SUT Journal of Mathematics
Vol. 50, No. 2 (2014), 167–203

A survey of sufficient descent conjugate gradient methods for unconstrained optimization

Yasushi Narushima and **Hiroshi Yabe**

(Received August 1, 2014; Revised February 24, 2015)

Abstract. In this decade, nonlinear conjugate gradient methods have been focused on as effective numerical methods for solving large-scale unconstrained optimization problems. Especially, nonlinear conjugate gradient methods with the sufficient descent property have been studied by many researchers. In this paper, we review sufficient descent nonlinear conjugate gradient methods.

AMS 2010 *Mathematics Subject Classification.* 90C30, 90C06, 65K05.

Key words and phrases. Unconstrained optimization, conjugate gradient method, sufficient descent condition, global convergence.

§1. Introduction

In 1952, the linear conjugate gradient (LCG) method was originally proposed by Hestenes and Stiefel [35] for solving symmetric positive definite linear systems of equations. Now the LCG method and its variants are major iterative methods for solving linear systems (see [38,51], for example). In 1964, based on the idea of the LCG method, Fletcher and Reeves [23] gave a nonlinear conjugate gradient (CG) method[1] for solving unconstrained optimization problems. In this decade, CG methods have been focused on as effective numerical methods for solving large-scale unconstrained optimization problems. Especially, CG methods with the sufficient descent property have been studied by many researchers.

In this paper, we review sufficient descent CG methods for solving the following unconstrained optimization problem:

$$\text{minimize } f(x), \tag{1.1}$$

[1] Although the CG method usually means the LCG method, we call the nonlinear conjugate gradient method the CG method in this paper.

where $f : \boldsymbol{R}^n \to \boldsymbol{R}$ is at least continuously differentiable and its gradient ∇f is denoted by g. The CG method is one of iterative methods that generate the sequence $\{x_k\}$ by

$$x_{k+1} = x_k + \alpha_k d_k \qquad \text{for } k \geq 0, \tag{1.2}$$

where α_k is a positive step size and d_k is a search direction. The search direction of the CG method is given by

$$d_k = \begin{cases} -g_k, & \text{if} \quad k = 0, \\ -g_k + \beta_k d_{k-1}, & \text{if} \quad k \geq 1, \end{cases} \tag{1.3}$$

where g_k denotes $\nabla f(x_k)$ and β_k is a scalar parameter that characterizes the method. Throughout the paper, we fix the initial direction by $d_0 = -g_0$.

In the first CG method given by Fletcher and Reeves [23], the parameter β_k is given by

$$\beta_k^{FR} = \frac{\|g_k\|^2}{\|g_{k-1}\|^2}.$$

We call the CG method with β_k^{FR} the FR method, and adopt the same usage for the other methods. Zoutendijk [68] proved the global convergence of the FR method with the exact line search. Al-Baali [1] extended this result to inexact line searches. Sorenson [53] applied the original Hestenes-Stiefel (namely the LCG) formula:

$$\beta_k^{HS} = \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}}$$

to general unconstrained optimization problems. Here, we define $y_{k-1} = g_k - g_{k-1}$. Polak and Ribière [47] gave another choice of parameter β_k:

$$\beta_k^{PR} = \frac{g_k^T y_{k-1}}{\|g_{k-1}\|^2}.$$

Powell [49,50] showed that the PR and HS methods can cycle infinitely without approaching a solution, and suggested the following modification:

$$\beta_k^{PR+} = \max\{\beta_k^{PR}, 0\}.$$

Gilbert and Nocedal [27] proved the global convergence of the PR+ method. Flectcher [22] gave a modification of the FR method

$$\beta_k^{CD} = \frac{\|g_k\|^2}{-g_{k-1}^T d_{k-1}}.$$

Note that CD stands for "Conjugate Descent". He showed that the CD method with some appropriate line search rule satisfies the descent condition:

$$g_k^T d_k < 0 \quad \text{for all } k. \tag{1.4}$$

Liu and Storey [42] proposed the following parameter:

$$\beta_k^{LS} = \frac{g_k^T y_{k-1}}{-g_{k-1}^T d_{k-1}}.$$

After that, Shi and Shen [52] proved the global convergence of the LS method with an Armijo-type line search. Dai and Yuan [17] proposed the CG method that generates a descent search direction at every iteration if the Wolfe conditions are satisfied, and they proved its global convergence. Their parameter is presented as

$$\beta_k^{DY} = \frac{\|g_k\|^2}{d_{k-1}^T y_{k-1}}.$$

Note that the above six methods (the FR, HS, PR, CD, LS, and DY methods) are known as typical CG methods, and are identical to the LCG method if the objective function f is a strictly convex quadratic function and if α_k is the exact one-dimensional minimizer. The above typical CG methods are usually classified by types of the numerators in the parameter β_k. The FR, CD and DY methods have the element $\|g_k\|^2$ in the numerator of the parameter β_k. Under some appropriate line search rule, these methods satisfy the following sufficient descent condition:

$$g_k^T d_k \leq -\bar{c}\|g_k\|^2 \quad \text{for all } k, \tag{1.5}$$

where $\bar{c}$ is a positive constant independent of k. the sufficient descent condition is stronger than the descent condition (1.4), and $\|g_k\|$ tends to zero if this condition holds and $g_k^T d_k \to 0$. Moreover, the sufficient descent condition plays an important role in establishing the global convergence of the method. On the other hand, the PR, HS, and LS methods have the element $g_k^T y_{k-1}$ in the numerator of the parameter β_k. When the iterates stagnate and the steps are too small, the element $g_k^T y_{k-1}$ is very small (because usually $\|y_{k-1}\| = O(\|x_k - x_{k-1}\|)$ holds) and the search direction is close to the steepest descent direction. Thus, these methods can automatically adjust β_k to avoid jamming and are more effective than the other three methods. However, these methods do not necessarily satisfy the descent condition.

Many other CG methods have been also proposed. For example, Iiduka and Narushima [37] proposed two new choices for β_k that incorporate the objective function values. Based on the modified secant condition by Zhang et al. [61, 62], Yabe and Sakaiwa [56] gave a modified DY method generating descent directions if the Wolfe conditions was imposed in the line search. In addition, many researchers have studied hybrid CG methods (see [4, 18, 19, 27, 36, 55], for example). On the other hand, Dai and Liao [16] proposed a method based on the secant condition of quasi-Newton methods, and later some researchers derived CG methods based on other secant conditions [8, 26, 39, 57, 67]. More recently, Babaie-Kafaki and Ghanbari [7] studied some suitable choices of parameters that was incorporated into Dai-Liao's method. Independently of Dai and Liao, Birgin and Martínez [9] proposed a scaled CG method (they called it the spectral CG method) based on the secant condition. Based on the memoryless BFGS quasi-Newton method, Andrei [3, 5, 6] proposed some three-term CG methods that generate descent directions under the Wolfe conditions.

Some CG methods introduced above satisfy the descent condition under certain line search rules, but other CG methods do not necessarily satisfy it. Recently, CG methods that satisfy the sufficient descent condition independently of line searches have been studied. By modifying the parameter β_k^{HS}, Hager and Zhang [30, 33] proposed a CG method that generates a sufficient descent direction. After that, following Hager-Zhang's modification scheme, some researchers proposed other sufficient descent CG methods [13, 40, 44, 58–60, 63]. Hager and Zhang [30–32] also developed a software CG-DESCENT based on the HZ method, and now it is one of the major softwares for solving large scale unconstrained optimization problems. Zhang, Zhou and Li [64–66] and Cheng [11] proposed scaled/three-term CG methods, which always satisfy $g_k^T d_k = -\|g_k\|^2$ for all k, independently of line search. Furthermore, Narushima, Yabe and Ford [46] proposed a three-term CG method that involves the above scaled/three-term CG methods.

In this paper, we survey CG methods satisfying the sufficient descent condition independent of line searches and their related topics. This paper is organized as follows. In Section 2, we give some preliminaries related with line searches and recall the properties of the typical CG methods. In Section 3, we review the HZ method and its variants. Scaled/three-term CG methods are introduced in Section 4. More recently, by combining Dai-Liao's idea and sufficient descent CG methods, some researchers proposed CG methods that satisfy the sufficient descent condition and are based on secant conditions. These methods are explained in Section 5. In Section 6, we introduce recent advances of the software CG-DESCENT. Finally, some numerical results are given in Section 7.

§2. Preliminaries and properties of the typical CG methods

In this section, we give some preliminaries related with line searches and recall the properties of the typical CG methods.

2.1. Line search conditions and their related properties

To achieve the global convergence of iterative methods, we need to choose an appropriate step size α_k. The most simple idea is the exact line search, namely

$$f(x_k + \alpha_k d_k) = \min_{\alpha>0} f(x_k + \alpha d_k).$$

However, it is too expensive or impossible to implement the exact line search in practice. Therefore, many practical line search rules to choose a step size have been proposed. Especially, the Wolfe conditions are well-known and these are given by

$$f(x_k + \alpha_k d_k) \leq f(x_k) + \delta \alpha_k g_k^T d_k, \tag{2.1}$$

$$g(x_k + \alpha_k d_k)^T d_k \geq \sigma_1 g_k^T d_k, \tag{2.2}$$

where $0 < \delta < \sigma_1 < 1$. In addition, for CG methods, the generalized strong Wolfe conditions: (2.1) and

$$-\sigma_2 g_k^T d_k \geq g(x_k + \alpha_k d_k)^T d_k \geq \sigma_1 g_k^T d_k \tag{2.3}$$

are often used, where $\sigma_2 > 0$. For the case $\sigma_1 = \sigma_2$, the generalized strong Wolfe conditions reduce to the strong Wolfe conditions: (2.1) and

$$|g(x_k + \alpha_k d_k)^T d_k| \leq -\sigma_1 g_k^T d_k. \tag{2.4}$$

The first condition of the Wolfe conditions (namely, (2.1)) is called the Armijo condition and it or its variant is often used alone.

We now recall some properties related with line searches. To the end, we make some assumptions for the objective function.

Assumption 1. *The objective function f is bounded below on $\boldsymbol{R}^n$ and is continuously differentiable in an open convex neighborhood $\mathcal{N}$ of the level set $\mathcal{L} = \{x | f(x) \leq f(x_0)\}$ at the initial point x_0. In addition, the gradient g is Lipschitz continuous in $\mathcal{N}$, i.e. there exists a positive constant L such that*

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

Assumption 2. *The level set $\mathcal{L}$ is bounded, namely, there exists a positive constant $\widehat{a}$ such that*

$$\|x\| \leq \widehat{a} \qquad \text{for all } x \in \mathcal{L}.$$

Throughout the paper, we assume that

$$g_k \neq 0$$

for all $k \geq 0$, otherwise a stationary point has been found.

The following lemma is known as the Zoutendijk condition, which is very critical to prove the global convergence of CG methods.

Lemma 1. [68] *Suppose that Assumption* 1 *is satisfied. Consider any iterative method of the form* (1.2) *such that the descent condition* (1.4) *and the Wolfe conditions* (2.1)–(2.2) *are satisfied. Then, the following holds*:

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

Under Assumption 1, we have the following lemma, which is easily obtained from the Zoutendijk condition. The proof of the lemma is given by [54], for example.

Lemma 2. *Suppose that Assumption* 1 *is satisfied. Consider any iterative method of the form* (1.2) *such that the sufficient descent condition* (1.5) *and the Wolfe conditions* (2.1)–(2.2) *are satisfied. If*

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

the following holds:

$$\liminf_{k\to\infty} \|g_k\| = 0. \tag{2.5}$$

Any CG method using the strong Wolfe line search possesses the following useful property. This was proved by Dai et al. [14] (see Theorem 2.3 and Corollary 2.4 in [14]).

Lemma 3. *Suppose that Assumption* 1 *holds. Consider any CG method of the form* (1.2) *and* (1.3) *such that the descent condition* (1.4) *and the generalized strong Wolfe conditions* (2.1) *and* (2.3) *are satisfied. If*

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

then (2.5) *holds.*

2.2. Properties of the FR, CD and DY methods

In this section, we recall properties of the FR, CD and DY methods. Note that these methods have the element $\|g_k\|^2$ in the numerator of the parameter β_k. If the step size α_k satisfies the generalized strong Wolfe conditions (2.1) and (2.3), the following properties are obtained.

Proposition 4. *The following statements hold*:

(a) *For the FR method, if* α_k *satisfies the generalized strong Wolfe conditions* (2.1) *and* (2.3) *with* $\sigma_1 + \sigma_2 < 1$*, then*

$$-\frac{1}{1-\sigma_1} \le \frac{g_k^T d_k}{\|g_k\|^2} \le -1 + \frac{\sigma_2}{1-\sigma_1}.$$

(b) *For the DY method, if* α_k *satisfies the generalized strong Wolfe conditions* (2.1) *and* (2.3)*, then*

$$-\frac{1}{1-\sigma_1} \le \frac{g_k^T d_k}{\|g_k\|^2} \le -\frac{1}{1+\sigma_2}.$$

(c) *For the CD method, if* α_k *satisfies the generalized strong Wolfe conditions* (2.1) *and* (2.3) *with* $\sigma_2 < 1$*, then*

$$-1-\sigma_1 \le \frac{g_k^T d_k}{\|g_k\|^2} \le -1 + \sigma_2$$

Proposition 4 implies that the FR, CD and DY methods satisfy the sufficient descent condition (1.5), dependent on line searches. The results (a) and (b) are simple extensions of the results in [47], and (c) is easily shown from (2.3). We now give the global convergence properties of the FR and DY methods, which were proven in [1] and [17], respectively.

Theorem 5. *Suppose that Assumption* 1 *holds. Let the sequence* $\{x_k\}$ *be generated by the CG method of the form* (1.2)–(1.3)*.*

(a) *If* $\beta_k = \beta_k^{FR}$ *and* α_k *satisfies the generalized strong Wolfe conditions* (2.1) *and* (2.3) *with* $\sigma_1 + \sigma_2 < 1$*, then* $\{x_k\}$ *converges globally in the sense that* (2.5) *holds.*

(b) *If* $\beta_k = \beta_k^{DY}$ *and* α_k *satisfies the Wolfe conditions* (2.1)–(2.2)*, then* d_k *satisfies the descent condition* (1.4) *and* $\{x_k\}$ *converges globally in the sense that* (2.5) *holds.*

Note that the CD method satisfies the sufficient descent condition under milder conditions than the FR method does. However, the global convergence of the CD method have not been established under the (generalized) strong Wolfe conditions. On the other hand, the global convergence and the sufficient descent properties of the DY method can be obtained under mild conditions.

2.3. Properties of the HS, PR and LS methods

In this section, we recall properties of the HS, PR and LS methods. Note that these methods have the element $g_k^T y_{k-1}$ in the numerator of the parameter β_k. We first introduce Property $\star$ for β_k given by Gilbert and Nocedal [27]. Property $\star$ implies that β_k is bounded and will be small when the step $s_{k-1} = x_k - x_{k-1}$ is small.

Property $\star$. *Consider the CG method* (1.2)–(1.3) *and suppose that there exists a positive constant ε such that*

$$\varepsilon \leq \|g_k\| \quad \textit{for all } k. \tag{2.6}$$

If there exist $b > 1$ and $\bar{\xi} > 0$ such that $|\beta_k| \leq b$ and

$$\|s_{k-1}\| \leq \bar{\xi} \quad \Longrightarrow \quad |\beta_k| \leq \frac{1}{2b},$$

then we say that the method has Property $\star$.

In order to prove that CG methods have Property $\star$, it suffices to show that there exists a positive constant c_1 such that

$$|\beta_k| \leq c_1 \|s_{k-1}\| \qquad \text{for all } k, \tag{2.7}$$

under the assumption (2.6). Then, by putting $\bar{\xi} = 1/(2bc_1)$, we have $|\beta_k| \leq \max\{1, 2\widehat{a}c_1\} \equiv b$ and

$$\|s_{k-1}\| \leq \bar{\xi} \quad \Longrightarrow \quad |\beta_k| \leq \frac{1}{2b},$$

which implies that Property $\star$ is satisfied. It is easily shown that (2.7) holds for the HS, PR and LS methods, and thus these methods have Property $\star$.

Next we give the global convergence theorem of CG methods satisfying Property $\star$. The proof of the theorem was first given in [27] and many researchers showed its variants (see [15, 16, 30], for example).

Theorem 6. *Suppose that Assumptions* 1 *and* 2 *hold. Let $\{x_k\}$ be the sequence generated by the CG method* (1.2)–(1.3) *that satisfies the following conditions*:

(C1) $\beta_k \geq \nu_k \equiv \min\{\nu_k^{(1)}, \nu_k^{(2)}, \nu_k^{(3)}\}$ *for all k, where*

$$\nu_k^{(1)} = \frac{-1}{\|d_{k-1}\| \min\{\bar{\nu}_1, \|g_{k-1}\|\}}, \quad \nu_k^{(2)} = \bar{\nu}_2 \frac{g_k^T d_{k-1}}{\|d_{k-1}\|^2}, \quad \nu_k^{(3)} = \bar{\nu}_3 \frac{g_{k-1}^T d_{k-1}}{\|d_{k-1}\|^2}$$

and $\bar{\nu}_1$, $\bar{\nu}_2$ and $\bar{\nu}_3$ are positive constants.

(C2) *The search direction satisfies the sufficient descent condition* (1.5).

(C3) *The Zoutendijk condition holds.*

(C4) *Property* $\star$ *holds.*

Then the sequence $\{x_k\}$ *converges globally in the sense that* (2.5) *holds.*

Since condition (C1) may not hold in certain cases, we modify the parameter β_k by

$$\beta_k^+ = \max\{\zeta_k, \beta_k\}, \tag{2.8}$$

where $\zeta_k \in [\nu_k, 0]$, so that $\beta_k^+ \geq \nu_k$. Note that the choices of $\zeta_k = 0$, $\zeta_k = \nu_k^{(1)}$, $\zeta_k = \nu_k^{(2)}$ and $\zeta_k = \nu_k^{(3)}$ reduce formula (2.8) to those proposed in [27], [30], [15], and [34] respectively. Although many CG methods use one of the above three modifications to show the global convergence, we consider the unified form (2.8) in this paper. For simplicity, we denote $\max\{\zeta_k, \beta_k^{HS}\}$ by β_k^{HS+} and call the CG method with β_k^{HS+} the HS+ method. Moreover, we use the same manner for all the other methods introduced in this paper.

We now give the global convergence results of the HS+, PR+ and LS+ methods.

Theorem 7. *Suppose that Assumptions* 1 *and* 2 *hold. Let* $\{x_k\}$ *be the sequence generated by the CG method* (1.2)–(1.3) *with* $\beta_k = \beta_k^{HS+}$, β_k^{PR+} *or* β_k^{LS+}. *If* d_k *and* α_k *satisfy the sufficient descent condition* (1.5) *and the Wolfe conditions* (2.1)–(2.2), *then the sequence* $\{x_k\}$ *converges globally in the sense that* (2.5) *holds.*

Note that the assumptions of Theorem 7 are stronger than those of Theorem 5. Specifically, Theorem 7 needs to assume the sufficient descent condition. Although the HS, PR and LS methods, as mentioned in Section 1, are more effective than the other typical CG methods in practise, the global convergence of the HS, PR and LS methods can be established only under the stronger conditions. Therefore, to overcome this weakness, many researchers have tried to develop robust and effective CG methods in this decade. In the subsequent sections, we will survey recent advances of such CG methods.

§3. Hager-Zhang's method and its variants

Hager and Zhang [30, 33] proposed a CG method in which the parameter β_k is given by

$$\beta_k^{HZ} = \beta_k^{HS} - \frac{\mu\|y_{k-1}\|^2}{(d_{k-1}^T y_{k-1})^2} g_k^T d_{k-1} = \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}} - \frac{\mu\|y_{k-1}\|^2}{(d_{k-1}^T y_{k-1})^2} g_k^T d_{k-1}, \tag{3.1}$$

where $\mu > 1/4$. Note that Hager and Zhang first gave (3.1) with $\mu = 2$ in [30], and after that they extended it to $\mu > 1/4$ in [33]. The search direction of the HZ method satisfies the sufficient descent condition (1.5) with $\bar{c} = 1 - (4\mu)^{-1}$, independently of line searches. The global convergence property of the HZ+ method can be obtained under the Wolfe conditions (2.1)–(2.2).

Later on, Yu, Guan and Chen [58] suggested that the CG methods with

$$\begin{aligned} \beta_k^{MFR} &= \beta_k^{FR} - \frac{\mu\|g_k\|^2}{\|g_{k-1}\|^4} g_k^T d_{k-1}, \\ \beta_k^{MPR} &= \beta_k^{PR} - \frac{\mu\|y_{k-1}\|^2}{\|g_{k-1}\|^4} g_k^T d_{k-1}, \\ \beta_k^{MDY} &= \beta_k^{DY} - \frac{\mu\|g_k\|^2}{(d_{k-1}^T y_{k-1})^2} g_k^T d_{k-1}, \\ \beta_k^{MCD} &= \beta_k^{CD} - \frac{\mu\|g_k\|^2}{(-g_{k-1}^T d_{k-1})^2} g_k^T d_{k-1}, \\ \beta_k^{MLS} &= \beta_k^{LS} - \frac{\mu\|y_{k-1}\|^2}{(-g_{k-1}^T d_{k-1})^2} g_k^T d_{k-1} \end{aligned}$$

also satisfy the sufficient descent condition (1.5) with $\bar{c} = 1 - (4\mu)^{-1}$, where $\mu > 1/4$. Yu, Guan and Li [59] showed that the MPR+ method is globally convergent under the assumption that the step size α_k satisfies an Armijo-type condition. Also, Yuan [60] proved the global convergence of the MPR method with the Wolfe conditions. However, Yuan assumed that the step size α_k was bounded away from zero, and this assumption is strong. In order to establish the global convergence of the MPR+ method under the Wolfe conditions, Zhang and Li [63] modified β_k^{MPR} and gave

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

where h is a positive constant. The ZL method converges globally under the Wolfe conditions (2.1)–(2.2). Li and Feng [40] proved the global convergence property of the MLS+ method under the strong Wolfe conditions (2.1) and (2.4). Dai and Wen [20], motivated by the MBFGS method of Li and Fukushima [41], proposed a modified HZ method:

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

where $v_{k-1} = y_{k-1} + h_{k-1} s_{k-1}$, $h_{k-1} = \bar{h} + \max\{-s_{k-1}^T y_{k-1}/\|s_{k-1}\|^2, 0\}$ and $\bar{h}$ is a nonnegative constant.

Dai [13] modified β_k of the form $\beta_k = g_k^T z_k$ by

$$\beta_k^{SD} = g_k^T z_k - \mu \|z_k\|^2 g_k^T d_{k-1}, \tag{3.2}$$

where $\mu > 1/4$ and $z_k \in \boldsymbol{R}^n$ is any vector, and showed that the SD method satisfies the sufficient descent condition (1.5). In fact, by using the inequality $u^T v \leq \frac{1}{2}(\|u\|^2 + \|v\|^2)$ for any vectors u and v, (3.2) yields

$$\begin{aligned} g_k^T d_k &= -\|g_k\|^2 + \beta_k^{SD} g_k^T d_{k-1} \\ &= -\|g_k\|^2 + \left(g_k^T z_k - \mu \|z_k\|^2 g_k^T d_{k-1}\right) g_k^T d_{k-1} \\ &= -\|g_k\|^2 + g_k^T z_k g_k^T d_{k-1} - \mu \|z_k\|^2 (g_k^T d_{k-1})^2 \\ &= -\|g_k\|^2 + \frac{g_k^T}{\sqrt{2\mu}} (\sqrt{2\mu} g_k^T d_{k-1} z_k) - \mu \|z_k\|^2 (g_k^T d_{k-1})^2 \\ &\leq -\|g_k\|^2 + \frac{1}{2} \left(\frac{\|g_k\|^2}{2\mu} + 2\mu \|z_k\|^2 (g_k^T d_{k-1})^2 \right) - \mu \|z_k\|^2 (g_k^T d_{k-1})^2 \\ &= -\left(1 - \frac{1}{4\mu}\right) \|g_k\|^2. \end{aligned}$$

Therefore, the SD method always satisfies the sufficient descent condition (1.5) with $\bar{c} = 1 - (4\mu)^{-1}$. We note that the SD method involves several methods mentioned in this section. For instance, if we set $z_k = y_{k-1}/(d_{k-1}^T y_{k-1})$, then we have $\beta_k^{SD} = \beta_k^{HZ}$.

Nakamura, Narushima and Yabe [44] introduced the following property and showed the global convergence of the SD+ method.

Property 1. *Consider the SD+ method. We assume that there exists a positive constant $\varepsilon > 0$ such that $\|g_k\| \geq \varepsilon$ holds for all k. Then we say that the method has Property 1 if there exist positive constants c_2 and c_3 such that*

$$\begin{aligned} \left|g_k^T z_k\right| &\leq c_2 \|s_{k-1}\|, \\ \|z_k\|^2 |g_k^T d_{k-1}| &\leq c_3 \|s_{k-1}\|^2 \end{aligned}$$

hold for all k.

We should note that if Property 1 is satisfied, the SD+ method has Property $\star$. Thus, the following theorem is obtained by Theorem 6.

Theorem 8. *Suppose that Assumptions* 1 *and* 2 *are satisfied. Assume that the sequence $\{x_k\}$ is generated by the SD+ method and that α_k satisfies the Wolfe conditions* (2.1)–(2.2). *If the method has Property 1, then the sequence $\{x_k\}$ converges globally in the sense that* (2.5) *holds.*

Theorem 8 plays an important role in establishing the global convergence property of CG methods with concrete β_k. For instance, the global convergence results in [30] (which is related with β_k^{HZ+}) and [40] (which is related with β_k^{MLS+}) are given as corollaries of Theorem 8.

Corollary 9. *Suppose that Assumptions* 1 *and* 2 *are satisfied. Assume that the sequence* $\{x_k\}$ *is generated by the SD+ method. Then the following statements hold:*

(a) *Assume that* α_k *satisfies the Wolfe conditions* (2.1)–(2.2)*. Then the CG method with* β_k^{HZ+} *converges in the sense that* (2.5) *holds.*

(b) *Assume that* α_k *satisfies the generalized strong Wolfe conditions* (2.1) *and* (2.3)*. Then the CG method with* β_k^{MLS+} *converges in the sense that* (2.5) *holds.*

Also Nakamura et al. [44] proposed two descent hybrid CG methods. The first one combines β_k^{HZ} and β_k^{MPR}, as follows:

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

The second one combines β_k^{HZ} and β_k^{MLS}, as follows:

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

In addition, they gave the global convergence of the proposed hybrid methods.

Corollary 10. *Suppose that Assumptions* 1 *and* 2 *are satisfied. Assume that the sequence* $\{x_k\}$ *is generated by the SD+ method and that* α_k *satisfies the Wolfe conditions* (2.1)–(2.2)*. Then, the CG methods with* β_k^{MHP+} *and* β_k^{MHL+} *converge in the sense that* (2.5) *holds, respectively.*

Recently, Dai and Kou [15] pointed out a relation between the BFGS quasi-Newton method and the HZ method and showed that the choice $\mu = 1$ is suitable. The search direction of the BFGS quasi-Newton method is given by

$$d_k^{QN} = -H_k g_k,$$

where H_k is an approximation matrix to $\nabla^2 f(x_k)$, and H_k is updated by

$$H_k = H_{k-1} - \frac{H_{k-1} y_{k-1} s_{k-1}^T + s_{k-1} y_{k-1}^T H_{k-1}}{s_{k-1}^T y_{k-1}} + \left(1 + \frac{y_{k-1}^T H_{k-1} y_{k-1}}{s_{k-1}^T y_{k-1}}\right) \frac{s_{k-1} s_{k-1}^T}{s_{k-1}^T y_{k-1}}.$$

By letting τ_k be a positive parameter and $H_{k-1} = \frac{1}{\tau_k} I$, the search direction $\tilde{d}_k^{QN} \equiv \tau_k d_k^{QN}$, which is multiplied by τ_k, is given by

$$\begin{aligned}\tilde{d}_k^{QN} = -g_k + \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}} d_{k-1} - &\left(\tau_k + \frac{\|y_{k-1}\|^2}{s_{k-1}^T y_{k-1}}\right) \\ &\times \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}} d_{k-1} + \frac{g_k^T d_{k-1}}{d_{k-1}^T y_{k-1}} y_{k-1}.\end{aligned}$$

In addition, Dai and Kou defined their search direction by the solution of the following minimization problem:

$$d_k = \arg\min_d \{\|d - \tilde{d}_k^{QN}\| \mid d = -g_k + \beta d_{k-1}, \beta \in \boldsymbol{R}\},$$

and they gave a search direction by (1.3) with

$$\beta_k^{DK} = \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}} - \left(\tau_k + \frac{\|y_{k-1}\|^2}{s_{k-1}^T y_{k-1}} - \frac{s_{k-1}^T y_{k-1}}{\|s_{k-1}\|^2}\right) \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}.$$

By taking into account the relation $\tau_k I \approx \nabla^2 f(x_{k-1})$, $\tau_k = s_{k-1}^T y_{k-1} / \|s_{k-1}\|^2$ is one of suitable choices, and then β_k^{DK} is identical to β_k^{HZ} with $\mu = 1$. Dai and Kou confirmed the good numerical performance of the HZ method with $\mu = 1$ and claimed that the choice $\mu = 1$ is superior not only in theory but also in practice.

§4. Scaled and three-term CG methods

Zhang, Zhou and Li [64] proposed the modified FR method by

$$d_k = -\bar{\theta}_k g_k + \beta_k^{FR} d_{k-1}, \qquad k \geq 1, \tag{4.1}$$

where $\bar{\theta}_k = d_{k-1}^T y_{k-1} / \|g_{k-1}\|^2$. Note that the search direction (4.1) can be rewritten by $d_k = \bar{\theta}_k(-g_k + \beta_k^{DY} d_{k-1})$, and hence it can be regarded as a scaled DY method. Cheng [11] gave the modified PR method:

$$d_k = -g_k + \beta_k^{PR} \left(I - \frac{g_k g_k^T}{g_k^T g_k}\right) d_{k-1} \quad k \geq 1. \tag{4.2}$$

Zhang, Zhou and Li proposed the three-term PR method [65] and the three-term HS method [66], which are respectively given by

$$d_k = -g_k + \beta_k^{PR} d_{k-1} - \theta_k^{(1)} y_{k-1}, \qquad k \geq 1, \tag{4.3}$$

$$d_k = -g_k + \beta_k^{HS} d_{k-1} - \theta_k^{(2)} y_{k-1}, \qquad k \geq 1, \tag{4.4}$$

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

Narushima, Yabe and Ford [46] proposed a three-term CG method:

$$d_k = -g_k + \beta_k (g_k^T p_k)^\dagger \{(g_k^T p_k) d_{k-1} - (g_k^T d_{k-1}) p_k\}, \quad k \geq 1, \tag{4.5}$$

where $p_k \in \boldsymbol{R}^n$ is a parameter vector and $\dagger$ denotes the generalized reciprocal such that

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

We emphasize that the method (1.2) and (4.5) always satisfies

$$g_k^T d_k = -\|g_k\|^2, \tag{4.6}$$

independently of choices of β_k, p_k and line searches. In addition, the relation (4.6) implies that the sufficient descent condition (1.5) holds with $\bar{c} = 1$. If $g_k^T p_k = 0$, (4.5) implies $d_k = -g_k$, otherwise (4.5) can be rewritten by

$$d_k = -g_k + \beta_k d_{k-1} - \beta_k \frac{g_k^T d_{k-1}}{g_k^T p_k} p_k = -g_k + \beta_k \left(I - \frac{p_k g_k^T}{g_k^T p_k} \right) d_{k-1}. \tag{4.7}$$

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

If we use the exact line search and p_k such that $g_k^T p_k \neq 0$, then (4.7) becomes the usual CG method (1.3). The most typical choices are $p_k = g_k$ and $p_k = y_{k-1}$. The choice $p_k = g_k$ yields

$$d_k = -\left(1 + \beta_k \frac{g_k^T d_{k-1}}{\|g_k\|^2} \right) g_k + \beta_k d_{k-1}, \qquad k \geq 1. \tag{4.8}$$

The direction (4.8) can be regarded as a scaled CG method. When $p_k = y_{k-1}$, $g_k^T p_k = g_k^T y_{k-1} = 0$ can occur. In this case, the direction (4.5) becomes the steepest descent direction $d_k = -g_k$, and then $g_k^T y_{k-1} = 0$ can be regarded as a restart criterion. On the other hand, if we choose $p_k = d_{k-1}$, then (4.5) implies $d_k = -g_k$ for all k.

We should note that the search direction (4.5) includes the search directions proposed in [11,64–66]. Since (4.1) satisfies $g_k^T d_k = -\|g_k\|^2$ for all k, (4.1) can be rewritten by the three-term form:

$$d_k = -g_k + \beta_k^{FR} d_{k-1} - \theta_k^{(3)} g_k,$$

where $\theta_k^{(3)} = g_k^T d_{k-1}/\|g_{k-1}\|^2$. Therefore, (4.5) with $\beta_k = \beta_k^{FR}$ and $p_k = g_k$ becomes (4.1). The search direction (4.5) with $\beta_k = \beta_k^{PR}$ and $p_k = g_k$ becomes (4.2). If $g_k^T y_{k-1} \neq 0$, (4.5) with $\beta_k = \beta_k^{PR}$ and $p_k = y_{k-1}$ becomes (4.3), and (4.5) with $\beta_k = \beta_k^{HS}$ and $p_k = y_{k-1}$ becomes (4.4).

Narushima et al. also showed the global convergence property of the method (1.2) and (4.5). Note that some straightforward calculations yield the following relation

$$\|d_k\|^2 \le \psi_k^2 \|d_{k-1}\|^2 + \|g_k\|^2$$

for all k, where ψ_k is defined by

$$\psi_k = \beta_k \|g_k\| \|p_k\| (g_k^T p_k)^\dagger. \tag{4.9}$$

Narushima et al. [46] introduced a property for ψ_k and gave the global convergence theorem as follows. These correspond to Property $\star$ and Theorem 6, respectively.

Property 2. *Consider the three-term CG method* (1.2) *and* (4.5), *and suppose that there exists a positive constant ε such that $\varepsilon \le \|g_k\|$ holds for all k. If there exist constants $b > 1$ and $\bar{\xi} > 0$ such that $|\psi_k| \le b$ and*

$$\|s_{k-1}\| \le \bar{\xi} \implies |\psi_k| \le \frac{1}{b}$$

for all k, then we say that the method has Property 2

Theorem 11. *Suppose that Assumptions* 1 *and* 2 *hold. Let $\{x_k\}$ be the sequence generated by the three-term CG method* (1.2) *and* (4.5) *that satisfies the following conditions:*

(C1) $\beta_k \ge \nu_k$ *for all k,*

(C2) *Property 2 holds.*

If α_k satisfies the generalized strong Wolfe conditions (2.1) *and* (2.3), *then the method converges in the sense that* (2.5) *holds.*

Theorem 11 plays an important role in establishing the global convergence of the three-term CG methods. For instance, the following results are given as a corollary of Theorem 11.

Corollary 12. *Suppose that Assumptions* 1 *and* 2 *are satisfied. Let $\{x_k\}$ be the sequence generated by the three-term CG method* (1.2) *and* (4.5), *where α_k satisfies the generalized strong Wolfe conditions* (2.1) *and* (2.3). *Then the following statements hold* :

(i) *The method with* $\beta_k = \beta_k^{PR+}$ *and* $p_k = y_{k-1}$ *(or* $p_k = g_k$*) converges in the sense that* (2.5) *holds.*

(ii) *The method with* $\beta_k = \beta_k^{HS+}$ *and* $p_k = y_{k-1}$ *(or* $p_k = g_k$*) converges in the sense that* (2.5) *holds.*

More recently, by extending the three-term CG method (4.5), Al-Baali, Narushima and Yabe [2] gave the following family of three-term CG methods:

$$d_k = \begin{cases} -g_k & \text{if } k = 0 \text{ or } |g_k^T p_k| \leq \bar{\theta}\|g_k\|\|p_k\|, \\ -g_k + \beta_k d_{k-1} + \eta_k p_k & \text{otherwise,} \end{cases} \tag{4.10}$$

where p_k is any nonzero vector, $0 < \bar{\theta} < 1$ is a constant and

$$\eta_k = -\frac{(\gamma_k - 1)\|g_k\|^2 + \beta_k g_k^T d_{k-1}}{g_k^T p_k}. \tag{4.11}$$

Here, $\gamma_k \in [\bar{\gamma}_1, \bar{\gamma}_2]$ is a parameter, where $0 < \bar{\gamma}_1 \leq 1 \leq \bar{\gamma}_2$. Note that the second case of (4.10) implies

$$g_k^T d_k = -\gamma_k \|g_k\|^2,$$

and hence the directional derivative $g_k^T d_k$ can be controlled by changing the parameter γ_k. Also note that (4.10) with $\gamma_k = 1$ reduces to (4.7). They defined a property for the proposed method similar to Property $\star$, and showed the global convergence of the method with such a property. In addition, by using this result, they proved the global convergence of the method with β_k^{HS+}, β_k^{PR+}, β_k^{LS+}, β_k^{HZ+}, β_k^{MPR+} and β_k^{MLS+} under the generalized strong Wolfe conditions. They also proposed several choices for γ_k, and recommended the following choice:

$$\gamma_k = \max\left\{\bar{\gamma}_1, \min\left\{\bar{\gamma}_2, 1 - \bar{\gamma}\frac{|\beta_k g_k^T d_{k-1}|}{\|g_k\|\|d_{k-1}\|}\right\}\right\}, \tag{4.12}$$

where $\bar{\gamma}$ is a nonnegative constant.

§5. Sufficient descent CG methods based on secant conditions

In order to accelerate CG methods, some researchers proposed the CG methods based on secant conditions [16, 26, 57, 67], and such methods are known as efficient CG methods. However, these methods do not necessarily generate descent directions. In order to overcome this weakness, some researchers recently proposed sufficient descent CG methods based on secant conditions.

In Section 5.1, we recall the CG methods based on secant conditions. In Section 5.2, we survey sufficient descent CG methods based on the SD method (recall (3.2)) and secant conditions. Furthermore, we review the sufficient descent CG methods based on the scaled/three-term CG method (recall (4.8) and (4.5)) and secant conditions in Section 5.3

5.1. CG methods based on secant conditions

In order to solve a symmetric positive definite system $Ax = b$ or equivalently minimize a strictly convex quadratic function $\frac{1}{2}x^T Ax - b^T x$, the LCG method generates search directions that satisfy the conjugacy condition:

$$d_i^T A d_j = 0, \qquad \forall i \neq j. \tag{5.1}$$

On the other hand, for general nonlinear functions, it follows from the mean value theorem that there exists some $\tau \in (0, 1)$ such that

$$d_k^T y_{k-1} = \alpha_{k-1} d_k^T \nabla^2 f(x_{k-1} + \tau\alpha_{k-1} d_{k-1}) d_{k-1}.$$

Therefore, it is reasonable to replace (5.1) by the following conjugacy condition for general objective functions:

$$d_k^T y_{k-1} = 0. \tag{5.2}$$

An extension of the conjugacy condition was studied by Perry [48]. Perry tried to incorporate the second-order information of the objective function into the CG method to accelerate it. Specifically, by using the secant condition and the search direction of the quasi-Newton methods, which are respectively defined by

$$B_k s_{k-1} = y_{k-1} \quad \text{and} \quad B_k d_k = -g_k, \tag{5.3}$$

the following relation is obtained:

$$d_k^T y_{k-1} = d_k^T (B_k s_{k-1}) = (B_k d_k)^T s_{k-1} = -g_k^T s_{k-1},$$

where B_k is a symmetric approximation matrix to the Hessian $\nabla^2 f(x_k)$. Then Perry replaced the conjugacy condition (5.2) by the following condition

$$d_k^T y_{k-1} = -g_k^T s_{k-1}. \tag{5.4}$$

Furthermore, Dai and Liao [16] incorporated a nonnegative parameter t into Perry's condition and gave

$$d_k^T y_{k-1} = -t g_k^T s_{k-1}. \tag{5.5}$$

For the case $t = 0$, (5.5) reduces to the usual conjugacy condition (5.2). On the other hand, for the case $t = 1$, (5.5) becomes Perry's condition (5.4). By substituting (1.3) into (5.5), Dai and Liao derived the following formula:

$$\beta_k^{DL} = \frac{g_k^T (y_{k-1} - t s_{k-1})}{d_{k-1}^T y_{k-1}}.$$

Later on, following Dai and Liao, several CG methods have been presented. Yabe and Takano [57] proposed a CG method based on the modified secant condition by Zhang, Deng and Chen [61] and Zhang and Xu [62]:

$$B_k s_{k-1} = y_{k-1}^{(1)}, \quad y_{k-1}^{(1)} = y_{k-1} + \rho_k \left(\frac{\phi_{k-1}}{s_{k-1}^T u_{k-1}} u_{k-1} \right), \tag{5.6}$$

where

$$\phi_{k-1} = 6(f(x_{k-1}) - f(x_k)) + 3(g_{k-1} + g_k)^T s_{k-1},$$

$\rho_k \geq 0$ is a scalar and $u_{k-1} \in \boldsymbol{R}^n$ is any vector such that $s_{k-1}^T u_{k-1} \neq 0$ holds. Yabe-Takano's formula for β_k is given by

$$\beta_k^{YT} = \frac{g_k^T (y_{k-1}^{(1)} - t s_{k-1})}{d_{k-1}^T y_{k-1}^{(1)}}.$$

On the other hand, Zhou and Zhang [67] proposed a CG method based on the MBFGS secant condition by Li and Fukushima [41]:

$$B_k s_{k-1} = y_{k-1}^{(2)}, \quad y_{k-1}^{(2)} = y_{k-1} + \Gamma \|g_{k-1}\|^q s_{k-1}, \tag{5.7}$$

where $\Gamma > 0$ and $q > 0$ are constants. Zhou-Zhang's formula for β_k is as follows

$$\beta_k^{ZZ} = \frac{g_k^T (y_{k-1}^{(2)} - t s_{k-1})}{d_{k-1}^T y_{k-1}^{(2)}}.$$

In addition, Ford, Narushima and Yabe [26] gave a CG method based on the multi-step secant condition by Ford and Moghrabi [24, 25]:

$$B_k s_{k-1}^{MS1} = y_{k-1}^{MS1}, \quad s_{k-1}^{MS1} = s_{k-1} - \xi_{k-1} s_{k-2}, \quad y_{k-1}^{MS1} = y_{k-1} - \xi_{k-1} y_{k-2}, \tag{5.8}$$

where

$$\xi_{k-1} = \frac{\delta_{k-1}^2}{1 + 2\delta_{k-1}}, \quad \delta_{k-1} = \kappa_k \frac{\|s_{k-1}\|}{\|s_{k-2}\|}, \tag{5.9}$$

and $\kappa_k \geq 0$ is a scaling factor. The formula for β_k is

$$\beta_k^{F1} = \frac{g_k^T(y_{k-1}^{MS1} - ts_{k-1}^{MS1})}{d_{k-1}^T y_{k-1}^{MS1}}.$$

In the case $\kappa_k = 0$, this condition reduces to the usual secant condition (5.3). Moreover, by using another multi-step secant condition:

(5.10)
$$B_k s_{k-1}^{MS2} = y_{k-1}^{MS2}, \quad s_{k-1}^{MS2} = s_{k-1} - \xi_{k-1}s_{k-2}, \quad y_{k-1}^{MS2} = y_{k-1} - t\xi_{k-1}y_{k-2},$$

they also proposed another formula

$$\beta_k^{F2} = \frac{g_k^T(y_{k-1}^{MS2} - ts_{k-1}^{MS2})}{d_{k-1}^T y_{k-1}^{MS2}}.$$

In order to unify the above secant conditions, we consider the following form:

(5.11)
$$B_k r_{k-1} = w_{k-1}.$$

In the case of $r_{k-1} = s_{k-1}$ and $w_{k-1} = y_{k-1}$, (5.11) reduces to the usual secant condition (5.3). The unified secant condition (5.11) derived the condition $d_{k-1}^T w_{k-1} = -tg_k^T r_{k-1}$, which is associated with (5.5), and then we have the following formula:

(5.12)
$$\beta_k = \frac{g_k^T(w_{k-1} - tr_{k-1})}{d_{k-1}^T w_{k-1}}.$$

Note that, if $d_{k-1}^T w_{k-1} = 0$, we set $\beta_k = 0$ in practice. In Table 1, we give w_{k-1} and r_{k-1} in (5.12) for the cases β_k^{DL}, β_k^{YT}, β_k^{ZZ}, β_k^{F1} and β_k^{F2}.

Table 1: w_{k-1} and r_{k-1} in (5.12)

β_k	w_{k-1}	r_{k-1}
β_k^{DL}	y_{k-1}	s_{k-1}
β_k^{YT}	$y_{k-1}^{(1)}$ in (5.6)	s_{k-1}
β_k^{ZZ}	$y_{k-1}^{(2)}$ in (5.7)	s_{k-1}
β_k^{F1}	y_{k-1}^{MS1} in (5.8)	s_{k-1}^{MS1} in (5.8)
β_k^{F2}	y_{k-1}^{MS2} in (5.10)	s_{k-1}^{MS2} in (5.10)

5.2. The SD method based on secant conditions

In order to establish the sufficient descent property of the CG method with (5.12), Narushima and Yabe [45] chose $z_k = \dfrac{w_{k-1} - tr_{k-1}}{d_{k-1}^T w_{k-1}}$ in (3.2) and proposed

$$\beta_k^{SSD} = \frac{g_k^T (w_{k-1} - tr_{k-1})}{d_{k-1}^T w_{k-1}} - \mu \frac{\|w_{k-1} - tr_{k-1}\|^2}{(d_{k-1}^T w_{k-1})^2} g_k^T d_{k-1}. \tag{5.13}$$

By Table 1, concrete formulae for β_k^{SSD} are respectively given by

$$\begin{aligned}
\beta_k^{SSDDL} &= \frac{g_k^T (y_{k-1} - ts_{k-1})}{d_{k-1}^T y_{k-1}} - \frac{\mu \|y_{k-1} - ts_{k-1}\|^2}{(d_{k-1}^T y_{k-1})^2} g_k^T d_{k-1}, \\
\beta_k^{SSDYT} &= \frac{g_k^T (y_{k-1}^{(1)} - ts_{k-1})}{d_{k-1}^T y_{k-1}^{(1)}} - \frac{\mu \|y_{k-1}^{(1)} - ts_{k-1}\|^2}{(d_{k-1}^T y_{k-1}^{(1)})^2} g_k^T d_{k-1}, \\
\beta_k^{SSDZZ} &= \frac{g_k^T (y_{k-1}^{(2)} - ts_{k-1})}{d_{k-1}^T y_{k-1}^{(2)}} - \frac{\mu \|y_{k-1}^{(2)} - ts_{k-1}\|^2}{(d_{k-1}^T y_{k-1}^{(2)})^2} g_k^T d_{k-1}, \\
\beta_k^{SSDF1} &= \frac{g_k^T (y_{k-1}^{MS1} - ts_{k-1}^{MS1})}{d_{k-1}^T y_{k-1}^{MS1}} - \frac{\mu \|y_{k-1}^{MS1} - ts_{k-1}^{MS1}\|^2}{(d_{k-1}^T y_{k-1}^{MS1})^2} g_k^T d_{k-1}, \\
\beta_k^{SSDF2} &= \frac{g_k^T (y_{k-1}^{MS2} - ts_{k-1}^{MS2})}{d_{k-1}^T y_{k-1}^{MS2}} - \frac{\mu \|y_{k-1}^{MS2} - ts_{k-1}^{MS2}\|^2}{(d_{k-1}^T y_{k-1}^{MS2})^2} g_k^T d_{k-1}.
\end{aligned}$$

Note that, in [45], they dealt with the parameter of the form:

$$\begin{aligned}
\beta_k^{SSD} &= g_k^T (w_{k-1} - tr_{k-1}) (d_{k-1}^T w_{k-1})^\dagger \\
&\quad - \mu \|w_{k-1} - tr_{k-1}\|^2 g_k^T d_{k-1} \{(d_{k-1}^T w_{k-1})^2\}^\dagger.
\end{aligned}$$

They proved the global convergence of the SSDZZ method as follows.

Theorem 13. *Suppose that Assumptions* 1 *and* 2 *hold. Let* $\{x_k\}$ *be the sequence generated by the SSDZZ method, where* α_k *satisfies the Wolfe conditions* (2.1)–(2.2). *Then the method converges globally in the sense that* (2.5) *holds.*

They also proved the global convergence of the SSD+ method, namely, the CG method with (1.2), (2.8) and (5.13). In order to establish the global convergence of the SSDYT+ method, they modified β_k^{SSDYT} and defined $\tilde{\beta}_k^{SSDYT+}$ by (2.8) and (5.13) with $r_{k-1} = s_{k-1}$ and

$$w_{k-1} = y_{k-1} + \rho_k \left(\frac{\max\{0, \phi_{k-1}\}}{s_{k-1}^T u_{k-1}} u_{k-1} \right). \tag{5.14}$$

Note that this modification yields $d_{k-1}^T w_{k-1} \geq d_{k-1}^T y_{k-1} > 0$ under the Wolfe conditions.

Theorem 14. *Suppose that Assumptions* 1 *and* 2 *hold. Let* $\{x_k\}$ *be the sequence generated by the SSD+ method, where* α_k *satisfies the Wolfe conditions* (2.1)–(2.2)*. Then the following statements hold* :

(i) *The SSDDL+ method converges globally in the sense that* (2.5) *holds.*

(ii) *Assume that* ρ_k *and* u_k *satisfy* $0 \leq \rho_k \leq \bar{\rho}$ *and*

$$|s_{k-1}^T u_{k-1}| \geq \bar{m}\|s_{k-1}\|\|u_{k-1}\|,$$

where $\bar{\rho}$ *is any fixed positive constant and* $\bar{m}$ *is some positive constant. Then the SSDYT+ method* (*which uses* $\tilde{\beta}_k^{SSDYT+}$) *converges globally in the sense that* (2.5) *holds.*

(iii) *Assume that there exists a positive constant* φ_1 *such that, for all* k,

$$\max\{|g_{k-1}^T d_{k-1}|,\ |g_k^T d_{k-1}|\} \leq \varphi_1 |d_{k-1}^T y_{k-1}^{MS1}| \tag{5.15}$$

holds. If κ_k *satisfies* $0 \leq \kappa_k \leq \bar{\kappa}$ *for any fixed positive constant* $\bar{\kappa}$*, then the SSDF1+ method converges globally in the sense that* (2.5) *holds.*

(iv) *Assume that there exists a positive constant* φ_2 *such that, for all* k,

$$\max\{|g_{k-1}^T d_{k-1}|,\ |g_k^T d_{k-1}|\} \leq \varphi_2 |d_{k-1}^T y_{k-1}^{MS2}| \tag{5.16}$$

holds. If κ_k *satisfies* $0 \leq \kappa_k \leq \bar{\kappa}$ *for any fixed positive constant* $\bar{\kappa}$*, then the SSDF2+ method converges globally in the sense that* (2.5) *holds.*

Assumptions (5.15)–(5.16) look like strong assumptions. However, Narushima and Yabe [45] claimed that these are reasonable if the generalized strong Wolfe conditions (2.1) and (2.3) with $\sigma_2 < 1$ are used. It follows from (2.3) that

$$|g_k^T d_{k-1}| \leq \max\{\sigma_1, \sigma_2\}|g_{k-1}^T d_{k-1}| \leq |g_{k-1}^T d_{k-1}|,$$

which implies that (5.15) holds if $|g_{k-1}^T d_{k-1}| \leq \varphi_1 |d_{k-1}^T y_{k-1}^{MS1}|$ is satisfied. From the definition of y_{k-1}^{MS1} in (5.8), we have

$$d_{k-1}^T y_{k-1}^{MS1} = d_{k-1}^T y_{k-1} - \xi_{k-1} d_{k-1}^T y_{k-2}. \tag{5.17}$$

If $s_{k-1}^T y_{k-2} \leq 0$, then (5.17), $\xi_{k-1} > 0$ and (2.3) yield

$$d_{k-1}^T y_{k-1}^{MS1} \geq d_{k-1}^T y_{k-1} \geq -(1-\sigma_1) g_{k-1}^T d_{k-1} \quad (> 0).$$

If $s_{k-1}^T y_{k-2} > 0$, we can control the magnitude of the last term in (5.17) by using the parameter κ_k in (5.9). Thus (5.15) is justified. The assumption (5.16) is also reasonable by the same reason as in (5.15).

5.3. Scaled/three-term CG methods based on secant conditions

Chen and Liu [12] and Livieris and Pintelas [43] respectively incorporated β_k^{YT} and a variant of β_k^{ZZ} into the scaled CG method (4.8). On the other hand, Sugiki, Narushima and Yabe [54] gave the three-term CG method (4.5) with the parameter β_k in (5.12) and $p_k = w_{k-1} - tr_{k-1}$. In addition, Sugiki et al. proved the global convergence of their method as follows.

Theorem 15. *Suppose that Assumptions* 1 *and* 2 *hold. Let* $\{x_k\}$ *be the sequence generated by the three-term CG method* (1.2) *and* (4.5) *with* β_k *in* (5.12) *and* $p_k = w_{k-1} - tr_{k-1}$, *where* α_k *satisfies the Wolfe conditions* (2.1)–(2.2)*. If there exist positive constants* c_4 *and* c_5 *such that* w_{k-1} *and* r_{k-1} *satisfy*

$$\begin{aligned} \|w_{k-1} - tr_{k-1}\| &\leq c_4\|s_{k-1}\|, \\ |d_{k-1}^T w_{k-1}| &\geq c_5\alpha_{k-1}\|d_{k-1}\|^2 \end{aligned}$$

for all k, *then the method converges globally in the sense that* (2.5) *holds.*

By using the above theorem, Sugiki et al. also showed the global convergence of the concrete methods under the assumption that the objective function is uniformly convex. We note that if f is a uniformly convex function on a convex set $\mathcal{N}$, then there exists a constant $\lambda > 0$ such that

$$(\nabla f(x) - \nabla f(\widetilde{x}))^T(x - \widetilde{x}) \geq \lambda\|x - \widetilde{x}\|^2, \quad \text{for all} \quad x, \widetilde{x} \in \mathcal{N}.$$

Theorem 16. *Suppose that Assumptions* 1 *and* 2 *hold and* f *is a uniformly convex function. Let* $\{x_k\}$ *be the sequence generated by the three-term CG method* (1.2) *and* (4.5) *with* β_k *in* (5.12) *and* $p_k = w_{k-1} - tr_{k-1}$, *where* α_k *satisfies the Wolfe conditions* (2.1)–(2.2)*. Let* x^* *be a unique optimal solution of the problem* (1.1)*. Then the following statements hold* :

(i) *The method with* β_k^{DL} *converges globally in the sense that* $\lim_{k\to\infty} x_k = x^*$.

(ii) *Assume that* ρ_k *and* u_k *satisfy* $0 \leq \rho_k \leq \bar{\rho}$ *and*

$$|s_{k-1}^T u_{k-1}| \geq \bar{m}\|s_{k-1}\|\|u_{k-1}\|,$$

where $\bar{\rho}$ *is a positive constant such that* $\bar{\rho} < \frac{\lambda}{3L}$, *and* $\bar{m}$ *is some positive constant. Then the method with* β_k^{YT} *converges globally in the sense that* $\lim_{k\to\infty} x_k = x^*$.

(iii) *If* κ_k *satisfies* $0 \leq \kappa_k \leq \bar{\kappa}$ *for some positive constant* $\bar{\kappa} < \frac{2\lambda}{L}$, *then the method with* β_k^{F1} *converges globally in the sense that* $\lim_{k\to\infty} x_k = x^*$.

(iv) *If κ_k satisfies $0 \le \kappa_k \le \bar{\kappa}$ for some positive constant $\bar{\kappa} < \frac{2\lambda}{Lt}$, then the method with β_k^{F2} converges globally in the sense that $\lim_{k\to\infty} x_k = x^*$.*

Sugiki et al. [54] also showed the following global convergence property for general objective functions.

Theorem 17. *Suppose that Assumptions 1 and 2 hold. Let $\{x_k\}$ be the sequence generated by the three-term CG method (1.2) and (4.5) with β_k^{ZZ} and $p_k = w_{k-1} - tr_{k-1}$, where α_k satisfies the Wolfe conditions (2.1)–(2.2). Then the method converges globally in the sense that (2.5) holds.*

Since Sugiki et al. did not prove the global convergence of the methods for general objective functions except for the method with β_k^{ZZ}, we now give the global convergence theorem of the method with β_k^{DL+}, β_k^{F1+} and β_k^{F2+} and its sketch of proof. In addition, to establish the global convergence of the method with β_k^{YT+}, similarly to the SSDYT+ method in Theorem 14, we need to modify the parameter β_k^{YT+} and define $\tilde{\beta}_k^{YT+}$ by (2.8) and (5.12) with $r_{k-1} = s_{k-1}$ and w_{k-1} given in (5.14).

Theorem 18. *Suppose that Assumptions 1 and 2 hold. Let $\{x_k\}$ be the sequence generated by the three-term CG method (1.2) and (4.5) with β_k in (5.12) and $p_k = w_{k-1} - tr_{k-1}$, where α_k satisfies the generalized strong Wolfe conditions (2.1) and (2.3). Then the following statements hold:*

(i) *The method with β_k^{DL+} converges globally in the sense that (2.5) holds.*

(ii) *Assume that ρ_k and u_k satisfy $0 \le \rho_k \le \bar{\rho}$ and*

$$|s_{k-1}^T u_{k-1}| \ge \bar{m}\|s_{k-1}\|\|u_{k-1}\|,$$

where $\bar{\rho}$ is any fixed positive constant and $\bar{m}$ is some positive constant. Then the method with $\tilde{\beta}_k^{YT+}$ converges globally in the sense that (2.5) holds.

(iii) *Assume that there exists a positive constant φ_3 such that, for all k,*

$$|g_{k-1}^T d_{k-1}| \le \varphi_3 |d_{k-1}^T y_{k-1}^{MS1}| \tag{5.18}$$

holds. If κ_k satisfies $0 \le \kappa_k \le \bar{\kappa}$ for any fixed positive constant $\bar{\kappa}$, then the method with β_k^{F1+} converges globally in the sense that (2.5) holds.

(iv) *Assume that there exists a positive constant φ_4 such that, for all k,*

$$|g_{k-1}^T d_{k-1}| \le \varphi_4 |d_{k-1}^T y_{k-1}^{MS2}| \tag{5.19}$$

holds. If κ_k satisfies $0 \le \kappa_k \le \bar{\kappa}$ for any fixed positive constant $\bar{\kappa}$, then the method with β_k^{F2+} converges globally in the sense that (2.5) holds.

Proof. By Theorem 11, we only need to prove that each method satisfies Property 2. Moreover, similarly to the case of Property $\star$, it suffices to show that there exists a positive constant c_6 such that

$$|\psi_k| \leq c_6 \|s_{k-1}\| \tag{5.20}$$

holds for all k under the assumption that $\varepsilon \leq \|g_k\|$ holds for all k and some positive constant ε.

By (4.9), (5.12) and $p_k = w_{k-1} - tr_{k-1}$, we have

$$\begin{aligned} |\psi_k| &= \left| \frac{g_k^T(w_{k-1} - tr_{k-1})}{d_{k-1}^T w_{k-1}} \right| \|g_k\| \|w_{k-1} - tr_{k-1}\| |g_k^T(w_{k-1} - tr_{k-1})|^\dagger \\ &\leq \frac{\|g_k\| \|w_{k-1} - tr_{k-1}\|}{|d_{k-1}^T w_{k-1}|}. \end{aligned} \tag{5.21}$$

Assumptions 1 and 2 yield that $\|g_k\|$ is bounded. In a similar way to the proof of Theorem 14 (namely, [45, Theorem 3.5]), we can show that there exist positive constants c_7 and c_8 such that $\|w_{k-1} - tr_{k-1}\| \leq c_7 \|s_{k-1}\|$ and $|d_{k-1}^T w_{k-1}| \geq c_8$ hold for each method. Therefore, (5.21) implies (5.20), and hence the proof is complete. □

Although the assumptions (5.18) and (5.19) look like strong assumptions, we can justify these by the same reason as in (5.15) and (5.16).

§6. CG-DESCENT

CG-DESCENT [30–32,34] is a software developed by Hager and Zhang, which is based on the HZ+ method, and now it is one of major software for solving large-scale unconstrained optimization problems. Until Version 5.3, CG-DESCENT implemented the usual HZ+ method with an efficient line search, and from Version 6.0, a subspace iteration and a preconditioning step techniques are inserted into the previous version. The latest version is 6.7. Codes of CG-DESCENT are written by Fortran or C, and are provided in Hager's web page [29].

Hager and Zhang improved the line search such that the HZ+ method becomes more effective. In the line search of each iteration, by using the bisection method and the quadratic and cubic interpolations, the step size α_k is obtained so that the Wolfe conditions (2.1)–(2.2) are satisfied. If the condition

$$|f(x_k + \alpha_k d_k) - f(x_k)| \leq \omega C_k$$

is satisfied, then CG-DESCENT switches permanently the Wolfe conditions to the condition

$$f(x_k + \alpha_k d_k) \le f(x_k) + \epsilon |f(x_k)|$$

and the approximate Wolfe conditions

$$-(1-2\delta) g_k^T d_k \ge g(x_k + \alpha_k d_k)^T d_k \ge \sigma_1 g_k^T d_k,$$

where $\epsilon > 0$ and $\omega > 0$ are small numbers, $0 \le \Delta \le 1$, and C_k and Q_k are updated by

$$\begin{aligned} C_k &= C_{k-1} + (|f(x_k)| - C_{k-1})/Q_k, & C_{-1} = 0, \\ Q_k &= 1 + \Delta Q_{k-1}, & Q_{-1} = 0. \end{aligned}$$

Moreover, from Version 6.0, the subspace iteration and the preconditioning step are used, which are given in Hager-Zhang's paper [34]. When $g_k \in \mathcal{S}_k = \mathrm{Span}\{d_{k-1}, \dots, d_{k-m}\}$ for some integer m, iterates may converge very slowly. In order to avoid this phenomenon, they considered the following subspace minimization problem:

$$\min_{z \in \mathcal{S}_k} f(x_k + z). \tag{6.1}$$

If z_k is a solution of this problem and $x_{k+1} = x_k + z_k$, then we have $g(x_{k+1})^T v = 0$ for all $v \in \mathcal{S}_k$ by the first order optimality condition of (6.1). Therefore, $g(x_{k+1}) \notin \mathcal{S}_k$ or $g(x_{k+1}) = 0$ holds. Furthermore, in order to accelerate the method, they used the following preconditioned HZ+ method:

$$\begin{aligned} d_k &= -P_k g_k + \beta_k^+ d_k, \quad \beta_k = \frac{g_k^T P_k y_{k-1}}{d_{k-1}^T y_{k-1}} - \mu \frac{y_{k-1}^T P_k y_{k-1}}{(d_{k-1}^T y_{k-1})^2} g_k^T d_{k-1}, \\ \beta_k^+ &= \max\left\{ \beta_k, \bar{\nu}_3 \frac{g_{k-1}^T d_{k-1}}{d_{k-1}^T P_k^- d_{k-1}} \right\}, \end{aligned} \tag{6.2}$$

where μ and $\bar{\nu}_3$ are constants such that $\mu > 1/4$ and $\bar{\nu}_3 > 0$, P_k is a preconditioner matrix made by using information obtained in the subspace minimization problem (6.1), and P_k^- is the pseudoinverse of P_k. The outline of the algorithm is given by the following procedures, where ϑ_1 and ϑ_2 are positive constants such that $0 < \vartheta_1 < \vartheta_2 < 1$ and $\mathrm{dist}\{x, \mathcal{S}_k\} = \inf\{\|y - x\| \mid y \in \mathcal{S}_k\}$.

Standard CG iteration. Perform the HZ+ method (CG-DESCENT 5.3) as long as $\mathrm{dist}\{g_k, \mathcal{S}_k\} > \vartheta_1 \|g_k\|$. When $\mathrm{dist}\{g_k, \mathcal{S}_k\} \le \vartheta_1 \|g_k\|$ is satisfied, branch to the subspace iteration.

Subspace iteration. Solve the subspace minimization problem (6.1) by using the preconditioned HZ+ method (6.2) with $P_k = Z\widehat{P}_k Z^T$, where Z is a matrix whose columns are an orthonormal basis for the subspace $\mathcal{S}_k$ and $\widehat{P}_k$ is a preconditioner in the subspace. Stop at the iteration where $\text{dist}\{g_{k+1}, \mathcal{S}_k\} \geq \vartheta_2 \|g_{k+1}\|$ is satisfied, and then branch to the preconditioning step.

Preconditioning step. When the subspace iteration terminates and we return to the full space standard CG iteration, we have found that the convergence can be accelerated by performing the preconditioned HZ+ method (6.2). Define σ_k by

$$\sigma_k = \max\left\{\sigma_{\min}, \min\left\{\sigma_{\max}, \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}}\right\}\right\},$$

where $\sigma_{\max}$ and $\sigma_{\min}$ are parameters such that $0 < \sigma_{\min} \leq \sigma_{\max} < \infty$. Let Z be a matrix whose columns are an orthonormal basis for the subspace $\mathcal{S}_k$, and set $P_k = Z\widehat{P}_k Z^T + \sigma_k(I - ZZ^T)$, where $\widehat{P}_k$ is a preconditioner defined in Subspace iteration. After completing the preconditioning iteration, return to the standard CG iteration.

CG-DESCENT (from Version 6.0) is implemented based on the above procedures. If the preconditioned HZ+ method (6.2) with $\mu = 1$ is preconditioned by the Hessian approximation gotten from a quasi-Newton method, then $\beta_k^+ = 0$, and the method reduces to the quasi-Newton method. Therefore, in the subspace iteration, the quasi-Newton method is used. More details of implementation of CG-DESCENT are given in [34]. We also find from the numerical results in [34] that the subspace iteration and the preconditioning step are very efficient. We note that it is expected that these techniques work efficiently for other CG methods.

§7. Numerical results

In this section, we present some numerical results of the CG methods surveyed in this paper. The programs were coded in C by modifying the software package CG-DESCENT Version 5.3 [30–32]. All computations were carried out on Lenovo G570 PC with Intel Core i5-2430M CPU (2.40GHz×2) and 8.0Gb RAM. We run virtual Linux OS Ubuntu 11 on Windows 7 by using VMware Player 4.04, and assigned one processor and 5.9Gb RAM to Ubuntu 11.

Our test problems consist of 132 tests used by Hager [29] and belong to the CUTEr library [10, 28] for unconstrained optimization. The names of these

Table 2: Test problems (names & dimensions); Collected by CUTEr

Name	n	Name	n	Name	n	Name	n
AKIVA	2	DIXMAANE	3000	HEART8LS	8	PALMER7C	8
ALLINITU	4	DIXMAANF	3000	HELIX	3	PALMER8C	8
ARGLINA	200	DIXMAANG	3000	HIELOW	3	PENALTY1	1000
ARGLINB	200	DIXMAANH	3000	HILBERTA	2	PENALTY2	200
ARWHEAD	5000	DIXMAANI	3000	HILBERTB	10	POWELLSG	5000
BARD	3	DIXMAANJ	3000	HIMMELBB	2	POWER	10000
BDQRTIC	5000	DIXMAANK	15	HIMMELBF	4	QUARTC	5000
BEALE	2	DIXMAANL	3000	HIMMELBG	2	ROSENBR	2
BIGGS6	6	DIXON3DQ	10000	HIMMELBH	2	S308	2
BOX3	3	DJTL	2	HUMPS	2	SCHMVETT	5000
BRKMCC	2	DQDRTIC	5000	JENSMP	2	SENSORS	100
BROWNAL	200	DQRTIC	5000	KOWOSB	4	SINEVAL	2
BROWNBS	2	EDENSCH	2000	LIARWHD	5000	SINQUAD	5000
BROWNDEN	4	EG2	1000	LOGHAIRY	2	SISSER	2
BROYDN7D	5000	ENGVAL1	5000	MANCINO	100	SNAIL	2
BRYBND	5000	ENGVAL2	3	MARATOSB	2	SPARSINE	5000
CHNROSNB	50	ERRINROS	50	MEXHAT	2	SPARSQUR	10000
CLIFF	2	EXPFIT	2	MOREBV	5000	SPMSRTLS	4999
COSINE	10000	EXTROSNB	1000	MSQRTALS	1024	SROSENBR	5000
CRAGGLVY	5000	FLETCBV2	5000	MSQRTBLS	1024	STRATEC	10
CUBE	2	FLETCHCR	1000	NONCVXU2	5000	TESTQUAD	5000
CURLY10	10000	FMINSRF2	5625	NONDIA	5000	TOINTGOR	50
CURLY20	10000	FMINSURF	5625	NONDQUAR	5000	TOINTGSS	5000
DECONVU	63	FREUROTH	5000	OSBORNEA	5	TOINTPSP	50
DENSCHNA	2	GENHUMPS	5000	OSBORNEB	11	TOINTQOR	50
DENSCHNB	2	GENROSE	500	OSCIPATH	10	TQUARTIC	5000
DENSCHND	3	GROWTHLS	3	PALMER1C	8	TRIDIA	5000
DENSCHNE	3	GULF	3	PALMER1D	7	VARDIM	200
DENSCHNF	2	HAIRY	2	PALMER2C	8	VAREIGVL	50
DIXMAANA	3000	HATFLDD	3	PALMER3C	8	WATSON	12
DIXMAANB	3000	HATFLDE	3	PALMER4C	8	WOODS	4000
DIXMAANC	3000	HATFLDFL	3	PALMER5C	6	YFITU	3
DIXMAAND	3000	HEART6LS	6	PALMER6C	8	ZANGWIL2	2

tests and their dimension n are given in Table 2. Hager [29] dealt with 145 test problems, while we did not consider the remaining tests here due to the fact that the memory of our PC was insufficient for some of them and different local solutions were obtained when different solvers were applied to those omitted problems.

Table 3 presents the methods used in our experiments, where the first column consists of abbreviation names of these methods.

As mentioned above, we have implemented all the methods under considerations on the basis of the software package CG-DESCENT Version 5.3. Although this version is not the most recent one, we used it for a fair comparison of the CG methods. In the line search, we used the default procedures of CG-DESCENT, which are described in Section 6. We used the parameters

Table 3: Tested methods

HS	The HS+ method
CGD 5	CG-DESCENT Version 5.3 (namely, the HZ+ method)
3THS	The three-term CG method (4.5) with $\beta_k = \beta_k^{HS+}$ and $p_k = g_k$
G3THS	The three-term CG method (4.10) with (4.12), $\beta_k = \beta_k^{HS+}$ and $p_k = g_k$
DL	The DL+ method
SSDDL	The SSDDL+ method
3TDL	The three-term CG method (4.5) with $\beta_k = \beta_k^{DL+}$ and $p_k = y_{k-1} - ts_{k-1}$

values of $\delta = 0.1$, $\sigma_1 = 0.9$ for the Wolfe and the approximate Wolfe conditions, and used $\epsilon = 10^{-6}$, $\omega = 10^{-3}$ and $\Delta = 0.7$ for the parameters of the switching. For the other parameters, we set $\mu = 2$ for CGD 5 and SSDDL, $t = 1$ for DL, SSDDL and 3TDL, and $\bar{\gamma}_1 = 0.01$, $\bar{\gamma}_2 = 100$, $\bar{\theta} = 10^{-12}$ and $\bar{\gamma} = 0.8$ for G3THS. Moreover, we used, for all methods, modification (2.8) with $\zeta_k = \nu_k^{(2)}$ and $\bar{\nu}_2 = 0.4$. Since HS and DL do not necessarily generate descent search directions, we used the restart strategy (namely, we set $d_k = -g_k$) when the descent condition (1.4) was not satisfied. We stopped the algorithm if either

$$\|g_k\|_\infty \leq 10^{-6}$$

held or the CPU time exceeded 600 seconds (10 minutes).

To compare performances among the tested methods, we adopt the performance profiles of Dolan and Moré [21]. For n_s solvers and n_p problems, the performance profile $P : \boldsymbol{R} \to [0, 1]$ is defined as follows:

Let $\mathcal{P}$ and $\mathcal{S}$ be the set of problems and the set of solvers, respectively. For each problem $p \in \mathcal{P}$ and for each solver $s \in \mathcal{S}$, we define $t_{p,s}$ = computing time (similarly for the number of iterations) required to solve problem p by solver s. The performance ratio is given by $r_{p,s} = t_{p,s} / \min_{s \in \mathcal{S}} t_{p,s}$. Then, the performance profile is defined by $P(\tau) = \frac{1}{n_p} \text{size}\{p \in \mathcal{P} | r_{p,s} \leq \tau\}$, for all $\tau > 0$, where size$A$, for any set A, stands for the number of the elements in that set. Note that $P(\tau)$ is the probability for solver $s \in \mathcal{S}$ such that the performance ratio $r_{p,s}$ is within a factor $\tau > 0$ of the best possible ratio. Note that $n_p = 132$ was used in each figure.

In Figures 1 and 2, we give the performance profiles based on the CPU time. In order to prevent a measurement error, we set the minimum of the measurement 0.2 seconds. We see from Figure 1 that G3THS is superior to the other methods, and 3THS also worked well. On the other hand, HS did not perform so well. Figure 2 shows that SSDDL outperforms CGD 5 a little, and DL is almost comparable with CGD 5. In this numerical experiment, 3TDL performed poorly. 3TDL was stopped for a few problems because the number of line search iterations exceeds the pre-given criterion number. This is the reason why the performance profile of 3TDL looks poor. For the other problems, 3TDL worked well.

As mentioned in Section 6, the CG-DESCENT Version 6.7 (stands for CGD 6) [34] is the latest one, which was superior to the other tested methods. Since we expect that the subspace iteration and the preconditioning step also work efficiently for other CG methods, we incorporate these procedures into G3THS and SSDDL. The resulting methods (referred to as G3THS 6 and SSDDL 6, respectively) differ from CG-DESCENT 6.7 in the following three points. First, in the standard CG iteration, we used the search direction of G3THS or SSDDL instead of Hager-Zhang's direction. Second, in the line search technique, we impose the generalized strong Wolfe conditions (2.1) and (2.3) with $\delta = 0.001$, $\sigma_1 = 0.2$ and $\sigma_2 = 0.6$, instead of the Wolfe conditions (2.1)–(2.2). Third, in the preconditioning step, we use the preconditioned steepest descent direction (namely, a kind of quasi-Newton direction $d_k = -P_k g_k$), instead of the direction (6.2). The performance profiles of these methods are given in Figure 3. We see from Figure 3 that CGD 6, SSDDL 6 and G3THS 6 are clearly superior to CGD 5, SSDDL and G3THS. This fact implies that the subspace iteration and the preconditioning step are very efficient. We also find that SSDDL 6 and G3THS 6 performed better than CGD 6.

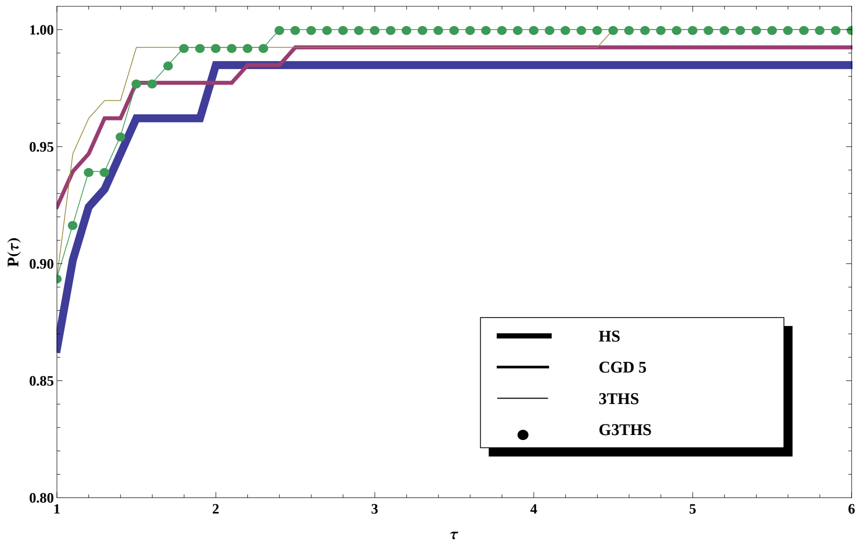

Figure 1: CPU Performance profile of HS, CGD 5, 3THS and G3THS.

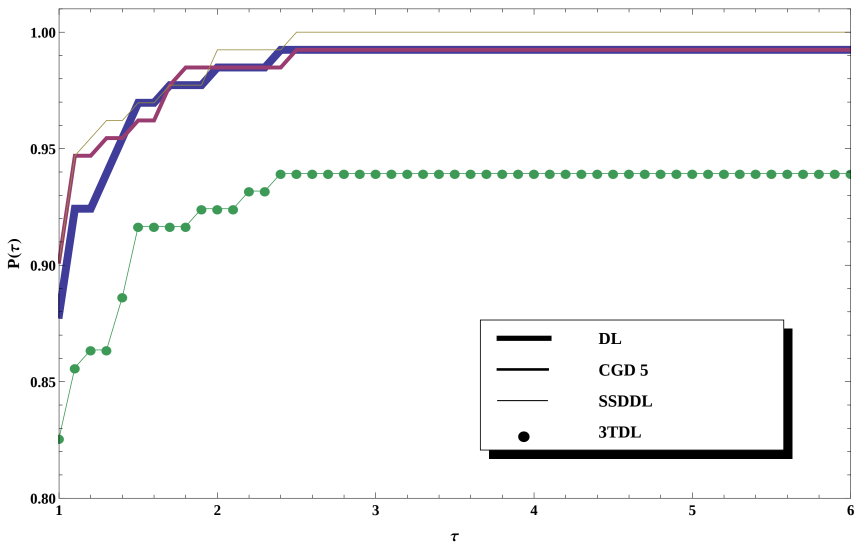

Figure 2: CPU Performance profile of DL, CGD 5, SSDDL and 3TDL.

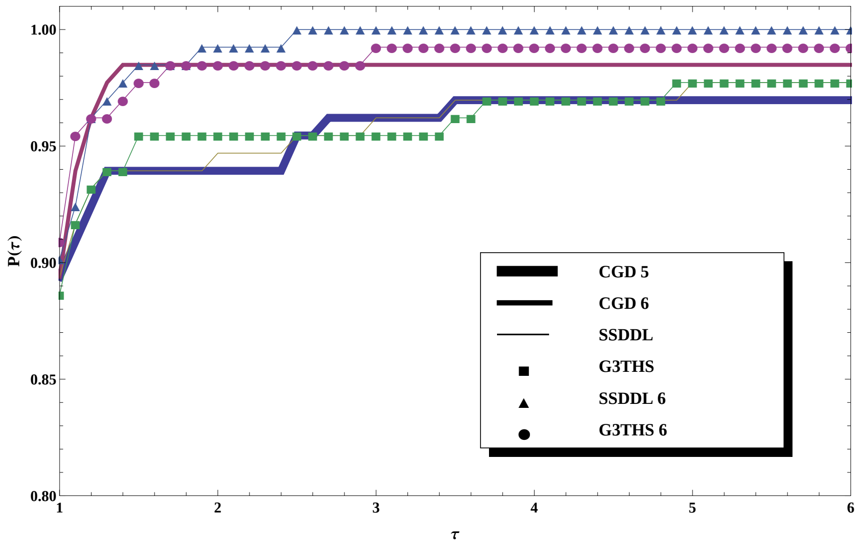

Figure 3: CPU Performance profile of usual CG methods and CG methods with the subspace iteration and the preconditioning step.

§8. Conclusions

In this decade, CG methods satisfying the sufficient descent property independent of line searches have been focused on by many researchers. In this paper, we have surveyed such sufficient descent CG methods. In order to establish the sufficient descent property, two kinds of strategies are well-known. The first one modifies the parameter β_k similarly to Hager-Zhang's method. The second one adds a term or incorporates a scaling factor to the search direction, which includes the three-term CG method by Narushima, Yabe and Ford. These methods overcome the weakness of the typical CG methods and work well in practice.

Moreover, CG methods based on secant conditions have been also studied. In this paper, we have introduced some sufficient descent CG methods based on secant conditions. CG-DESCENT is a software based on Hager-Zhang's CG method, and it is one of major software for solving large-scale unconstrained optimization problems. We have reviewed recent advances of CG-DESCENT.

We have confirmed performances of some sufficient descent CG methods. Moreover, we have incorporated the acceleration techniques into sufficient descent CG methods, and have seen that the resulting methods are very effective.

Acknowledgments

The authors are supported in part by the Grant-in-Aid for Scientific Research (C) 25330030 of Japan Society for the Promotion of Science.

References

[1] M. Al-Baali, Descent property and global convergence of the Fletcher-Reeves method with inexact line search, *IMA Journal of Numerical Analysis*, **5** (1985), 121–124.

[2] M. Al-Baali, Y. Narushima and H. Yabe, A family of three-term conjugate gradient methods with sufficient descent property for unconstrained optimization, *Computational Optimization and Applications*, **60** (2015), 89–110.

[3] N. Andrei, Scaled conjugate gradient algorithms for unconstrained optimization, *Computational Optimization and Applications*, **38** (2007), 401–416.

[4] N. Andrei, Hybrid conjugate gradient algorithm for unconstrained optimization, *Journal of Optimization Theory and Applications*, **141** (2009), 249–264.

[5] N. Andrei, A modified Polak-Ribière-Polyak conjugate gradient algorithm for unconstrained optimization, *Optimization*, **60** (2011), 1457–1471.

[6] N. Andrei, A simple three-term conjugate gradient algorithm for unconstrained optimization, *Journal of Computational and Applied Mathematics*, **241** (2013), 19–29.

[7] S. Babaie-Kafaki and R. Ghanbari, The Dai-Liao nonlinear conjugate gradient method with optimal parameter choices, *European Journal of Operational Research*, **234** (2014), 625–630.

[8] S. Babaie-Kafaki, R. Ghanbari and N. Mahdavi-Amiri, Two new conjugate gradient methods based on modified secant equations, *Journal of Computational and Applied Mathematics*, **234** (2010), 1374–1386.

[9] E.G. Birgin and J.M. Martínez, A spectral conjugate gradient method for unconstrained optimization, *Applied Mathematics and Optimization*, **43** (2001), 117–128.

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

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

[12] W. Cheng and Q. Liu, Sufficient descent nonlinear conjugate gradient methods with conjugacy condition, *Numerical Algorithms*, **53** (2010), 113-131.

[13] Y.H. Dai, Nonlinear conjugate gradient methods, in "Wiley Encyclopedia of Operations Research and Management Science" (eds. J. J. Cochran, L. A. Cox, Jr., P. Keskinocak, J. P. Kharoufeh and J. C. Smith), John Wiley & Sons, (2011).

[14] Y.H. Dai, J.Y. Han, G.H. Liu, D.F. Sun, H.X. Yin and Y. Yuan, Convergence properties of nonlinear conjugate gradient methods, *SIAM Journal on Optimization*, **10** (1999), 345–358.

[15] Y.-H. Dai and C.-X. Kou, A nonlinear conjugate gradient algorithm with an optimal property and an improved Wolfe line search, *SIAM Journal on Optimization*, **23** (2013), 296–320.

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

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

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

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

[20] Z. Dai and F. Wen, Global convergence of a modified Hestenes-Stiefel nonlinear conjugate gradient method with Armijo line search, *Numerical Algorithms*, **59** (2012), 79–93.

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

[22] R. Fletcher, Practical Methods of Optimization (Second Edition), John Wiley & Sons, 1987.

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

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

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

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

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

[28] N.I.M. Gould, D. Orban and P.L. Toint, CUTEr and SifDec: A constrained and unconstrained testing environment, revisited, *ACM Transactions on Mathematical Software*, **29** (2003), 373–394.

[29] W.W. Hager, Hager's web page, *http://people.clas.ufl.edu/hager/*, the last access date is June 24, 2014.

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

[31] W.W. Hager and H. Zhang, CG_DESCENT Version 1.4 User's Guide, University of Florida, November 2005.

[32] W.W. Hager and H. Zhang, Algorithm 851: CG_DESCENT, A conjugate gradient method with guaranteed descent, *ACM Transactions on Mathematical Software*, **32** (2006), 113–137.

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

[34] W.W. Hager and H. Zhang, The limited memory conjugate gradient method, *SIAM Journal on Optimization*, **23** (2013), 2150–2168.

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

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

[37] H. Iiduka and Y. Narushima, Conjugate gradient methods using value of objective function for unconstrained optimization, *Optimization Letters*, **6** (2012), 941–955.

[38] C.T. Kelly, Iterative methods for Linear and Nonlinear Equations, SIAM, 1995.

[39] M. Kobayashi, Y. Narushima and H. Yabe, Nonlinear conjugate gradient methods with structured secant condition for nonlinear least squares problems. *Journal of Computational and Applied Mathematics*, **234** (2010), 375–397.

[40] M. Li and H. Feng, A sufficient descent LS conjugate gradient method for unconstrained optimization problems, *Applied Mathematics and Computation*, **218** (2011), 1577–1586.

[41] D.H. Li and M. Fukushima, A modified BFGS method and its global convergence in nonconvex minimization, *Journal of Computational and Applied Mathematics*, **129** (2001), 15–35.

[42] Y. Liu and C. Storey, Efficient generalized conjugate gradient algorithms, Part1: Theory, *Journal of Optimization Theory and Applications*, **69** (1991), 129–137.

[43] I.E. Livieris and P. Pintelas, Globally convergent modified Perry's conjugate gradient method, *Applied Mathematics and Computation*, **218** (2012), 9197–9207.

[44] W. Nakamura, Y. Narushima and H. Yabe, Nonlinear conjugate gradient methods with sufficient descent properties for unconstrained optimization, *Journal of Industrial and Management Optimization*, **9** (2013), 595–619.

[45] Y. Narushima and H. Yabe, Conjugate gradient methods based on secant conditions that generate descent search directions for unconstrained optimization, *Journal of Computational and Applied Mathematics*, **236** (2012), 4303–4317.

[46] Y. Narushima, H. Yabe and J.A. Ford, A three-term conjugate gradient method with sufficient descent property for unconstrained optimization, *SIAM Journal on Optimization*, **21** (2011), 212–230.

[47] J. Nocedal and S.J. Wright, Numerical Optimization (2nd ed.), Springer Series in Operations Research, Springer-Verlag, New York, 2006.

[48] A. Perry, A modified conjugate gradient algorithm, *Operations Research*, **26** (1978), 1073–1078.

[49] M.J.D. Powell, Nonconvex minimization calculations and the conjugate gradient method, *in Lecture Notes in Mathematics*, no. 1066, Springer-Verlag, Berlin, 1984, 122–141.

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

[51] Y. Saad, Iterative methods for sparse linear systems (2nd ed.), SIAM, 2003.

[52] Z.-J. Shi and J. Shen, Convergence of Liu-Storey conjugate gradient method, *European Journal of Operational Research*, **182** (2007), 552–560.

[53] H.W. Sorenson, Comparison of some conjugate direction procedures for function minimization, *Journal of the Franklin Institute*, **288** (1969), 421–441.

[54] K. Sugiki, Y. Narushima and H. Yabe, Globally convergent three-term conjugate gradient methods that use secant conditions and generate descent search directions for unconstrained optimization, *Journal of Optimization Theory and Applications*, **153** (2012), 733–757.

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

[56] H. Yabe and N. Sakaiwa, A new nonlinear conjugate gradient method for unconstrained optimization, *Journal of the Operations Research Society of Japan*, **48** (2005), 282–296.

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

[58] G. Yu, L. Guan and W. Chen, Spectral conjugate gradient methods with sufficient descent property for large-scale unconstrained optimization, *Optimization Methods and Software*, **23** (2008), 275–293.

[59] G. Yu, L. Guan and G. Li, Global convergence of modified Polak-Ribière-Polyak conjugate gradient methods with sufficient descent property, *Journal of Industrial and Management Optimization*, **4** (2008), 565–579.

[60] G. Yuan, Modified nonlinear conjugate gradient methods with sufficient descent property for large-scale optimization problems, *Optimization Letters*, **3** (2009), 11–21.

[61] J.Z. Zhang, N.Y. Deng, and L.H. Chen, New quasi-Newton equation and related methods for unconstrained optimization, *Journal of Optimization Theory and Applications*, **102** (1999), 147–167.

[62] J.Z. Zhang and C.X. Xu, Properties and numerical performance of quasi-Newton methods with modified quasi-Newton equations, *Journal of Computational and Applied Mathematics*, **137** (2001), 269–278.

[63] L. Zhang and J. Li, A new globalization technique for nonlinear conjugate gradient methods for nonconvex minimization, *Applied Mathematics and Computation*, **217** (2011), 10295–10304.

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

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

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

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

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

Yasushi Narushima
Department of Management System Science, Yokohama National University,
79-4, Tokiwadai, Hodogaya-ku, Yokohama 240-8501, Japan.
E-mail: narushima(at)ynu.ac.jp

Hiroshi Yabe
Department of Mathematical Information Science, Tokyo University of Science,
1-3, Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan.
E-mail: yabe(at)rs.kagu.tus.ac.jp