \frac{a+bt}{c+dt}, \quad t \in R^1, \tag{2.10}$$

where a, b, c and $d \neq 0$ are given real numbers. If

$$bc - ad > 0, \tag{2.11}$$

$\rho(t)$ is strictly monotonically increasing for $t < -\frac{c}{d}$ and $t > -\frac{c}{d}$; otherwise, if

$$bc - ad < 0, \tag{2.12}$$

$\rho(t)$ is strictly monotonically decreasing for $t < -\frac{c}{d}$ and $t > -\frac{c}{d}$.

3. Convergence properties

By (1.2), (1.3) and (1.4), direct calculations show that

$$g_k^T d_k = -\frac{\lambda(\|g_{k-1}\|^2 - g_k^T d_{k-1}) + (1-\lambda)(-g_{k-1}^T d_{k-1})}{\lambda\|g_{k-1}\|^2 + (1-\lambda)d_{k-1}^T y_{k-1}}\|g_k\|^2. \tag{3.1}$$

Substituting (3.1) into (2.4), we obtain

$$\phi_k = \frac{\lambda\|g_{k-1}\|^2 + (1-\lambda)(-g_{k-1}^T d_{k-1})}{\lambda\|g_{k-1}\|^2 + (1-\lambda)d_{k-1}^T y_{k-1}}\|g_k\|^2. \tag{3.2}$$

Define $\overline{r}_k = -\frac{g_k^T d_k}{\|g_k\|^2}$ and $l_k = \frac{g_{k+1}^T d_k}{g_k^T d_k}$. Then we have by (3.1) that

$$\overline{r}_k = \frac{\lambda + (1-\lambda+\lambda l_{k-1})\overline{r}_{k-1}}{\lambda + (1-\lambda)(1-l_{k-1})\overline{r}_{k-1}}. \tag{3.3}$$

Using this relation, we can show the following convergence result for the extended class of methods with $\lambda > 1$.

Theorem 3.1 *Suppose that x_1 is a starting point for which Assumption 2.1 holds. Consider the method (1.2)–(1.4), where $\lambda > 1$ and α_k satisfies the line search conditions (1.5) and (1.6). If the scalars σ_1 and σ_2 in (1.6) are such that*

$$\sigma_1 + \sigma_2 \leq \lambda^{-1}, \tag{3.4}$$

then we have for all $k \geq 1$,

$$\overline{r}_k > 0. \tag{3.5}$$

Further, the method converges in the sense that (2.8) is true.

Proof First, we show that

$$0 < \overline{r}_k < (1-\sigma_1)^{-1} \tag{3.6}$$

for all $k \geq 1$. Since $d_1 = -g_1$ and hence $\overline{r}_1 = 1$, (3.6) clearly holds for $k = 1$. We now suppose that (3.6) is true for some $k-1$. The right hand side of (3.3) can be regarded as a function of λ, l_{k-1} and $\overline{r}_{k-1}$. For convenience, we denote it to be $\psi(\lambda, l_{k-1}, \overline{r}_{k-1})$. The search condition (1.6) implies that

$$-\sigma_2 \leq l_{k-1} \leq \sigma_1. \tag{3.7}$$

Then by Lemma 2.5, we get that

$$\overline{r}_k \le \psi(\lambda, \sigma_1, \overline{r}_{k-1}) < \psi(\lambda, \sigma_1, (1-\sigma_1)^{-1}) = (1-\sigma_1)^{-1}. \tag{3.8}$$

On the other hand, by Lemma 2.5 and relation (3.4), it follows that

$$\overline{r}_k \ge \psi(\lambda, -\sigma_2, \overline{r}_{k-1}) > \psi(\lambda, -\sigma_2, (1-\sigma_1)^{-1}) \ge 0. \tag{3.9}$$

Thus (3.6) is true for k. By induction, (3.6) holds for all $k \ge 1$.

To show the truth of (2.8), by Lemma 2.3, it suffices to prove that

$$\max\{r_{k-1}, r_k\} \ge c_1 \tag{3.10}$$

for all $k \ge 2$ and some constant $c_1 > 0$. In fact, if

$$\overline{r}_{k-1} \le 1, \tag{3.11}$$

we can get by Lemma 2.5 that

$$\overline{r}_k \ge \psi(\lambda, -\sigma_2, 1). \tag{3.12}$$

It follows from (3.4) that for $\lambda > 1$,

$$\psi(\lambda, -\sigma_2, 1) = \frac{1-\lambda\sigma_2}{1+\sigma_2-\lambda\sigma_2} \ge \frac{\sigma_1\lambda}{\sigma_2+\sigma_1\lambda} \ge \frac{\sigma_1}{\sigma_1+\sigma_2}. \tag{3.13}$$

Therefore we always have

$$\max\{\overline{r}_{k-1}, \overline{r}_k\} \ge \frac{\sigma_1}{\sigma_1+\sigma_2} \tag{3.14}$$

for all $k \ge 2$. Further, by the definition of r_k and relation (2.4), we get that

$$r_k = \frac{\overline{r}_k}{\lambda + (1-\lambda)\overline{r}_k}, \tag{3.15}$$

which, with (3.14) and Lemma 2.5, implies that (3.10) is also true with $c_1 = \frac{\sigma_1}{\sigma_1+\sigma_2}$. This completes our proof. □

The above theorem shows that for $\lambda > 1$, if the parameters satisfy (3.4), then the method (1.2)–(1.4) produces a descent direction at every iteration and converges globally. For the case that $\lambda < 0$, we can prove the following theorem.

Theorem 3.2 *Suppose that x_1 is a starting point for which Assumption 2.1 holds. Consider the method (1.2)–(1.4), where $\lambda < 0$ and α_k satisfies the line search conditions (1.5) and (1.6). If the scalars σ_1 and σ_2 in (1.6) are such that*

$$(\sigma_1+\sigma_2)\lambda \ge \sigma_1 - 1, \tag{3.16}$$

then we have for all $k \ge 1$,

$$\overline{r}_k > (1+\sigma_2)^{-1}. \tag{3.17}$$

Further, the method converges in the sense that (2.8) is true.

Proof First, we show by induction that (3.17) holds for all $k \geq 1$. In fact, since $d_1 = -g_1$, we clearly have that $\overline{r}_1 = 1 > (1+\sigma_2)^{-1}$. Suppose that

$$\overline{r}_{k-1} > (1+\sigma_2)^{-1}. \tag{3.18}$$

It follows from the fact that $\lambda < 0$, (3.7), (3.18) and (3.16) that

$$\begin{aligned} \lambda + (1-\lambda)(1-l_{k-1})\overline{r}_{k-1} &> \lambda + (1-\lambda)(1-\sigma_1)(1+\sigma_2)^{-1} \\ &= [(1-\sigma_1) + (\sigma_1+\sigma_2)\lambda](1+\sigma_2)^{-1} \geq 0. \end{aligned} \tag{3.19}$$

The above relation implies that the formula (3.3) of $\overline{r}_k$ has a denominator greater than zero. Thus the β_k given by (1.4) is well defined. Similarly to (3.19), we can verify

$$\lambda + (1-\lambda)\overline{r}_{k-1} > \lambda + (1-\lambda)(1+\sigma_2)^{-1} > 0. \tag{3.20}$$

Using this relation and the fact that $-\lambda\sigma_2 > 0$, we can get from Lemma 2.5 that

$$\overline{r}_k > \psi(-\sigma_2, \overline{r}_{k-1}, \lambda) > \psi(-\sigma_2, (1+\sigma_2)^{-1}), \lambda = (1+\sigma_2)^{-1}, \tag{3.21}$$

where $\psi(\lambda, l_{k-1}, \overline{r}_{k-1})$ still denotes the function given by (3.3). Thus by induction, (3.17) holds for all $k \geq 1$.

In the case that $\lambda < 0$, we know from Lemma 2.5, (3.17) and (3.20) that the r_k given in (3.15) is monotonically decreasing as $\overline{r}_k \to \infty$. Thus

$$r_k \geq \lim_{\overline{r}_k \to \infty} \frac{\overline{r}_k}{\lambda + (1-\lambda)\overline{r}_k} = \frac{1}{1-\lambda}, \tag{3.22}$$

which shows that (2.7) holds. Therefore by Lemma 2.3, (2.8) holds. □

Combining the above two theorems, Theorem 2.3 in [1] and Theorem 4.2 in [2], we can state the following unified result for the extended class of conjugate gradient methods.

Theorem 3.3 *Suppose that x_1 is a starting point for which Assumption 2.1 holds. Consider the method (1.2)–(1.4), where $\lambda \in (-\infty, +\infty)$ and α_k satisfies the line search conditions (1.5) and (1.6). If the scalars σ_1 and σ_2 in (1.6) are such that*

$$\sigma_1 - 1 \leq (\sigma_1+\sigma_2)\lambda \leq 1, \tag{3.23}$$

then the method produces a descent search direction at every iteration. Further, the method converges in the sense that (2.8) holds.

The above theorem can easily be extended to the hybrid methods related to the extended class of conjugate gradient methods. The proof of the following theorem is not difficult and hence omitted here.

Theorem 3.4 *Suppose that x_1 is a starting point for which Assumption 2.1 holds. Consider the method (1.2), (1.3) and*

$$\tilde{\beta}_k = \frac{\max\{0, \min\{g_k^T y_{k-1}, ||g_{k-1}||^2\}\}}{\lambda||g_{k-1}||^2 + (1-\lambda)d_{k-1}^T y_{k-1}}, \tag{3.24}$$

where $\lambda \in (-\infty, +\infty)$ and α_k satisfies (1.5) and (1.6). If (3.23) holds, then the method produces a descent direction at every iteration. Further, the method converges in the sense that (2.8) holds.

4. Numerical experiments

We tested the extended class of conjugate gradient methods with double precisions on an SGI Indigo workstation. The test problems and the used starting points were taken from Morè, Garbow and Hillstrom [6]. The stopping condition is

$$\|g_k\| \leq 10^{-6}, \tag{4.1}$$

which is stronger than those normally used in real applications. The upper bound for the number of function evaluations is set to 9999. We also terminate the calculation if the function value change is too small,

$$|f(x_k) - f(x_{k+1})|/[1 + |f(x_k)|]) \leq 10^{-16}, \tag{4.2}$$

in which case we denote it by a superscript "*". In addition, we write "*Failed*" if d_k is so large that a numerical overflow happens while the algorithm tries to compute $f(x_k + d_k)$. We use "$\uparrow$" to denote the case whenever a descent search direction is not produced by the algorithm.

The scalar δ in the line search condition (1.5) is always set to 0.01. The parameters σ_1 and σ_2 are both 0.5. From (3.23), we see that λ can be chosen from $[-0.5, 1]$. We tested for λ=-0.5, -0.25, 0, 0.5, and 1 respectively.

The numerical results are listed in Table 4.1. They are written in the form of I/F/G, where I, F, G 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. Our numerical computations show that the algorithm always produces a descent direction at every iteration for all the problems, and it solves most of the 18 test problems. Therefore our numerical experiments fully demonstrate the convergence results, namely, Theorems 3.1 and 3.2, of the extended class of conjugate gradient methods.

Table 4.1. Numerical results

P	N	$\lambda = -\frac{1}{2}$	$\lambda = -\frac{1}{4}$	$\lambda = 0$	$\lambda = \frac{1}{2}$	$\lambda = 1$
1	3	85/274/97	31/106/40	139/434/151	70/226/80	57/191/68
2	6	526/1463/631	1360/4071/1363	1088/3255/1092	248/465/298	1766/5290/1774
3	3	3/7/5	3/7/5	3/7/5	3/7/5	3/7/5
4	2	>9999	>9999	>9999	>9999	>9999
5	3	132/250/152	46/91/59	63/116/81	27/57/40	74/139/89
6	6	4/17/7	4/17/7	4/17/7	4/17/7	4/17/7
7	9	1441/4333/1457	1480/4457/1493	1130/3400/1139	1236/3709/1245	1412/4247/1419
8	8	32/77/45	43/93/59	35/75/48	49/110/87	32/62/50
9	3	16/38/21	17/37/20	18/38/19	30/74/38	17/37/19
10	2	>9999	>9999	>9999	>9999	>9999
11	4	59/264/75*	*Failed*	*Failed*	*Failed*	*Failed*
12	3	*Failed*	287/2389/814	400/1234/430	480/1462/493	205/609/242
13	20	>9999	>9999	>9999	>9999	1797/5320/1815
14	14	72/264/92	54/194/66	86/284/96	52/183/74	88/289/109
15	16	>9999	>9999	>9999	932/2796/937	178/528/188
16	2	50/131/57	31/81/36	46/116/54	57/160/61	30/78/40
17	4	1385/5584/1416*	99/305/104	53/164/58	70/222/77	41/128/43
18	8	325/987/338	247/752/259	295/893/303	231/700/239	244/743/253

5. Some discussions

This paper extends the class of conjugate gradient methods in [2]. The extended class can be regarded as some kind of linear combination of the Fletcher-Reeves method and the conjugate gradient method in [1]. A unified condition, namely, (3.23), is established, which ensures the descent property and global convergence of the extended class of methods under mild assumptions on the objective function. The numerical experiments made in Section 4 further demonstrated the result.

As mentioned in Section 2, the Wolfe conditions are corresponding to (1.5)–(1.6) with $\sigma_2 = +\infty$. From (3.23), we see that the scalar σ_2 must be less than infinity if λ is not equal to zero. Therefore among the extended class of methods (1.4), the conjugate gradient method in [1], which corresponds with $\lambda = 0$, is the unique method which can guarantee the descent property and global convergence provided that the line search satisfies the Wolfe conditions.

This paper is based on a recent report [3]. Almost simultaneously, L. Nazareth [7] proposed a two-parameter family of conjugate gradient methods. Later, Dai and Yuan [4] proposed a three-parameter family of conjugate gradient methods. Our numerical results reported in the paper set λ to different constants. However, it is also possible to choose λ depending on informations at the current iteration. For example, we tried the formular $\lambda_k = \frac{3}{2} l_{k-1} + \frac{1}{4}$, which produced encouraging results. More researches are needed on this issue.

References

[1] Y. H. Dai and Y. Yuan, *A Nonlinear Conjugate Gradient Method with A Strong Global Convergence Property*, SIAM Journal on Optimization 10 : 1 (1999), pp. 177-182.

[2] Y. H. Dai and Y. Yuan, *A class of globally convergent conjugate gradient methods*, Research report ICM-98-030, Institute of Computational Mathematics and Scientific/Engineering Computing, Chinese Academy of Sciences, 1998.

[3] Y. H. Dai and Y. Yuan, *An extended class of globally convergent conjugate gradient methods*, Research report ICM-98-049, Institute of Computational Mathematics and Scientific/Engineering Computing, Chinese Academy of Sciences, 1998.

[4] Y. H. Dai and Y. Yuan, *A three-parameter family of nonlinear conjugate gradient methods*, Mathematics of Computation 70 (2001), pp. 1155-1167.

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

[6] J. J. Morè, B. S. Garbow and K. E. Hillstrom, *Testing unconstrained optimization software*, ACM Transactions on Mathematical Software 7 (1981), pp. 17-41.

[7] L. Nazareth, *Conjugate-Gradient Methods*, Encyclopedia of Optimization: Volume I (C. Floudas and P. Pardalos, eds.), Kluwer Academic Publishers, Boston, USA and Dordrecht, The Netherlands, 2001, pp. 319-323.

[8] P. Wolfe, *Convergence conditions for ascent methods*, SIAM Review **11** (1969), pp. 226-235.