Computational Approaches in Large-Scale Unconstrained Optimization

Saman Babaie-Kafaki

Abstract As a topic of great significance in nonlinear analysis and mathematical programming, unconstrained optimization is widely and increasingly used in engineering, economics, management, industry and other areas. Unconstrained optimization also arises in reformulation of the constrained optimization problems in which the constraints are replaced by some penalty terms in the objective function. In many big data applications, solving an unconstrained optimization problem with thousands or millions of variables is indispensable. In such situations, methods with the important feature of low memory requirement are helpful tools. Here, we study two families of methods for solving large-scale unconstrained optimization problems: conjugate gradient methods and limited-memory quasi-Newton methods, both of them are structured based on the line search. Convergence properties and numerical behaviors of the methods are discussed. Also, recent advances of the methods are reviewed. Thus, new helpful computational tools are supplied for engineers and mathematicians engaged in solving large-scale unconstrained optimization problems.

Keywords Unconstrained optimization · Large-scale optimization · Line search · Memoryless quasi-Newton method · Conjugate gradient method

1 Introduction

We consider the minimization of a smooth nonlinear function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, that is,

$$\min_{x\in\mathbb{R}^n} f(x), \tag{1}$$

in the case where the number of variables n is large and analytic expressions for the function f and its gradient ∇f are available. Although the minimizer of f is a solution

S. Babaie-Kafaki (✉)
Department of Mathematics, Faculty of Mathematics, Statistics and Computer Science, Semnan University, P.O. Box: 35195-363, Semnan, Iran
e-mail: sbk@semnan.ac.ir

A. Emrouznejad (ed.), *Big Data Optimization: Recent Developments and Challenges*, Studies in Big Data 18, DOI 10.1007/978-3-319-30265-2_17

of the system $\nabla f(x) = 0$, solving this generally nonlinear and complicated system is not practical.

Among the most useful tools for solving large-scale cases of (1) there are the conjugate gradient methods and the limited-memory quasi-Newton methods, because the amount of memory storage required by the methods is low. In addition, the methods possess the attractive features of simple iterative formula and strong global convergence as well as applying the Hessian information. The methods can also be straightly employed in penalty function methods, a class of efficient methods for solving constrained optimization problems.

Generally, iterations of the above-mentioned methods are in the following form:

$$x_0 \in \mathbb{R}^n,\ x_{k+1} = x_k + s_k,\ s_k = \alpha_k d_k,\ k = 0, 1, \ldots, \tag{2}$$

where d_k is a search direction to be computed by a few inner products and α_k is a step length to be determined by a line search procedure. The search direction d_k should be a descent direction, i.e.,

$$g_k^T d_k < 0, \tag{3}$$

where $g_k = \nabla f(x_k)$, to ensure that the function f can be reduced along the search direction d_k. The most reduction is achieved when the exact (optimal) line search is used in which

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

Hence, in the exact line search the step length α_k can be considered as a solution of the following equation:

$$\nabla f(x_k + \alpha d_k)^T d_k = 0. \tag{4}$$

Since the exact line search is not computationally tractable, inexact line search techniques have been developed [74, 86], most of them structured based on quadratic or cubic polynomial interpolations of the one-dimensional function $\varphi(\alpha) = f(x_k + \alpha d_k)$. Finding minimizers of the polynomial approximations of $\varphi(\alpha)$, inexact line search procedures try out a sequence of candidate values for the step length, stopping to accept one of these values when certain conditions are satisfied.

Among the stopping conditions for the inexact line search procedures, the so-called Wolfe conditions [91, 92] have attracted especial attention in convergence analyses and implementations of the unconstrained optimization algorithms, requiring that

$$f(x_k + \alpha_k d_k) - f(x_k) \leq \delta \alpha_k g_k^T d_k, \tag{5}$$

$$\nabla f(x_k + \alpha_k d_k)^T d_k \geq \sigma g_k^T d_k, \tag{6}$$

where $0 < \delta < \sigma < 1$. The first condition, called the Armijo condition, ensures adequate reduction of the objective function value while the second condition, called the curvature condition, ensures unacceptably of the short step lengths. However, a step length may fulfill the Wolfe conditions without being sufficiently close to a minimizer of $\varphi(\alpha)$. To overcome this problem, the strong Wolfe conditions have been proposed which consist of (5) and the following strengthened version of (6):

$$|\nabla f(x_k + \alpha_k d_k)^T d_k| \leq -\sigma g_k^T d_k. \tag{7}$$

Considering (4), if $\sigma \to 0$, then the step length which satisfies the strong Wolfe conditions (5) and (7) tends to the optimal step length.

In practical computations, the Wolfe condition (5) may never be satisfied due to the existence of numerical errors. This computational drawback of the Wolfe conditions was carefully analyzed in [59] on a one-dimensional quadratic function. Based on the insight gained by the numerical example of [59], one of the most accurate and efficient inexact line search algorithms has been proposed in [59, 60], using a quadratic interpolation scheme and the following approximate Wolfe conditions:

$$\sigma g_k^T d_k \leq \nabla f(x_k + \alpha_k d_k)^T d_k \leq (2\delta - 1) g_k^T d_k, \tag{8}$$

where $0 < \delta < \frac{1}{2}$ and $\delta \leq \sigma < 1$. The line search algorithm of [60] has been further improved in [42].

In what follows, at first we discuss several basic choices for the search direction d_k in (2) corresponding to the steepest descent method, Newton method, conjugate direction methods and quasi-Newton methods, together with their advantages and disadvantages as well as their relationships. Then, we focus on the conjugate gradient methods and the limited-memory quasi-Newton methods which are proper algorithms for large-scale unconstrained optimization problems. For all of these methods, the line search procedure of [60] can be applied efficiently. Also, a popular stopping criterion for the iterative method (2) is given by

$$||g_k|| < \varepsilon,$$

in which ε is a small positive constant and $||.||$ stands for the Euclidean norm.

2 Basic Unconstrained Optimization Algorithms

Here, we briefly study basic algorithms in the field of unconstrained optimization, all of them are iterative in the form of (2) with especial choices for the search direction d_k. A detailed discussion can be found in [86].

2.1 Steepest Descent Method

One of the simplest and most fundamental methods for solving the unconstrained optimization problem (1) is the steepest descent (or the gradient) method [39] in which the search direction is computed by

$$d_k = -g_k,$$

that is trivially a descent direction. Although the steepest descent method is globally convergent under a variety of inexact line search conditions, the method performs poorly, converges linearly and is badly affected by ill conditioning [1, 55].

2.2 Newton Method

Based on a quadratic interpolation of the objective function at the kth iteration, search direction of the Newton method can be computed by

$$d_k = -\nabla^2 f(x_k)^{-1} g_k,$$

where $\nabla^2 f$ is the Hessian matrix of the objective function f. If $\nabla^2 f(x_k)$ is a positive definite matrix, then the Newton search direction is a descent direction and in such situation, it can be effectively computed by solving the following linear system using Cholesky decomposition [88]:

$$\nabla^2 f(x_k) d_k = -g_k.$$

In the Newton method, the Hessian information is employed in addition to the gradient information. Also, if the starting point x_0 is adequately close to the optimal point x^*, then the sequence $\{x_k\}_{k\geq 0}$ generated by the Newton method converges to x^* with a quadratic rate. However, since in the Newton method it is necessary to compute and save the Hessian matrix $\nabla^2 f(x_k) \in \mathbb{R}^{n\times n}$, the method is not proper for large-scale problems. Moreover, far from the solution, the Hessian $\nabla^2 f(x_k)$ may not be a positive definite matrix and consequently, the Newton search direction may not be a descent direction. To overcome this problem, a variant of modified Newton methods have been proposed in the literature [86].

2.3 Conjugate Direction Methods

Consider the problem of minimizing a strictly convex quadratic function, i.e.,

$$\min_{x\in\mathbb{R}^n} q(x), \tag{9}$$

in which

$$q(x) = \frac{1}{2}x^T Ax - b^T x, \tag{10}$$

where the Hessian $A \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix and $b \in \mathbb{R}^n$. To find the optimal solution x^*, the following system of linear equations can be solved:

$$\nabla q(x) = Ax - b = 0, \tag{11}$$

or equivalently,

$$Ax = b.$$

Although the problem can be solved by Cholesky decomposition, conjugate direction methods are a class of efficient algorithms for finding the minimizer of a strictly convex quadratic function in large-scale cases.

Definition 1 Let $A \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix and $\{d_k\}_{k=1}^m$, $m \leq n$, be a set of nonzero vectors in $\mathbb{R}^n$. If

$$d_i^T Ad_j = 0, \ \forall i \neq j,$$

then the vectors $\{d_k\}_{k=1}^m$ are called A-conjugate, or simply called conjugate.

Exercise 1 *(i) Show that a set of conjugate vectors are linearly independent.*

(ii) Assume that a symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$ and a set of linearly independent vectors $\{d'_k\}_{k=1}^m \subseteq \mathbb{R}^n$ are available. Describe how a set of A-conjugate vectors $\{d_k\}_{k=1}^m$ can be constructed from $\{d'_k\}_{k=1}^m$.

(Hint: Use the Gram-Schmidt orthogonalization scheme [88].)

In each iteration of a conjugate direction method for solving (9), the function $q(x)$ given by (10) is minimized along the search direction d_k for which we have

$$d_k^T Ad_i = 0, \ \forall i < k.$$

Here, since the objective function is quadratic, the exact line search can be used. The following theorem shows that under the exact line search, the conjugate direction methods have quadratic termination property which means that the methods terminate in at most n steps when they are applied to a strictly convex quadratic function.

Theorem 1 *For a quadratic function with the positive definite Hessian A, the conjugate direction method terminates in at most n exact line searches. Also, each x_{k+1} is the minimizer in subspace $\mathscr{S}_k$ generated by x_0 and the directions $\{d_i\}_{i=0}^k$, that is, $\mathscr{S}_k = x_0 + span\{d_0, \dots, d_k\}$.*

Exercise 2 *Prove Theorem 1.*
(Hint: By induction show that $\nabla q(x_{k+1}) \perp d_i, \ i = 0, 1, \dots, k$.)

2.4 Quasi-Newton Methods

As known, quasi-Newton methods are of particular performance for solving unconstrained optimization problems since they do not require explicit expressions of the second derivatives and their convergence rate is often superlinear [86]. The methods are sometimes referred to variable metric methods.

In the quasi-Newton methods, the search direction is often calculated by

$$d_k = -H_k g_k, \tag{12}$$

in which $H_k \in \mathbb{R}^{n\times n}$ is an approximation of the inverse Hessian; more precisely, $H_k \approx \nabla^2 f(x_k)^{-1}$. The methods are characterized by the fact that H_k is effectively updated to achieve a new matrix H_{k+1} as an approximation of $\nabla^2 f(x_{k+1})^{-1}$, in the following general form:

$$H_{k+1} = H_k + \varDelta H_k,$$

where $\varDelta H_k$ is a correction matrix. The matrix H_{k+1} is imposed with the scope of satisfying a particular equation, namely secant (quasi-Newton) equation, which includes the second order information. The most popular equation is the standard secant equation, that is,

$$H_{k+1} y_k = s_k, \tag{13}$$

in which $y_k = g_{k+1} - g_k$. Note that the standard secant equation is obtained based on the mean-value theorem, or equivalently, the following approximation:

$$\nabla^2 f(x_{k+1}) s_k \approx y_k,$$

which holds exactly for the quadratic objective functions.

Among the well-known quasi-Newton update formulas there are the BFGS (Broyden-Fletcher-Goldfarb-Shanno) and DFP (Davidon-Fletcher-Powell) updates [86] given by

$$H_{k+1}^{BFGS} = H_k - \frac{s_k y_k^T H_k + H_k y_k s_k^T}{s_k^T y_k} + \left(1 + \frac{y_k^T H_k y_k}{s_k^T y_k}\right)\frac{s_k s_k^T}{s_k^T y_k}, \tag{14}$$

and

$$H_{k+1}^{DFP} = H_k + \frac{s_k s_k^T}{s_k^T y_k} - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k},$$

in which the initial approximation H_0 can be considered as an arbitrary positive definite matrix. In a generalization scheme, the BFGS and DFP updates have been combined linearly and the Broyden class of quasi-Newton update formulas [86] has been proposed as follows:

$$\begin{aligned} H^{\phi}_{k+1} &= (1-\phi)H^{DFP}_{k+1} + \phi H^{BFGS}_{k+1} \\ &= H_k + \frac{s_k s_k^T}{s_k^T y_k} - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} + \phi v_k v_k^T, \end{aligned} \tag{15}$$

in which ϕ is a real parameter and

$$v_k = \sqrt{y_k^T H_k y_k}\left(\frac{s_k}{s_k^T y_k} - \frac{H_k y_k}{y_k^T H_k y_k}\right). \tag{16}$$

It can be seen that if H_k is a positive definite matrix and the line search ensures that $s_k^T y_k > 0$, then H^{ϕ}_{k+1} with $\phi \geq 0$ is also a positive definite matrix [86] and consequently, the search direction $d_{k+1} = -H^{\phi}_{k+1} g_{k+1}$ is a descent direction. Moreover, for a strictly convex quadratic objective function, search directions of a quasi-Newton method with the update formulas of the Broyden class are conjugate directions. So, in this situation the method possesses the quadratic termination property. Also, under convexity assumption on the objective function and when $\phi \in [0, 1]$, it has been shown that the method is globally and locally superlinearly convergent [86]. It is worth noting that among the quasi-Newton update formulas of the Broyden class, the BFGS update is superior with respect to the computational performance. A nice survey on the quasi-Newton methods has been provided in [93].

Similar to the quasi-Newton approximations $\{H_k\}_{k\geq 0}$ for the inverse Hessian satisfying (13), quasi-Newton approximations $\{B_k\}_{k\geq 0}$ for the Hessian can be proposed for which the following equivalent version of the standard secant equation (13) should be satisfied:

$$B_{k+1} s_k = y_k. \tag{17}$$

In such situation, considering (12), search directions of the quasi-Newton method can be computed by solving the following linear system:

$$B_k d_k = -g_k. \tag{18}$$

Exercise 3 *(i) Prove that if the search direction d_k is a descent direction and the line search fulfills the Wolfe conditions (5) and (6), then $s_k^T y_k > 0$.*

(ii) For the Broyden class of update formulas (15), prove that if H_k is a positive definite matrix, $s_k^T y_k > 0$ and $\phi \geq 0$, then H^{ϕ}_{k+1} is also a positive definite matrix.

Exercise 4 *(i) (Sherman-Morrison Theorem) Let $A \in \mathbb{R}^{n\times n}$ be a nonsingular matrix and $u, v \in \mathbb{R}^n$ be arbitrary vectors. Prove that if $1 + v^T A^{-1} u \neq 0$, then the rank-one update $A + uv^T$ of A is nonsingular, and*

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1} u}.$$

(ii) Compute $H^{DFP^{-1}}_{k+1}$ and find its relationship with H^{BFGS}_{k+1}.

2.4.1 Scaled Quasi-Newton Updates

In order to achieve an ideal distribution of the eigenvalues of quasi-Newton updates of the Broyden class, improving the condition number of successive approximations of the inverse Hessian and consequently, increasing the numerical stability in the iterative method (2), the scaled quasi-Newton updates have been developed [86]. In this context, replacing H_k by $\theta_k H_k$ in (15), where $\theta_k > 0$ is called the scaling parameter, the scaled Broyden class of quasi-Newton updates can be achieved as follows:

$$H_{k+1}^{\phi,\theta_k} = \left(H_k - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} + \phi v_k v_k^T \right) \theta_k + \frac{s_k s_k^T}{s_k^T y_k}, \tag{19}$$

where v_k is defined by (16). The most effective choices for θ_k in (19) have been proposed by Oren and Spedicato [75, 77],

$$\theta_k = \frac{s_k^T y_k}{y_k^T H_k y_k}, \tag{20}$$

and, Oren and Luenberger [75, 76],

$$\theta_k = \frac{s_k^T H_k^{-1} s_k}{s_k^T y_k}. \tag{21}$$

A scaled quasi-Newton update in the form of (19) with one of the parameters (20) or (21) is called a self-scaling quasi-Newton update.

Although the self-scaling quasi-Newton methods are numerically efficient, as an important defect the methods need to save the matrix $H_k \in \mathbb{R}^{n\times n}$ in each iteration, being improper for solving large-scale problems. Hence, in a simple modification in the sense of replacing H_k by the identity matrix in (19), self-scaling memoryless update formulas of the Broyden class have been proposed as follows:

$$\tilde{H}_{k+1}^{\phi,\theta_k} = \left(I - \frac{y_k y_k^T}{y_k^T y_k} + \phi \tilde{v}_k \tilde{v}_k^T \right) \theta_k + \frac{s_k s_k^T}{s_k^T y_k},$$

where

$$\tilde{v}_k = \sqrt{y_k^T y_k} \left(\frac{s_k}{s_k^T y_k} - \frac{y_k}{y_k^T y_k} \right).$$

Similarly, memoryless version of the scaling parameters (20) and (21) can be respectively written as:

$$\theta_k = \frac{s_k^T y_k}{||y_k||^2}, \tag{22}$$

and

$$\theta_k = \frac{||s_k||^2}{s_k^T y_k}. \tag{23}$$

The scaling parameter (23) can also be determined based on a two-point approximation of the standard secant equation (13) [35].

Exercise 5 *(i) Find all the eigenvalues of the scaled memoryless BFGS update formula with the parameters (22) or (23).*

(ii) Assume that ∇f is Lipschitz continuous on a nonempty open convex set $\mathscr{N}$; that is, there exists a positive constant L such that

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y||, \ \forall x, y \in \mathscr{N}. \tag{24}$$

Prove that if the objective function f is uniformly convex, then there exists a positive constant c such that for the sequence $\{x_k\}_{k\geq 0}$ generated by the scaled memoryless BFGS method with the parameter (23) we have

$$g_k^T d_k \leq -c||g_k||^2, \ \forall k \geq 0. \tag{25}$$

(Hint: Note that a differentiable function f is said to be uniformly (or strongly) convex on a nonempty open convex set $\mathscr{S}$ if and only if there exists a positive constant μ such that

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \mu ||x - y||^2, \ \forall x, y \in \mathscr{S} \text{ [86].)}$$

Definition 2 Inequality (25) is called the sufficient descent condition.

2.4.2 Modified Secant Equations

The standard secant equation (13), or its equivalent form (17), only uses the gradient information and ignores the function values. So, efforts have been made to modify the Eq. (17) such that more available information be employed and consequently, better approximations for the (inverse) Hessian be obtained (see [16] and the references therein).

Assume that the objective function f is smooth enough and let $f_k = f(x_k)$, $\forall k \geq 0$. From Taylor's theorem we get

$$f_k = f_{k+1} - s_k^T g_{k+1} + \frac{1}{2} s_k^T \nabla^2 f(x_{k+1}) s_k - \frac{1}{6} s_k^T (T_{k+1} s_k) s_k + O(||s_k||^4), \tag{26}$$

where

$$s_k^T (T_{k+1} s_k) s_k = \sum_{i,j,l=1}^{n} \frac{\partial^3 f(x_{k+1})}{\partial x^i \partial x^j \partial x^l} s_k^i s_k^j s_k^l. \tag{27}$$

So, after some algebraic manipulations it can be seen that

$$s_k^T \nabla^2 f(x_{k+1}) s_k = s_k^T y_k + 2(f_k - f_{k+1}) + s_k^T (g_k + g_{k+1}) + \frac{1}{3} s_k^T (T_{k+1} s_k) s_k + O(||s_k||^4).$$

Hence, the following approximation can be proposed:

$$s_k^T \nabla^2 f(x_{k+1}) s_k \approx s_k^T y_k + \vartheta_k,$$

where

$$\vartheta_k = 2(f_k - f_{k+1}) + s_k^T (g_k + g_{k+1}), \tag{28}$$

which leads to the following modified secant equation [89, 90]:

$$B_{k+1} s_k = z_k, \ z_k = y_k + \frac{\vartheta_k}{s_k^T u_k} u_k, \tag{29}$$

where $u_k \in \mathbb{R}^n$ is a vector parameter satisfying $s_k^T u_k \neq 0$ (see also [98, 99]).

Again, from Taylor's theorem we can write:

$$s_k^T g_k = s_k^T g_{k+1} - s_k^T \nabla^2 f(x_{k+1}) s_k + \frac{1}{2} s_k^T (T_{k+1} s_k) s_k + O(||s_k||^4). \tag{30}$$

Now, considering (26) and (30), by canceling the terms which include tensor we get

$$s_k^T \nabla^2 f(x_{k+1}) s_k = s_k^T y_k + 3\vartheta_k + O(||s_k||^4),$$

where ϑ_k is defined by (28). Hence, the following secant equation can be proposed [100]:

$$B_{k+1} s_k = w_k, \ w_k = y_k + \frac{3\vartheta_k}{s_k^T u_k} u_k, \tag{31}$$

where $u_k \in \mathbb{R}^n$ is a vector parameter satisfying $s_k^T u_k \neq 0$.

For a quadratic objective function f, we have $\vartheta_k = 0$, and consequently, the modified secant equations (29) and (31) reduce to the standard secant equation. For the vector parameter u_k, we can simply let $u_k = s_k$, or $u_k = y_k$ provided that the line search fulfills the Wolfe conditions. To guarantee positive definiteness of the successive quasi-Newton approximations for the (inverse) Hessian obtained based on the modified secant equations (29) and (31) we should respectively have $s_k^T z_k > 0$ and $s_k^T w_k > 0$ which may not be necessarily satisfied for general functions. To overcome this problem, in a simple modification we can replace ϑ_k in (29) and (31) by $\max\{\vartheta_k, 0\}$. The modified secant equations (29) and (31) are justified by the following theorem [89, 100, 104], demonstrating their accuracy in contrast to the standard secant equation (17).

Theorem 1 *If f is sufficiently smooth and $||s_k||$ is small enough, then the following estimating relations hold:*

$$
\begin{aligned}
s_k^T(\nabla^2 f(x_{k+1})s_k - y_k) &= \frac{1}{2}s_k^T(T_{k+1}s_k)s_k + O(||s_k||^4),\\
s_k^T(\nabla^2 f(x_{k+1})s_k - z_k) &= \frac{1}{3}s_k^T(T_{k+1}s_k)s_k + O(||s_k||^4),\\
s_k^T(\nabla^2 f(x_{k+1})s_k - w_k) &= O(||s_k||^4),
\end{aligned}
$$

where T_{k+1} is defined by (27).

Convexity assumption on the objective function plays an important role in convergence analysis of the quasi-Newton methods with secant equations (17), (29) and (31). However, in [64] a modified BFGS method has been proposed which is globally and locally superlinearly convergent for nonconvex objective functions (see also [58, 65]). The method has been designed based on the following modified secant equation:

$$B_{k+1}s_k = \bar{y}_k,\ \bar{y}_k = y_k + h_k||g_k||^r s_k, \tag{32}$$

where r is a positive constant and $h_k > 0$ is defined by

$$h_k = C + \max\{-\frac{s_k^T y_k}{||s_k||^2}, 0\}||g_k||^{-r},$$

with some positive constant C. As an interesting property, for the modified secant equation (32) we have $s_k^T\bar{y}_k > 0$, independent of the line search conditions and the objective function convexity, which guarantees heredity of positive definiteness for the related BFGS updates. Recently, scaled memoryless BFGS methods have been proposed based on the modified secant equations (29), (31) and (32) which possess the sufficient descent property (25) [17–19, 22, 28]. In addition to the modified secant equations (29), (31) and (32) which apply information of the current iteration, the multi-step secant equations have been developed by Ford et al. [51–53] based on the polynomial interpolation using available data from the m recent steps.

3 Conjugate Gradient Methods

Conjugate gradient methods comprise a class of algorithms which are between the steepest descent method and the Newton method. Utilizing the Hessian information implicitly, the methods deflect the steepest descent direction by adding to it a multiple of the direction used in the last step, that is,

$$d_{k+1} = -g_{k+1} + \beta_k d_k,\ k = 0, 1, \ldots, \tag{33}$$

with $d_0 = -g_0$, where β_k is a scalar called the conjugate gradient (update) parameter. Although the methods only require the first-order derivatives, they overcome the slow convergence of the steepest descent method. Also, the methods need not to save and compute the second-order derivatives which are needed in the Newton method. Hence, they are widely used to solve large-scale optimization problems.

Different conjugate gradient methods mainly correspond to different choices for the conjugate gradient parameter [61]. Although the conjugate gradient methods are equivalent in the linear case, that is, when f is a strictly convex quadratic function and the line search is exact, their behavior for general functions may be quite different [3, 45, 82]. It is worth noting that search directions of the linear conjugate gradient methods are conjugate directions. In what follows, we deal with several essential conjugate gradient methods.

3.1 The Hestenes-Stiefel Method

Conjugate gradient methods were originally developed in the 1950s by Hestenes and Stiefel [62] (HS) as an alternative to factorization methods for solving linear systems. Conjugate gradient parameter of the HS method is given by

$$\beta_k^{HS} = \frac{g_{k+1}^T y_k}{d_k^T y_k}.$$

From the mean-value theorem, there exists some $\xi \in (0, 1)$ such that

$$d_{k+1}^T y_k = d_{k+1}^T (g_{k+1} - g_k) = \alpha_k d_{k+1}^T \nabla^2 f(x_k + \xi \alpha_k d_k) d_k.$$

Hence, the condition

$$d_{k+1}^T y_k = 0, \tag{34}$$

can be considered as a conjugacy condition for the nonlinear objective functions since it shows that search directions d_k and d_{k+1} are conjugate directions. As an attractive feature, considering (33) it can be seen that search directions of the HS method satisfy the conjugacy condition (34), independent of the line search and the objective function convexity.

Perry [78] noted that the search direction d_{k+1} of the HS method can be written as:

$$d_{k+1} = -\left(I - \frac{s_k y_k^T}{s_k^T y_k}\right) g_{k+1}. \tag{35}$$

Then, he made a modification on the search direction (35) as follows:

$$d_{k+1} = -\underbrace{\left(I - \frac{s_k y_k^T}{s_k^T y_k} + \frac{s_k s_k^T}{s_k^T y_k}\right)}_{P_{k+1}} g_{k+1} = -P_{k+1} g_{k+1}.$$

Perry justified the addition of the correction term $\frac{s_k s_k^T}{s_k^T y_k}$ by noting that the matrix P_{k+1} satisfies the following equation:

$$y_k^T P_{k+1} = s_k^T,$$

which is similar, but not identical, to the standard secant equation (13). To improve Perry's approach, Shanno [84] modified the matrix P_{k+1} as follows:

$$P_{k+1}^S = I - \frac{s_k y_k^T + y_k s_k^T}{s_k^T y_k} + \left(1 + \frac{y_k^T y_k}{s_k^T y_k}\right) \frac{s_k s_k^T}{s_k^T y_k}.$$

Thus, the related conjugate gradient method is precisely the BFGS method in which the approximation of the inverse Hessian is restarted as the identity matrix at every step and so, no significant storage is used to develop a better approximation for the inverse Hessian. Hence, the HS method can be extended to the memoryless BFGS method. This idea was also discussed by Nazareth [71] and Buckley [38]. A nice survey concerning the relationship between conjugate gradient methods and the quasi-Newton methods has been provided in [72].

Although the HS method is numerically efficient, its search directions generally fail to satisfy the descent condition (3), even for strictly convex objective functions [40]. It is worth noting that when in an iteration of a conjugate gradient method the search direction does not satisfy the descent condition (3), i.e., when encountering with an uphill search direction, the steepest descent direction can be used. This popular scheme for the conjugate gradient methods is called the restart procedure. In another approach, Powell [82] suggested to restart the conjugate gradient method if the following inequality is violated:

$$g_k^T g_{k-1} \leq \varsigma ||g_k||^2,$$

where ς is a small positive constant (see also [56]).

As another defect of the HS method that will be discussed in the next parts of this section, it can be stated that the method lacks global convergence in certain circumstances in the sense of cycling infinitely [82].

3.2 The Fletcher-Reeves Method

Since solving a linear system is equivalent to minimizing a quadratic function, in the 1960s Fletcher and Reeves [50] (FR) modified the HS method and developed a conjugate gradient method for unconstrained minimization with the following parameter:

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

Although search directions of the FR method generally are not descent directions, convergence analysis of the method has been appropriately developed. As a brief review, at first Zoutendijk [105] established a convergence result for the FR method under the exact line search. Then, Al-Baali [2] dealt with convergence of the FR method when the line search fulfills the strong Wolfe conditions (5) and (7), with $0 < \delta < \sigma < 1/2$. Liu et al. [68] extended the Al-Baali's result for $\sigma = 1/2$. A comprehensive study on the convergence of the FR method has been made by Gilbert and Nocedal [56]. Notwithstanding the strong convergence properties, numerical performance of the FR method is essentially affected by jamming [56, 81], i.e., generating many short steps without making significant progress to the solution because the search directions became nearly orthogonal to the gradient.

3.3 The Polak-Ribière-Polyak Method

One of the efficient conjugate gradient methods has been proposed by Polak et al. [79, 80] (PRP) where its parameter is computed by

$$\beta_k^{PRP} = \frac{g_{k+1}^T y_k}{||g_k||^2}.$$

It is important that when the iterations jam, the step s_k is small. So, the factor y_k in the numerator of β_k^{PRP} tends to zero and consequently, β_k^{PRP} becomes small. Therefore, the search direction d_{k+1} tends to the steepest descent direction and an automatic restart occurs. This favorable numerical feature of jamming prevention also occurs for the HS method.

In spite of numerical efficiency of the PRP method, the method lacks the descent property. Also, Powell [82] constructed a three-dimensional counter example with the exact line search, demonstrating the method can cycle infinitely without convergence to a solution. Nevertheless, based on the insight gained by his counter example, Powell [82] suggested the following truncation of β_k^{PRP}:

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

which yields a globally convergent conjugate gradient method [56], being also computationally efficient [3].

Since under the exact line search the PRP and the HS methods are equivalent, the cycling phenomenon may occur for the HS method. The following truncation of β_k^{HS} has been shown to lead to a globally convergent conjugate gradient method [43, 56]:

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

which is also more efficient than the HS method [3].

3.4 The Dai-Yuan Method

Another essential conjugate gradient method has been proposed by Dai and Yuan [47] (DY) with the following parameter:

$$\beta_k^{DY} = \frac{||g_{k+1}||^2}{d_k^T y_k}.$$

It is notable that under mild assumptions on the objective function, the DY method has been shown to be globally convergent under a variety of inexact line search conditions. Also, in addition to the generation of descent search directions when $d_k^T y_k > 0$, as guaranteed by the Wolfe conditions (5) and (6), the DY method has been proved to have a certain self-adjusting property, independent of the line search and the objective function convexity [41]. More exactly, if there exist positive constants γ_1 and γ_2 such that $\gamma_1 \leq ||g_k|| \leq \gamma_2$, for all $k \geq 0$, then, for any $p \in (0, 1)$, there exists a positive constant c such that the sufficient descent condition $g_i^T d_i \leq -c||g_i||^2$ holds for at least $\lfloor pk \rfloor$ indices $i \in [0, k]$, where $\lfloor j \rfloor$ denotes the largest integer less than or equal to j. However, similar to the FR method, in spite of strong theoretical properties the DY method has a poor computational performance due to the jamming phenomenon.

3.5 The Dai-Liao Method

In order to employ quasi-Newton aspects in the conjugacy condition (34), Dai and Liao [43] (DL) noted that considering $B_{k+1} \in \mathbb{R}^{n \times n}$ as an approximation of $\nabla^2 f(x_{k+1})$ given by a quasi-Newton method, from the standard secant equation (17) and the linear system (18) we can write

$$d_{k+1}^T y_k = d_{k+1}^T (B_{k+1} s_k) = -g_{k+1}^T s_k. \tag{36}$$

If the line search is exact, then $g_{k+1}^T s_k = 0$, and consequently (36) reduces to (34). However, in practice the algorithms normally adopt inexact line searches. Hence, the following extension of the conjugacy condition (34) has been proposed in [43]:

$$d_{k+1}^T y_k = -t g_{k+1}^T s_k, \tag{37}$$

where t is a nonnegative parameter. If $t = 0$ or the line search is exact, then (37) reduces to (34), and if $t = 1$, then (37) reduces to (36) which implicitly contains the effective standard secant equation (17). Also, for small values of t, the conjugacy condition (37) tends to the conjugacy condition (34). Thus, the conjugacy condition (37) can be regarded as a generalization of the conjugacy conditions (34) and (36).

Taking inner product of (33) with y_k and using (37), Dai and Liao [43] obtained the following formula for the conjugate gradient parameter:

$$\beta_k^{DL} = \frac{g_{k+1}^T y_k}{d_k^T y_k} - t \frac{g_{k+1}^T s_k}{d_k^T y_k}, \tag{38}$$

shown to be globally convergent for uniformly convex objective functions. Theoretical and numerical features of the DL method is very dependent on the parameter t for which there is no any optimal choice [15]. It is worth noting that if

$$t = 2\frac{||y_k||^2}{s_k^T y_k}, \tag{39}$$

then the conjugate gradient parameter proposed by Hager and Zhang [59] is achieved. Also, the choice

$$t = \frac{||y_k||^2}{s_k^T y_k}, \tag{40}$$

yields another conjugate gradient parameter suggested by Dai and Kou [42]. The choices (39) and (40) are effective since they guarantee the sufficient descent condition (25), independent of the line search and the objective function convexity, and lead to numerically efficient conjugate gradient methods [21, 42, 60]. Recently, Babaie-Kafaki and Ghanbari [25, 27, 32] dealt with other proper choices for the parameter t in the DL method.

Based on Powell's approach of nonnegative restriction of the conjugate gradient parameters [82], Dai and Liao proposed the following modified version of β_k^{DL}:

$$\beta_k^{DL+} = \beta_k^{HS+} - t \frac{g_{k+1}^T s_k}{d_k^T y_k},$$

and showed that the DL+ method is globally convergent for general objective functions [43]. In several other attempts to make modifications on the DL method, modified secant equations have been applied in the Dai-Liao approach. In this context, in order to employ the objective function values in addition to the gradient information, Yabe and Takano [94] used the modified secant equation (31). Also, Li et al. [66] used the modified secant equation (29). Babaie-Kafaki et al. [33] applied a revised form of the modified secant equation (31), and the modified secant equation proposed in [98]. Ford et al. [54] employed the multi-step quasi-Newton equations proposed by Ford and Moghrabi [51]. In another attempt to achieve global convergence without convexity assumption on the objective function, Zhou and Zhang [104] applied the modified secant equation (32).

Exercise 6 *For the DL method, assume that* $s_k^T y_k > 0$ *and* $t > 0$.

(i) *Find the matrix* Q_{k+1} *for which search directions of the DL method can be written as* $d_{k+1} = -Q_{k+1} g_{k+1}$. *The matrix* Q_{k+1} *is called the search direction matrix.*
(ii) *Find all the eigenvalues of the matrix* $A_{k+1} = \dfrac{Q_{k+1}^T + Q_{k+1}}{2}$.
(iii) *Prove that if*

$$t > \frac{1}{4}\left(\frac{||y_k||^2}{s_k^T y_k} - \frac{s_k^T y_k}{||s_k||^2}\right),$$

then search directions of the DL method satisfy the descent condition (3).

Exercise 7 *For the DL method, assume that* $s_k^T y_k > 0$ *and* $t > 0$.

(i) *Prove that the search direction matrix* Q_{k+1} *is nonsingular. Then, find the inverse of* Q_{k+1}.
(ii) *Find* $||Q_{k+1}||_F^2$ *and* $||Q_{k+1}^{-1}||_F^2$, *where* $||.||_F$ *stands for the Frobenius norm.*
(iii) *Prove that if* $n \to \infty$, *then* $t^* = \sqrt{\dfrac{||y_k||(s_k^T y_k)}{||s_k||^3}}$ *is the minimizer of* $\kappa_F(Q_{k+1}) = ||Q_{k+1}||_F ||Q_{k+1}^{-1}||_F$.

3.6 The CG-Descent Algorithm

In an attempt to make a modification of the HS method in order to achieve the sufficient descent property, Hager and Zhang [59] proposed the following conjugate gradient parameter:

$$\beta_k^N = \frac{1}{d_k^T y_k}\left(y_k - 2d_k \frac{||y_k||^2}{d_k^T y_k}\right)^T g_{k+1} = \beta_k^{HS} - 2\frac{||y_k||^2}{d_k^T y_k}\frac{d_k^T g_{k+1}}{d_k^T y_k},$$

which can be considered as an adaptive version of β_k^{DL} given by (38). The method has been shown to be globally convergent for uniformly convex objective functions. In order to achieve the global convergence for general functions, the following truncation of β_k^N has been proposed in [59]:

$$\bar{\beta}_k^N = \max\{\beta_k^N, \eta_k\}, \ \eta_k = \frac{-1}{||d_k|| \min\{\eta, ||g_k||\}},$$

where η is a positive constant. A conjugate gradient method with the parameter $\bar{\beta}_k^N$ in which the line search fulfills the approximate Wolfe conditions given by (8) is called the CG-Descent algorithm [60]. Search directions of the CG-Descent algorithm satisfy the sufficient descent condition (25) with $c = \frac{7}{8}$. The CG-Descent algorithm is one of the most efficient and popular conjugate gradient methods, widely used by engineers and mathematicians engaged in solving large-scale unconstrained optimization problems.

Based on the Hager-Zhang approach [59], Yu et al. [96] proposed a modified form of β_k^{PRP} as follows:

$$\beta_k^{DPRP} = \beta_k^{PRP} - C\frac{||y_k||^2}{||g_k||^4}g_{k+1}^T d_k,$$

with a constant $C > \frac{1}{4}$, guaranteeing the sufficient descent condition (25) (see also [20]). Afterwards, several other descent extensions of the PRP method have been proposed in [26, 97], using the conjugate gradient parameter β_k^{DPRP}.

Exercise 8 *Prove that if $d_k^T y_k > 0$, $\forall k \geq 0$, then search directions of a conjugate gradient method with the following parameter:*

$$\beta_k^{\tau} = \beta_k^{HS} - \tau_k\frac{||y_k||^2(g_{k+1}^T d_k)}{(d_k^T y_k)^2},$$

in which $\tau_k \geq \bar{\tau}$, for some positive constant $\bar{\tau} > \frac{1}{4}$, satisfy the sufficient descent condition (25).

Exercise 9 *Prove that search directions of the DPRP method with $C > \frac{1}{4}$ satisfy the sufficient descent condition (25).*

3.7 Hybrid Conjugate Gradient Methods

Essential conjugate gradient methods generally can be divided into two categories. In the first category, all the conjugate gradient parameters have the common numerator $g_{k+1}^T y_k$; such as the HS and PRP methods, and also, the conjugate gradient method proposed by Liu and Storey [69] (LS) with the following parameter:

$$\beta_k^{LS} = -\frac{g_{k+1}^T y_k}{d_k^T g_k}.$$

In the second category, all the conjugate gradient parameters have the common numerator $||g_{k+1}||^2$; such as the FR and DY methods, and also, the conjugate descent

(CD) method proposed by Fletcher [49] with the following conjugate gradient parameter:

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

There are some advantages and disadvantages for the conjugate gradient methods in each category. As mentioned before, generally the methods of the first category are numerically efficient because of an automatic restart feature which avoids jamming while the methods of the second category are theoretically strong in the sense of (often) generating descent search directions and being globally convergent under a variety of line search conditions and some mild assumptions. To attain good computational performance and to maintain the attractive feature of strong global convergence, researchers paid especial attention to hybridize the conjugate gradient parameters of the two categories. Hybrid conjugate gradient methods are essentially designed based on an adoptive switch from a conjugate gradient parameter in the second category to one in the first category when the iterations jam. Well-known hybrid conjugate gradient parameters can be listed as follows:

- $\beta_k^{HuS} = \max\{0, \min\{\beta_k^{PRP}, \beta_k^{FR}\}\}$, proposed by Hu and Storey [63];
- $\beta_k^{TaS} = \begin{cases} \beta_k^{PRP}, \ 0 \leq \beta_k^{PRP} \leq \beta_k^{FR}, \\ \beta_k^{FR}, \quad \text{otherwise}, \end{cases}$ which has been proposed by Touati-Ahmed and Storey [87];
- $\beta_k^{GN} = \max\{-\beta_k^{FR}, \min\{\beta_k^{PRP}, \beta_k^{FR}\}\}$, proposed by Gilbert and Nocedal [56];
- $\beta_k^{hDYz} = \max\{0, \min\{\beta_k^{HS}, \beta_k^{DY}\}\}$, proposed by Dai and Yuan [48];
- $\beta_k^{hDY} = \max\{-\frac{1-\sigma}{1+\sigma}\beta_k^{DY}, \min\{\beta_k^{HS}, \beta_k^{DY}\}\}$, with the positive constant σ used in the Wolfe condition (6) [48];
- $\beta_k^{LS-CD} = \max\{0, \min\{\beta_k^{LS}, \beta_k^{CD}\}\}$, proposed by Andrei [7] (see also [95]).

In all of the above hybridization schemes, discrete combinations of the conjugate gradient parameters of the two categories have been considered. Recently, Andrei [8, 9, 11, 12] dealt with convex combinations of the conjugate gradient parameters of the two categories which are continuous hybridizations. More exactly, in [8] the following hybrid conjugate gradient method has been proposed:

$$\beta_k^C = (1-\mu_k)\beta_k^{HS} + \mu_k\beta_k^{DY},$$

in which $\mu_k \in [0, 1]$ is called the hybridization parameter. As known, if the point x_{k+1} is close enough to a local minimizer x^*, then a good direction to follow is the Newton direction, that is, $d_{k+1} = -\nabla^2 f(x_{k+1})^{-1} g_{k+1}$. So, considering search directions of the hybrid conjugate gradient method with the parameter β_k^C we can write:

$$-\nabla^2 f(x_{k+1})^{-1} g_{k+1} = -g_{k+1} + (1-\mu_k)\frac{g_{k+1}^T y_k}{s_k^T y_k} s_k + \mu_k \frac{g_{k+1}^T g_{k+1}}{s_k^T y_k} s_k.$$

After some algebraic manipulations we get

$$\mu_k = \frac{s_k^T \nabla^2 f(x_{k+1}) g_{k+1} - s_k^T g_{k+1} - \dfrac{g_{k+1}^T y_k}{s_k^T y_k} s_k^T \nabla^2 f(x_{k+1}) s_k}{\dfrac{g_{k+1}^T g_k}{s_k^T y_k} s_k^T \nabla^2 f(x_{k+1}) s_k}.$$

Due to the essential property of low memory requirement for the conjugate gradient methods, Andrei applied the secant equations in order to avoid exact computation of $\nabla^2 f(x_{k+1}) s_k$ [8, 9, 12] (see also [24, 34]). In a different approach, recently Babaie-Kafaki and Ghanbari [30, 31] proposed two other continuous hybrid conjugate gradient methods in which the hybridization parameter is computed in a way to make the search directions of the hybrid method as closer as possible to the search directions of the descent three-term conjugate gradient methods proposed by Zhang et al. [102, 103].

3.8 Spectral Conjugate Gradient Methods

In the stream of overcoming drawbacks of the steepest descent method, Barzilai and Borwein [35] developed the two-point stepsize gradient algorithms in which the search directions are computed by

$$d_0 = -g_0, \ d_{k+1} = -\theta_k g_{k+1}, \ k = 0, 1, \ldots,$$

where the positive parameter θ_k, called the scaling parameter, is computed by solving the following least-squares problem:

$$\min_{\theta \geq 0} ||\frac{1}{\theta} s_k - y_k||, \tag{41}$$

being a two-point approximation of the standard secant equation (17). After some algebraic manipulations, it can be seen that the solution of (41) is exactly the scaling parameter θ_k given by (23), used in the scaled memoryless quasi-Newton methods. Convergence of the two-point stepsize gradient algorithms has been studied in [44]. Using a nonmonotone line search procedure [57], Raydan [83] showed that the two-point stepsize gradient algorithms can be regarded as an efficient approach for solving large-scale unconstrained optimization problems. In [23, 46] the modified secant equations (29), (31) and (32) have been employed in the two-point stepsize gradient algorithms.

Combining search directions of the conjugate gradient methods and the two-point stepsize gradient algorithms, the spectral conjugate gradient methods [37] have been proposed in which the search directions are given by

$$d_{k+1} = -\theta_k g_{k+1} + \beta_k d_k, \; k = 0, 1, \ldots,$$

with $d_0 = -g_0$ and the scaling parameter θ_k often computed by (23) (see also [4–6, 10, 13]).

3.9 Three-Term Conjugate Gradient Methods

Although the concept of three-term conjugate gradient methods has been originally developed in 1970s [36, 70], recently researchers dealt with them in order to achieve the sufficient descent property. As known, some of the conjugate gradient methods such as HS, FR and PRP generally can not guarantee the descent condition (3). To overcome this problem, three-term versions of the HS, FR and PRP methods have been proposed respectively with the following search directions [101–103]:

$$d_{k+1} = -g_{k+1} + \beta_k^{HS} d_k - \frac{g_{k+1}^T d_k}{d_k^T y_k} y_k,$$

$$d_{k+1} = -g_{k+1} + \beta_k^{FR} d_k - \frac{g_{k+1}^T d_k}{||g_k||^2} g_{k+1},$$

$$d_{k+1} = -g_{k+1} + \beta_k^{PRP} d_k - \frac{g_{k+1}^T d_k}{||g_k||^2} y_k,$$

for all $k \geq 0$, with $d_0 = -g_0$. When the line search is exact, the above three-term conjugate gradient methods respectively reduce to the HS, FR and PRP methods. Also, for all of these methods we have the sufficient descent condition $d_k^T g_k = -||g_k||^2$, $\forall k \geq 0$, independent of the line search and the objective function convexity. A nice review of different three-term conjugate gradient methods has been presented in [85] (see also [14, 29]).

4 Limited-Memory Quasi-Newton Methods

As known, since quasi-Newton methods save an $n \times n$ matrix as an approximation of the inverse Hessian, they are not useful for solving large-scale unconstrained optimization problems. However, limited-memory quasi-Newton methods maintain a compact approximation of the inverse Hessian, saving only a few vectors of length n available from a certain number of the most recent iterations, and so, being useful in large-scale cases [74]. Convergence properties of the methods are often acceptable [13, 67, 73]. Although various limited-memory quasi-Newton methods have been proposed in the literature, here we deal with the limited-memory BFGS method, briefly called the L-BFGS method.

Note that the BFGS updating formula (14) can be written as:

$$H_{k+1}^{BFGS} = V_k^T H_k V_k + \rho_k s_k s_k^T, \tag{42}$$

where

$$\rho_k = \frac{1}{s_k^T y_k}, \text{ and } V_k = I - \rho_k y_k s_k^T.$$

In the limited-memory approach, a modified version of H_{k+1} is implicitly stored, saving a set of vector pairs $\{s_i, y_i\}$ available from the $m > 1$ recent iterations. More precisely, by repeated application of the formula (42), we get

$$\begin{aligned} H_{k+1}^{L-BFGS} = & (V_k^T \cdots V_{k-m+1}^T) H_{k-m+1} (V_{k-m+1} \cdots V_k) \\ & + \rho_{k-m+1} (V_k^T \cdots V_{k-m+2}^T) s_{k-m+1} s_{k-m+1}^T (V_{k-m+2} \cdots V_k) \\ & + \cdots \\ & + \rho_k s_k s_k^T, \end{aligned}$$

in which in order to use a low memory storage, H_{k-m+1} is computed by

$$H_{k-m+1} = \theta_k I,$$

where θ_k is often calculated by (22), proved to be practically effective [67]. Also, the search direction $d_{k+1} = -H_{k+1}^{L-BFGS} g_{k+1}$ can be effectively computed by the following recursive procedure [74].

Algorithm 1 (*Computing search directions of the L-BFGS method*)

```
$q = g_{k+1}$;
for $i = k, k-1, \ldots, k-m+1$
    $\gamma_i \leftarrow \rho_i s_i^T q$;
    $q \leftarrow q - \gamma_i y_i$;
end
$r \leftarrow \theta_k q$;
for $i = k-m+1, k-m+2, \ldots, k$
    $\xi \leftarrow \rho_i y_i^T r$;
    $r \leftarrow r + s_i(\gamma_i - \xi)$;
end
$d_{k+1} = -r$.
```

Remark 1 Practical experiences have shown that the values of m between 3 and 20 often produce satisfactory numerical results [74].

5 Conclusions

Recent line search-based approaches in large-scale unconstrained optimization have been studied. Especially, the conjugate gradient methods and the memoryless quasi-Newton methods have been focused on. At first, after introducing the essential unconstrained optimization algorithms, merits and demerits of the classical conjugate gradient methods have been reviewed. Then, their descent extensions, their hybridizations based on the secant equations, and their three-term versions with sufficient descent property have been discussed. Finally, a limited-memory quasi-Newton method has been presented. So, recent efficient tools for big data applications have been provided.

Acknowledgments This research was supported by Research Council of Semnan University. The author is grateful to Professor Ali Emrouznejad for his valuable suggestions helped to improve the presentation.

References

1. Akaike, H.: On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method. Ann. Inst. Statist. Math. Tokyo **11**(1), 1–16 (1959)
2. Al-Baali, M.: Descent property and global convergence of the Fletcher-Reeves method with inexact line search. IMA J. Numer. Anal. **5**(1), 121–124 (1985)
3. Andrei, N.: Numerical comparison of conjugate gradient algorithms for unconstrained optimization. Stud. Inform. Control **16**(4), 333–352 (2007)
4. Andrei, N.: A scaled BFGS preconditioned conjugate gradient algorithm for unconstrained optimization. Appl. Math. Lett. **20**(6), 645–650 (2007)
5. Andrei, N.: Scaled conjugate gradient algorithms for unconstrained optimization. Comput. Optim. Appl. **38**(3), 401–416 (2007)
6. Andrei, N.: Scaled memoryless BFGS preconditioned conjugate gradient algorithm for unconstrained optimization. Optim. Methods Softw. **22**(4), 561–571 (2007)
7. Andrei, N.: 40 conjugate gradient algorithms for unconstrained optimization—a survey on their definition. ICI Technical Report No. 13/08 (2008)
8. Andrei, N.: Another hybrid conjugate gradient algorithm for unconstrained optimization. Numer. Algorithms **47**(2), 143–156 (2008)
9. Andrei, N.: A hybrid conjugate gradient algorithm for unconstrained optimization as a convex combination of Hestenes-Stiefel and Dai-Yuan. Stud. Inform. Control **17**(1), 55–70 (2008)
10. Andrei, N.: A scaled nonlinear conjugate gradient algorithm for unconstrained optimization. Optimization **57**(4), 549–570 (2008)
11. Andrei, N.: Hybrid conjugate gradient algorithm for unconstrained optimization. J. Optim. Theory Appl. **141**(2), 249–264 (2009)
12. Andrei, N.: Accelerated hybrid conjugate gradient algorithm with modified secant condition for unconstrained optimization. Numer. Algorithms **54**(1), 23–46 (2010)
13. Andrei, N.: Accelerated scaled memoryless BFGS preconditioned conjugate gradient algorithm for unconstrained optimization. Eur. J. Oper. Res. **204**(3), 410–420 (2010)
14. Andrei, N.: A modified Polak-Ribière-Polyak conjugate gradient algorithm for unconstrained optimization. Optimization **60**(12), 1457–1471 (2011)
15. Andrei, N.: Open problems in conjugate gradient algorithms for unconstrained optimization. B. Malays. Math. Sci. So. **34**(2), 319–330 (2011)

16. Babaie-Kafaki, S.: A modified BFGS algorithm based on a hybrid secant equation. Sci. China Math. **54**(9), 2019–2036 (2011)
17. Babaie-Kafaki, S.: A note on the global convergence theorem of the scaled conjugate gradient algorithms proposed by Andrei. Comput. Optim. Appl. **52**(2), 409–414 (2012)
18. Babaie-Kafaki, S.: A modified scaled memoryless BFGS preconditioned conjugate gradient method for unconstrained optimization. 4OR **11**(4):361–374 (2013)
19. Babaie-Kafaki, S.: A new proof for the sufficient descent condition of Andrei's scaled conjugate gradient algorithms. Pac. J. Optim. **9**(1), 23–28 (2013)
20. Babaie-Kafaki, S.: An eigenvalue study on the sufficient descent property of a modified Polak-Ribière-Polyak conjugate gradient method. Bull. Iranian Math. Soc. **40**(1), 235–242 (2014)
21. S. Babaie-Kafaki. On the sufficient descent condition of the Hager-Zhang conjugate gradient methods. 4OR **12**(3):285–292 (2014)
22. Babaie-Kafaki, S.: Two modified scaled nonlinear conjugate gradient methods. J. Comput. Appl. Math. **261**(5), 172–182 (2014)
23. Babaie-Kafaki, S., Fatemi, M.: A modified two-point stepsize gradient algorithm for unconstrained minimization. Optim. Methods Softw. **28**(5), 1040–1050 (2013)
24. Babaie-Kafaki, S., Fatemi, M., Mahdavi-Amiri, N.: Two effective hybrid conjugate gradient algorithms based on modified BFGS updates. Numer. Algorithms **58**(3), 315–331 (2011)
25. Babaie-Kafaki, S., Ghanbari, R.: The Dai-Liao nonlinear conjugate gradient method with optimal parameter choices. Eur. J. Oper. Res. **234**(3), 625–630 (2014)
26. Babaie-Kafaki, S., Ghanbari, R.: A descent extension of the Polak-Ribière-Polyak conjugate gradient method. Comput. Math. Appl. **68**(12), 2005–2011 (2014)
27. Babaie-Kafaki, S., Ghanbari, R.: A descent family of Dai-Liao conjugate gradient methods. Optim. Methods Softw. **29**(3), 583–591 (2014)
28. Babaie-Kafaki, S., Ghanbari, R.: A modified scaled conjugate gradient method with global convergence for nonconvex functions. Bull. Belg. Math. Soc. Simon Stevin **21**(3), 465–477 (2014)
29. Babaie-Kafaki, S., Ghanbari, R.: Two modified three-term conjugate gradient methods with sufficient descent property. Optim. Lett. **8**(8), 2285–2297 (2014)
30. Babaie-Kafaki, S., Ghanbari, R.: A hybridization of the Hestenes-Stiefel and Dai-Yuan conjugate gradient methods based on a least-squares approach. Optim. Methods Softw. **30**(4), 673–681 (2015)
31. Babaie-Kafaki, S., Ghanbari, R.: A hybridization of the Polak-Ribière-Polyak and Fletcher-Reeves conjugate gradient methods. Numer. Algorithms **68**(3), 481–495 (2015)
32. Babaie-Kafaki, S., Ghanbari, R.: Two optimal Dai-Liao conjugate gradient methods. Optimization **64**(11), 2277–2287 (2015)
33. Babaie-Kafaki, S., Ghanbari, R., Mahdavi-Amiri, N.: Two new conjugate gradient methods based on modified secant equations. J. Comput. Appl. Math. **234**(5), 1374–1386 (2010)
34. Babaie-Kafaki, S., Mahdavi-Amiri, N.: Two modified hybrid conjugate gradient methods based on a hybrid secant equation. Math. Model. Anal. **18**(1), 32–52 (2013)
35. Barzilai, J., Borwein, J.M.: Two-point stepsize gradient methods. IMA J. Numer. Anal. **8**(1), 141–148 (1988)
36. Beale, E.M.L.: A derivation of conjugate gradients. In: Lootsma, F.A. (ed.) Numerical Methods for Nonlinear Optimization, pp. 39–43. Academic Press, NewYork (1972)
37. Birgin, E., Martínez, J.M.: A spectral conjugate gradient method for unconstrained optimization. Appl. Math. Optim. **43**(2), 117–128 (2001)
38. Buckley, A.G.: Extending the relationship between the conjugate gradient and BFGS algorithms. Math. Program. **15**(1), 343–348 (1978)
39. Cauchy, A.: Méthodes générales pour la résolution des systèmes déquations simultanées. C. R. Acad. Sci. Par. **25**(1), 536–538 (1847)
40. Dai, Y.H.: Analyses of conjugate gradient methods. Ph.D. Thesis, Mathematics and Scientific/Engineering Computing, Chinese Academy of Sciences (1997)
41. Dai, Y.H.: New properties of a nonlinear conjugate gradient method. Numer. Math. **89**(1), 83–98 (2001)

42. Dai, Y.H., Kou, C.X.: A nonlinear conjugate gradient algorithm with an optimal property and an improved Wolfe line search. SIAM J. Optim. **23**(1), 296–320 (2013)
43. Dai, Y.H., Liao, L.Z.: New conjugacy conditions and related nonlinear conjugate gradient methods. Appl. Math. Optim. **43**(1), 87–101 (2001)
44. Dai, Y.H., Liao, L.Z.: R-linear convergence of the Barzilai and Borwein gradient method. IMA J. Numer. Anal. **22**(1), 1–10 (2002)
45. Dai, Y.H., Ni, Q.: Testing different conjugate gradient methods for large-scale unconstrained optimization. J. Comput. Math. **22**(3), 311–320 (2003)
46. Dai, Y.H., Yuan, J., Yuan, Y.X.: Modified two-point stepsize gradient methods for unconstrained optimization. Comput. Optim. Appl. **22**(1), 103–109 (2002)
47. Dai, Y.H., Yuan, Y.X.: A nonlinear conjugate gradient method with a strong global convergence property. SIAM J. Optim. **10**(1), 177–182 (1999)
48. Dai, Y.H., Yuan, Y.X.: An efficient hybrid conjugate gradient method for unconstrained optimization. Ann. Oper. Res. **103**(1–4), 33–47 (2001)
49. Fletcher, R.: Practical Methods of Optimization. Wiley, New York (1987)
50. Fletcher, R., Reeves, C.M.: Function minimization by conjugate gradients. Comput. J. **7**(2), 149–154 (1964)
51. Ford, J.A., Moghrabi, I.A.: Multi-step quasi-Newton methods for optimization. J. Comput. Appl. Math. **50**(1–3), 305–323 (1994)
52. Ford, J.A., Moghrabi, I.A.: Minimum curvature multistep quasi-Newton methods. Comput. Math. Appl. **31**(4–5), 179–186 (1996)
53. Ford, J.A., Moghrabi, I.A.: Using function-values in multi-step quasi-Newton methods. J. Comput. Appl. Math. **66**(1–2), 201–211 (1996)
54. Ford, J.A., Narushima, Y., Yabe, H.: Multi-step nonlinear conjugate gradient methods for unconstrained minimization. Comput. Optim. Appl. **40**(2), 191–216 (2008)
55. Forsythe, G.E.: On the asymptotic directions of the *s*-dimensional optimum gradient method. Numer. Math. **11**(1), 57–76 (1968)
56. Gilbert, J.C., Nocedal, J.: Global convergence properties of conjugate gradient methods for optimization. SIAM J. Optim. **2**(1), 21–42 (1992)
57. Grippo, L., Lampariello, F., Lucidi, S.: A nonmonotone line search technique for Newton's method. SIAM J. Numer. Anal. **23**(4), 707–716 (1986)
58. Guo, Q., Liu, J.G., Wang, D.H.: A modified BFGS method and its superlinear convergence in nonconvex minimization with general line search rule. J. Appl. Math. Comput. **28**(1–2), 435–446 (2008)
59. Hager, W.W., Zhang, H.: A new conjugate gradient method with guaranteed descent and an efficient line search. SIAM J. Optim. **16**(1), 170–192 (2005)
60. Hager, W.W., Zhang, H.: Algorithm 851: CG-Descent, a conjugate gradient method with guaranteed descent. ACM Trans. Math. Softw. **32**(1), 113–137 (2006)
61. Hager, W.W., Zhang, H.: A survey of nonlinear conjugate gradient methods. Pac. J. Optim. **2**(1), 35–58 (2006)
62. Hestenes, M.R., Stiefel, E.: Methods of conjugate gradients for solving linear systems. J. Research Nat. Bur. Standards **49**(6), 409–436 (1952)
63. Hu, Y.F., Storey, C.: Global convergence result for conjugate gradient methods. J. Optim. Theory Appl. **71**(2), 399–405 (1991)
64. Li, D.H., Fukushima, M.: A modified BFGS method and its global convergence in nonconvex minimization. J. Comput. Appl. Math. **129**(1–2), 15–35 (2001)
65. Li, D.H., Fukushima, M.: On the global convergence of the BFGS method for nonconvex unconstrained optimization problems. SIAM J. Optim. **11**(4), 1054–1064 (2001)
66. Li, G., Tang, C., Wei, Z.: New conjugacy condition and related new conjugate gradient methods for unconstrained optimization. J. Comput. Appl. Math. **202**(2), 523–539 (2007)
67. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large-scale optimization. Math. Program. **45**(3, Ser. B), 503–528 (1989)
68. Liu, G.H., Han, J.Y., Yin, H.X.: Global convergence of the Fletcher-Reeves algorithm with an inexact line search. Appl. Math. J. Chin. Univ. Ser. B **10**(1), 75–82 (1995)

69. Liu, Y., Storey, C.: Efficient generalized conjugate gradient algorithms. I. Theory. J. Optim. Theory Appl. **69**(1), 129–137 (1991)
70. Nazareth, J.L.: A conjugate direction algorithm without line searches. J. Optim. Theory Appl. **23**(3), 373–387 (1977)
71. Nazareth, J.L.: A relationship between the BFGS and conjugate gradient algorithms and its implications for the new algorithms. SIAM J. Numer. Anal. **16**(5), 794–800 (1979)
72. Nazareth, J.L.: Conjugate gradient methods less dependent on conjugacy. SIAM Rev. **28**(4), 501–511 (1986)
73. Nocedal, J.: Updating quasi-Newton matrices with limited storage. Math. Comput. **35**(151), 773–782 (1980)
74. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, New York (2006)
75. Oren, S.S.: Self-scaling variable metric (SSVM) algorithms. II. Implementation and experiments. Manage. Sci. **20**(5), 863–874 (1974)
76. S.S. Oren and D.G. Luenberger. Self-scaling variable metric (SSVM) algorithms. I. Criteria and sufficient conditions for scaling a class of algorithms. Manage. Sci. **20**(5), 845–862 (1973/1974)
77. Oren, S.S., Spedicato, E.: Optimal conditioning of self-scaling variable metric algorithms. Math. Program. **10**(1), 70–90 (1976)
78. Perry, A.: A modified conjugate gradient algorithm. Oper. Res. **26**(6), 1073–1078 (1976)
79. Polak, E., Ribière, G.: Note sur la convergence de méthodes de directions conjuguées. Rev. Française Informat. Recherche Opérationnelle **3**(16), 35–43 (1969)
80. Polyak, B.T.: The conjugate gradient method in extreme problems. USSR Comput. Math. Math. Phys. **9**(4), 94–112 (1969)
81. Powell, M.J.D.: Restart procedures for the conjugate gradient method. Math. Program. **12**(2), 241–254 (1977)
82. Powell, M.J.D.: Nonconvex minimization calculations and the conjugate gradient method. In: Griffiths, D.F. (ed.) Numerical Analysis (Dundee, 1983), Lecture Notes in Mathematics, vol. 1066, pp. 122–141. Springer, Berlin (1984)
83. Raydan, M.: The Barzilai and Borwein gradient method for the large-scale unconstrained minimization problem. SIAM J. Optim. **7**(1), 26–33 (1997)
84. Shanno, D.F.: Conjugate gradient methods with inexact searches. Math. Oper. Res. **3**(3), 244–256 (1978)
85. Sugiki, K., Narushima, Y., Yabe, H.: Globally convergent three-term conjugate gradient methods that use secant conditions and generate descent search directions for unconstrained optimization. J. Optim. Theory Appl. **153**(3), 733–757 (2012)
86. Sun, W., Yuan, Y.X.: Optimization Theory and Methods: Nonlinear Programming. Springer, New York (2006)
87. Touati-Ahmed, D., Storey, C.: Efficient hybrid conjugate gradient techniques. J. Optim. Theory Appl. **64**(2), 379–397 (1990)
88. Watkins, D.S.: Fundamentals of Matrix Computations. Wiley, New York (2002)
89. Wei, Z., Li, G., Qi, L.: New quasi-Newton methods for unconstrained optimization problems. Appl. Math. Comput. **175**(2), 1156–1188 (2006)
90. Wei, Z., Yu, G., Yuan, G., Lian, Z.: The superlinear convergence of a modified BFGS-type method for unconstrained optimization. Comput. Optim. Appl. **29**(3), 315–332 (2004)
91. Wolfe, P.: Convergence conditions for ascent methods. SIAM Rev. **11**(2), 226–235 (1969)
92. Wolfe, P.: Convergence conditions for ascent methods. II. Some corrections. SIAM Rev. **13**(2), 185–188 (1971)
93. Xu, C., Zhang, J.Z.: A survey of quasi-Newton equations and quasi-Newton methods for optimization. Ann. Oper. Res. **103**(1–4), 213–234 (2001)
94. Yabe, H., Takano, M.: Global convergence properties of nonlinear conjugate gradient methods with modified secant condition. Comput. Optim. Appl. **28**(2), 203–225 (2004)
95. Yang, X., Luo, Z., Dai, X.: A global convergence of LS-CD hybrid conjugate gradient method. Adv. Numer. Anal. Article ID 517452 (2013)

96. Yu, G., Guan, L., Li, G.: Global convergence of modified Polak-Ribière-Polyak conjugate gradient methods with sufficient descent property. J. Ind. Manage. Optim. **4**(3), 565–579 (2008)
97. Yuan, G.L.: Modified nonlinear conjugate gradient methods with sufficient descent property for large-scale optimization problems. Optim. Lett. **3**(1), 11–21 (2009)
98. Yuan, Y.X.: A modified BFGS algorithm for unconstrained optimization. IMA J. Numer. Anal. **11**(3), 325–332 (1991)
99. Yuan, Y.X., Byrd, R.H.: Non-quasi-Newton updates for unconstrained optimization. J. Comput. Math. **13**(2), 95–107 (1995)
100. Zhang, J.Z., Deng, N.Y., Chen, L.H.: New quasi-Newton equation and related methods for unconstrained optimization. J. Optim. Theory Appl. **102**(1), 147–167 (1999)
101. Zhang, L., Zhou, W., Li, D.H.: A descent modified Polak-Ribière-Polyak conjugate gradient method and its global convergence. IMA J. Numer. Anal. **26**(4), 629–640 (2006)
102. Zhang, L., Zhou, W., Li, D.H.: Global convergence of a modified Fletcher-Reeves conjugate gradient method with Armijo-type line search. Numer. Math. **104**(4), 561–572 (2006)
103. Zhang, L., Zhou, W., Li, D.H.: Some descent three-term conjugate gradient methods and their global convergence. Optim. Methods Softw. **22**(4), 697–711 (2007)
104. Zhou, W., Zhang, L.: A nonlinear conjugate gradient method based on the MBFGS secant condition. Optim. Methods Softw. **21**(5), 707–714 (2006)
105. Zoutendijk, G.: Nonlinear programming computational methods. In: Abadie, J. (ed.) Integer and Nonlinear Programming, pp. 37–86. North-Holland, Amsterdam (1970)

Author Biography

Saman Babaie-Kafaki is an Associate Professor of Mathematics in Semnan University, Iran. He received his B.Sc. degree in Applied Mathematics from Mazandaran University, Iran, in 2003, and his M.Sc. and Ph.D. degrees in Applied Mathematics from Sharif University of Technology, Iran, in 2005 and 2010, respectively, under the supervision of Professor Nezam Mahdavi-Amiri. His research interests include numerical optimization, matrix computations and heuristic algorithms.