A SURVEY OF NONLINEAR CONJUGATE GRADIENT METHODS *

WILLIAM W. HAGER† AND HONGCHAO ZHANG‡

Abstract. This paper reviews the development of different versions of nonlinear conjugate gradient methods, with special attention given to global convergence properties.

Key words. Nonlinear conjugate gradient methods, Unconstrained optimization, Nonlinear programming

AMS subject classifications. 90C06, 90C26, 65Y20

1. Introduction. Conjugate gradient (CG) methods comprise a class of unconstrained optimization algorithms which are characterized by low memory requirements and strong local and global convergence properties. CG history, surveyed by Golub and O'Leary in [48], begins with research of Cornelius Lanczos and Magnus Hestenes and others (Forsythe, Motzkin, Rosser, Stein) at the Institute for Numerical Analysis (National Applied Mathematics Laboratories of the United States National Bureau of Standards in Los Angeles), and with independent research of Eduard Stiefel at Eidg. Technische Hochschule Zürich. In the seminal 1952 paper [59] of Hestenes and Stiefel, the algorithm is presented as an approach to solve symmetric, positive-definite linear systems.

In this survey, we focus on conjugate gradient methods applied to the nonlinear unconstrained optimization problem

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

where $f : \mathbb{R}^n \mapsto \mathbb{R}$ is a continuously differentiable function, bounded from below. A nonlinear conjugate gradient method generates a sequence $\mathbf{x}_k$, $k \geq 1$, starting from an initial guess $\mathbf{x}_0 \in \mathbb{R}^n$, using the recurrence

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k, \tag{1.2}$$

where the positive step size α_k is obtained by a line search, and the directions $\mathbf{d}_k$ are generated by the rule:

$$\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + \beta_k \mathbf{d}_k, \quad \mathbf{d}_0 = -\mathbf{g}_0. \tag{1.3}$$

Here β_k is the CG update parameter and $\mathbf{g}_k = \nabla f(\mathbf{x}_k)^\mathsf{T}$, where the gradient $\nabla f(\mathbf{x}_k)$ of f at $\mathbf{x}_k$ is a row vector and $\mathbf{g}_k$ is a column vector. Different CG methods correspond to different choices for the scalar β_k.

Let $\|\cdot\|$ denote the Euclidean norm and define $\mathbf{y}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$. Table 1.1 provides a chronological list of some choices for the CG update parameter. The 1964 formula of Fletcher and Reeves is usually considered the first nonlinear CG algorithm since their paper [45] focuses on nonlinear optimization, while the 1952 paper [59] of Hestenes and Stiefel focuses on symmetric, positive-definite linear systems.

* February 7, 2005. Revised August 22, 2005. This material is based upon work supported by the National Science Foundation under Grant No. 0203270.

† hager@math.ufl.edu, http://www.math.ufl.edu/∼hager, PO Box 118105, Department of Mathematics, University of Florida, Gainesville, FL 32611-8105. Phone (352) 392-0281. Fax (352) 392-8357.

‡hzhang@math.ufl.edu, http://www.math.ufl.edu/∼hzhang, PO Box 118105, Department of Mathematics, University of Florida, Gainesville, FL 32611-8105.

$\beta_k^{HS} = \dfrac{\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{y}_k}{\mathbf{d}_k^{\mathsf{T}}\mathbf{y}_k}$	(1952)	in the original (linear) CG paper of Hestenes and Stiefel [59]
$\beta_k^{FR} = \dfrac{\|\mathbf{g}_{k+1}\|^2}{\|\mathbf{g}_k\|^2}$	(1964)	first nonlinear CG method, proposed by Fletcher and Reeves [45]
$\beta_k^{D} = \dfrac{\mathbf{g}_{k+1}^{\mathsf{T}}\nabla^2 f(\mathbf{x}_k)\mathbf{d}_k}{\mathbf{d}_k^{\mathsf{T}}\nabla^2 f(\mathbf{x}_k)\mathbf{d}_k}$	(1967)	proposed by Daniel [39], requires evaluation of the Hessian $\nabla^2 f(\mathbf{x})$
$\beta_k^{PRP} = \dfrac{\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{y}_k}{\|\mathbf{g}_k\|^2}$	(1969)	proposed by Polak and Ribière [84] and by Polyak [85]
$\beta_k^{CD} = \dfrac{\|\mathbf{g}_{k+1}\|^2}{-\mathbf{d}_k^{\mathsf{T}}\mathbf{g}_k}$	(1987)	proposed by Fletcher [44], CD stands for "Conjugate Descent"
$\beta_k^{LS} = \dfrac{\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{y}_k}{-\mathbf{d}_k^{\mathsf{T}}\mathbf{g}_k}$	(1991)	proposed by Liu and Storey [67]
$\beta_k^{DY} = \dfrac{\|\mathbf{g}_{k+1}\|^2}{\mathbf{d}_k^{\mathsf{T}}\mathbf{y}_k}$	(1999)	proposed by Dai and Yuan [27]
$\beta_k^{N} = \left(\mathbf{y}_k - 2\mathbf{d}_k\dfrac{\|\mathbf{y}_k\|^2}{\mathbf{d}_k^{\mathsf{T}}\mathbf{y}_k}\right)^{\mathsf{T}}\dfrac{\mathbf{g}_{k+1}}{\mathbf{d}_k^{\mathsf{T}}\mathbf{y}_k}$	(2005)	proposed by Hager and Zhang [53]

TABLE 1.1
Various choices for the CG update parameter

Daniel's choice for the update parameter, which is fundamentally different from the other choices, is not discussed in this paper. For large-scale problems, choices for the update parameter that do not require the evaluation of the Hessian matrix are often preferred in practice over methods that require the Hessian in each iteration. In the remaining methods of Table 1.1, except for the new method at the end, the numerator of the update parameter β_k is either $\|\mathbf{g}_{k+1}\|^2$ or $\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{y}_k$ and the denominator is either $\|\mathbf{g}_k\|^2$ or $\mathbf{d}_k^{\mathsf{T}}\mathbf{y}_k$ or $-\mathbf{d}_k^{\mathsf{T}}\mathbf{g}_k$. The 2 possible choices for the numerator and the 3 possible choices for the denominator lead to 6 different choices for β_k shown above.

If f is a strongly convex quadratic, then in theory, all 8 choices for the update parameter in Table 1.1 are equivalent with an exact line search. For non-quadratic cost functions, each choice for the update parameter leads to different performance. Some of today's best performing CG algorithms are hybrid methods, which dynamically adjust the formula for β_k as the iterations evolve, and a method based on the recent update parameter β_k^N, with close connections to memoryless quasi-Newton methods. In numerical experiments reported in [54], using CUTEr test problems [6], the top performance relative to CPU time was obtained by a code based on β_k^N, while the second best performance was obtained by either a code based on a hybrid DY/HS scheme [37] or a code based on the L-BFGS scheme of Nocedal [79] and Liu and Nocedal [66]. In all these codes, the best performance was obtained using the approximate Wolfe line search developed in [53, 54].

In this paper, we focus on global convergence properties of CG methods; consequently, n-step quadratic convergence results [14, 60], which should be taken into account (see [54]) in the design of efficient CG algorithms, are not discussed. In Sec-

tion 2 we discuss classical line search criteria based on the Wolfe conditions [100, 101], and we present the related Zoutendijk condition. In Section 3, we briefly summarize how the initial search direction affects global convergence. Section 4 discusses the global convergence properties of CG methods with numerator $\|\mathbf{g}_{k+1}\|^2$ for the update parameter, while Section 5 considers methods with $\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{y}_k$ in the numerator. The convergence theory for the methods of Section 4 is more highly developed than the theory for the methods of Section 5. On the other hand, the methods of Section 5 often perform better in practice. Section 6 introduces hybrid methods obtained by dynamically adjusting the formula for β_k as the iterations evolve, making use of methods from both Sections 4 and 5. Section 7 discusses the CG method associated with the new choice β_k^N for the update parameter, called CG_DESCENT in [54]. An important feature of this scheme, which distinguishes it from the other schemes, is that $\mathbf{d}_{k+1}$ is always a descent direction for any stepsize $\alpha_k > 0$, as long as $\mathbf{d}_k^{\mathsf{T}}\mathbf{y}_k \neq 0$. Finally, Section 8 discusses preconditioning techniques for CG algorithms.

2. Line search and Zoutendijk type conditions. In each CG iteration, the stepsize α_k is chosen to yield an approximate minimum for the problem:

$$\min_{\alpha \geq 0} \quad f(\mathbf{x}_k + \alpha \mathbf{d}_k). \tag{2.1}$$

Since $\alpha \geq 0$, the direction $\mathbf{d}_k$ should satisfy the *descent condition*

$$\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k < 0, \tag{2.2}$$

for all $k \geq 0$. If there exists a constant $c > 0$ such that

$$\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k < -c\|\mathbf{g}_k\|^2 \tag{2.3}$$

for all $k \geq 0$, then the search directions satisfy the *sufficient descent condition*.

The termination conditions for the CG line search are often based on some version of the Wolfe conditions. The *standard Wolfe conditions* [100, 101] are

$$f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) - f(\mathbf{x}_k) \leq \delta \alpha_k \mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k, \tag{2.4}$$

$$\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{d}_k \geq \sigma \mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k, \tag{2.5}$$

where $\mathbf{d}_k$ is a descent direction and $0 < \delta \leq \sigma < 1$. The *strong Wolfe conditions* consists of (2.4) and the following strengthened version of (2.5):

$$|\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{d}_k| \leq -\sigma \mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k. \tag{2.6}$$

In the *generalized Wolfe conditions* [24], the absolute value in (2.6) is replaced by a pair of inequalities:

$$\sigma_1 \mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k \leq \mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{d}_k \leq -\sigma_2 \mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k, \tag{2.7}$$

where $0 < \delta < \sigma_1 < 1$ and $\sigma_2 \geq 0$. The special case $\sigma_1 = \sigma_2 = \sigma$ corresponds to the strong Wolfe conditions. Ideally, we would like to terminate the line search in a CG algorithm when the standard Wolfe conditions are satisfied. For some CG algorithms, however, stronger versions of the Wolfe conditions are needed to ensure convergence and to enhance stability.

Recently, we introduced the *approximate Wolfe conditions*

$$\sigma \mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k \leq \mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{d}_k \leq (2\delta - 1)\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k, \tag{2.8}$$

where $0 < \delta < 1/2$ and $\delta < \sigma < 1$. The first inequality in (2.8) is the same as (2.5). The second inequality in (2.8) is equivalent to (2.4) when f is quadratic. In general, when $\phi(\alpha) = f(\mathbf{x}_k + \alpha \mathbf{d}_k)$ is replaced by a quadratic interpolant $q(\cdot)$ that matches $\phi(\alpha)$ at $\alpha = 0$ and $\phi'(\alpha)$ at $\alpha = 0$ and $\alpha = \alpha_k$, (2.4) reduces to the second inequality in (2.8). Observe that the approximate Wolfe conditions have the same form as the generalized Wolfe condition (2.7), but with a special choice for σ_2. Note though that the decay condition (2.4) is one component of the generalized Wolfe conditions, while in the approximate Wolfe conditions, the decay condition is approximately enforced through the second inequality in (2.8).

The standard or generalized or strong Wolfe conditions are used to prove convergence of CG algorithms. The approximate Wolfe conditions are used in efficient, high accuracy implementations of CG algorithms for which there is no convergence theory, but the practical performance is often much better than that of the rigorous implementations. As shown in [53], the first Wolfe condition (2.4) limits the accuracy of a CG algorithm to the order of the square root of the machine precision, while with the approximation contained in (2.8), we can achieve accuracy on the order of the machine precision. As explained further in [54], we often achieve faster convergence when using the approximate Wolfe conditions since a local minimizer of ϕ satisfies (2.8), while a point satisfying the standard or strong Wolfe conditions is obtained by computing a local minimizer of the approximating function ψ introduced in [72]:

$$\psi(\alpha) = \phi(\alpha) - \phi(0) - \alpha\delta\phi'(0).$$

When using the approximate Wolfe conditions, we minimize the function f along the search direction $\mathbf{d}_k$ rather than an approximation ψ to f.

Either of the following assumptions are often utilized in convergence analysis for CG algorithms:

Lipschitz Assumption: In some neighborhood $\mathcal{N}$ of the level set

$$\mathcal{L} = \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \le f(\mathbf{x}_0)\},$$

the gradient $\nabla f(\mathbf{x})$ is Lipschitz continuous. That is, there exists a constant $L < \infty$ such that

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le L\|\mathbf{x} - \mathbf{y}\| \qquad \text{for all} \quad \mathbf{x}, \mathbf{y} \in \mathcal{N}.$$

Boundedness Assumption: The level set $\mathcal{L}$ is bounded. That is, there exists a constant $B < \infty$ such that

$$\|x\| \le B \quad \text{for all} \quad \mathbf{x} \in \mathcal{L}.$$

The conclusion of the following theorem, often called the *Zoutendijk condition*, is used to prove the global convergence of nonlinear CG methods; it was originally given by Zoutendijk [108] and Wolfe [100, 101].

THEOREM 2.1. *Consider any iterative method of the form* (1.2) *where* $\mathbf{d}_k$ *satisfies the descent condition* (2.2) *and* α_k *satisfies the standard Wolfe conditions. If the Lipschitz Assumption holds, then*

$$\sum_{k=0}^{\infty} \frac{(\mathbf{g}_k^{\mathsf{T}} \mathbf{d}_k)^2}{\|\mathbf{d}_k\|^2} < +\infty. \tag{2.9}$$

Global convergence proofs for CG methods are often based on the Zoutendijk condition combined with analysis showing that

(a) the sufficient descent condition $\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k \le -c\|\mathbf{g}_k\|^2$ holds and
(b) there exists a constant β such that $\|\mathbf{d}_k\|^2 \le \beta k$.

Combining (a), (b), and (2.9), we have

$$\liminf_{k\to\infty} \|\mathbf{g}_k\| = 0. \tag{2.10}$$

Throughout this survey, the statement that a CG method converges globally means either $\mathbf{g}_k = \mathbf{0}$ *for some* k *or* (2.10) *holds.*

Another result related to the Zoutendijk condition, found in [21] (also see [57]), is the following: assuming the search directions are descent,

THEOREM 2.2. *Consider any iterative method of the form* (1.2)–(1.3) *where* $\mathbf{d}_k$ *satisfies the descent condition* (2.2) *and* α_k *satisfies the strong Wolfe conditions. If the Lipschitz Assumption holds, then either*

$$\liminf_{k\to\infty} \|\mathbf{g}_k\| = 0$$

or

$$\sum_{k=1}^{\infty} \frac{\|\mathbf{g}_k\|^4}{\|\mathbf{d}_k\|^2} < \infty.$$

Notice that in Theorems 2.1 and 2.2, we assume both the descent condition and the Wolfe conditions. These two requirements are essentially independent of each other. Hence, when implementing a line search, we need to satisfy some version of the Wolfe conditions, and we need to ensure that the new search direction is a descent direction. In recent versions [27, 37] of the CG algorithms associated with the choice β_k^{DY}, the descent condition holds automatically, when the line search satisfies the standard Wolfe conditions. And in the very recent CG_DESCENT [53, 54], sufficient descent holds for $\mathbf{x}_{k+1}$ with $c = 7/8$ if $\mathbf{d}_k^{\mathsf{T}}\mathbf{y}_k \ne 0$ (the second Wolfe condition (2.5) implies that $\mathbf{d}_k^{\mathsf{T}}\mathbf{y}_k > 0$).

3. Starting search direction. It is critical to take $\mathbf{d}_0 = -\mathbf{g}_0$ in a CG algorithm. In 1972 Crowder and Wolfe [16] gave a 3-dimensional example showing that the convergence rate is linear if the initial search direction is not the steepest descent direction, even for a strongly convex quadratic. In 1976 Powell [87] obtained an even stronger result; he showed that if the objective function is a convex quadratic and if the initial search direction is an arbitrary descent direction, then either optimality is achieved in at most $n+1$ iterations or the rate of convergence is only linear. Moreover, by analyzing the relationship between $\mathbf{x}_0$ and $\mathbf{d}_0$, it follows that linear convergence is more common than finite convergence.

In order to achieve finite convergence for an arbitrary initial search direction, Nazareth [73] proposed a CG algorithm based on a three-term recurrence:

$$\mathbf{d}_{k+1} = -\mathbf{y}_k + \frac{\mathbf{y}_k^{\mathsf{T}}\mathbf{y}_k}{\mathbf{d}_k^{\mathsf{T}}\mathbf{y}_k}\mathbf{d}_k + \frac{\mathbf{y}_{k-1}^{\mathsf{T}}\mathbf{y}_k}{\mathbf{d}_{k-1}^{\mathsf{T}}\mathbf{y}_{k-1}}\mathbf{d}_{k-1}, \tag{3.1}$$

with $\mathbf{d}_{-1} = \mathbf{0}$ and $\mathbf{d}_0$ an arbitrary descent direction. If f is a convex quadratic, then for any stepsize α_k, the search directions generated by (3.1) are conjugate relative to the Hessian of f. However, this interesting innovation has not seen significant use in practice.

4. Methods with $\|\mathbf{g}_{k+1}\|^2$ in the numerator of β_k. The FR, DY and CD methods all have the common numerator $\|\mathbf{g}_{k+1}\|^2$. One theoretical difference between these methods, and the other choices for the update parameter, is that the global convergence theorems only require the Lipschitz Assumption, not the Boundedness Assumption. The first global convergence result for the FR method was given by Zoutendijk [108] in 1970. He proved the FR method converges globally when α_k is an exact solution of the problem (2.1); in other words, global convergence is achieved when the line search is exact. In 1977 Powell [88] pointed out that the FR method, with exact line search, was susceptible to jamming. That is, the algorithm could take many short steps without making significant progress to the minimum. The poor performance of the FR method in applications was often attributed to this jamming phenomenon.

The first global convergence result of the FR method for an inexact line search was given by Al-Baali [1] in 1985. Under the strong Wolfe conditions with $\sigma < 1/2$, he proved the FR method generates sufficient descent directions. More precisely, he proved that

$$\frac{1-2\sigma+\sigma^{k+1}}{1-\sigma} \le \frac{-\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k}{\|\mathbf{g}_k\|^2} \le \frac{1-\sigma^{k+1}}{1-\sigma},$$

for all $k \ge 0$. As a consequence, global convergence was established using the Zoutendijk condition. For $\sigma = 1/2$, $\mathbf{d}_k$ is a descent direction, however, the analysis did not establish sufficient descent.

In Liu *et al.* [65], the global convergence proof of Al-Baali is extended to the case $\sigma = 1/2$. Dai and Yuan [24] analyzed this further, and showed that in consecutive FR iterations, at least one iteration satisfies the sufficient descent property. In other words,

$$\max\{\frac{-\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k}{\|\mathbf{g}_k\|^2}, \frac{-\mathbf{g}_{k-1}^{\mathsf{T}}\mathbf{d}_{k-1}}{\|\mathbf{g}_{k-1}\|^2}\} \ge \frac{1}{2}.$$

The more recent result Theorem 2.2 can also be used to obtain a global convergence result for FR implemented with a strong Wolfe line search and $\sigma \le 1/2$ since the search directions are always descent directions.

In [35], Dai and Yuan show that with the FR scheme, the strong Wolfe conditions may not yield a direction of descent when $\sigma > 1/2$, even for the function $f(\mathbf{x}) = \lambda\|\mathbf{x}\|^2$, where $\lambda > 0$ is a constant. Hence, the constraint $\sigma \le 1/2$ must be imposed to ensure descent. In typical implementations of the Wolfe conditions, it is often most efficient to choose σ close to 1. Hence, the constraint $\sigma \le 1/2$, needed to ensure descent, represents a significant restriction in the choice of the line search parameters. On the other hand, Dai and Yuan show in [29] that when $\sigma > 1/2$ and $\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k > 0$, $-\mathbf{d}_k$ can be used for a search direction; if $\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k = 0$, then the line search can be skipped by setting $\mathbf{x}_{k+1} = \mathbf{x}_k$. If there exists a constant γ such that $\|\mathbf{g}_k\| \le \gamma$, then under the Lipschitz Assumption, the FR method, with a standard Wolfe line search and with these special adjustments when $\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k \ge 0$, is globally convergent.

In [24] the strong Wolfe line search is relaxed to a generalized Wolfe line search. Global convergence is obtained when $\sigma_1 + \sigma_2 \le 1$. For a strong Wolfe line search, $\sigma_1 = \sigma_2 = \sigma$, in which case the constraint $\sigma_1 + \sigma_2 \le 1$ implies that $\sigma \le 1/2$. Hence, the condition $\sigma_1 + \sigma_2 \le 1$ is weaker than the strong Wolfe constraint $\sigma \le 1/2$. And it is possible to take σ_1 close to 1, by taking σ_2 close to 0.

The CD method of Fletcher [44] is closely related to the FR method. With an exact line search, $\beta_k^{FR} = \beta_k^{CD}$. One important difference between FR and CD is that with CD, sufficient descent (2.3) holds for a strong Wolfe line search (the constraint $\sigma \leq 1/2$ that arose with FR, is not needed for CD). Moreover, for a line search that satisfies the generalized Wolfe conditions (2.7) with $\sigma_1 < 1$ and $\sigma_2 = 0$, it can be shown that $0 \leq \beta_k^{CD} \leq \beta_k^{FR}$. Consequently, from the analysis in [1] or by Theorem 2.2, global convergence is achieved. On the other hand, if $\sigma_1 \geq 1$ or $\sigma_2 > 0$, Dai and Yuan [25] construct examples where $\|\mathbf{d}_k\|^2$ increases exponentially and the CD method converges to a point where the gradient does not vanish. In particular, the CD method may not converge to a stationary point for a strong Wolfe line search.

The DY method, first developed in [27], is fundamentally different from either the FR or the CD method. With a standard Wolfe line search, the DY method always generates descent directions. Furthermore, there is global convergence when the Lipschitz Assumption holds. In the paper [19], Dai analyzed the DY method further and established the following remarkable property, relating the descent directions generated by DY to the sufficient descent condition:

THEOREM 4.1. *Consider the method* (1.2)–(1.3)*, where* $\beta_k = \beta_k^{DY}$*. If the DY method is implemented with any line search for which the search directions are descent directions, and if there exist constants* γ_1 *and* γ_2 *such that* $\gamma_1 \leq \|\mathbf{g}_k\| \leq \gamma_2$ *for all* $k \geq 0$*, then for any* $p \in (0,1)$*, there exists a constant* $c > 0$ *such that the sufficient descent condition*

$$\mathbf{g}_i^{\mathsf{T}} \mathbf{d}_i \leq -c\|\mathbf{g}_i\|^2$$

holds for at least $\lfloor pk \rfloor$ *indices* $i \in [0,k]$*, where* $\lfloor r \rfloor$ *denotes the largest integer* $\leq r$*.*

In the process of analyzing the DY method, Dai and Yuan also established a convergence result applicable to any method for which β_k can be expressed as a ratio:

$$\beta_k = \frac{\Phi_{k+1}}{\Phi_k}. \tag{4.1}$$

The FR method corresponds to the choice $\Phi_k = \|\mathbf{g}_k\|^2$. Utilizing (1.3), β_k^{DY} can be rewritten as

$$\beta_k^{DY} = \frac{\mathbf{g}_{k+1}^{\mathsf{T}} \mathbf{d}_{k+1}}{\mathbf{g}_k^{\mathsf{T}} \mathbf{d}_k}.$$

Hence, the DY method has the form (4.1) with $\Phi_k = \mathbf{g}_k^{\mathsf{T}} \mathbf{d}_k$. The following result was established [36, 38]:

THEOREM 4.2. *Consider any iterative method of the form* (1.2)–(1.3) *where* β_k *has the form* (4.1)*,* $\mathbf{d}_k$ *satisfies the descent condition* (2.2)*, and the Lipschitz Assumption holds. If the Zoutendijk condition holds and if*

$$\sum_{k=0}^{\infty} \frac{(\mathbf{g}_k^{\mathsf{T}} \mathbf{d}_k)^2}{\Phi_k^2} = \infty \quad \textit{or} \quad \sum_{k=0}^{\infty} \frac{\|\mathbf{g}_k\|^2}{\Phi_k^2} = \infty \quad \textit{or} \quad \sum_{k=1}^{\infty} \prod_{i=1}^{k} \beta_i^{-2} = \infty,$$

then the iterates are globally convergent.

As a corollary of this result, the DY method is globally convergent when implemented with a standard Wolfe line search since

$$\sum_{k=0}^{N} \frac{(\mathbf{g}_k^{\mathsf{T}} \mathbf{d}_k)^2}{\Phi_k^2} = N + 1 \quad \text{when} \quad \Phi_k = \mathbf{g}_k^{\mathsf{T}} \mathbf{d}_k.$$

FR is globally convergent when implemented with a strong Wolfe line search with $\sigma \leq 1/2$ since

$$\sum_{k=0}^{N} \frac{\|\mathbf{g}_k\|^2}{\Phi_k^2} = N+1 \quad \text{when} \quad \Phi_k = \|\mathbf{g}_k\|^2.$$

Notice that a general CG method can be expressed in the form (4.1) by taking $\Phi_0 = 1$, and

$$\Phi_k = \prod_{j=1}^{k} \beta_k, \qquad \text{for } k > 0.$$

5. Methods with $\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{y}_k$ in the numerator of β_k. Despite the strong convergence theory that has been developed for methods with $\|\mathbf{g}_{k+1}\|^2$ in the numerator of β_k, these methods are all susceptible to jamming, as discussed earlier. That is, they begin to take small steps without making significant progress to the minimum. The PRP, HS and LS methods, which share the common numerator $\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{y}_k$, possess a built-in restart feature that addresses the jamming problem: When the step $\mathbf{x}_{k+1}-\mathbf{x}_k$ is small, the factor $\mathbf{y}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$ in the numerator of β_k tends to zero. Hence, β_k becomes small and the new search direction $\mathbf{d}_{k+1}$ is essentially the steepest descent direction $-\mathbf{g}_{k+1}$. In a sense, the the PRP, HS, and LS method automatically adjust β_k to avoid jamming; in general, the performance of these methods is better than the performance of methods with $\|\mathbf{g}_{k+1}\|^2$ in the numerator of β_k.

In [84] global convergence of the PRP method is established when f is strongly convex and the line search is exact. Powell [88] proved that for a general nonlinear function, if

(a) the step size $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ approaches zero,
(b) the line search is exact, and
(c) the Lipschitz Assumption holds,

then the PRP method is globally convergent. On the other hand, Powell showed later [89], using a 3 dimensional example, that with an exact line search, the PRP method could cycle infinitely, without converging to a stationary point. Hence, the assumption that the stepsize tends to zero is needed for convergence.

Under the assumption that the search direction is a descent direction, Yuan [104] established the global convergence of the PRP method for strongly convex objective functions and a Wolfe line search. However, for a strong Wolfe line search, Dai [17] gave an example which showed that even when the objective function is strongly convex and $\sigma \in (0,1)$ is sufficiently small, the PRP method may still fail by generating an ascent search direction.

In summary, the convergence of the PRP method for general nonlinear function is uncertain; Powell's example shows that when the function is not strongly convex, the PRP method may not converge, even with an exact line search. And Dai's example shows that even for a strongly convex function, the PRP method may not generate a descent direction with an inexact line search. Based on insight gained from Powell's example, he suggested [89] the following modification in the update parameter for the PRP method:

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

In [47] Gilbert and Nocedal proved the convergence of the PRP+ method.

The analysis of Gilbert and Nocedal applies to a class of CG algorithms which have the following property:

> *Consider a method of the form* (1.2)–(1.3)*, and suppose that* $0 < \gamma \leq \|\mathbf{g}_k\| \leq \bar{\gamma}$*, for all* $k \geq 0$*, where* γ *and* $\bar{\gamma}$ *are two positive constants. Property* $(\star)$ *holds if there exist constants* $b > 1$ *and* $\lambda > 0$ *such that for all* k*,*
>
> $$|\beta_k| \leq b \quad \text{and} \quad \|s_k\| \leq \lambda \quad \text{implies} \quad |\beta_k| \leq \frac{1}{2b}.$$

The following result is proved in [47]

THEOREM 5.1. *Consider any CG method* (1.2)–(1.3) *that satisfies the following conditions:*

(a) $\beta_k \geq 0$.
(b) *The search directions satisfy the sufficient descent condition* (2.3).
(c) *The Zoutendijk condition holds.*
(d) *Property* $(\star)$ *holds.*

If the Lipschitz and Boundedness Assumptions hold, then the iterates are globally convergent.

As a corollary of this result, when the search direction satisfies the sufficient descent condition and when a standard Wolfe line search is employed, the PRP+ method is globally convergent. In [26], Dai and Yuan gave examples to show that the Boundedness Assumption is really needed to obtain the global convergence of Theorem 5.1; moreover, the constraint $\beta_k \geq 0$ can not be relaxed to $\max\{\beta_k^{PRP}, -\epsilon\}$ for any choice of $\epsilon > 0$. In [21] it is shown that the sufficient descent condition in Theorem 5.1 can be relaxed to a descent condition if a strong Wolfe line search is used.

The PRP+ method was introduced to rectify the convergence failure of the PRP method when implemented with a Wolfe line search. Another approach for rectifying the convergence failure is to retain the PRP update formula, but modify the line search. In particular, Grippo and Lucidi [49] proposed a new Armijo type line search of the following form:

$$\alpha_k = \max\{\lambda^j \frac{\tau |\mathbf{g}_k^\mathsf{T} \mathbf{d}_k|}{\|\mathbf{d}_k\|^2}\}$$

where $j \geq 0$ is the smallest integer with the property that

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \delta \alpha_k^2 \|\mathbf{d}_k\|^2, \tag{5.1}$$

and

$$-c_1 \|\mathbf{g}_{k+1}\|^2 \leq \mathbf{g}_{k+1}^\mathsf{T} \mathbf{d}_{k+1} \leq c_2 \|\mathbf{g}_{k+1}\|^2, \tag{5.2}$$

where $0 < c_2 < 1 < c_1$, $0 < \lambda < 1$ and $\tau > 0$ are constants. With this new line search, they prove global convergence of the PRP method. In the more recent paper [50], they combine this line search with a "trust region" technique.

In another avenue of research, it is shown in [35] that the PRP method is globally convergent when the line search employs a constant stepsize $\alpha_k = \eta < 1/4L$, where L is a Lipschitz constant for ∇f. In [97] Sun and Zhang give a global convergence result for the choice $\alpha_k = -\delta \frac{\mathbf{g}_k^\mathsf{T} \mathbf{d}_k}{\mathbf{d}_k^\mathsf{T} Q_k \mathbf{d}_k}$, where Q_k is some positive definite matrix with

smallest eigenvalue $\nu_{min} > 0$, $\delta \in (0, \nu_{min}/L)$, and L is a Lipschitz constant for ∇f. For these stepsize choices, the search directions are no longer conjugate when f is a quadratic. Hence, these methods should be viewed as steepest descent methods, rather than conjugate gradient methods.

The HS method has the property that the conjugacy condition

$$\mathbf{d}_{k+1}^{\mathsf{T}} \mathbf{y_k} = 0 \tag{5.3}$$

always holds, independent of line search. For an exact line search, $\beta_k^{HS} = \beta_k^{PRP}$. Hence, the convergence properties of the HS method should be similar to the convergence properties of the PRP method. In particular, by Powell's example [89], the HS method with an exact line search may not converge for a general nonlinear function. It can be verified that if the search directions satisfy the sufficient descent condition and if a standard Wolfe line search is employed, then the HS method satisfies Property ($\star$). Similar to the PRP+ method, if we let

$$\beta_k^{HS+} = \max\{\beta_k^{HS}, 0\},$$

then it follows from Theorem 5.1, that the HS+ method is globally convergent.

The LS method is also identical to the the PRP method for an exact line search. Although not much research has been done on this choice for the update parameter, except for the paper [67], we expect that the techniques developed for the analysis of the PRP method should apply to the LS method.

6. Hybrid CG methods and parametric families. As we have seen, the first set of methods FR, DY, and CD have strong convergence properties, but they may not perform well in practice due to jamming. On the other hand, although the second set of methods PRP, HS, and LS may not converge in general, they often perform better than the first set of methods. Consequently, combinations of methods have been proposed to try to exploit attractive features of each set. Touati-Ahmed and Storey [98] suggested the following hybrid method:

$$\beta_k = \begin{cases} \beta_k^{PRP} & \text{if} \quad 0 \le \beta_k^{PRP} \le \beta_k^{FR}, \\ \beta_k^{FR} & \text{otherwise.} \end{cases}$$

Thus, when the iterations jam, the PRP update parameter is used. By the same motivations, Hu and Storey [61] suggested to take

$$\beta_k = \max\{0, \min\{\beta_k^{PRP}, \beta_k^{FR}\}\}.$$

In [47] it is pointed out that β_k^{PRP} can be negative, even for strongly convex functions. In an effort to extend the allowed choices for the PRP update parameter, while retaining global convergence, Nocedal and Gilbert [47] suggested taking

$$\beta_k = \max\{-\beta_k^{FR}, \min\{\beta_k^{PRP}, \beta_k^{FR}\}\}.$$

With this hybrid method, β_k can be negative since β_k^{FR} is always nonnegative. Note that in the numerical results reported in [47], the performance of this hybrid method was not better than that of PRP+.

Convergence results for CG methods that can be bounded in terms of the FR method are developed in [47, 57, 61]. In particular, the following result is established in [57]:

THEOREM 6.1. *Consider any CG method of the form* (1.2)–(1.3) *which employs a strong Wolfe line search with* $\sigma \leq 1/2$. *If the Lipschitz Assumption holds and* $2\sigma|\beta_k| \leq \beta_k^{FR}$, *then the iterates are globally convergent, and the search directions are always descent directions.*

In [47, 61], global convergence and sufficient descent are established when $2\sigma|\beta_k| < \beta_k^{FR}$, while Theorem 6.1 claims global convergence and descent when $2\sigma|\beta_k| \leq \beta_k^{FR}$.

Recall that the DY method has even better global convergence properties than the FR method. As a result, Dai and Yuan [37] studied the possibility of combining DY with other CG methods. For a standard Wolfe line search and for $\beta_k \in [-\eta\beta_k^{DY}, \beta_k^{DY}]$, where $\eta = (1-\sigma)/(1+\sigma)$, they establish global convergence when the Lipschitz Assumption holds. The following two hybrid method were proposed in [37]:

$$\beta_k = \max\{-(\frac{1-\sigma}{1+\sigma})\beta_k^{DY}, \min\{\beta_k^{HS}, \beta_k^{DY}\}\}$$

and

$$\beta_k = \max\{0, \min\{\beta_k^{HS}, \beta_k^{DY}\}\}.$$

The numerical experiments in [23] indicated that the second hybrid method gave the best result, performing better than the PRP+ method.

Another hybrid method, proposed by Dai in [20], employs either the DY scheme or the CD scheme:

$$\beta_k = \frac{\|\mathbf{g}_{k+1}\|^2}{\max\{\mathbf{d}_k^\mathsf{T}\mathbf{y}_k, -\mathbf{g}_k^\mathsf{T}\mathbf{d}_k\}}$$

He shows that this hybrid scheme generates descent directions, independent of the line search. This descent property is stronger than that of the DY scheme itself, where descent holds for a Wolfe line search. Dai shows that for this hybrid scheme, $\beta_k \in [0, \beta_k^{DY}]$. This property, together with the descent property of the hybrid scheme, imply convergence for typical line search methods.

In the same way that quasi-Newton methods have been combined together by introducing parameters, as in the Broyden [7] family, CG methods can be combined together. In [33, 38], Dai and Yuan proposed a one-parameter family of CG methods with

$$\beta_k = \frac{\|\mathbf{g}_{k+1}\|^2}{\lambda_k\|\mathbf{g}_k\|^2 + (1-\lambda_k)\mathbf{d}_k^\mathsf{T}\mathbf{y}_k},$$

where $\lambda_k \in [0,1]$ is a parameter. The FR method corresponds to $\lambda_k = 1$, while the DY method corresponds to $\lambda_k = 0$. In [34], this family is extended by considering $\lambda_k \in (-\infty, \infty)$; if the Lipschitz Assumption holds, then there is global convergence for each member of the family when a generalized Wolfe line search is employed with

$$\sigma_1 - 1 \leq (\sigma_1 + \sigma_2)\lambda_k \leq 1.$$

By considering convex combinations of the numerators and denominators of β_k^{FR} and β^{HS}, Nazareth [78] independently proposes a two-parameter family of CG methods:

$$\beta_k = \frac{\mu_k\|\mathbf{g}_{k+1}\|^2 + (1-\mu_k)\mathbf{g}_{k+1}^\mathsf{T}\mathbf{y}_k}{\lambda_k\|\mathbf{g}_k\|^2 + (1-\lambda_k)\mathbf{d}_k^\mathsf{T}\mathbf{y}_k},$$

where $\lambda_k, \mu_k \in [0,1]$. This two-parameter family includes FR, DY, PRP, and HS methods as extreme cases.

Observing that the six standard CG methods share two numerators and three denominators, Dai and Yuan [36] considered an even wider family of CG methods by introducing one more parameter; they chose

$$\beta_k = \frac{\mu_k\|\mathbf{g}_{k+1}\|^2 + (1-\mu_k)\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{y}_k}{(1-\lambda_k-\omega_k)\|\mathbf{g}_k\|^2 + \lambda_k\mathbf{d}_k^{\mathsf{T}}\mathbf{y}_k - \omega_k\mathbf{d}_k^{\mathsf{T}}\mathbf{g}_k},$$

where $\lambda_k, \mu_k \in [0,1]$ and $\omega_k \in [0, 1-\lambda_k]$. This three-parameter family includes the six standard CG methods, the previous one-parameter and two-parameter families, and many hybrid methods as special cases. In order to ensure that the search directions generated by this family will be descent directions, Powell's [88] restart criterion is employed: set $\mathbf{d}_k = -\mathbf{g}_k$ if

$$|\mathbf{g}_k^{\mathsf{T}}\mathbf{g}_{k-1}| > \xi\|\mathbf{g}_k\|^2,$$

where $\xi > 0$ is some fixed constant. For a strong Wolfe line search where

$$(1+\xi)\sigma \le \frac{1}{2},$$

Dai and Yuan show that the search directions $\mathbf{d}_k$ are descent directions. Global convergence results are also established in [36].

In [22] Dai and Liao modify the numerator of the HS update parameter to obtain

$$\beta_k^{DL} = \frac{\mathbf{g}_{k+1}^{\mathsf{T}}(\mathbf{y}_k - t\mathbf{s}_k)}{\mathbf{d}_k^{\mathsf{T}}\mathbf{y}_k} = \beta_k^{HS} - t\frac{\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{s}_k}{\mathbf{d}_k^{\mathsf{T}}\mathbf{y}_k},$$

where $t > 0$ is some constant. For an exact line search, $\mathbf{g}_{k+1}$ is orthogonal to $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k = \alpha_k\mathbf{d}_k$. Hence, for an exact line search, the DL method reduces to the HS and PRP methods. Again, due to Powell's example, the DL method may not converge for an exact line search. Similar to the PRP+ method, Dai and Liao also modified their formula in the following way to ensure convergence:

$$\beta_k^{DL+} = \max\{\frac{\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{y}_k}{\mathbf{d}_k^{\mathsf{T}}\mathbf{y}_k}, 0\} - t\frac{\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{s}_k}{\mathbf{d}_k^{\mathsf{T}}\mathbf{y}_k}.$$

If the Lipschitz and Boundedness Assumptions hold and if $\mathbf{d}_k$ satisfies the sufficient descent condition (2.3), it is shown in [22] that DL+, implemented with a strong Wolfe line search, is globally convergent.

Very recently, in a further development of this update strategy, Yabe and Takano [103] derive the following choice for the update parameter, based on a modified secant condition given by Zhang *et al.* [106, 107]:

$$\beta_k^{YT} = \frac{\mathbf{g}_{k+1}^{\mathsf{T}}(\mathbf{z}_k - t\mathbf{s}_k)}{\mathbf{d}_k^{\mathsf{T}}\mathbf{z}_k},$$

where

$$\mathbf{z}_k = \mathbf{y}_k + \left(\frac{\rho\theta_k}{\mathbf{s}_k^{\mathsf{T}}\mathbf{u}_k}\right)\mathbf{u}_k,$$
$$\theta_k = 6(f_k - f_{k+1}) + 3(\mathbf{g}_k + \mathbf{g}_{k+1})^{\mathsf{T}}\mathbf{s}_k,$$

$\rho \geq 0$ is a constant and $\mathbf{u}_k \in \mathbb{R}^n$ satisfies $\mathbf{s}_k^\mathsf{T}\mathbf{u}_k \neq 0$; for example, $\mathbf{u}_k = \mathbf{d}_k$. Again, similar to PRP+, Yabe and Takano modified their formula in the following way to ensure convergence:

$$\beta_k^{YT+} = \max\{\frac{\mathbf{g}_{k+1}^\mathsf{T}\mathbf{z}_k}{\mathbf{d}_k^\mathsf{T}\mathbf{z}_k}, 0\} - t\frac{\mathbf{g}_{k+1}^\mathsf{T}\mathbf{s}_k}{\mathbf{d}_k^\mathsf{T}\mathbf{z}_k}.$$

They show that the YT+ scheme is globally convergent if the Lipschitz and Boundedness Assumptions hold, $\mathbf{d}_k$ satisfies the sufficient descent condition (2.3), and a strong Wolfe line search is employed with

$$0 \leq \rho < \frac{1-\sigma}{3(1+\sigma-2\delta)}.$$

Preliminary numerical results reported for both the DL+ and YT+ schemes indicate these CG method are efficient with a proper choice of the parameters. However, for different choices of the parameters, the performance of the methods can be quite different. Notice that both the DL and YT schemes are not scale invariant. That is, if f is multiplied by a positive scalar, then the values of β_k^{DL} and β_k^{YT} typically change. Hence, if a proper choice of t is found for some f, then t must be changed when f is rescaled. In contrast, all the methods of Table 1.1 are scale invariant.

7. CG_DESCENT. In the previous discussion, we have seen that the global convergence of CG methods is closely connected to the descent conditions (2.2) and (2.3) – we must not only employ a Wolfe type line search, but we must constrain α_k so that $\mathbf{d}_{k+1}$ a direction of descent. The final method of Table 1.1 was devised in order to ensure sufficient descent, independent of the accuracy of the line search. We obtain this method by modifying the HS method in the following way:

$$\beta_k^\theta = \beta_k^{HS} - \theta_k\left(\frac{\|\mathbf{y}_k\|^2\mathbf{g}_{k+1}^\mathsf{T}\mathbf{d}_k}{(\mathbf{d}_k^\mathsf{T}\mathbf{y}_k)^2}\right), \tag{7.1}$$

where $\theta_k \geq 0$. (We assume that $\mathbf{d}_k^\mathsf{T}\mathbf{y}_k \neq 0$ so that β_k^{HS} is defined. With a standard Wolfe line search, $\mathbf{d}_k^\mathsf{T}\mathbf{y}_k > 0$ when $\mathbf{g}_k \neq \mathbf{0}$).

An attractive feature of the HS method is that the conjugacy condition (5.3) always holds, independent of the line search. Consequently, if θ_k in (7.1) is near zero, then the conjugacy condition holds approximately. Also, the HS method is not susceptible to jamming. The expression multiplying θ_k in (7.1) has the following properties:

(a) it is scale invariant (β_k^θ does not change when f is multiplied by a positive scalar),
(b) it goes to zero when the iterates jam, and
(c) it enhances descent.

With regard to (b), observe that for an approximate Wolfe line search, we have

$$\frac{|\mathbf{g}_{k+1}^\mathsf{T}\mathbf{d}_k|}{\mathbf{d}_k^\mathsf{T}\mathbf{y}_k} \leq \frac{\max\{\sigma, 1-2\delta\}|\mathbf{d}_k^\mathsf{T}\mathbf{g}_k|}{(1-\sigma)|\mathbf{d}_k^\mathsf{T}\mathbf{g}_k|} = \frac{\max\{\sigma, 1-2\delta\}}{(1-\sigma)}.$$

If the search direction $\mathbf{d}_k$ satisfies the sufficient descent condition (2.3), it follows that

$$\frac{\|\mathbf{y}_k\|^2|\mathbf{g}_{k+1}^\mathsf{T}\mathbf{d}_k|}{(\mathbf{d}_k^\mathsf{T}\mathbf{y}_k)^2} \leq \left(\frac{\max\{\sigma, 1-2\delta\}}{c(1-\sigma)}\right)\left(\frac{\|\mathbf{y}_k\|}{\|\mathbf{g}_k\|}\right)^2.$$

When the iterates jam, $\mathbf{y}_k$ becomes tiny while $\|\mathbf{g}_k\|$ is bounded away from zero. Consequently, when the iterates jam, the θ_k term in (7.1) becomes negligible.

To see the effect of the new term on descent, we multiply the update formula (1.3) by $\mathbf{g}_{k+1}^{\mathsf{T}}$ to obtain

$$\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{d}_{k+1} = -\|\mathbf{g}_{k+1}\|^2 + \left(\frac{\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{y}_k(\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{d}_k)}{\mathbf{d}_k^{\mathsf{T}}\mathbf{y}_k}\right) - \theta_k\left(\frac{(\|\mathbf{y}_k\|^2(\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{d}_k)^2}{(\mathbf{d}_k^{\mathsf{T}}\mathbf{y}_k)^2}\right). \tag{7.2}$$

Due to the minus in front of the θ_k term above, it follows that the θ_k modification of the HS formula has enhanced descent. In fact, the only term on the right side of (7.2) that could be positive is the middle term, associated with β_k^{HS}.

An upper bound for the middle term in (7.2) is obtained using the inequality

$$\mathbf{u}_k^{\mathsf{T}}\mathbf{v}_k \le \frac{1}{2}\left(\|\mathbf{u}_k\|^2 + \|\mathbf{v}_k\|^2\right)$$

with the choice

$$\mathbf{u}_k = \frac{1}{\sqrt{2\theta_k}}(\mathbf{d}_k^{\mathsf{T}}\mathbf{y}_k)\mathbf{g}_{k+1} \quad \text{and} \quad \mathbf{v}_k = \sqrt{2\theta_k}(\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{d}_k)\mathbf{y}_k.$$

We have

$$\begin{aligned}\frac{\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{y}_k(\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{d}_k)}{\mathbf{d}_k^{\mathsf{T}}\mathbf{y}_k} &= \frac{\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{y}_k(\mathbf{d}_k^{\mathsf{T}}\mathbf{y}_k)(\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{d}_k)}{(\mathbf{d}_k^{\mathsf{T}}\mathbf{y}_k)^2} \\ &\le \frac{1}{4\theta_k}\|\mathbf{g}_{k+1}\|^2 + \theta_k\left(\frac{(\|\mathbf{y}_k\|^2(\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{d}_k)^2}{(\mathbf{d}_k^{\mathsf{T}}\mathbf{y}_k)^2}\right).\end{aligned}$$

Combining this with (7.2) gives

$$\mathbf{g}_{k+1}^{\mathsf{T}}\mathbf{d}_{k+1} \le -\left(1 - \frac{1}{4\theta_k}\right)\|\mathbf{g}_{k+1}\|^2. \tag{7.3}$$

Hence, if $\theta_k = \bar{\theta} > 1/4$ for each k, then the scheme (7.1) satisfies the sufficient descent condition (2.3) with $c = 1 - (4\bar{\theta})^{-1}$.

The parameter θ_k essentially controls the relative weight placed on conjugacy versus descent. As θ_k becomes small, the last term in (7.1) goes to zero, and the update parameter approaches the HS parameter, which satisfies the conjugacy condition (5.3). As θ_k tends to infinity, the sufficient descent parameter $c = 1 - (4\bar{\theta})^{-1}$ increases, approaching 1. The update parameter β_k^N given in Table 1.1 corresponds to β_k^θ and the choice $\theta_k = 2$. As discussed in [53], for strongly convex functions and a relatively accurate line search, the search directions generated by $\theta_k = 2$ are approximately multiples of the search directions produced by the memoryless quasi-Newton method of Perry [83] and Shanno [93].

In order to obtain global convergence for general nonlinear functions, we need to truncate β_k^θ, similar to PRP+. In [53] we introduce the following truncation:

$$\beta_k^{\theta+} = \max\left\{\beta_k^N, \eta_k\right\}, \quad \eta_k = \frac{-1}{\|\mathbf{d}_k\|\min\{\eta, \|\mathbf{g}_k\|\}}, \tag{7.4}$$

where $\eta > 0$ is a constant. In [53] and [54] we give numerical results for the update parameter β_k^N of Tabel 1.1, corresponding to $\theta_k = 2$. For a broad set of large-scale

unconstrained optimization problems in the CUTEr library [6], the new scheme, called CG_DESCENT in [54], performed better than either PRP+ or L-BFGS.

As in [53, Thm 1.1], the update parameter $\beta_k^{\theta+}$ satisfies the sufficient descent condition (7.3), when $\theta_k \geq \bar{\theta} > 1/4$ for all k. Similar to the result established in [53] for the $N+$ scheme (corresponding to $\theta_k = 2$ in $\beta_k^{\theta+}$), we have the following result for the more general scheme based on $\beta_k^{\theta+}$ with θ_k bounded away from 1/4:

THEOREM 7.1. *Suppose the Lipschitz and the Boundedness Assumptions hold, and the line search satisfies the standard Wolfe conditions* (2.4) *and* (2.5)*. Then the CG method* (1.2) *and* (1.3)*, with* $\beta_k = \beta_k^{\theta+}$ *and* $\theta_k \geq \bar{\theta} > 1/4$ *for all* k*, is globally convergent.*

In Theorem 7.1, $\theta_k \geq \bar{\theta} > 1/4$. Global convergence also holds when θ_k approaches 1/4 from the right, but with a strong Wolfe line search instead of the standard Wolfe line search. The proof of this result is based on Theorem 2.2, not the approach in [53]. Although global convergence for a standard Wolfe line search is still an open question, we have the following result: If we perform a restart, taking $\mathbf{d}_k = -\mathbf{g}_k$ whenever

$$\mathbf{g}_k^{\mathsf{T}} \mathbf{g}_{k-1} \leq -\zeta \|\mathbf{g}_k\| \|\mathbf{g}_{k-1}\|, \tag{7.5}$$

where $\zeta \in (0, 1)$, then for a standard Wolfe line search and for $\theta_k > 1/4$, we obtain global convergence. The condition (7.5) is a one-sided version of the restart condition suggested by Powell [88] for Beale's 3-term CG method. Note that ζ should be taken close to 1 to reduce the number of restarts.

8. Preconditioning. The idea behind preconditioning is to make a change of variables $\mathbf{x} = \mathbf{S}\mathbf{y}$ where $\mathbf{S}$ is an invertible matrix chosen to speedup the convergence. After writing the conjugate gradient algorithm in the transformed variable $\mathbf{y}$ and converting back to the $\mathbf{x}$ variable, we obtain the iteration:

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{x}_k + \alpha_k \mathbf{d}_k, \\ \mathbf{d}_{k+1} &= \mathbf{P}\mathbf{g}_{k+1} + \bar{\beta}_k \mathbf{d}_k, \quad \mathbf{d}_0 = \mathbf{P}\mathbf{g}_0, \end{aligned} \tag{8.1}$$

where $\mathbf{P} = \mathbf{S}\mathbf{S}^{\mathsf{T}}$. The update parameter $\bar{\beta}_k$ is the same as β_k except that $\mathbf{g}_k$ and $\mathbf{d}_k$ are replaced by $\mathbf{S}^{\mathsf{T}}\mathbf{g}_k$ and $\mathbf{S}^{-1}\mathbf{d}_k$ respectively. As illustrations, we have

$$\bar{\beta}_k^{FR} = \frac{\mathbf{g}_{k+1}^{\mathsf{T}} \mathbf{P} \mathbf{g}_{k+1}}{\mathbf{g}_k^{\mathsf{T}} \mathbf{P} \mathbf{g}_k} \quad \text{and} \quad \bar{\beta}_k^{CD} = \frac{\mathbf{g}_{k+1}^{\mathsf{T}} \mathbf{P} \mathbf{g}_{k+1}}{-\mathbf{d}_k^{\mathsf{T}} \mathbf{g}_k}.$$

To obtain insights into the effect of preconditioning, we examine how the convergence speed of CG depends on the eigenvalues of the Hessian. Suppose that f is quadratic:

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathsf{T}} \mathbf{Q} \mathbf{x} + \mathbf{b}^{\mathsf{T}} \mathbf{x}, \tag{8.2}$$

where $\mathbf{Q}$ is a symmetric matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$. With an exact line search, the error in the k-th CG iterate satisfies the following bound [95]:

$$(\mathbf{x}_k - \mathbf{x}^*)^{\mathsf{T}} \mathbf{Q} (\mathbf{x}_k - \mathbf{x}^*) \leq \min_{p \in \mathcal{P}_{k-1}} \max_{1 \leq i \leq n} (1 + \lambda_i p(\lambda_i))^2 \ (\mathbf{x}_0 - \mathbf{x}^*)^{\mathsf{T}} \mathbf{Q} (\mathbf{x}_0 - \mathbf{x}^*),$$

where $\mathcal{P}_k$ denotes the set of polynomials of degree at most k. Given some integer $l \in [1, k]$, it follows that if $p \in \mathcal{P}_{k-1}$ is chosen so that the degree k polynomial

$1 + \lambda p(\lambda)$ vanishes with multiplicity 1 at λ_i, $1 \le i \le l-1$, and with multiplicity $k-l+1$ at $(\lambda_l + \lambda_n)/2$, then we have

$$(\mathbf{x}_k - \mathbf{x}^*)^\mathsf{T}\mathbf{Q}(\mathbf{x}_k - \mathbf{x}^*) \le \left(\frac{\lambda_l - \lambda_n}{\lambda_l + \lambda_n}\right)^{2(k-l+1)} (\mathbf{x}_0 - \mathbf{x}^*)^\mathsf{T}\mathbf{Q}(\mathbf{x}_0 - \mathbf{x}^*). \tag{8.3}$$

After the change of variables $\mathbf{x} = \mathbf{S}\mathbf{y}$ in (8.2), we obtain

$$f(\mathbf{S}\mathbf{y}) = \frac{1}{2}\mathbf{y}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{Q}\mathbf{S}\mathbf{y} + \mathbf{b}^\mathsf{T}\mathbf{S}\mathbf{y}.$$

The matrix $\mathbf{S}^\mathsf{T}\mathbf{Q}\mathbf{S}$ associated with the quadratic in $\mathbf{y}$ is similar to the matrix $\mathbf{Q}\mathbf{S}\mathbf{S}^\mathsf{T} = \mathbf{Q}\mathbf{P}$. Hence, the best preconditioner is $\mathbf{P} = \mathbf{Q}^{-1}$, which leads to convergence in one step since the eigenvalues of $\mathbf{S}^\mathsf{T}\mathbf{Q}\mathbf{S}$ are all 1.

When f is a general nonlinear function, a good preconditioner is any matrix that approximates $\nabla^2 f(\mathbf{x}^*)^{-1}$. As an illustration showing how to choose $\mathbf{P}$, let us consider a nonlinear, equality constrained optimization problem

$$\min \ \ f(\mathbf{x}) \text{ subject to } \mathbf{h}(\mathbf{x}) = \mathbf{0},$$

where $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^m$. The quadratic penalty approximation is the unconstrained problem

$$\min \ \ f(\mathbf{x}) + \frac{\pi}{2}\|\mathbf{h}(\mathbf{x})\|^2, \tag{8.4}$$

where the penalty parameter π is a relatively large number. The Hessian of the penalized problem (8.4), evaluated at a local minimizer $\mathbf{x}^*$, is

$$\mathbf{H}(\pi) = \nabla^2 f(\mathbf{x}^*) + \pi \nabla \mathbf{h}(\mathbf{x}^*)^\mathsf{T} \nabla \mathbf{h}(\mathbf{x}^*).$$

If the rank of $\nabla \mathbf{h}(\mathbf{x}^*)$ is m, then by a result of Loewner [68], $\mathbf{H}(\pi)$ has m eigenvalues that approach $+\infty$ as π tends to $+\infty$. Hence, the ratio between the largest and the smallest eigenvalue of $\mathbf{H}(\pi)$ tends to ∞ as π tends to ∞ if $m < n$.

A suitable preconditioner for the penalized problem is presented in [51] where the following result is established:

THEOREM 8.1. *If $\mathbf{C}$ is a symmetric, positive-definite matrix and the rows of $\mathbf{B}$ are linearly independent, then for any $\mathbf{A}$, we have*

$$\lim_{\pi \to \infty} (\mathbf{A} + \pi \mathbf{B}^\mathsf{T}\mathbf{B})(\mathbf{C} + \pi \mathbf{B}^\mathsf{T}\mathbf{B})^{-1} = \mathbf{D},$$

where

$$\mathbf{D} = \mathbf{A}\mathbf{C}^{-1} + (\mathbf{I} - \mathbf{A}\mathbf{C}^{-1})\mathbf{B}^\mathsf{T}(\mathbf{B}\mathbf{C}^{-1}\mathbf{B}^\mathsf{T})^{-1}\mathbf{B}\mathbf{C}^{-1}.$$

Moreover, $\mathbf{D}$ is nonsingular if and only if

$$\min_{\substack{\mathbf{B}\mathbf{y} = \mathbf{0} \\ \|\mathbf{y}\| = 1}} \ \max_{\substack{\mathbf{B}\mathbf{x} = \mathbf{0} \\ \|\mathbf{x}\| = 1}} \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{y} > 0.$$

Thus, if $\mathbf{B}$ is an m by n rank m matrix, the matrix $\mathbf{A} + \pi\mathbf{B}^\mathsf{T}\mathbf{B}$ has m eigenvalues that tend to ∞ as π tends to ∞. For the preconditioner $\mathbf{P} = (\mathbf{C} + \pi\mathbf{B}^\mathsf{T}\mathbf{B})^{-1}$, the

eigenvalues of the product $(\mathbf{A}+\pi\mathbf{B}^\mathsf{T}\mathbf{B})(\mathbf{C}+\pi\mathbf{B}^\mathsf{T}\mathbf{B})^{-1}$ approach a finite limit as π tends to ∞. Consequently, when the rows of $\nabla\mathbf{h}(\mathbf{x}^*)$ are linearly independent and the second order sufficient optimality conditions hold at $\mathbf{x}^*$, a suitable preconditioner for CG methods applied to (8.4) is $\mathbf{P}=(\mathbf{I}+\pi\nabla\mathbf{h}(\mathbf{x}_0)\nabla\mathbf{h}(\mathbf{x}_0))^{-1}$. In practice, we would periodically update the preconditioner and restart the CG method as the iterations converge.

A possible CG preconditioning strategy for a general nonlinear function f is to take $\mathbf{P}=\mathbf{D}_k$ where $\mathbf{D}_k$ is an approximation to $\nabla^2 f(\mathbf{x}^*)^{-1}$ generated by a quasi-Newton update formula, such as the Broyden family:

$$\mathbf{D}_{k+1}=\left(\mathbf{I}-\frac{\mathbf{s}_k\mathbf{y}_k^\mathsf{T}}{\mathbf{s}_k^\mathsf{T}\mathbf{y}_k}\right)\mathbf{D}_k\left(\mathbf{I}-\frac{\mathbf{y}_k\mathbf{s}_k^\mathsf{T}}{\mathbf{s}_k^\mathsf{T}\mathbf{y}_k}\right)+\frac{\mathbf{s}_k\mathbf{s}_k^\mathsf{T}}{\mathbf{s}_k^\mathsf{T}\mathbf{y}_k}+\alpha\mathbf{v}_k\mathbf{v}_k^\mathsf{T} \tag{8.5}$$

where $\alpha\geq 0$ is a parameter, and

$$\mathbf{v}_k=(\mathbf{y}_k^\mathsf{T}\mathbf{D}_k\mathbf{y}_k)^{1/2}\left(\frac{\mathbf{D}_k\mathbf{y}_k}{\mathbf{y}_k^\mathsf{T}\mathbf{D}_k\mathbf{y}_k}-\frac{\mathbf{s}_k}{\mathbf{s}_k^\mathsf{T}\mathbf{y}_k}\right).$$

This idea was first discussed by Nazareth [74] and Buckley [8]. In [74], Nazareth showed that when the objective function is quadratic and an exact line search is employed, preconditioned CG with a fixed preconditioner $\mathbf{P}=\mathbf{D}_0$ is identical to preconditioned CG with $\mathbf{P}=\mathbf{D}_k$ at iteration k provided $\mathbf{D}_k$ is generated by the BFGS formula (corresponding to $\alpha=0$ in (8.5)). Moreover, Buckley shows [8] that if the quasi-Newton preconditioner $\mathbf{D}_k$ is randomly updated by the BFGS formula, the iterates are identical to preconditioned CG with fixed preconditioner $\mathbf{P}=\mathbf{D}_0$. Although there would appear to be no benefit from utilizing a preconditioner generated by a quasi-Newton update, at least in the special case of BFGS and a quadratic cost function, it is expected that for inexact arithmetic or for a general nonlinear function, the quasi-Newton preconditioner will improve the problem conditioning.

In [9] Buckley considers infrequent quasi-Newton updates. A quasi-Newton step was performed and the preconditioner was updated when

$$\left|\frac{\mathbf{g}_k^\mathsf{T}\mathbf{P}\mathbf{g}_{k+1}}{\mathbf{g}_k\mathbf{P}\mathbf{g}_k}\right|\leq\rho,$$

where $\rho\in(0,1)$ is a constant. With infrequent quasi-Newton updates, he could store the vectors used to generate the quasi-Newton matrix rather than the matrix itself. Buckley reports [9] that these infrequent updates led to improvements over the unpreconditioned CG. Another general preconditioning strategy is to use the matrix generated from a limited memory update such as Liu and Nocedal's L-BFGS formula [66]. For a nice survey concerning the relationship between preconditioned CG and quasi-Newton methods, see [75]. The development of effective ways to precondition optimization problems remains an area of interest.

9. Conclusion. The conjugate gradient method has been the subject of intense analysis for more than 50 years. It started out as an algorithm for solving symmetric, positive-definite linear systems of equations. It was soon extended to nonlinear unconstrained optimization. As seen Table 1.1, various choices for the CG update parameter have been proposed. Problems with the early choices concern jamming or loss of descent or convergence failure. Recent choices, such as that of Dai and Yuan provides descent for a Wolfe line search, while the scheme of Hager and Zhang

provides sufficient descent independent of the line search. Hybrid schemes exploit advantages of several different methods, leading to excellent performance in practice. The development of efficient CG preconditioners is an area of active research. Today, the conjugate gradient method is an important component of general constrained optimization algorithms that are based on the iterative solution of unconstrained or bounded constrained problems ([15, 52, 55]).

REFERENCES

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

[2] P. Baptist and J. Stoer, *On the relation between quadratic termination and convergence properties of minimization algorithms, Part II, Applications*, Numer. Math., 28 (1977), pp. 367–392.

[3] E. M. L. Beale, *On an iterative method of finding a local minimum of a function of more than one variable*, Tech. Rep. No. 25, Statistical Techniques Research Group, Princeton Univ., Princeton, N. J., 1958.

[4] E. M. L. Beale, *A derivative of conjugate gradients*, in Numerical Methods for Nonlinear Optimization, F. A. Lootsma, ed., Academic Press, London, 1972, pp. 39–43.

[5] M. C. Biggs, *Minimization algorithms making use of non-quadratic properties of the objective function,* J. Inst. Math. Appl., 8 (1971), pp. 315–327.

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

[7] C. G. Broyden, *The convergence of a class of double-rank minimization algorithms 1. general considerations*, J. Inst. Math. Appl., 6 (1970), pp. 76–90.

[8] A. G. Buckley, *Extending the relationship between the conjugate gradient and BFGS algorithms*, Math. Prog., 15 (1978), pp. 343–348.

[9] A. G. Buckley, *A combined conjugate-gradient quasi-Newton minimization algorithm*, Math. Prog., 15 (1978), pp. 200–210.

[10] A. Buckley, *Conjugate gradient methods*, in Nonlinear Optimization 1981, M. J. D. Powell, ed., Academic Press, London, 1982, pp. 17–22.

[11] R. H. Byrd, J. Nocedal, and Y. Yuan, *Global convergence of a class of quasi-Newton methods on convex problems*, SIAM J. Numer. Anal., 24 (1987), pp. 1171–1190.

[12] R. H. Byrd and J. Nocedal, *A Tool for the Analysis of Quasi-Newton Methods with Application To Unconstrained Minimization*, SIAM J. Numer. Anal., 26 (1989), pp. 727–739.

[13] A. Cauchy, *Méthode générale pour la résolution des systèms d'équations simulatnées*, Comptes Rendus de L'Acadmia Des Sciences, 25 (1847), pp. 536–538.

[14] A. Cohen, *Rate of convergence of several conjugate gradient algorithms*, SIAM J. Numer. Anal., 9 (1972), pp. 248–259.

[15] A. R. Conn, N. I. M. Gould and Ph. L. Toint, *LANCELOT: a Fortran package for large-scale nonlinear optimization*, Springer Series in Computational Mathematics, Vol. 17, Springer Verlag, Heidelberg, 1992.

[16] H. P. Crowder and P. Wolfe, *Linear convergence of the conjugate gradient method*, IBM J. Res. Dev., 16 (1969), pp. 431–433.

[17] Y. H. Dai, *Analyses of conjugate gradient methods*, Ph.D. thesis, Institute of Computational Mathematics and Scientific/Engineering Computing, Chinese Academy of Sciences, 1997.

[18] Y. H. Dai, *Further insight into the convergence of the Fletcher-Reeves method*, Sci. China Ser. A, 42 (1999), pp. 905–916.

[19] Y. H. Dai, *New properties of a nonlinear conjugate gradient method*, Numer. Math., 89 (2001), pp. 83–98.

[20] Y. H. Dai, *A nonmonotone conjugate gradient algorithm for unconstrained optimization*, J. Syst. Sci. Complex., 15 (2002), pp. 139–145.

[21] 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 J. Optim., 10 (1999), pp. 345–358.

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

[23] Y. H. Dai and Q. Ni, *Testing different conjugate gradient methods for large-scale unconstrained optimization*, J. Comput. Math., 21 (2003), pp. 311–320.

[24] Y. H. Dai and Y. Yuan, *Convergence properties of the Fletcher-Reeves method*, IMA J.

Numer. Anal., 16 (1996), pp. 155–164.
[25] Y. H. Dai and Y. Yuan, *Convergence properties of the conjugate descent method*, Adv. Math. (China), 26 (1996), pp. 552–562.
[26] Y. H. Dai and Y. Yuan, *Further studies on the Polak-Ribière-Polyak method*, Research report ICM-95-040, Institute of Computational Mathematics and Scientific/Engineering Computing, Chinese Academy of Sciences, 1995.
[27] Y. H. Dai and Y. Yuan, *A nonlinear conjugate gradient method with a strong global convergence property*, SIAM J. Optim., 10 (1999), pp. 177–182.
[28] Y. H. Dai and Y. Yuan, *Global convergence of the method of shortest residuals*, Numer. Math., 83 (1999), pp. 581–598.
[29] Y. H. Dai and Y. Yuan, *Convergence of the Fletcher-Reeves method under a generalized Wolfe wearch*, J. Comput. Math., 2 (1996), pp. 142–148.
[30] Y. H. Dai and Y. Yuan, *Convergence properties of the Beale-Powell restart algorithm*, Sci. China Ser. A, 41 (1998), pp. 1142–1150.
[31] Y. H. Dai and Y. Yuan, *Some properties of a new conjugate gradient method*, in Advances in Nonlinear Programming, Y. Yuan ed., Kluwer Publications, Boston, 1998, pp. 251–262.
[32] Y. H. Dai and Y. Yuan, *A note on the nonlinear conjugate gradient method*, J. Comput. Math., 20 (2002), pp. 575–582.
[33] 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.
[34] Y. H. Dai and Y. Yuan, *Extension of a class of nonlinear conjugate gradient methods*, Research report ICM-98-049, Institute of Computational Mathematics and Scientific/Engineering Computing, Chinese Academy of Sciences, 1998.
[35] Y. H. Dai and Y. Yuan, *Nonlinear Conjugate Gradient Methods*, Shanghai Science and Technology Publisher, Shanghai, 2000.
[36] Y. H. Dai and Y. Yuan, *A three-parameter family of hybrid conjugate gradient method*, Math. Comp., 70 (2001), pp. 1155–1167.
[37] Y. H. Dai and Y. Yuan, *An efficient hybrid conjugate gradient method for unconstrained optimization*, Ann. Oper. Res., 103 (2001), pp. 33–47.
[38] Y. H. Dai and Y. Yuan, *A class of globally convergent conjugate gradient methods*, Sci. China Ser. A, 46 (2003), pp. 251–261.
[39] J. W. Daniel, The conjugate gradient method for linear and nonlinear operator equations, SIAM J. Numer. Anal., 4 (1967), pp. 10–26.
[40] J. W. Daniel, A correction concerning the convergence rate for the conjugate gradient method, SIAM J. Numer. Anal., 7 (1970), pp. 277–280.
[41] N. Deng and Z. Li, Global convergence of three terms conjugate gradient methods. Optim. Methods Softw., 4 (1995), pp. 273–282.
[42] R. Fletcher, *Function minimization without evaluating derivatives–A review*, Comput. J., 8 (1965), pp. 33–41.
[43] R. Fletcher, *A FORTRAN subroutine for minimization by the method of conjugate gradients*, Atomic Energy Research Establishment, Harwell, Oxfordshire, England, Report No. R-7073, 1972.
[44] R. Fletcher, *Practical Methods of Optimization vol. 1: Unconstrained Optimization*, John Wiley & Sons, New York, 1987.
[45] R. Fletcher and C. Reeves, *Function minimization by conjugate gradients*, Comput. J., 7 (1964), pp. 149–154.
[46] R. W. Freund, G. H. Golub, and N. M. Nachtigal, *Iterative solution of linear systems*, Acta Numerica, (1991), pp. 57–100.
[47] J. C. Gilbert and J. Nocedal, *Global convergence properties of conjugate gradient methods for optimization*, SIAM J. Optim., 2 (1992), pp. 21–42.
[48] G. H. Golub and D. P. O'Leary, *Some history of the conjugate gradient methods and Lanczos algorithms: 1948 - 1976*, SIAM Rev., 31 (1989), pp. 50–100.
[49] L. Grippo and S. Lucidi, *A globally convergent version of the Polak-Ribiére conjugate gradient method*, Math. Prog., 78 (1997), pp. 375–391.
[50] L. Grippo and S. Lucidi, *Convergence conditions, line search algorithms and trust region implementations for the Polak-Ribiére conjugate gradient method*, Technical Report 25-03, Dipartimento di Informatica e Sistemistica, Università di Roma "La Sapienza", November 2003 (to appear on Optim. Methods Softw.).
[51] W. W. Hager, *Dual techniques for constrained optimization*, J. Optim. Theory Appl., 55 (1987), pp. 37–71.
[52] W. W. Hager, *Analysis and implementation of a dual algorithm for constrained optimization*,

J. Optim. Theory Appl., 79 (1993), pp. 427–462.
[53] W. W. Hager and H. Zhang, *A new conjugate gradient method with guaranteed descent and an efficient line search*, November 17, 2003 (to appear in SIAM J. Optim.).
[54] W. W. Hager and H. Zhang, *CG_DESCENT, a conjugate gradient method with guaranteed descent*, January 15, 2004 (to appear in ACM TOMS).
[55] W. W. Hager and H. Zhang, *A new active set algorithm for box constrained optimization*, July 4, 2005.
[56] J. Y. Han, G. H. Liu, D. F. Sun, and H. X. Yin, *Two fundamental convergence theorems for nonlinear conjugate gradient methods and their applications*, Acta Math. Appl. Sinica, 17 (2001), pp. 38–46.
[57] J. Y. Han, G. H. Liu, and H. X. Yin, *Convergence properties of conjugate gradient methods with strong Wolfe linesearch*, Systems Sci. Math. Sci., 11 (1998), pp. 112–116.
[58] M. R. Hestenes, *Conjugate direction methods in optimization*, Springer-Verlag, New York, 1980.
[59] M. R. Hestenes and E. L. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436.
[60] H. Hirst, *n-step quadratic convergence in the conjugate gradient method*, PhD Dissertation, Department of Mathematics, Pennsylvania State University, State College, PA, 1989.
[61] Y. F. Hu and C. Storey, *Global convergence result for conjugate gradient methods*, J. Optim. Theory Appl., 71 (1991), pp. 399–405.
[62] H. Y. Huang, *Unified approach to quadratically convergent algorithms for function minimization*, J. Optim. Theory Appl., 5 (1970), pp. 405–423.
[63] T. G. Kolda, D. P. O'Leary, and L. Nazareth, *BFGS with update skipping and varying memory*, SIAM J. Optim., 8 (1998), pp. 1060–1083.
[64] C. Lemaréchal, *A view of line searches*, in Optimization and Optimal Control, A. Auslander, W. Oettli, and J. Stoer, eds., Lecture Notes in Control and Information Science, Vol. 30, Springer-Verlag, Berlin, 1981, pp. 59–78.
[65] G. H. Liu, J. Y. Han and H. X. Yin, *Global convergence of the Fletcher-Reeves algorithm with an inexact line search*, Appl. Math. J. Chinese Univ. Ser. B, 10 (1995), pp. 75–82.
[66] D. C. Liu and J. Nocedal, *On the limited memory BFGS method for large scale optimization*, Math. Prog., 45 (1989), pp. 503–528.
[67] Y. Liu and C. Storey, *Efficient generalized conjugate gradient algorithms, Part 1: Theory*, J. Optim. Theory Appl., 69 (1991), pp. 129–137.
[68] C. Loewner, *Über monotone Matrixfunctionen*, Math. Zeir., 38 (1934), pp. 177–216.
[69] A. Miele and J. W. Cantrell, *Study on a memory gradient method for the minization of functions*, J. Optim. Theory Appl., 3 (1969), pp. 459–185.
[70] G. P. McCormick and K. Ritter, *Alternative Proofs of the convergence properties of the conjugate-gradient method*, J. Optim. Theory Appl., 13 (1975), pp. 497–518.
[71] M. F. McGuire and P. Wolfe, *Evaluating a restart procedure for conjugate gradients*, Report RC-4382, IBM Research Center, Yorktown Heights, 1973.
[72] J. J. Moré and D. J. Thuente, *Line search algorithms with guaranteed sufficient decrease*, ACM Trans. Math. Software, 20 (1994), pp. 286–307.
[73] J. L. Nazareth, *A conjugate direction algorithm without line searches*, J. Optim. Theory Appl., 23 (1977), pp. 373–387.
[74] J. L. Nazareth, *A relationship between the BFGS and conjugate gradient algorithms and its implications for the new algorithms*, SIAM J. Numer. Anal., 16 (1979), pp. 794–800.
[75] J. L. Nazareth, *Conjugate gradient methods less dependent on conjugacy*, SIAM Review, 28 (1986), pp. 501–511.
[76] J. L. Nazareth, *The Newton-Cauchy framework: A unified Approach to unconstrained nonlinear optimization*, LNCS 769, Springer-Verlags, Berlin, 1994.
[77] J. L. Nazareth, *A view of conjugate gradient-related algorithms for nonlinear optimization*, Proceedings of the AMS-IMS-SIAM Summer Research Conference on *Linear and Nonlinear Conjugate Gradient-Related Methods*, University of Washington, Seattle, WA (July 9–13, 1995).
[78] J. L. Nazareth, *Conjugate-gradient methods*, Encyclopedia of Optimization, C. Floudas and P. Pardalos, eds., Kluwer Academic Publishers, Boston, 1999.
[79] J. Nocedal, *Updating quasi-Newton matrices with limited storage*, Math. Comput., 35 (1980), pp. 773–782.
[80] J. Nocedal, *Theory of Algorithm for Unconstrained Optimization*, Acta Numerica, Cambridge University Press, 1991, pp. 199–242.
[81] J. Nocedal, *Conjugate Gradient Methods and Nonlinear Optimization*, Proceedings of the AMS-IMS-SIAM Summer Research Conference on *Linear and Nonlinear Conjugate*

Gradient-Related Methods, University of Washington, Seattle, WA (July 9–13, 1995).
[82] J. Nocedal, *Large scale unconstrained optimization*, in State of the Art in Numerical Analysis, A. Watson and I. Duff, eds., Oxford University Press, 1997, pp. 311–338.
[83] J. M. Perry, *A class of conjugate gradient algorithms with a two-step variable-metric memory*, Discussion Paper 269, Center for Mathematical Studies in Economics and Management Sciences, Northwestern University, Evanston, Illinois, 1977.
[84] E. Polak and G. Ribière, *Note sur la convergence de directions conjugées*, Rev. Francaise Informat Recherche Opertionelle, 3e Année 16 (1969), pp. 35–43.
[85] B. T. Polyak, *The conjugate gradient method in extreme problems*, USSR Comp. Math. Math. Phys., 9 (1969), pp. 94–112.
[86] M. J. D. Powell, *A hybrid method for nonlinear equations*, In: Numerical Methods for Nonlinear Algebraic Equations, P. Rabinowitz (Ed.), Gordon & Breach, New York, 1970, pp. 87–114.
[87] M. J. D. Powell, *Some convergence properties of the conjugate gradient method*, Math. Prog., 11 (1976), pp. 42–49.
[88] M. J. D. Powell, *Restart procedures of the conjugate gradient method*, Math. Prog., 2 (1977), pp. 241–254.
[89] M. J. D. Powell, *Nonconvex minimization calculations and the conjugate gradient method*, Numerical Analysis (Dundee, 1983), Lecture Notes in Mathematics, Vol. 1066, Springer-Verlag, Berlin, 1984, pp. 122–141.
[90] M. J. D. Powell, *Convergence properties of algirithms for nonlinear optimization*, SIAM Review, 28 (1986), pp. 487–500.
[91] R. Pytlak, *On the convergence of conjugate gradient algorithm*, IMA J. Numer. Anal; 14 (1989), pp. 443–460.
[92] K. Ritter, *On the rate of superlinear convergence of a class of variable metric methods*, Numer. Math., 35 (1980), pp. 293–313.
[93] D. F. Shanno, *On the convergence of a new conjugate gradient algorithm*, SIAM J. Numer. Anal., 15 (1978), pp. 1247–1257.
[94] D. F. Shanno, *Conjugate gradient methods with inexact searches*, Math. Oper. Res., 3 (1978), pp. 244–256.
[95] E. L. Stiefel, *Kernel polynomials in linear algebra and their numerical applications*, Nat. Bur. Standards, Appl. Math. Ser., 49 (1958), pp. 1–22.
[96] J. Stoer, *On the relation between quadratic termination and convergence properties of minimization algorithms*, Numer. Math., 28 (1977), pp. 343–366.
[97] J. Sun and J. Zhang, *Global convergence of conjugate gradient methods without line search*, Ann. Oper. Res., 163 (2001), pp. 161–173
[98] D. Touati-Ahmed and C. Storey, *Efficient hybrid conjugate gradient techniques*, J. Optim. Theory Appl., 64 (1990), pp. 379–397.
[99] C. Y. Wang and Y. Z. Zhang, *Global convergence properties of s-related conjugate gradient methods*, Chinese Science Bulletin, 43 (1998), pp. 1959–1965.
Report, Qufu Normal University, 1996.
[100] P. Wolfe, *Convergence conditions for ascent methods*, SIAM Review, 11 (1969), pp. 226–235.
[101] P. Wolfe, *Convergence conditions for ascent methods. II: Some corrections*, SIAM Review, 13 (1971), pp. 185–188.
[102] P. Wolfe P, *A method of conjugate subgradients for minimizing nondifferentiable functions*, Math. Program. Study, 3 (1975), pp. 145–173.
[103] H. Yabe and M. Takano, *Global convergence properties of nonlinear conjugate gradient methods with modified secant condition*, Comput. Optim. Appl., 28, (2004), pp. 203–225.
[104] Y. Yuan, *Analysis on the conjugate gradient method*, Optim. Methods Softw., 2 (1993), pp. 19–29.
[105] Y. Yuan and J. Stoer, *A subspace study on conjugate algorithms*, ZAMM Z. Angew. Math. Mech., 75 (1995), pp. 69–77.
[106] J. Z. Zhang, N. Y. Deng, and L. H. Chen, *new quasi-Newton equation and related methods for unconstrained optimization*, J. Optim. Theory Appl., 102 (1999), pp. 147–167.
[107] J. Z. Zhang and C. X. Xu, *Properties and numerical performance of quasi-Newton methods with modified quasi-Newton equations*, J. Comput. Appl. Math., 137 (2001), pp. 269–278.
[108] G. Zoutendijk, *Nonlinear Programming, Computational Methods*, in Integer and Nonlinear Programming, J. Abadie, ed., North-Holland, Amsterdam, 1970, pp. 37–86.