Chapter 2
Conjugate Gradient Methods for Nonconvex Problems

2.1 Introduction

It is worthwhile to notice that when interests in conjugate gradient algorithms for quadratic problems subsided their versions for nonconvex differentiable problems were proposed. These propositions relied on the simplicity of their counterparts for quadratic problems. As we have shown in the previous chapter a conjugate gradient algorithm is an iterative process which requires at each iteration the current gradient and the previous direction. The simple scheme for calculating the current direction was easy to extend to a nonquadratic problem

$$\min_{x \in \mathscr{R}^n} f(x). \tag{2.1}$$

Since we assume that f is a nonconvex function the extension of a conjugate gradient algorithm to these functions needs the specification of the optimality conditions which would serve as a stopping criterion in new methods. Throughout Chap. 1 we dealt with quadratic functions defined by symmetric positive definite matrices. For these quadratic functions the condition that the gradient is equal to zero are both necessary and sufficient conditions of optimality of a global solution. In the case of general nonconvex functions the considered extensions of a conjugate gradient algorithm are aimed at finding a local solution to the problem (2.1). A stopping criterion of these methods refers to the necessary optimality conditions of the problem (2.1). If at points fulfilling a stopping criterion the sufficient optimality conditions are satisfied these methods find also local minimizers. The formal statement of these conditions requires several definitions beginning from the definition of a local solution to the problem (2.1) and ending up with the sufficient optimality conditions for a local solution.

Definition 2.1. The point $\bar{x} \in \mathscr{R}^n$ is a local solution to the problem (2.1) if there exists a neighborhood $\mathscr{N}$ of $\bar{x}$ such that

$$f(x) \geqslant f(\bar{x})$$

for all $x \in \mathscr{N}$. If $f(x) > f(\bar{x})$, $\forall x \in \mathscr{N}$ with $x \neq \bar{x}$, then $\bar{x}$ is a strict local solution.

R. Pytlak, *Conjugate Gradient Algorithms in Nonconvex Optimization,*
Nonconvex Optimization and Its Applications.

The necessary optimality conditions for a local solution $\bar{x}$ are given in the following lemma.

Lemma 2.1. *Suppose that f is a continuously differentiable function in some neighborhood of $\bar{x}$ which solves the problem (2.1), then*

$$g(\bar{x}) = 0. \tag{2.2}$$

Proof. The proof of the lemma follows directly from the Taylor's expansion theorem. For any $d \in \mathscr{R}^n$ there exists $\alpha \in (0,1)$ such that (cf. Appendix A)

$$f(\bar{x}+d) = f(\bar{x}) + g(\bar{x}+\alpha d)^T d. \tag{2.3}$$

If $g(\bar{x}) \neq 0$ then we can assume $\bar{d} = -g(\bar{x})$ and due to the continuous differentiability of f we can take some small $\bar{\alpha} \in (0,1)$ such that

$$g(\bar{x}+\bar{\alpha}\bar{d})^T \bar{d} < 0 \tag{2.4}$$

and (by substituting $\bar{\alpha}\bar{d}$ for d in (2.3))

$$f(\bar{x}+\bar{\alpha}\bar{d}) = f(\bar{x}) + g(\bar{x}+\tilde{\alpha}\bar{d})^T \bar{d} \tag{2.5}$$

for some $\tilde{\alpha} \in (0,\bar{\alpha})$. Equation (2.5) together with (2.4) contradict our assumption that $\bar{x}$ is a local minimizer of f. □

A point $\bar{x}$ satisfying (2.2) is called a *stationary point*. If at a stationary point $\bar{x}$ the Hessian matrix of f is positive definite, or in other words that *sufficient optimality conditions* are satisfied:

$$g(\bar{x}) = 0 \tag{2.6}$$

$$z^T \nabla^2 f(\bar{x}) z > 0, \;\; \forall z \in \mathscr{R}^n \quad \text{with } z \neq 0, \tag{2.7}$$

then $\bar{x}$ is a local solution.

Lemma 2.2. *Suppose that f is twice continuously differentiable and at a point $\bar{x}$ (2.6)–(2.7) hold. Then $\bar{x}$ is a strict local solution of the problem (2.1).*

Proof. Since f is twice continuously differentiable there exists a neighborhood of $\bar{x}$ – $\mathscr{N}$ such that $\nabla^2 f(x)$ is positive definite for all $x \in \mathscr{N}$. There exists a positive number α such that for any normalized direction d we have $\bar{x}+\alpha d \in \mathscr{N}$. Using the Taylor's expansion theorem, for any $\bar{\alpha} \in [0,\alpha]$, we can write

$$f(\bar{x}+\bar{\alpha}d) = f(\bar{x}) + \frac{1}{2} d^T \nabla^2 f(z) d > f(\bar{x}) \tag{2.8}$$

since $z \in \mathscr{N}$. This implies that $\bar{x}$ is a local solution to the problem (2.1). □

Having the necessary optimality conditions (2.2) the straightforward extension of a conjugate gradient algorithm, stated in Chap. 1 for the quadratic problem, would be as follows.

Algorithm 2.1. (The conjugate gradient algorithm with exact line search)

1. Choose an arbitrary $x_1 \in \mathscr{R}^n$. Set
$$d_1 = -g_1 \tag{2.9}$$
and $k = 1$.
2. Find α_k which minimizes f on the line $\mathscr{L}_k = \{x \in \mathscr{R}^n : \ x = x_k + \alpha d_k, \ \alpha \in (0,\infty)\}$. Substitute $x_k + \alpha_k d_k$ for x_{k+1}.
3. If $\|g_{k+1}\| = 0$ then STOP, otherwise calculate d_{k+1} according to
$$d_{k+1} = -g_{k+1} + \beta_{k+1} d_k.$$
4. Increase k by one and go to Step 2.

The above algorithm is an example of an iterative optimization algorithm which generates the sequence $\{x_k\}$ in the following way

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

where α_k is chosen to decrease value of f:

$$f(x_{k+1}) < f(x_k). \tag{2.10}$$

We will demand from an iterative optimization procedure that either at some point x_l $g_l = 0$, or there exists a subsequence $\{x_{k_l}\}$ such that

$$\lim_{k_l \to \infty} g(x_{k_l}) = 0.$$

In Step 2 condition (2.10) is satisfied by requiring that α_k is a minimum of f along a direction d_k. Contrary to the quadratic case this is not a simple algebraic problem. In general we have to apply an iterative procedure to find an approximate solution to the problem

$$\min_{\alpha > 0} \left[\phi(\alpha) = f(x_k + \alpha d_k)\right]. \tag{2.11}$$

2.2 Line Search Methods

In order to avoid solving problem (2.11) several rules for finding α_k, which guarantee the global convergence of an optimization algorithm, have been proposed. These rules related to the problem (2.11) are called the *directional minimization rules*.

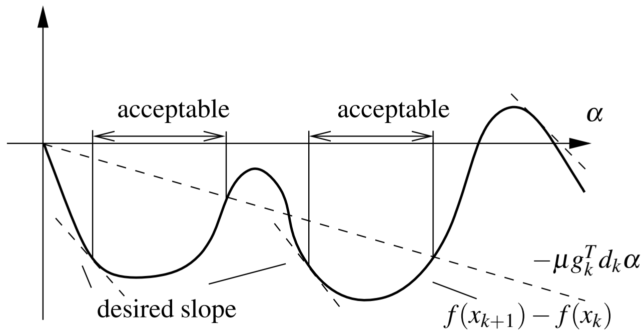

Fig. 2.1 The Wolfe line search rule

The most popular rule is expressed by the Wolfe conditions [207], [208]:

$$f(x_k + \alpha_k d_k) - f(x_k) \leqslant \mu \alpha_k g_k^T d_k \tag{2.12}$$

$$g(x_k + \alpha_k d_k)^T d_k \geqslant \eta g_k^T d_k, \tag{2.13}$$

where $0 < \mu < \eta < 1$ (cf. Fig. 2.1).

Condition (2.12) is sometimes referred as *Armijo condition* although this name is often reserved for another directional minimization rule which will be discussed later. Condition (2.12) stipulates a decrease of f along d_k. Notice that if

$$g_k^T d_k < 0, \tag{2.14}$$

i.e. d_k is the *direction of descent*, then there exists, under some conditions imposed on f, $\tilde{\alpha}$ such that for all $\alpha \in [0, \tilde{\alpha})$ (2.12) is satisfied:

$$f(x_k + \alpha d_k) - f(x_k) \leqslant \mu \alpha g_k^T d_k$$

since $0 < \mu < 1$.

Condition (2.13) is called the *curvature condition* and its role is to force α_k to be sufficiently far away from zero which could happen if only condition (2.12) were to be used. Notice that there exists $\hat{\alpha} > 0$ such that for any $\alpha \in [0, \hat{\alpha}]$ (2.13) will not be satisfied:

$$g(x_k + \alpha d_k)^T d_k < \eta g_k^T d_k$$

for any $\alpha \in [0, \hat{\alpha}]$ since $0 < \eta < 1$.

If we wish to find a point α_k, which is closer to a solution of problem (2.11) than a point satisfying (2.12)–(2.13), we can impose on α_k the *strong Wolfe conditions*:

$$f(x_k + \alpha_k d_k) - f(x_k) \leqslant \mu \alpha_k g_k^T d_k \tag{2.15}$$

$$|g(x_k + \alpha_k d_k)^T d_k| \leqslant \eta |g_k^T d_k| \tag{2.16}$$

with $0 < \mu < \eta < 1$. In contrast to the Wolfe conditions $g(x_k + \alpha_k d_k)^T d_k$ cannot be arbitrarily large.

The directional minimization rule which does not refer to the curvature condition, called the *Armijo rule*, (the corresponding procedure is often called *backtracking line search*) is defined by the following algorithm:

Algorithm 2.2. (The Armijo line search algorithm)

Parameters: $s > 0$, $\mu,\ \beta \in (0,1)$.

1. Set $\alpha = s$.
2. If

$$f(x_k + \alpha d_k) - f(x_k) \leqslant \mu \alpha g_k^T d_k \tag{2.17}$$

substitute α for α_k and STOP.
3. Substitute $\beta\alpha$ for α, go to Step 2.

The essence of the Armijo rule is that α_k is accepted only if α_k/β does not satisfy the Armijo condition therefore we avoid too small values of α_k which could hamper the global convergence. As we will show later, at each iteration initial s can assume different values, provided they lie in some set $[s_{min}, s_{max}]$ and $0 < s_{min} < s_{max}$.

The Wolfe conditions can be modified by taking, as the accuracy reference in the Armijo and the curvature conditions, $-\|d_k\|^2$ instead of $g_k^T d_k$. Assume that

$$g_k^T d_k \leqslant -\|d_k\|^2$$

then the *modified Wolfe conditions* are

$$f(x_k + \alpha_k d_k) - f(x_k) \leqslant -\mu \alpha_k \|d_k\|^2 \tag{2.18}$$

$$g(x_k + \alpha_k d_k)^T d_k \geqslant -\eta \|d_k\|^2, \tag{2.19}$$

where $0 < \mu < \eta < 1$. Notice that condition (2.18) is weaker in comparison to the Armijo condition while condition (2.19) is stronger than the curvature condition. The equivalence of these conditions holds when $g_k^T d_k = -\|d_k\|^2$.

The above mentioned line search methods have any practical meaning if we can guarantee finding α_k satisfying them in a finite number of operations. However, first of all we have to show that such α_k exist.

Lemma 2.3. *Assume that f is continuously differentiable and that it is bounded from below along the line $\mathscr{L}_k = \{x \in \mathscr{R}^n :\ x = x_k + \alpha d_k,\ \alpha \in (0,\infty)\}$. Suppose also that d_k is a direction of descent ((2.14) is satisfied). If $0 < \mu < \eta < 1$ then there exist nonempty intervals of step lengths satisfying the Wolfe and the strong Wolfe conditions.*

Proof. The proof is standard one – we follow that given in [146]. Consider the line $l(\alpha) = f(x_k) + \mu\alpha g_k^T d_k$. Since $0 < \mu < 1$ it must intersect the graph of $\phi(\alpha)$. Suppose that $\tilde{\alpha}$ is the smallest intersection point. Therefore, we have

$$f(x_k + \tilde{\alpha} d_k) = f(x_k) + \mu\tilde{\alpha} g_k^T d_k.$$

Furthermore, from the mean value theorem there exists $\hat{\alpha} \in (0, \tilde{\alpha})$ such that

$$f(x_k + \tilde{\alpha} d_k) = f(x_k) + \tilde{\alpha} g(x_k + \hat{\alpha} d_k)^T d_k$$

and since $0 < \mu < \eta < 1$ we also have

$$g(x_k + \hat{\alpha} d_k)^T d_k = \mu g_k^T d_k > \eta g_k^T d_k.$$

Due to our smoothness assumption (2.12)–(2.13) are satisfied by all α in some neighborhood of $\hat{\alpha}$. The interval on which the strong Wolfe conditions hold is the same since $f(x_k + \tilde{\alpha} d_k) - f(x_k) < 0$ and

$$|g(x_k + \hat{\alpha} d_k)^T d_k| < \eta |g_k^T d_k|.$$

It completes the proof. □

In order to show that there exists a nonempty interval of step lengths satisfying the modified Wolfe conditions we will provide a procedure which determines α_k in a finite number of operations.

Lemma 2.4. *Assume that f is continuously differentiable and that it is bounded from below along the line $\mathscr{L}_k = \{x \in \mathscr{R}^n : \; x = x_k + \alpha d_k, \; \alpha \in (0, \infty)\}$. Suppose also that d_k is a direction of descent ((2.14) is satisfied). If $0 < \mu < \eta < 1$ then there exists a procedure which finds α_k satisfying (2.18)–(2.19) in a finite number of operations.*

Proof. The proof is based on the paper [133]. Let $\varepsilon > 0$ be such that $\eta > \mu + \varepsilon$. Let α^0 be an arbitrary positive number and set $\alpha_N^0 = \infty, \; \alpha_\mu^0 = 0$. The procedure, which we propose, produces a sequence of points $\{\alpha^l\}$ constructed in the following way. If

$$f(x_k + \alpha^l d_k) - f(x_k) \leqslant -(\mu + \varepsilon)\alpha^l \parallel d_k \parallel^2 \tag{2.20}$$

we set $\alpha_\mu^{l+1} = \alpha^l, \; \alpha_N^{l+1} = \alpha_N^l$ and $\alpha_\mu^{l+1} = \alpha_\mu^l, \; \alpha_N^{l+1} = \alpha^l$ otherwise. Next, we substitute α^{l+1} by $(\alpha_\mu^l + \alpha_N^l)/2$ if $\alpha_N^{l+1} < \infty$, or by $2\alpha^{l+1}$ if $\alpha_N^{l+1} = \infty$.

If for every α^l generated by the procedure we have that α^l satisfies (2.20) then α_N^l will remain equal to infinity and $f(x_k + \alpha^l d_k) \to -\infty$. Therefore, let us suppose that there exists α^l such that (2.20) is not fulfilled. This means that at some iteration doubling has started and we have the sequences $\{\alpha_N^l\}, \; \{\alpha_\mu^l\}$ such that $\alpha_N^l - \alpha_\mu^l \to 0, \; \alpha_N^l \to \hat{\alpha}$, because either α_N^{l+1} or α_μ^{l+1} is set to $(\alpha_N^l + \alpha_\mu^l)/2$. Moreover, let us assume that for every l, α^l does not satisfy (2.19) and (2.20) (otherwise we have found the desired coefficient α_k). In this case $\alpha_\mu^l \to \hat{\alpha}$ and $\hat{\alpha}$ satisfies (2.20). If we have (2.19) for α^l infinitely often then the procedure will terminate after finite number of iterations, because $\alpha^l \to \hat{\alpha}$, f is continuous and (2.18) will have to be satisfied.

Now, let us suppose that (2.19) holds only for the finite number of α^l, thus

$$g(x_k + \alpha^l d_k)^T d_k < -\eta \|d_k\|^2$$

for infinitely many times. This leads to

$$g(x_k + \hat{\alpha} d_k)^T d_k \leqslant -\eta \|d_k\|^2.$$

Because

$$f(x_k + \alpha_N^l d_k) - f(x_k) > -(\mu + \varepsilon)\alpha_N^l \|d_k\|^2, \tag{2.21}$$

thus

$$f(x_k + \alpha_N^l d_k) - f(x_k + \hat{\alpha} d_k) > -(\mu + \varepsilon)(\alpha_N^l - \hat{\alpha})\|d_k\|^2 \tag{2.22}$$

and

$$\begin{aligned} -\eta \|d_k\|^2 &\geqslant \lim_{l\to\infty} g(x_k + \alpha^l d_k)^T d_k = g(x_k + \hat{\alpha} d_k)^T d_k \\ &= \lim_{l\to\infty} \frac{f(x_k + \alpha_N^l d_k) - f(x_k + \hat{\alpha} d_k)}{\alpha_N^l - \hat{\alpha}} \geqslant -(\mu + \varepsilon)\|d_k\|^2, \end{aligned} \tag{2.23}$$

but this is impossible since $\eta > \mu + \varepsilon$. $\square$

Since the Wolfe conditions are a special case of the modified Wolfe conditions we have also shown how to find α_k, in a finite number of operations, which satisfies (2.12)–(2.13).

Following the first part of the proof of Lemma 2.3 we can easily show that the backtracking procedure will find an α_k which fulfills the Armijo condition in a finite number of operations provided that d_k is a direction of descent.

We have specified the main elements of algorithms which belong to the class of line search methods. An algorithm from the class can be stated as follows.

Algorithm 2.3. (The line search method)

1. Choose an arbitrary $x_1 \in \mathscr{R}^n$ and a descent direction d_1. Set $k = 1$.
2. Find $\alpha_k > 0$ which approximately solves the problem

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

   Substitute $x_k + \alpha_k d_k$ for x_{k+1}.
3. If $\|g_{k+1}\| = 0$ then STOP, otherwise find d_{k+1}, a direction of descent.
4. Increase k by one and go to Step 2.

2.3 General Convergence Results

The main convergence result, which will be extensively exploited in the context of several algorithms, is due to Zoutendijk [214]. It refers to the angle θ_k between d_k and the *steepest descent direction* $-g_k$ defined by

$$\cos\theta_k = \frac{-g_k^T d_k}{\|g_k\|\|d_k\|}. \tag{2.24}$$

Essentially the Zoutendijk's result states that if the θ_k is always less than π then the algorithm is globally convergent in the sense that

$$\lim_{k\to\infty} \|g_k\| = 0. \tag{2.25}$$

Equation (2.25) is a straightforward consequence of the following theorem.

Theorem 2.1. *Suppose that* $\{x_k\}$ *is generated by* Algorithm 2.3 *and*

(i) α_k *satisfies the Wolfe conditions,*
(ii) f *is bounded below in* $\mathscr{R}^n$*,*
(iii) f *is continuously differentiable in an open set* $\mathscr{N}$ *containing the level set* $\mathscr{M} = \{x \in \mathscr{R}^n : f(x) \leqslant f(x_1)\}$*,*
(iv) $g(x)$ *is Lipschitz continuous – there exists* $L > 0$ *such that*

$$\|g(y) - g(x)\| \leqslant L\|y - x\| \tag{2.26}$$

for all $x,\ y \in \mathscr{N}$.

Then

$$\sum_{k\geqslant 0} \cos^2\theta_k \|g_k\|^2 < \infty.$$

Proof. From the curvature condition

$$(g_{k+1} - g_k)^T d_k \geqslant (\eta - 1)\, g_k^T d_k \tag{2.27}$$

and since g is Lipschitz continuous we also have

$$|\,(g_{k+1} - g_k)^T d_k| \leqslant \alpha_k L \|d_k\|^2. \tag{2.28}$$

Combining (2.27) with (2.28) we obtain

$$\alpha_k \geqslant \frac{\eta - 1}{L\|d_k\|^2} g_k^T d_k \tag{2.29}$$

which together with the Armijo condition results in

$$f(x_{k+1}) - f(x_k) \leqslant -\mu \frac{1-\eta}{L} \frac{\left(g_k^T d_k\right)^2}{\|d_k\|^2}. \tag{2.30}$$

Taking into account (2.24) we rewrite (2.30) as

$$f(x_{k+1}) - f(x_k) \leqslant -c\cos^2\theta_k \|g_k\|^2$$

with $c = \mu(1-\eta)/L$. After summing this expression over all indices from one to k we come to

$$f(x_{k+1}) - f(x_1) \leqslant -c\sum_{l=1}^{k}\cos^2\theta_l\|g_l\|^2.$$

Since f is bounded below we must have

$$\sum_{k=1}^{\infty}\cos^2\theta_k\|g_k\|^2 < \infty. \tag{2.31}$$

□

Observe that if there exists $m > 0$ such that

$$\cos\theta_k \geqslant m \tag{2.32}$$

for all k then from (2.31) we have

$$\lim_{k\to\infty}\cos^2\theta_k\|g_k\|^2 = 0$$

and

$$\lim_{k\to\infty}\|g_k\| = 0$$

which guarantees global convergence of the algorithm as we have already stated.

If the modified Wolfe conditions or the Armijo rule are used in general only weaker convergence results can be established.

Theorem 2.2. *Suppose that assumptions of* Theorem 2.1 *hold and* α_k *satisfies the modified Wolfe conditions. Then*

$$\lim_{k\to\infty}\|d_k\| = 0.$$

Proof. From the modified curvature condition we have

$$(g_{k+1} - g_k)^T d_k \geqslant -(\eta - 1)\|d_k\|^2. \tag{2.33}$$

On the other hand, from the Lipschitz condition, we come to

$$|(g_{k+1} - g_k)^T d_k| \leqslant \alpha_k L\|d_k\|^2$$

which combined with (2.33) give the lower bound on α_k:

$$\alpha_k \geqslant \frac{1-\eta}{L}$$

for all k.

Using arguments similar to those in the proof of Theorem 2.1 we can show that

$$f(x_{k+1}) - f(x_1) \leqslant -\mu c \sum_{l=1}^{k} \|d_l\|^2$$

with $c = (1-\eta)/L$. Since f is bounded below we must have

$$\sum_{k=1}^{\infty} \|d_k\|^2 < \infty$$

from which it follows

$$\lim_{k\to\infty} \|d_k\| = 0.$$

□

Theorem 2.2 is useful in proving global convergence if we can show that there exists $M < \infty$ such that

$$\|g_k\| \leqslant M\|d_k\|$$

for all k. Suppose, for example, that

$$d_k = -B_k^{-1} g_k \tag{2.34}$$

and there exists $M < \infty$ such that

$$\|B_k\|_2 \leqslant M$$

where $\|\cdot\|_2$ is the matrix norm induced by the Euclidean norm. Then, we will have

$$\|g_k\| \leqslant M\|d_k\|$$

and from Theorem 2.2 global convergence follows

$$\lim_{k\to\infty} \|g_k\| = 0.$$

Yet another convergence result can be provided for algorithms with the Armijo rule (cf. [9]), or the *modified Armijo rule* that is defined by Algorithm 2.2 in which the condition (2.17) is replaced by the condition

$$f(x_k + \alpha d_k) - f(x_k) \leqslant -\mu\alpha\|d_k\|^2 \tag{2.35}$$

and it is assumed that

$$g_k^T d_k \leqslant -\|d_k\|^2. \tag{2.36}$$

Theorem 2.3. *Suppose that assumptions (ii)–(iii) of* Theorem 2.1 *are satisfied and* α_k *is determined by the Armijo rule, or the modified Armijo rule. If for any convergent subsequence* $\{x_k\}_{k\in K}$ *such that*

$$\lim_{k\to\infty,\ k\in K} g_k \neq 0$$

we have

$$\limsup_{k\to\infty, k\in K} \|d_k\| < \infty,$$

$$\liminf_{k\to\infty, k\in K} \left|g_k^T d_k\right| > 0$$

then,

$$\lim_{k\to\infty, k\in K} g_k^T d_k = 0 \tag{2.37}$$

in the case of the Armijo rule, or

$$\lim_{k\to\infty, k\in K} \|d_k\| = 0 \tag{2.38}$$

in the case of the modified Armijo rule.

Proof. The proof for the Armijo rule is given in [9], therefore we concentrate on the proof of the part of theorem related to the modified Armijo rule.

As in the proof of the previous theorem we can show that

$$f(x_{k+1}) - f(x_1) \leqslant -\mu \sum_{l=1}^{k} \alpha_l \|d_l\|^2, \tag{2.39}$$

in the case of the modified Armijo rule. On the basis on the assumption *(iii)*, (2.39) implies that

$$\lim_{k\to\infty} \alpha_k \|d_k\|^2 = 0. \tag{2.40}$$

Consider now any convergent subsequence $\{x_k\}_{k\in K}$ and let its accumulation point be $\bar{x}$. Suppose also that (2.38) is not satisfied, therefore, from (2.39), we must have

$$\lim_{k\to\infty, k\in K} \alpha_k = 0. \tag{2.41}$$

The above means that there exists $k_1 \in K$ such that for all $k \in K,\ k \geqslant k_1$

$$f(x_k + (\alpha_k/\beta)d_k) - f(x_k) > -\mu(\alpha_k/\beta)\|d_k\|^2 \tag{2.42}$$

since otherwise we would have $\alpha_k = s$ for k sufficiently large contradicting (2.41).

Since the sequence $\{d_k\}_{k\in K}$ is bounded there exists convergent subsequence of $\{d_k\}_{k\in K}$ which for the simplicity of presentation will be denoted in the same way. Now (2.42) can be rewritten as

$$\frac{f(x_k+\bar{\alpha}_k d_k)-f(x_k)}{\bar{\alpha}_k} > -\mu\|d_k\|^2$$

where $\bar{\alpha}_k = \alpha_k/\beta$.

If we take limits of both sides of the above expression and notice that f is continuously differentiable we obtain

$$g(\bar{x})^T\bar{d} \geqslant -\mu\|\bar{d}\|^2, \tag{2.43}$$

where

$$\lim_{k\to\infty, k\in K} d_k = \bar{d}.$$

On the other hand, d_k satisfies (2.36) thus

$$g(\bar{x})^T\bar{d} \leqslant -\|\bar{d}\|^2$$

which contradicts (2.43). This means that our assumption (2.41) has been wrong and the proof of (2.38) is concluded. □

As the possible application of Theorem 2.3 consider the algorithm with the direction defined by (2.34). However, now we assume that the condition numbers of B_k are bounded (cf. Appendix C):

$$\kappa(B_k) = \|B_k\|_2\|B_k^{-1}\|_2 \leqslant M < \infty$$

for all k. In this case eigenvalues of B_k^{-1} lie in the subinterval $[\gamma,\Gamma]$ where γ, Γ are smallest and largest eigenvalues of B_k^{-1}. We have

$$-g_k^T d_k = g_k^T B_k^{-1} g_k \geqslant \gamma\|g_k\|^2$$
$$\|d_k\| \leqslant \|B_k^{-1}\|\|g_k\| \leqslant \Gamma\|g_k\|.$$

and all assumptions of Theorem 2.3 are satisfied. Therefore, for any convergent subsequence $\{x_k\}_{k\in K}$ its limit point $\bar{x}$ satisfies $g(\bar{x}) = 0$.

All convergence theorems stated so far refer only to stationary points. In order to show global convergence to local minimum points we need the assumption that every stationary point satisfies sufficient optimality conditions.

The following theorem is due to Bertsekas [9] – its extension to constrained problems is given in [176].

Theorem 2.4. *Assume that*

(i) f is twice continuously differentiable,
(ii) the sequence $\{x_k\}$ constructed by Algorithm 2.3 *is such that*

$$\begin{aligned} f(x_{k+1}) &\leqslant f(x_k) \\ \alpha_k &\leqslant s \\ \|d_k\| &\leqslant c\|g_k\| \end{aligned}$$

for all k, where $s > 0$, $c > 0$,
(iii) every convergent subsequence of $\{x_k\}$ is convergent to a stationary point.

Then, for every local minimum point $\bar{x}$ of f at which $\nabla^2 f(\bar{x})$ is positive definite there exists a neighborhood of $\bar{x}$ – $\mathcal{N}$ such that, if $x_{\bar{k}} \in \mathcal{N}$ for some $\bar{k}$ then $x_k \in \mathcal{N}$ for all $k \geqslant \bar{k}$ and $x_k \to_{k\to\infty} \bar{x}$.

Proof. We follow the proof stated in [9]. Since $\nabla^2 f(\bar{x})$ is positive definite exists $\bar{\varepsilon} > 0$ such that $\nabla^2 f(x)$ is positive definite for all x satisfying $\|x - \bar{x}\| \leqslant \bar{\varepsilon}$. Introduce

$$\begin{aligned} \gamma &= \min_{\|z\|=1,\ \|x-\bar{x}\|\leqslant\bar{\varepsilon}} z^T \nabla^2 f(x) z \\ \Gamma &= \max_{\|z\|=1,\ \|x-\bar{x}\|\leqslant\bar{\varepsilon}} z^T \nabla^2 f(x) z \end{aligned}$$

which will satisfy $\Gamma \geqslant \gamma > 0$.

Define now an open set

$$\mathcal{N} = \left\{ x \in \mathscr{R}^n :\ \|x - \bar{x}\| < \bar{\varepsilon},\ f(x) - f(\bar{x}) < \frac{1}{2}\gamma \left(\frac{\bar{\varepsilon}}{1 + sc\Gamma} \right)^2 \right\}.$$

We will show that the set $\mathcal{N}$ is the one stated in the thesis, i.e. if $x_{\bar{k}} \in \mathcal{N}$ then $x_k \in \mathcal{N}$ for all $k \geqslant \bar{k}$.

Assume that $x_{\bar{k}} \in \mathcal{N}$, then from the mean value theorem (cf. Appendix A) the following relation holds

$$\frac{1}{2}\gamma\|x_{\bar{k}} - \bar{x}\| \leqslant f(x_{\bar{k}}) - f(\bar{x}) \leqslant \frac{1}{2}\gamma \left(\frac{\bar{\varepsilon}}{1 + sc\Gamma} \right)^2$$

which implies that

$$\|x_{\bar{k}} - \bar{x}\| \leqslant \frac{\bar{\varepsilon}}{1 + sc\Gamma}. \tag{2.44}$$

Furthermore, from the iterative rule

$$\begin{aligned} \|x_{\bar{k}+1} - \bar{x}\| &= \|x_{\bar{k}} - \alpha_{\bar{k}} d_{\bar{k}} - \bar{x}\| \\ &\leqslant \|x_{\bar{k}} - \bar{x}\| + sc\|g_{\bar{k}}\|. \end{aligned}$$

Using again the mean value theorem but for g we can write

$$\|x_{\bar{k}+1} - \bar{x}\| \leqslant (1 + sc\Gamma)\,\|x_{\bar{k}} - \bar{x}\| \tag{2.45}$$

since $g(\bar{x}) = 0$.

Taking into account (2.44) and (2.45) we conclude that

$$\|x_{\bar{k}+1} - \bar{x}\| < \bar{\varepsilon}$$

and since

$$f(x_{\bar{k}+1}) \leqslant f(x_{\bar{k}}) < \frac{1}{2}\gamma\left(\frac{\bar{\varepsilon}}{1 + sc\Gamma}\right)^2$$

thus we also have $x_{\bar{k}+1} \in \mathscr{L}$.

If we denote by $\bar{\mathscr{N}}$ the closure of $\mathscr{N}$ then $x_k \in \bar{\mathscr{N}}$ for all $k \geqslant \bar{k}$. $\bar{\mathscr{N}}$ is a compact set thus $\{x_k\}_{k \geqslant \bar{k}}$ has at least one accumulation point which, by the assumption, must be the stationary point. Since $\bar{x}$ is the only stationary point in $\bar{\mathscr{N}}$ we have shown that the sequence $\{x_k\}_{k \geqslant \bar{k}}$ converges to $\bar{x}$. □

Theorem 2.4 has many applications. We can use it to prove that the steepest descent method with the Armijo rule is convergent to local minimum points if all of them satisfy sufficient optimality conditions. Moreover, the quasi-Newton methods defined by

$$d_k = -D_k g_k$$

with bounded positive definite matrices D_k and with the Wolfe rules are also globally convergent to local minimum points provided that the step selection procedure determines bounded α_k. As we will see in the next section practical step selection algorithms usually fulfill this requirement.

2.4 Moré–Thuente Step-Length Selection Algorithm

In this section we discuss a procedure which finds a step-length parameter α_k, in a finite number of iterations, that fulfills the strong Wolfe conditions. The strong Wolfe conditions guarantee global convergence of several line search methods. They are employed in quasi-Newton methods (cf. [29, 166]) and in conjugate gradient methods (cf. [2, 43, 78, 125]).

From mathematical point of view we aim at the approximate minimization of the function

$$\phi(\alpha) = f(x_k + \alpha d_k) : [0, \infty) \to \mathscr{R}.$$

By an approximate minimizing point of $\phi(\cdot)$ we mean an $\alpha > 0$ such that

$$\phi(\alpha) \leqslant \phi(0) + \mu\dot{\phi}(0)\alpha \tag{2.46}$$

$$|\dot{\phi}(\alpha)| \leqslant \eta|\dot{\phi}(0)|. \tag{2.47}$$

Since we assume that d_k is a direction of descent we also have that $\dot{\phi}(0) < 0$. The condition (2.47) is called the curvature condition since it implies that the average curvature of ϕ on $[0,\alpha]$ is positive

$$\dot{\phi}(\alpha) - \dot{\phi}(0) \geqslant (1-\eta)|\dot{\phi}(0)|.$$

In addition to conditions (2.46)–(2.47) additional requirements are imposed on α:

$$0 < \alpha_{min} \leqslant \alpha \leqslant \alpha_{max} < \infty.$$

The main reason for requiring a lower bound α_{min} is to avoid too small values of α which could cause some numerical problems. The upper bound α_{max} is needed if the function ϕ is unbounded below.

In this section we present the step-length selection procedure proposed by Moré and Thuente [136]. We describe this procedure since it has been verified in professional optimization software, it has guaranteed convergence in a finite number of operations. As far as the other step-length algorithms are concerned we have to mention that suggested by Fletcher [70] which led to algorithms developed by Al-Baali and Fletcher [3], Moré and Sorensen [135]. Furthermore, we could recommend implementations of algorithms by Dennis and Schnabel [55], by Lemaréchal [120] and by Fletcher [71]. In the next section we introduce to a new line search procedure proposed by Hager and Zhang in [91]. The interesting features of the procedure are the precautions taken to cope with a finite precision calculations.

The procedure given by Moré and Thuente requires that $0 < \mu < \eta < 1$ although value of $\mu < 1/2$ is often assumed, because if ϕ is a quadratic function with $\dot{\phi}(0) < 0$ and $\ddot{\phi}(0) > 0$, then the global minimizer $\bar{\alpha}$ of ϕ satisfies

$$\phi(\bar{\alpha}) = \phi(0) + \frac{1}{2}\bar{\alpha}\ddot{\phi}(0) \tag{2.48}$$

which means that $\bar{\alpha}$ satisfies (2.46) only if $\mu \leqslant 1/2$. Furthermore, if $\mu < 1/2$ is used $\alpha = 1$ is accepted at final iterations of the Newton and quasi-Newton methods.

The procedure by Moré and Thuente finds a point α which belongs to the set

$$T(\mu) = \left\{\alpha : \ \phi(\alpha) \leqslant \phi(0) + \alpha\mu\dot{\phi}(0), \ |\dot{\phi}(\alpha)| \leqslant \mu|\dot{\phi}(0)|\right\}$$

thus the parameter η is used only in the convergence test.

The iteration of the procedure consists of three major steps.

Algorithm 2.4. (Iteration of the Moré–Thuente step-length selection algorithm)

Input and output data: $I_i \subset [0,\infty]$.

1. Choose a trial point $\alpha_t \in I_i \bigcap [\alpha_{min}, \alpha_{max}]$.
2. Test for convergence.
3. Update I_i.

Initially we assume that $I_0 = [0,\infty]$, and then we find $I_i = [\alpha_l^i, \alpha_u^i]$ such that α_l^i, α_u^i satisfy

$$\psi(\alpha_l^i) \leqslant \psi(\alpha_u^i),\ \psi(\alpha_l^i) \leqslant 0,\ \dot{\psi}(\alpha_l^i)\left(\alpha_u^i - \alpha_l^i\right) < 0, \tag{2.49}$$

where

$$\psi(\alpha) = \phi(\alpha) - \phi(0) - \mu\dot{\phi}(0)\alpha.$$

The endpoints of I_i are chosen in such a way that I_i always contains α from $T(\mu)$ – the following lemma gives basis for that [136].

Lemma 2.5. *Let I be a closed interval with endpoints $[\alpha_l,\ \alpha_u]$ which satisfy (2.49). Then there exists an $\bar{\alpha}$ in I with $\psi(\bar{\alpha}) \leqslant \psi(\alpha_l)$ and $\dot{\psi}(\bar{\alpha}) = 0$, in particular $\bar{\alpha} \in I \bigcap T(\mu)$.*

The updated interval I_{i+1} is determined, on the basis of the trial point α_t^i, by the following procedure.

1. If $\psi(\alpha_t^i) > \psi(\alpha_l^i)$, then $\alpha_l^{i+1} = \alpha_l^i$, $\alpha_u^{i+1} = \alpha_t^i$.
2. If $\psi(\alpha_t^i) \leqslant \psi(\alpha_l^i)$ and $\psi'(\alpha_t^i)\left(\alpha_l^i - \alpha_t^i\right) > 0$, then $\alpha_l^{i+1} = \alpha_t^i$ and $\alpha_u^{i+1} == \alpha_u^i$.
3. If $\psi(\alpha_t^i) \leqslant \psi(\alpha_l^i)$ and $\psi'(\alpha_t^i)\left(\alpha_l^i - \alpha_t^i\right) < 0$, then $\alpha_l^{i+1} = \alpha_t^i$ and $\alpha_u^{i+1} = \alpha_l^i$.
4. Substitute α_t^i for α_i.

The aim of the updating procedure is to choose I_i in such a way that $I_i \bigcap T(\mu)$ is not empty and then to move to I_{i+1} guaranteeing that $I_{i+1} \bigcap T(\mu)$ is also non empty.

In general, with the exception of some very rare degenerate cases (cf. [136]) the above step-length selection procedure either terminates at α_{min}, α_{max} or an interval $I_i \subset [\alpha_{min}, \alpha_{max}]$ is generated.

We can show that the procedure terminates at α_{min} if the sequence $\{\alpha_i\}$ is decreasing and the following holds

$$\psi(\alpha_i) > 0 \quad \text{or} \quad \dot{\psi}(\alpha_i) \geqslant 0,\ \ k = 0,1,\ldots. \tag{2.50}$$

On the other hand $\{\alpha_i\}$ terminates at α_{max} if

$$\psi(\alpha_i) \leqslant 0 \quad \text{and} \quad \dot{\psi}(\alpha_i) < 0,\ \ k = 0,1,\ldots, \tag{2.51}$$

– then the sequence of trial points is increasing.

Therefore, the finite operation convergence of the above step-length selection procedure requires assumptions imposed on α_{min} and α_{max}. Namely, we require that

$$\psi(\alpha_{min}) \leqslant 0 \ \text{ and } \ \dot{\psi}(\alpha_{max}) < 0, \tag{2.52}$$

$$\psi(\alpha_{max}) > 0 \ \text{ and } \ \dot{\psi}(\alpha_{max}) \geqslant 0. \tag{2.53}$$

Under these conditions we can prove the following convergence result.

Theorem 2.5. *If the bounds α_{min} and α_{max} satisfy (2.52)–(2.53), then the step-length selection algorithm terminates in a finite number of iterations with $\alpha_i \in T(\mu)$, or the sequence $\{\alpha_i\}$ converges to some $\bar{\alpha} \in T(\mu)$ with $\dot{\psi}(\bar{\alpha}) = 0$. If the search algorithm does not terminate in a finite number of steps, then there is an index i_0 such that the endpoints of the interval I_i satisfy $\alpha_l^i < \bar{\alpha} < \alpha_u^i$. Moreover, if $\psi(\bar{\alpha}) \geqslant 0$, then $\dot{\psi}$ changes sign on $\left[\alpha_l^i, \bar{\alpha}\right]$ for all $i \geqslant i_0$, while if $\psi(\bar{\alpha}) < 0$, then $\dot{\psi}$ changes sign on $\left[\alpha_l^i, \bar{\alpha}\right]$ or $\left[\bar{\alpha}, \alpha_u^i\right]$ for all $i \geqslant i_0$.*

The algorithm for the selection of a trial point α_t contains several safeguards. In order to guarantee termination at α_{min}, provided that (2.50) holds, we use

$$\alpha_t^{i+1} \in \left[\alpha_{min}, \max\left[\delta_{min}\alpha_t^i, \alpha_{min}\right]\right]$$

for some factor $\delta_{min} < 1$.

If (2.51) holds then taking

$$\alpha_t^{i+1} \in \left[\min\left[\delta_{max}\alpha_t^i, \alpha_{max}\right]\right]$$

for some $\delta_{max} > 1$, ensures that eventually $\alpha_t = \alpha_{max}$.

In the other case α_t is determined by using quadratic or cubic interpolation – the details are given in [136].

The step-length selection procedure is especially important for conjugate gradient algorithms where good approximation of a minimum point of ϕ, obtained through few function and gradient evaluations, is needed. In general we cannot expect that $\alpha_k = 1$ is accepted close to a minimum point of the problem (2.1). In this situation the first trial point α_t^0 can have big influence on the whole efficiency of the line search method.

In practice two choices for α_t^0 are employed. In the first approach we assume that the first-order change in the function at iterate x_k will be the same as that obtained at the previous step. This implies that $\alpha_{k-1}g_{k-1}^T d_{k-1} = \alpha_t^0 g_k^T d_k$, therefore

$$\alpha_t^0 = \alpha_{k-1}\frac{g_{k-1}^T d_{k-1}}{g_k^T d_k}.$$

Another choice is to take as α_t^0 the minimum point of the quadratic defined by $f(x_{k-1})$, $f(x_k)$ and $\dot{\phi}(0) = g_k^T d_k$. This leads to

$$\alpha_t^0 = \frac{2\left(f(x_k) - f(x_{k-1})\right)}{\dot{\phi}(0)}. \tag{2.54}$$

One advantage of using (2.54) is that if $x_k \to \bar{x}$ superlinearly (cf. Sect. 2 in Chap. 4) then the ratio in (2.54) converges to 1. Therefore, taking

$$\alpha_t^0 = \min[1, 1.01\alpha_t^0]$$

ensures that $\alpha_k = 1$ will be tried and accepted for large k.

2.5 Hager–Zhang Step-Length Selection Algorithm

The step-length selection procedure proposed by Hager and Zhang [91] is also intended for optimization algorithms which use the Wolfe conditions. The procedure aims at finding the point α which satisfies both (2.46) and (2.47). The drawback of Moré–Thuente step-length selection algorithm is that it is designed for computer with an infinite precision arithmetic. If the method is used on computers with a finite precision arithmetic then the algorithm can fail to find, in a finite number of operations (or in any number of operations), a point α satisfying (2.46)–(2.47).

Hager and Zhang give in [91] the following example. Suppose that we want to minimize the function $f(x) = 1 - 2x + x^2$ in a neighborhood of the point $x = 1$. In Fig. 2.2 we present the graph generated by a Matlab using an Intel equipped PC. The solid line is obtained by generating 10,000 values of x on the interval $[1.0 - 2.5 \times 10^{-8}, 1 + 2.5 \times 10^{-8}]$ and then by connecting computed values of f by straight lines.

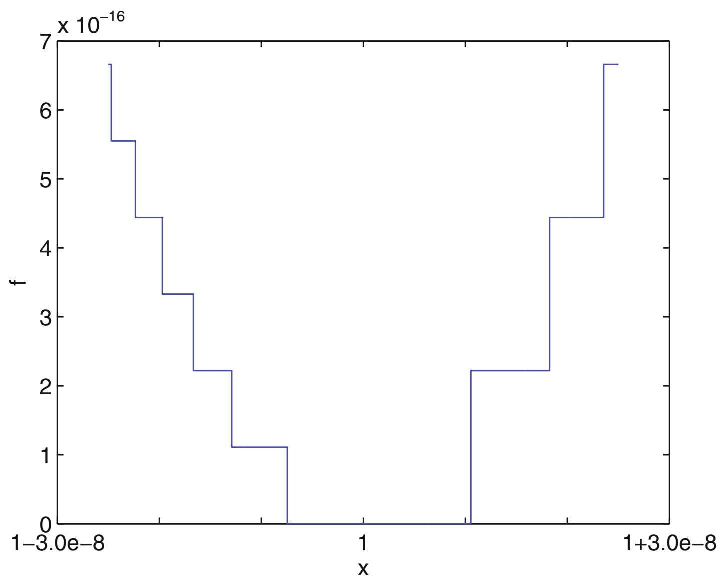

Fig. 2.2 Plot of $f(x) = 1 - 2x + x^2$

The striking feature of the solid line is a flat section on the subinterval $[1.0 - 0.9 \times 10^{-8}, 1 + 0.9 \times 10^{-8}]$ to which a zero value is assigned. Notice that its width can be estimated as 1.8×10^{-8} which is much wider then the machine precision $\varepsilon = 2.2 \times 10^{-16}$. The phenomenon is easily explained by a well-known effect of the reduction of significant digits. Near $x = 1$, both $1 - 2x$ and x^2 have almost equal values close to 1 in absolute terms but with opposite signs. As the result in order to evaluate f at the points close to 1 we have to subtract numbers of almost equal values. The error of these operations is equal to the square root of the machine precision [90].

The same phenomenon does not occur for the derivative of f: $\dot{f} = 2(x-1)$ in which case a flat section near the point $x = 1$ has the width of 1.6×10^{-16} (cf. Fig. 2.3). It implies that if we base our line search procedure on the derivative of f then the accuracy of determining its minimum point is close to the arithmetic precision 2.2×10^{-16} while the accuracy of the minimum value of f with the order of the square root of the arithmetic precision.

This suggested to Hager and Zhang to use instead of (2.46) the criterion based on the derivative of ϕ. This can be achieved by using a quadratic approximation of ϕ. The quadratic interpolating polynomial q that matches $\phi(\alpha)$ at $\alpha = 0$, and $\dot{\phi}(\alpha)$ at $\alpha = 0$ and $\alpha = \alpha_k$ (which is unknown) is given by

$$q(\alpha) = \phi(0) + \dot{\phi}(0)\alpha + \frac{\dot{\phi}(\alpha_k) - \dot{\phi}(0)}{2\alpha_k}\alpha^2.$$

Now, if we replace ϕ by q in the first Wolfe condition (2.46) we will obtain

$$q(\alpha_k) - q(0) \leqslant \mu \dot{q}(\alpha_k)$$

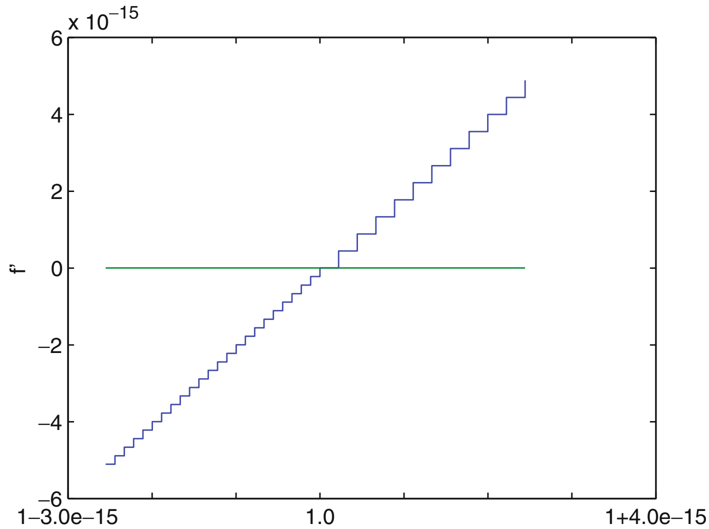

Fig. 2.3 Plot of the derivative $\dot{f}(x) = 2(x-1)$

which is translated into the condition

$$\frac{\dot{\phi}(\alpha_k) - \dot{\phi}(0)}{2}\alpha_k + \dot{\phi}(0)\alpha_k \leqslant \mu \alpha_k \dot{\phi}(0)$$

which can also be restated as

$$\dot{\phi}(\alpha_k) \leqslant (2\delta - 1)\dot{\phi}(0) \tag{2.55}$$

where δ must satisfy $\delta \in (0, \min[0.5, \eta]]$.

The procedure aims at finding α_k which either satisfies the Wolfe conditions: (2.46) and

$$\dot{\phi}(\alpha) \geqslant \eta \dot{\phi}(0)$$

– we call these criteria **LSR1**; or the condition (2.55) together with

$$\begin{aligned} \phi(\alpha) &\leqslant \phi(0) + \varepsilon_k \\ \varepsilon_k &= \varepsilon \, |f(x_k)| \end{aligned}$$

with ε to be some small positive values – we call these criteria **LSR2**.

The major iterations of the procedure are as follows.

Algorithm 2.5. (The Hager–Zhang step-length selection algorithm)

Parameter: $\gamma \in (0,1)$.

1. Choose $[a_0, b_0]$ and set $k = 0$.
2. If either **LSR1**, or **LSR2** are satisfied at a_k then STOP.
3. Define a new interval $[a,b]$ by using *secant* procedure: $[a,b] = secant(a_k, b_k)$.
4. If $b - a > \gamma(b_k - a_k)$, then $c = (a+b)/2$ and use the *update* procedure: $[a,b] = update(a,b,c)$.
5. Increase k by one, substitute $[a,b]$ for $[a_k, b_k]$, go to Step 2.

The essential part of the algorithm is the *update* function which changes the current bracketing interval $[a,b]$ into a new one $[\bar{a}, \bar{b}]$ with help of the additional point which is either obtained by a bisection step, or a secant step.

Algorithm 2.6. (The *update* procedure)

Parameters: $\theta \in (0,1)$, ε_k.
Input data: a, b, c.
Output data: $\bar{a}$, $\bar{b}$.

1. If $c \notin (a,b)$, then $\bar{a} = a$, $\bar{b} = b$ and STOP.
2. If $\dot{\phi}(c) \geqslant 0$, then $\bar{a} = a$, $\bar{b} = c$ and STOP.
3. If $\dot{\phi}(c) < 0$ and $\phi(c) \leqslant \phi(0) + \varepsilon_k$, then $\bar{a} = c$, $\bar{b} = b$ and STOP.

4. If $\dot{\phi}(c) < 0$ and $\phi(c) > \phi(0) + \varepsilon_k$, then $\bar{a} = a$, $\bar{b} = c$ and perform the steps:
   (a) Calculate $d = (1-\theta)\bar{a} + \theta\bar{b}$; if $\dot{\phi}(d) \geqslant 0$, then set $\bar{b} = d$ and STOP.
   (b) If $\dot{\phi}(d) < 0$ and $\phi(d) \leqslant \phi(0) + \varepsilon_k$, then set $\bar{a} = d$ and go to step (a).
   (c) If $\dot{\phi}(d) < 0$ and $\phi(d) > \phi(0) + \varepsilon_k$, then set $\bar{b} = d$ and go to step (a).

The *update* procedure finds the interval $[a,b]$ such that

$$\phi(a) < \phi(0) + \varepsilon_k, \;\; \dot{\phi}(a) < 0, \;\; \phi^{'}(b) \geqslant 0. \tag{2.56}$$

Eventually, a nested sequence of intervals $[a_k, b_k]$ is constructed which converges to the point satisfying either **LSR1**, or **LSR2**.

Another part of the Hager–Zhang algorithm is the *secant* procedure which updates interval with the help of secant steps. Suppose that $\dot{\phi}(\alpha)$ is a convex function on the interval $[a,b]$ with the property: $\dot{\phi}(a) < 0$, $\dot{\phi}(b) > 0$. Then in order to approximate a point at which $\dot{\phi}$ vanishes we construct a linear approximation of $\dot{\phi}$ on $[a,b]$:

$$\dot{\phi}_{la}(\alpha) = \frac{\dot{\phi}(b) - \dot{\phi}(a)}{b-a}(\alpha - a) + \dot{\phi}(a).$$

Under the assumption that $\dot{\phi}$ is convex, the point at which $\dot{\phi}_{la}$ is equal to zero lies on the right of a and is given by formula:

$$\alpha_0 = \frac{a\dot{\phi}(b) - b\dot{\phi}(a)}{\dot{\phi}(b) - \dot{\phi}(a)}.$$

Once α_0 is obtained $\bar{a}$ is set to α_0 and α_0 is used to perform the second secant step this time based on a and $\bar{\alpha}$ yielding a point $\bar{b}$ which must lie on the left of b provided that $\dot{\phi}$ is convex. Similarly secant steps are found in the case of concave $\dot{\phi}$ – details and the full description of the *secant* procedure is given in [91].

The Hang–Zhang line search procedure finds α_k fulfilling either **LSR1** or **LSR2** in a finite number of operations as stated in the following theorem proved in [91].

Theorem 2.6. *Suppose that ϕ is continuously differentiable on an interval $[a_0, b_0]$, where (2.56) holds. If $\delta \in (0, 0.5)$, then the Hager–Zhang step-length selection algorithm terminates at a point satisfying either* **LSR1***, or* **LSR2***.*

In [91] it is also shown that under some additional assumptions the interval width $|b_k - a_k|$ tends to zero with root convergence order $1 + \sqrt{2}$.

The Hager–Zhang step-length selection procedure was successfully applied, among others, in the conjugate gradient algorithm which is discussed in Sect. 10.

2.6 Nonlinear Conjugate Gradient Algorithms

The first conjugate gradient algorithm for nonconvex problems was proposed by Fletcher and Reeves in 1964 [73]. Its direction of descent is

$$d_k = -g_k + \beta_k d_{k-1} \tag{2.57}$$

with

$$\beta_k = \frac{\|g_k\|^2}{\|g_{k-1}\|^2}. \tag{2.58}$$

The algorithm is a straightforward extension of a conjugate gradient algorithm for the quadratic with the particular choice of a coefficient β_k. We know that the other equivalents for β_k are

$$\beta_k = \frac{(g_k - g_{k-1})^T g_k}{\|g_{k-1}\|^2}, \tag{2.59}$$

$$\beta_k = \frac{(g_k - g_{k-1})^T g_k}{d_{k-1}^T (g_k - g_{k-1})}, \tag{2.60}$$

$$\beta_k = \frac{\|g_k\|^2}{d_{k-1}^T (g_k - g_{k-1})}. \tag{2.61}$$

Although these formulae are equivalent for the quadratic the performance of a nonlinear conjugate gradient algorithm strongly depends on coefficients β_k. In general, formula (2.59) proposed by Polak and Ribière [161], is preferred over formula (2.58). On the other hand, formula (2.60) which is equivalent to formula (2.59) when directional minimization is exact, is used in several, more sophisticated conjugate gradient algorithms which will be discussed later.

Yet, the Fletcher–Reeves formula has an attractive property as far as convergence analysis is concerned. It appeared, after several failed attempts to prove global convergence of any conjugate gradient algorithm, that positiveness of a coefficient β_k is crucial as far as convergence is concerned.

Before presenting global convergence of the Fletcher–Reeves algorithm with inexact directional minimization and for nonconvex problems we provide the analysis of the behavior of conjugate gradient algorithms with different coefficients β_k. We follow Powell [168] who provided for the first time systematic arguments in favor of the Polak–Ribière algorithm.

Consider a conjugate gradient algorithm in which α_k is determined according to the exact directional minimization rule

$$\alpha_k = \arg\min_{\alpha>0} f(x_k + \alpha d_k) \tag{2.62}$$

with d_k defined by (2.57). If we denote by θ_k the angle between d_k and $-g_k$ then the following relation holds

$$\|d_k\| = \sec\theta_k \|g_k\|. \tag{2.63}$$

Furthermore, using the fact that $g_{k+1}^T d_k = 0$ we can state

$$\beta_{k+1}\|d_k\| = \tan\theta_{k+1}\|g_{k+1}\| \tag{2.64}$$

and combining these two relations we have

$$\begin{aligned} \tan\theta_{k+1} &= \sec\theta_k \frac{\|g_{k+1}\|}{\|g_k\|} \\ &> \tan\theta_k \frac{\|g_{k+1}\|}{\|g_k\|} \end{aligned} \tag{2.65}$$

if β_k is defined by (2.58) since $\sec\theta > \tan\theta$ for $\theta \in [0, \pi/2)$.

Powell'a arguments are as follows. If θ_k is close to $\pi/2$ then d_k is not a good direction of descent and the iteration takes very small step which implies that $g_{k+1} - g_k$ is small. This results in $\|g_{k+1}\|/\|g_k\|$ close to one and, from (2.65), θ_{k+1} is also close to $\pi/2$. Thus, the Fletcher–Reeves algorithm does not have self-correcting mechanism with this respect.

Consider now the Polak–Ribière version of a conjugate gradient algorithm. If, due to a small step, $g_{k+1} - g_k$ is close to zero, then according to rule (2.59) the coefficient β_{k+1} is also close to zero

$$\beta_{k+1} \leqslant \frac{\|g_{k+1}\|\|g_{k+1} - g_k\|}{\|g_k\|^2}$$

and the next direction is taken as the steepest direction: $d_{k+1} = -g_{k+1}$ since

$$\tan\theta_{k+1} \leqslant \sec\theta_k \frac{\|g_{k+1} - g_k\|}{\|g_k\|} \tag{2.66}$$

from (2.63)–(2.64). This self-correcting behavior of the Polak–Ribière algorithm makes it more efficient method for most problems. This analysis suggests the following restarting criteria in any conjugate gradient algorithms. If

$$|g_{k+1}^T g_k| \geqslant 0.2\|g_{k+1}\|^2 \tag{2.67}$$

holds then an algorithm should assume $d_{k+1} = -g_{k+1}$ [168].

Powell's observation led him to one of the first global convergence result on a conjugate gradient algorithm for a general nonconvex function. Although he assumed that the directional minimization is exact his result is important nonetheless, since he noticed that the prerequisite for the global convergence of the Polak–Ribière algorithm is

$$\lim_{k\to\infty} \|x_{k+1} - x_k\| = 0. \tag{2.68}$$

This condition will appear in the analysis of the global convergence of the Polak–Ribière versions of a conjugate gradient algorithm.

Theorem 2.7. *Assume that*

(i) the set

$$\mathscr{M} = \{x \in \mathscr{R}^n : \ f(x) \leqslant f(x_1)\} \tag{2.69}$$

is bounded,

(ii) g is a Lipschitz continuous function,

(iii) condition (2.68) is satisfied.

Then, the Polak–Ribière version of Algorithm 2.1 *(β_k defined by (2.59)) generates the sequence $\{x_k\}$ such that*

$$\liminf_{k\to\infty} \|g_k\| = 0. \tag{2.70}$$

Proof. Assume the opposite to (2.70) – there exists a positive ε and an integer k_1 such that

$$\|g_k\| \geqslant \varepsilon \tag{2.71}$$

for all $k \geqslant k_1$. Furthermore, due to (2.68), there exists $k_2 \geqslant k_1$ such that

$$\|g_{k+1} - g_k\| \leqslant \frac{1}{2}\varepsilon \tag{2.72}$$

for all $k \geqslant k_2$. Combining (2.66), (2.71)–(2.72) and the fact that

$$\sec\theta \leqslant 1 + \tan\theta$$

holds for all $\theta \in [0, \pi/2)$, we come to the relation

$$\tan\theta_{k+1} \leqslant \frac{1}{2}\left(1 + \tan\theta_k\right) \tag{2.73}$$

for all $k \geqslant k_2$.

Applying (2.73) recursively m times results in

$$\begin{aligned} \tan\theta_{k+1} &\leqslant \frac{1}{2} + \frac{1}{4} + \cdots \left(\frac{1}{2}\right)^{k_2+m+1} \left(1 + \tan\theta_{k_2}\right) \\ &\leqslant 1 + \tan\theta_{k_2}. \end{aligned}$$

This means that θ_k is bounded away from $\pi/2$. If we now apply Theorem 2.1 we come to the contradiction with (2.71). This completes the proof. □

The conjugate gradient algorithms considered so far have relied on the exact directional minimization, from now on we will concentrate on the convergence analysis of the methods which require only approximate minimization of f on the line $\mathscr{L}_k$.

Algorithm 2.7. (The standard conjugate gradient algorithm for nonconvex problems)

1. Choose an arbitrary $x_1 \in \mathscr{R}^n$. Set

$$d_1 = -g_1$$

and $k = 1$.
2. Find α_k which satisfies the Wolfe conditions (2.12)–(2.13), or the strong Wolfe conditions (2.15)–(2.16). Substitute $x_k + \alpha_k d_k$ for x_{k+1}.
3. If $\|g_{k+1}\| = 0$ then STOP, otherwise evaluate β_{k+1} and calculate d_{k+1} according to

$$d_{k+1} = -g_{k+1} + \beta_{k+1} d_k.$$

4. Increase k by one and go to Step 2.

2.7 Global Convergence of the Fletcher–Reeves Algorithm

Although, as explained in the previous section, the Fletcher–Reeves algorithm is inferior to the Polak–Ribière version of a conjugate gradient procedure, it was the method for which the global convergence properties were established first. The credit goes to Al-Baali who using ingenious analysis showed that the Fletcher–Reeves method is globally convergent for general nonconvex problems even if the directional minimization is not exact. His arguments are given below since they influenced research on nonlinear conjugate gradients algorithms for some time suggesting that the global convergence of conjugate gradient algorithms requires some assumptions on parameters in the step-length selection procedure based on the strong Wolfe conditions.

The global convergence analysis which we present in this section follows some unified pattern. First of all we use Theorem 2.1 to obtain a final conclusion about convergence. Observe that if a directional minimization is exact, i.e.

$$g_k^T d_{k-1} = 0,$$

then we can easily show that

$$\cos\theta_k = \frac{\|g_k\|}{\|d_k\|}.$$

This implies, from Theorem 2.1, that

$$\sum_{k=1}^{\infty} \frac{\|g_k\|^4}{\|d_k\|^2} < \infty \tag{2.74}$$

and, if one can show that $\{\cos\theta_k\}$ is bounded from zero, that

$$\lim_{k\to\infty} \|g_k\| = 0. \tag{2.75}$$

Unfortunately, this can be shown only under strong assumptions concerning function f, for example when f is a strictly convex function (cf. [161]). If we cannot bound the sequence $\{\|g_k\|/\|d_k\|\}$ from zero then we are content with the result weaker than (2.75), namely

$$\liminf_{k\to\infty} \|g_k\| = 0. \tag{2.76}$$

The standard way of reasoning to obtain (2.76) is to assume that (2.76) does not hold, i.e. we suppose that there exists $\varepsilon > 0$ such that

$$\|g_k\| \geqslant \varepsilon \tag{2.77}$$

for all k. Then, under this assumption we show that we can find $c > 0$ with the property

$$\cos\theta_k \geqslant c\frac{\|g_k\|}{\|d_k\|} \tag{2.78}$$

for all k. Equation (2.74) together with (2.77) and (2.78) imply that

$$\sum_{k=1}^{\infty} \frac{1}{\|d_k\|^2} < \infty. \tag{2.79}$$

Eventually, if one can show that $\|d_k\|^2$ grows at most linearly, i.e.

$$\|d_k\|^2 \leqslant Mk$$

for some constant $M < \infty$, then we come to the contradiction with (2.79).

It is interesting to observe that (2.78) is, due to the definition of θ_k, equivalent to the condition

$$g_k^T d_k \leqslant -c\|d_k\|^2 \tag{2.80}$$

for all k.

The main difficulty in proving global convergence of a conjugate gradient algorithm is to show that d_k is a direction of descent under mild assumption on the directional minimization. If we assume that the directional minimization is exact, i.e.

$$g_k^T d_{k-1} = 0 \tag{2.81}$$

for all k, then d_{k+1} is the steepest descent direction since

$$\begin{aligned} g_k^T d_k &= -\|g_k\|^2 + \beta_k g_k^T d_{k-1} \\ &= -\|g_k\|^2. \end{aligned}$$

Powell, using this property and assuming that the set (2.69) is bounded, and the function f is twice continuously differentiable, showed that the Fletcher–Reeves method generates a sequence $\{x_k\}$ satisfying (2.76).

Al-Baali showed that the descent property holds for all k if α_k is determined by using the strong Wolfe conditions. On that basis he then proved the method fulfills (2.76).

Theorem 2.8. *Assume that α_k is calculated in such a way that (2.13) is satisfied with $\eta \in (0, \frac{1}{2})$ for all k with $g_k \neq 0$. Then the descent property of the Fletcher–Reeves method holds for all k.*

Proof. Al-Baali proved by induction that the following inequalities hold for all k with $g_k \neq 0$:

$$-\sum_{j=0}^{k-1} \eta^j \leqslant \frac{g_k^T d_k}{\|g_k\|^2} \leqslant -2 + \sum_{j=0}^{k-1} \eta^j. \tag{2.82}$$

Equation (2.82) is obviously satisfied for $k = 1$, we assume that (2.82) is valid for $k > 1$ – we must show it for $k + 1$. We have

$$\sum_{j=0}^{k-1} \eta^j < \sum_{j=0}^{\infty} \eta^j = \frac{1}{1-\eta},$$

thus, from (2.82), d_k is a direction of descent for $\eta \in (0, \frac{1}{2})$. Moreover, from (2.58) we also have

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

hence from the strong Wolfe curvature condition

$$-1 - \eta \frac{g_k^T d_k}{\|g_k\|^2} \leqslant \frac{g_{k+1}^T d_{k+1}}{\|g_{k+1}\|^2} \leqslant -1 + \eta \frac{g_k^T d_k}{\|g_k\|^2}.$$

Applying the induction hypothesis (2.82) we eventually get

$$\begin{aligned} -\sum_{j=0}^{k} \eta^j = -1 - \eta \sum_{j=0}^{k-1} \eta^j &\leqslant \frac{g_{k+1}^T d_{k+1}}{\|g_{k+1}\|^2} \\ &\leqslant -1 + \eta \left(-2 + \sum_{j=0}^{k-1} \eta^j \right) \\ &\leqslant -2 + \sum_{j=0}^{k} \eta^j \end{aligned}$$

which completes the proof of (2.82). Since

$$-2+\sum_{j=0}^{\infty}\eta^{j} = -2+\frac{1}{(1-\eta)}, \tag{2.83}$$

thus, from (2.82), d_k is a direction of descent. □

The main convergence result given by Al-Baali refers to Theorem 2.1 by using the fact that d_k is a direction of descent.

Theorem 2.9. *Assume that* $\{x_k\}$ *is generated by* Algorithm 2.7 *with* β_k *defined by the Fletcher–Reeves formula (2.58). Furthermore, suppose that the assumptions (ii)–(iv) of* Theorem 2.1 *are satisfied and* α_k *is determined by the strong Wolfe conditions with* $0<\mu<\eta<\frac{1}{2}$. *Then, the sequence* $\{x_k\}$ *is such that*

$$\liminf_{k\to\infty}\|g_k\| = 0. \tag{2.84}$$

Proof. Notice that

$$\begin{aligned}\|d_k\|^2 &= \|g_k\|^2 - 2\beta_k g_k^T d_{k-1} + \beta_k^2\|d_{k-1}\|^2\\ &\leqslant \|g_k\|^2 - 2\eta\beta_k g_k^T d_{k-1} + \beta_k^2\|d_{k-1}\|^2\end{aligned} \tag{2.85}$$

from the curvature condition. However, form (2.82), we also have

$$\begin{aligned}\eta g_{k-1}^T d_{k-1} \leqslant g_k^T d_{k-1} &\leqslant -\eta g_k^T d_{k-1}\\ &\leqslant \frac{\eta}{(1-\eta)}\|g_k\|^2\end{aligned}$$

which together with (2.85) imply

$$\|d_k\|^2 \leqslant \frac{(1+\eta)}{(1-\eta)}\|g_k\|^2 + \beta_k^2\|d_{k-1}\|^2. \tag{2.86}$$

Notice that

$$\begin{aligned}\|d_k\|^2 &\leqslant \frac{(1+\eta)}{(1-\eta)}\|g_k\|^2 + \beta_k^2\left(\frac{(1+\eta)}{(1-\eta)}\|g_{k-1}\|^2 + \beta_{k-1}^2\|d_{k-2}\|^2\right)\\ &= \frac{(1+\eta)}{(1-\eta)}\|g_k\|^2 + \frac{(1+\eta)}{(1-\eta)}\frac{\|g_k\|^4}{\|g_{k-1}\|^4}\|g_{k-1}\|^2 +\\ &\quad \beta_{k-1}^2\beta_{k-2}^2\|d_{k-2}\|^2\\ &= \frac{(1+\eta)}{(1-\eta)}\frac{\|g_k\|^4}{\|g_k\|^2} + \frac{(1+\eta)}{(1-\eta)}\frac{\|g_k\|^4}{\|g_{k-1}\|^2} +\\ &\quad \frac{\|g_k\|^4}{\|g_{k-2}\|^2}\|d_{k-2}\|^2\end{aligned} \tag{2.87}$$

which, when applied recursively, gives us

$$\|d_k\|^2 \leqslant \frac{(1+\eta)}{(1-\eta)}\|g_k\|^4 \sum_{j=1}^{k} \|g_j\|^{-2}. \tag{2.88}$$

The rest of the proof is typical. We consider θ_k, the angle between $-g_k$ and d_k. Assuming that the thesis is not valid we will show that θ_k are bounded away from $\pi/2$ and this contradicts Theorem 2.1.

Notice that from (2.82), (2.83) we have

$$\cos\theta_k \geqslant \frac{(1-2\eta)}{(1-\eta)}\frac{\|g_k\|}{\|d_k\|}. \tag{2.89}$$

Furthermore, since x_k belong to a bounded set and g is continuous, there exists a positive number c such that $c < \infty$ and

$$\|d_k\|^2 \leqslant ck \tag{2.90}$$

(from (2.88)). Assume now that the thesis is false, thus we can find a positive number c_1 such that

$$\|g_k\|^2 \geqslant c_1 \tag{2.91}$$

for all k. Taking into account (2.89)–(2.91) we come to the relation

$$\cos^2\theta_k \geqslant \frac{c_1}{ck}.$$

However, this also means that

$$\sum_{k=1}^{\infty} \cos^2\theta_k \geqslant \frac{c_1}{c}\sum_{k=1}^{\infty}\frac{1}{k} = \infty. \tag{2.92}$$

From Theorem 2.1 we know that if (2.92) holds we cannot have (2.91). This concludes the proof. □

Al-Baali introduces the condition on η under which the Fletcher–Reeves conjugate gradient algorithm is globally convergent. Assuming that $\eta \in (0, \frac{1}{2})$ imposes strict curvature condition therefore it is not surprising that there were efforts to weaken this condition. Here, we state some results given by Dai and Yuan [42] (see also [44]).

Dai and Yuan showed that the condition on η can be weakened. They propose the following curvature condition

$$\eta_1 g_k^T d_k \leqslant g(x_k + \alpha_k d_k)^T d_k \leqslant -\eta_2 g_k^T d_k \tag{2.93}$$

with $0 < \mu < \eta_1 < 1$ and $\eta_2 > 0$. If $\eta_1 = \eta_2 \in (0, \frac{1}{2})$ then we have the curvature condition considered so far. Furthermore, (2.93) is the Wolfe curvature condition if $\eta_1 = \eta \in (0,1)$ and $\eta_2 = \infty$.

Their step-length conditions are general but according to them the global convergence of the Fletcher–Reeves algorithm holds if

$$\eta_1 + \eta_2 \leqslant 1. \tag{2.94}$$

Theorem 2.10. *Assume that* $\{x_k\}$ *is generated by* Algorithm 2.7 *with* β_k *defined by the Fletcher–Reeves formula (2.58). Furthermore, suppose that the assumptions (ii)–(iv) of* Theorem 2.1 *are satisfied,* α_k *is determined by conditions (2.12), (2.93) with* $0 < \mu < \eta_1 < 1$, $\eta_2 > 0$ *and gives a direction of descent. Then, the sequence* $\{x_k\}$ *is such that*

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

Proof. Due to the definition of the Fletcher–Reeves algorithm

$$-\frac{g_{k+1}^T d_{k+1}}{\|g_{k+1}\|^2} = 1 - \frac{g_{k+1}^T d_k}{\|g_k\|^2} \tag{2.95}$$

which together with (2.93) leads to

$$1 - \eta_2 \rho_k \leqslant \rho_{k+1} \leqslant 1 + \eta_1 \rho_k, \tag{2.96}$$

where

$$\rho_k = -\frac{g_k^T d_k}{\|g_k\|^2}. \tag{2.97}$$

The proof stated in [42] (which we follow) is similar to that proposed by Al-Baali in [2]. The main ingredients of the proof are steps which show that $\|d_k\|^2$ have a linear growth rate and that it leads to the contradiction with the thesis of Theorem 2.1. In order to persuade the first step assume that

$$g_k^T d_k \leqslant 0 \tag{2.98}$$

for all k. In order to guarantee the feasibility of the step-length selection procedure we have to consider the case: $g_k^T d_k = 0$. We assume that in this case we take $\alpha_k = 0$ which implies $x_{k+1} = x_k$ and $d_{k+1} = -g_{k+1} + d_k$. Therefore, at the point x_{k+1} (2.98) holds with the strict inequality.

From the first inequality in (2.96) we deduce that

$$\eta_2 \rho_k + \rho_{k+1} \geqslant 1 \tag{2.99}$$

for all k, which with the help of the Hölder inequality (cf. Appendix A) leads to

$$\rho_k^2 + \rho_{k+1}^2 \geqslant \frac{1}{1+\eta_2^2} = c_1 \tag{2.100}$$

for all k. On the other hand, the second inequality in (2.96) gives us

$$\rho_{k+1} \leqslant 1 + \eta_1 \rho_k \leqslant 1 + \eta_1 \left(1 + \eta_1 \rho_{k-1}\right)$$

thus

$$\rho_k < \frac{1}{1-\eta_1} \tag{2.101}$$

for all k.

Before proceeding further we have to show that condition (2.98) is satisfied for all k. Since for $k = 1$ the condition is obviously fulfilled we assume that it holds for all $k = 1, \ldots, \bar{k}$. If this happens then (2.96), (2.101) are true, and

$$\begin{aligned}\rho_{k+1} &\geqslant 1 - \eta_2 \rho_k > 1 - \eta_2 \frac{1}{1-\eta_1} \\ &= \frac{1-\eta_1-\eta_2}{1-\eta_1} \geqslant 0\end{aligned}$$

which means that (2.98) also holds for $k = \bar{k} + 1$ and by induction arguments we have shown that d_k satisfy (2.98).

As in the proof of Al-Baali's theorem now we look for the upper bound on $\|d_k\|^2$. From the Fletcher–Reeves rule we have

$$\begin{aligned}\|d_{k+1}\|^2 &= \|g_{k+1}\|^2 - 2\beta_{k+1} g_{k+1}^T d_k + \beta_{k+1}^2 \|d_k\|^2 \\ &\leqslant \|g_{k+1}\|^2 + \frac{\|g_{k+1}\|^4}{\|g_k\|^4} \|d_k\|^2 + 2\eta_1 \rho_k \|g_{k+1}\|^2 \\ &\leqslant \frac{\|g_{k+1}\|^4}{\|g_k\|^4} \|d_k\|^2 + \frac{1+2\eta_1}{1-\eta_1} \|g_{k+1}\|^2. \end{aligned} \tag{2.102}$$

Introducing

$$t_k = \frac{\|d_k\|^2}{\|g_k\|^4}$$

we can transform (2.102) into

$$t_{k+1} \leqslant t_k + \frac{1+2\eta_1}{1-\eta_1} \frac{1}{\|g_{k+1}\|^2} \tag{2.103}$$

Now, if the thesis is not satisfied, there exists $\varepsilon > 0$ such that

$$\|g_k\|^2 \geqslant \varepsilon \tag{2.104}$$

for all k. Taking into account (2.103)–(2.104) we come to

$$t_{k+1} \leqslant t_1 + c_2 k \tag{2.105}$$

with

$$c_2 = \frac{1}{\varepsilon} \frac{1+2\eta_1}{1-\eta_1}.$$

Next the contradiction to the thesis of Theorem 2.1 has to be established. It is not so straightforward as in the proof of Al-Baali's theorem due to more general step-length conditions. We consider separately odd and even cases of k. From (2.105) we have

$$\begin{aligned}\frac{1}{t_{2k}} &\geqslant \frac{1}{c_2(2k-1)+t_1} \\ \frac{1}{t_{2k-1}} &\geqslant \frac{1}{c_2(2k-2)+t_1} > \frac{1}{c_2(2k-1)+t_1}.\end{aligned}$$

On that basis

$$\begin{aligned}\sum_{k=1}^{\infty} cos^2\theta_k \|g_k\|^2 &= \sum_{k=1}^{\infty} \frac{\left(g_k^T d_k\right)^2}{\|d_k\|^2} \\ &= \sum_{k=1}^{\infty} \frac{\rho_k^2}{t_k} = \sum_{k=1}^{\infty} \left(\frac{\rho_{2k-1}^2}{t_{2k-1}} + \frac{\rho_{2k}^2}{t_{2k}} \right) \\ &\geqslant \sum_{k=1}^{\infty} \frac{\rho_{2k-1}^2 + \rho_{2k}^2}{c_2(2k-1)+t_1} \\ &\geqslant \sum_{k=1}^{\infty} \frac{c_1}{c_2(2k-1)+t_1} = \infty\end{aligned}$$

which is impossible. □

If condition (2.94) is not satisfied then the Fletcher–Reeves method with line search rules (2.12) may not converge. Dai and Yuan construct the counter-example for the case

$$\eta_1 + \eta_2 > 1. \tag{2.106}$$

Suppose that the curvature condition (2.93) is satisfied with

$$\begin{aligned}g_{k+1}^T d_k &= \eta_1 g_k^T d_k, \ \ k = 1, \ldots, N \\ g_{N+1}^T d_N &= -\eta_2 g_N^T d_N.\end{aligned}$$

Equation (2.95) implies that for the corresponding sequence $\{\rho_k\}$ the following holds

$$\begin{aligned}\rho_{k+1} &= 1 + \eta_1 \rho_k, \ \ k = 2, \ldots, N \\ \rho_{N+1} &= 1 - \eta_2 \rho_N.\end{aligned}$$

By taking into account $\rho_1 = 1$ this leads to

$$\rho_k = \frac{1-\eta_1^k}{1-\eta_1}, \quad k = 2, \ldots, N \tag{2.107}$$

$$\rho_{N+1} = 1 - \eta_2 \frac{1-\eta_1^N}{1-\eta_1} = \frac{1-\eta_1-\eta_2+\eta_2\eta_1^N}{1-\eta_1}. \tag{2.108}$$

Since we have assumed (2.106) we can choose N in such a way that

$$1 + \eta_2\eta_1^N < \eta_1 + \eta_2 \tag{2.109}$$

which together with (2.108) lead to the conclusion that $\rho_{N+1} < 0$ which, from (2.97), means that d_{N+1} is not a direction of descent. Therefore, we have shown that under (2.106) applying step-length selection procedure based on (2.12), (2.93) does not prevent us from increasing value of f.

The example is based on the function

$$f(x) = \frac{1}{2}x^2 \tag{2.110}$$

where $x \in \mathscr{R}$. We claim that it is possible to find step-sizes α_k which satisfy (2.12), (2.93), they are generated by the Fletcher–Reeves algorithm and lead to the sequence

$$x_k = \eta_1^k x_1, \quad k = 1, \ldots, N, \tag{2.111}$$

$$x_{N+1} = -\eta_2\eta_1^N x_1. \tag{2.112}$$

If $\{x_k\}$ is defined by (2.111)–(2.112) and f by (2.110) then

$$d_k = -\frac{1-\eta_1^k}{1-\eta_1} g_k, \quad k = 1, \ldots, N. \tag{2.113}$$

Equation (2.113) is obviously true for $k = 1$. Now assume that it holds for $k > 1$, we show that it is valid also for $k+1$. Indeed,

$$\begin{aligned} d_{k+1} &= -g_{k+1} + \eta_1^2 d_k = -g_{k+1} - \eta_1^2 \frac{1-\eta_1^k}{1-\eta_1} g_k \\ &= -g_{k+1} - \eta_1 \frac{1-\eta_1^k}{1-\eta_1} g_{k+1} \\ &= -\frac{1-\eta_1^{k+1}}{1-\eta_1} g_{k+1}. \end{aligned}$$

It remains to show that $\{x_k\}$ is generated in such a way that conditions (2.12), (2.93) are satisfied. First, we verify condition (2.12). Due to the fact that $g_k = x_k$ we can write

$$\begin{aligned} f(x_{k+1}) - f(x_k) &= \frac{1}{2}(x_{k+1} - x_k)(x_{k+1} - x_k) \\ &= \frac{1}{2} \frac{x_k + x_{k+1}}{g_k} \alpha_k d_k g_k \end{aligned}$$

$$= \begin{cases} \frac{1}{2}(1+\eta_1)\alpha_k d_k g_k & \text{if } k = 1,\ldots,N-1 \\ \frac{1}{2}(1-\eta_2)\alpha_k d_k g_k & \text{if } k = N \end{cases}$$
$$\leqslant \frac{1}{2}(1-\eta_2)\alpha_k d_k g_k \leqslant \mu \alpha_k d_k g_k.$$

Furthermore, from (2.111)–(2.113) we have

$$g_{k+1} d_k = \eta_1 d_k g_k, \quad k = 1,\ldots,N-1$$
$$g_{k+1} d_k = -\eta_2 d_k g_k.$$

thus conditions (2.93) are also fulfilled. Therefore, from (2.108)–(2.109) we conclude that d_{N+1} is a direction of ascent.

2.8 Other Versions of the Fletcher–Reeves Algorithm

The Fletcher–Reeves algorithm has one of the strongest convergence properties among all standard conjugate gradient algorithms. Yet it requires the strong Wolfe conditions for the directional minimization. The question is whether we can provide a version of the Fletcher–Reeves algorithm which is globally convergent under the Wolfe conditions. In some sense such a method has been proposed by Dai and Yuan [43].

First, Dai and Yuan look at the Fletcher–Reeves and Hestenes–Stiefel versions of conjugate gradient algorithms defined by the coefficients β_{k+1}:

$$\beta_{k+1}^{FR} = \frac{\|g_{k+1}\|^2}{\|g_k\|^2}$$
$$\beta_{k+1}^{HS} = \frac{g_k^T y_k}{d_k^T y_k}$$

where $y_k = g_{k+1} - g_k$. Observe that if the directional minimization is exact and f is a quadratic function then these two formulae lead to the same iterative algorithm.

Now, consider β_{k+1} defined by

$$\beta_{k+1}^{DY} = \frac{\|g_{k+1}\|^2}{d_k^T y_k} \tag{2.114}$$

as proposed by Dai and Yuan. It is also equivalent to β_{k+1}^{FR} and β_{k+1}^{HS} if f is quadratic. However, notice that if the directional minimization is exact then $\beta_{k+1}^{FR} = \beta_{k+1}^{DY}$ for all k since in this case $d_k^T(g_{k+1} - g_k) = \|g_k\|^2$. Therefore, the method discussed can be regarded as a version of the Fletcher–Reeves algorithm.

Before stating the main convergence result of this section notice that β_{k+1}^{YD} can be expressed differently:

$$\beta_{k+1}^{YD} = \frac{g_{k+1}^T d_{k+1}}{g_k^T d_k}, \tag{2.115}$$

which follows from the relation

$$\begin{aligned} g_{k+1}^T d_{k+1} &= -\|g_{k+1}\|^2 + \beta_{k+1}^{YD} g_{k+1}^T d_k \\ &= \beta_{k+1}^{YD} \left(-y_k^T d_k + g_{k+1}^T d_k\right) \\ &= \beta_{k+1}^{YD} g_k^T d_k. \end{aligned} \tag{2.116}$$

Theorem 2.11. *Suppose that the assumptions (ii)–(iv) of* Theorem 2.1 *are satisfied and the step-length* α_k *is determined by the Wolfe conditions with* $0 < \mu < \eta < 1$. *Then,* Algorithm 2.7 *with* β_k *defined by (2.114) generates the sequence* $\{x_k\}$ *such that*

$$\liminf_{k\to\infty} \|g_k\| = 0. \tag{2.117}$$

Proof. Assume that (2.117) is not satisfied, thus there exists $\varepsilon > 0$ such that

$$\|g_k\| \geqslant \varepsilon \tag{2.118}$$

for all k.

First, we show that d_k is a direction of descent. Since α_k satisfies the Wolfe curvature condition we have

$$(g_{k+1} - g_k)^T d_k \geqslant (\eta - 1) g_k^T d_k.$$

Because

$$g_k^T d_k < 0 \tag{2.119}$$

holds for $k = 1$, assuming that (2.119) is true for $k > 1$ we have to show that it is also satisfied for $k+1$. From (2.116) and (2.119) we have that d_{k+1} is also a direction of descent.

In order to complete the proof we aim at showing that (2.118) leads to the contradiction with Theorem 2.1. Indeed, the rule for d_{k+1} can be written as

$$d_{k+1} + g_{k+1} = \beta_{k+1}^{DY} d_k.$$

Squaring it gives us

$$\|d_{k+1}\|^2 = \left(\beta_{k+1}^{DY}\right)^2 \|d_k\|^2 - 2 g_{k+1}^T d_{k+1} - \|g_{k+1}\|^2. \tag{2.120}$$

This result is then used to show that

$$\sum_{k=0}^{\infty} \frac{\left(g_{k+1}^T d_{k+1}\right)^2}{\|d_{k+1}\|^2} = \infty$$

which contradicts the thesis of Theorem 2.1.

Due to (2.115) and (2.120) we have

$$\begin{aligned}
\frac{\|d_{k+1}\|^2}{\left(g_{k+1}^T d_{k+1}\right)^2} &= \frac{\|d_k\|^2}{\left(g_k^T d_k\right)^2} - \frac{2}{g_{k+1}^T d_{k+1}} \\
&\quad - \frac{\|g_{k+1}\|^2}{\left(g_{k+1}^T d_{k+1}\right)^2} \\
&= \frac{\|d_k\|^2}{\left(g_k^T d_k\right)^2} - \left(\frac{1}{\|g_{k+1}\|} + \frac{\|g_{k+1}\|}{g_{k+1}^T d_{k+1}}\right)^2 \\
&\quad + \frac{1}{\|g_{k+1}\|^2} \\
&\leqslant \frac{\|d_k\|^2}{\left(g_k^T d_k\right)^2} + \frac{1}{\|g_{k+1}\|^2}. \qquad (2.121)
\end{aligned}$$

From (2.121) follows immediately that

$$\frac{\|d_k\|^2}{\left(g_k^T d_k\right)^2} \leqslant \sum_{i=1}^{k} \frac{1}{\|g_k\|^2}$$

and, since (2.118) holds,

$$\frac{\|d_k\|^2}{\left(g_k^T d_k\right)^2} \leqslant \frac{k}{\varepsilon^2}$$

which leads to

$$\sum_{k=1}^{\infty} \frac{\left(g_k^T d_k\right)^2}{\|d_k\|^2} = \infty.$$

Due to Theorem 2.1 this completes the proof. □

Eventually, we have provided a version of the Fletcher–Reeves algorithm which is globally convergent under the Wolfe conditions in the directional minimization and without assuming that $\eta < \frac{1}{2}$.

2.9 Global Convergence of the Polak–Ribière Algorithm

The work of Al-Baali inspired other researches to provide new versions of conjugate gradient algorithms. Gilbert and Nocedal [78] consider a conjugate gradient algorithm with coefficients β_k defined by

$$|\beta_k| \leqslant \beta_k^{FR} = \frac{\|g_k\|^2}{\|g_{k-1}\|^2}. \qquad (2.122)$$

The example of the sequence (2.122) is the following one

$$\beta_k = \begin{cases} -\beta_k^{FR} & \text{if } \beta_k^{PR} < -\beta_k^{FR} \\ \beta_k^{PR} & \text{if } |\beta_k^{PR}| \leqslant \beta_k^{FR} \\ \beta_k^{FR} & \text{if } \beta_k^{PR} > \beta_k^{FR} \end{cases}$$

where

$$\beta_k^{PR} = \frac{(g_k - g_{k-1})^T g_k}{\|g_{k-1}\|^2}. \tag{2.123}$$

Observe that β_k defined in that way leads to a conjugate gradient algorithm which does not have the drawback of the Fletcher–Reeves algorithm mentioned in Sect. 2.6. If $x_{k+1} \approx x_k$ then $\beta_k \approx 0$ instead of being close to one.

The global convergence of the conjugate gradient algorithm based on (2.123) follows from the familiar Al-Baali's analysis. First, we have to show that d_k is always a direction of descent if the strong Wolfe conditions are applied with $\eta < \frac{1}{2}$.

Lemma 2.6. *Suppose that the assumptions (ii)–(iv) of* Theorem 2.1 *hold and the step-length α_k satisfies the strong Wolfe conditions with* $0 < \eta < \frac{1}{2}$ *Then,* Algorithm 2.7 *with β_k defined by (2.122) generates directions d_k with the property*

$$-\frac{1}{1-\eta} \leqslant \frac{g_k^T d_k}{\|g_k\|^2} \leqslant \frac{2\eta - 1}{1-\eta} \tag{2.124}$$

for all k.

Proof. Equation (2.124) obviously holds for $k = 1$. Assume now that (2.124) is true for some $k > 1$. In order to apply the induction arguments we have to show that it also holds for $k+1$.

First, observe that since $0 < \eta < \frac{1}{2}$ we have

$$g_k^T d_k < 0.$$

Furthermore, the following holds

$$\begin{aligned} \frac{g_{k+1}^T d_{k+1}}{\|g_{k+1}\|^2} &= -1 + \beta_{k+1} \frac{g_{k+1}^T d_k}{\|g_{k+1}\|^2} \\ &= -1 + \frac{\beta_{k+1}}{\beta_{k+1}^{FR}} \frac{g_{k+1}^T d_k}{\|g_k)\|^2}. \end{aligned} \tag{2.125}$$

From the line search conditions we come to

$$|\beta_{k+1} g_{k+1}^T d_k| \leqslant -\eta |\beta_{k+1}| g_k^T d_k$$

which combined with (2.125) imply

$$-1+\eta\frac{|\beta_{k+1}|}{\beta_{k+1}^{FR}}\frac{g_k^T d_k}{\|d_k\|^2} \leqslant \frac{g_{k+1}^T d_{k+1}}{\|g_{k+1}\|^2}$$

$$\leqslant -1-\eta\frac{|\beta_{k+1}|}{\beta_{k+1}^{FR}}\frac{g_k^T d_k}{\|g_k\|^2}.$$

The induction hypothesis leads to

$$-1+\frac{|\beta_{k+1}|}{\beta_{k+1}^{FR}}\frac{\eta}{1-\eta} \leqslant \frac{g_{k+1}^T d_{k+1}}{\|g_{k+1}\|^2}$$

$$\leqslant -1-\frac{|\beta_{k+1}|}{\beta_{k+1}^{FR}}\frac{\eta}{1-\eta}$$

which, due to (2.122), proves (2.124) for $k+1$. □

Lemma 2.6 shows that (2.78) holds with $c=(1-2\eta)/(1-\eta)$ and that

$$c_1\frac{\|g_k\|}{\|d_k\|} \leqslant \cos\theta_k \leqslant c_2\frac{\|g_k\|}{\|d_k\|} \tag{2.126}$$

for all k with

$$c_1=\frac{1-2\eta}{1-\eta},\ \ c_2=\frac{1}{1-\eta}. \tag{2.127}$$

Equation (2.127) is essential in proving the global convergence of conjugate gradient algorithm if we can show that $\|d_k\|^2$ grows at most linearly when the gradients g_k are not too small.

Theorem 2.12. *Suppose that assumptions of* Lemma 2.6 *are fulfilled. then*

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

Proof. From Lemma 2.6 and the strong Wolfe curvature condition we deduce that

$$|g_k^T d_{k-1}| \leqslant -\eta g_{k-1}^T d_{k-1} \leqslant \frac{\eta}{1-\eta}\|g_{k-1}\|^2.$$

As in the proof of Theorem 2.9 we can show that (cf. (2.86))

$$\|d_k\|^2 \leqslant \left(\frac{1+\eta}{1-\eta}\right)\|g_k\|^2+\beta_k^2\|d_{k-1}\|^2.$$

Using this, as in (2.87)–(2.88), we come to the relation

$$\|d_k\|^2 \leqslant \gamma \|g_k\|^4 \sum_{j=1}^{k} \|g_j\|^{-2}$$

with $\gamma = (1+\eta)/(1-\eta)$. If we assume now that there exists $\varepsilon > 0$ such that

$$\|g_k\| \geqslant \varepsilon$$

for all k, then

$$\|d_k\|^2 \leqslant \gamma \varepsilon^2 k. \tag{2.128}$$

The rest of the proof follows the appropriate lines of the proof of Theorem 2.9. □

2.10 Hestenes–Stiefel Versions of the Standard Conjugate Gradient Algorithm

Another globally convergent algorithm was proposed in [91]. The direction d_k is determined according to the formula:

$$d_{k+1} = -g_{k+1} + \beta_{k+1}^{HZ} d_k \tag{2.129}$$

$$\beta_{k+1}^{HZ} = \frac{1}{d_k^T y_k} \left(y_k - 2 d_k \frac{\|y_k\|^2}{d_k^T y_k} \right)^T g_{k+1}. \tag{2.130}$$

The formula is related to the Hestenes–Stiefel version of the standard conjugate gradient algorithm. Assume that the line search is exact. Then we have $d_k^T g_{k+1} = 0$ and β_{k+1}^{HZ} becomes

$$\beta_{k+1}^{HZ} = \frac{g_{k+1}^T y_k}{d_k^T y_k}$$

which is the coefficient in the Hestenes–Stiefel formula (cf. (2.60)).

The formula (2.130) is also related to the formerly proposed scheme by Dai and Liao [41]:

$$\beta_{k+1}^{DL} = \frac{1}{d_k^T y_k} (y_k - t s_k)^T g_{k+1} \tag{2.131}$$

with $t > 0$ as a constant parameter. Equation (2.130) can be regarded as an adaptive version of (2.131) with $t = 2\|y_k\|^2 / s_k^T d_k$.

More interesting is that the conjugate gradient algorithm defined by (2.129)–(2.130) can be treated as a version of the Perry–Shanno memoryless quasi-Newton method as discussed in the next chapter (cf. (3.19)). If d_{k+1}^{PS} is a direction according to the Perry–Shanno formula and d_{k+1}^{HZ} is given by (2.129)–(2.130), then

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

and

$$d_{k+1}^{HZ} = \gamma_k \left(d_{k+1}^{PS} + \frac{d_k^T g_{k+1}}{d_k^T y_k} y_k\right). \tag{2.132}$$

Indeed we have to show that

$$\begin{aligned} &-\gamma_k g_{k+1} + \gamma_k \left[\frac{1}{d_k^T y_k}\left(y_k - 2d_k \frac{1}{\gamma_k}\right)^T g_{k+1}\right]^T d_k + \gamma_k \frac{d_k^T g_{k+1}}{d_k^T y_k} \\ &= -\gamma_k g_{k+1} - \left[\left(1+\gamma_k \frac{1}{\gamma_k}\right)\frac{d_k^T g_{k+1}}{d_k^T y_k} - \gamma_k \frac{y_k^T g_{k+1}}{d_k^T y_k}\right]^T d_k + \gamma_k \frac{d_k^T g_{k+1}}{p_k^T y_k} \end{aligned}$$

which obviously holds.

Thus, if the second term in the bracket of (2.132) can be neglected then the direction d_k^{HZ} is a multiple of d_k^{PS}. Hager and Zhang show in [91] that this is the case if f is strongly convex and the cosine of the angle between g_{k+1} and d_k is small.

The direction d_{k+1} is a direction of descent if instead of β_{k+1}^{HZ} in (2.129) we use $\hat{\beta}_{k+1}^{HZ}$ that avoids low nonnegative values of β_k^{HZ} – the new algorithm is as follows:

$$d_{k+1} = -g_{k+1} + \hat{\beta}_{k+1}^{HZ} d_k \tag{2.133}$$

$$\hat{\beta}_{k+1}^{HZ} = \max\left[\beta_{k+1}^{HZ}, \eta_k\right], \quad \eta_k = \frac{-1}{\|d_k\| \min\left[\eta, \|g_k\|\right]} \tag{2.134}$$

where $\eta > 0$. Notice that when $\{d_k\}$ is bounded and $\|g_k\|$ assume small values then η_k have big negative values and we regard $\hat{\beta}_{k+1}^{HZ}$ and β_{k+1}^{HZ} as equal.

The introduction of $\hat{\beta}_{k+1}^{HZ}$ is motivated in [91] by the need of having the scheme which guarantees global convergence properties of algorithm (2.130), (2.133)–(2.134) with the Wolfe line search rules. The convergence analysis given in [91] breaks down when β_{k+1}^{HZ} assume negative values. We could overcome the problem by taking

$$\hat{\beta}_{k+1}^{HZ} = \max[\beta_{k+1}^{HZ}, 0]$$

however that would impair the efficiency of the method. The other formula which comes to mind is

$$\hat{\beta}_{k+1}^{HZ} = |\beta_{k+1}^{HZ}| \tag{2.135}$$

but in this case the essential result in [91]–Theorem 2.13 stated below couldn't be proved. Notice that this choice for coefficients β_k is natural since coefficients β_k are always positive when f is quadratic and nonlinear conjugate gradient algorithms rely on the quadratic approximation of f near a solution. The presentation of the Hager–Zhang algorithm is based on the formula (2.134).

The global convergence properties of algorithm (2.130), (2.133)–(2.134) refer to the important property of d_k being always a direction of descent.

Theorem 2.13. *Suppose that $d_k^T y_k \neq 0$ and*

$$d_{k+1} = -g_{k+1} + \tau d_k, \;\; d_1 = -g_1,$$

for any $\tau \in \left[\beta_{k+1}^{HZ}, \max\left[\beta_{k+1}^{HZ}, 0\right]\right]$, then

$$g_{k+1}^T d_{k+1} \leqslant -\frac{7}{8}\|g_{k+1}\|^2. \tag{2.136}$$

Proof. We follow the proof stated in [91]. Suppose first that $\tau = \beta_{k+1}^{HZ}$. Since for $k = 1$ (2.136) is obviously satisfied, assume that for $k > 1$. Then, we have

$$\begin{aligned}
g_{k+1}^T d_{k+1} &= -\|g_{k+1}\|^2 + \beta_{k+1}^{HZ} g_{k+1}^T d_k \\
&= -\|g_{k+1}\|^2 + g_{k+1}^T d_k \left(\frac{y_k^T g_{k+1}}{d_k^T y_k} - 2\frac{\|y_k\|^2 g_{k+1}^T d_k}{\left(d_k^T y_k\right)^2} \right) \\
&= \frac{y_k^T g_{k+1}\left(d_k^T y_k\right)\left(g_{k+1}^T d_k\right) - \|g_{k+1}\|^2 \left(d_k^T y_k\right)^2 - 2\|y_k\|^2 \left(g_{k+1}^T d_k\right)^2}{\left(d_k^T y_k\right)^2} \\
&= \frac{u^T v - 4\|u\|^2 - \frac{1}{2}\|v\|^2}{\left(d_k^T y_k\right)^2}
\end{aligned}$$

where

$$u = \frac{1}{2}\left(d_k^T y_k\right) g_{k+1}, \;\; v = 2\left(g_{k+1}^T d_k\right) y_k. \tag{2.137}$$

Taking into account the inequality

$$u^T v \leqslant \frac{1}{2}\left(\|u\|^2 + \|v\|^2\right),$$

from (2.137) we obtain

$$g_{k+1}^T d_{k+1} \leqslant -\frac{7}{2}\frac{\|u\|^2}{\left(d_k^T y_k\right)^2} = -\frac{7}{8}\|g_{k+1}\|^2.$$

The case when $\tau \neq \beta_{k+1}^{HZ}$ and $\beta_{k+1}^{HZ} \leqslant \tau < 0$ can be proved by noting that

$$g_{k+1}^T d_{k+1} = -\|g_{k+1}\|^2 + \tau g_{k+1}^T d_k \leqslant -\|g_{k+1}\|^2 + \beta_{k+1}^{HZ} g_{k+1}^T d_k \qquad (2.138)$$

provided that $g_{k+1}^T d_k < 0$. If, on the other hand $g_{k+1}^T d_k > 0$, then (2.136) follows directly from the equality in (2.138). $\square$

Obviously, for $\tau = \beta_{k+1}^{HZ}$, d_k is a direction of descent satisfying (2.136) as is the case for the scheme defined by (2.133), (2.130) and (2.134) since

$$\hat{\beta}_{k+1}^{HZ} = \max\left[\beta_{k+1}^{HZ}, \eta_k\right] \in \left[\hat{\beta}_{k+1}^{HZ}, \max\left[\beta_{k+1}^{HZ}, 0\right]\right]$$

which follows from the fact that η_k is negative.

Hager and Zhang prove in [91] global convergence of their method in the sense of condition (2.84). The quite elaborate proof borrows from the analysis by Gilbert and Nocedal given in [78].

Theorem 2.14. *Suppose that the assumptions (i)–(ii) of* Theorem 2.7 *are satisfied. Then,* Algorithm 2.7 *with β_k defined by (2.130), (2.133)–(2.134) and with the Wolfe line search rules generates the sequence $\{x_k\}$ such that*

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

Instead of providing the proof of the theorem we show, following [91], the analysis of the algorithm applied to strongly convex functions.

Theorem 2.15. *Suppose that f is strongly convex and Lipschitz continuous on the level set defined in (i) of* Theorem 2.7. *That is, there exist constants $L < \infty$ and $\nu > 0$ such that (2.26) holds and*

$$(g(x) - g(y))^T (x-y) \geqslant \nu \|x-y\| \qquad (2.139)$$

for all x, $y \in \mathscr{M}$. Then for the scheme (2.129)–(2.130) and with the Wolfe line search rules

$$\lim_{k\to\infty} \|g_k\| = 0.$$

Proof. Following the proof of Theorem 2.1 we can show that

$$\alpha_k \geqslant \frac{(1-\eta)}{L} \frac{\left\|g_k^T d_k\right\|}{\|d_k\|^2} \qquad (2.140)$$

(cf. (2.29)).

Then, by the strong convexity assumption we have:

$$y_k^T d_k \leqslant \nu \alpha_k \|d_k\|^2. \qquad (2.141)$$

Theorem 2.13 implies that if $g_k \neq 0$ then $d_k \neq 0$, and (2.141) shows that $y_k^T d_k > 0$. From the strong convexity assumption we know that f is bounded from below. Therefore, from the first Wolfe condition (2.46) we conclude that

$$\sum_{k=1}^{\infty} \alpha_k g_k^T d_k > -\infty. \tag{2.142}$$

Now, we take into account (2.142), (2.140) and (2.136) to obtain

$$\sum_{k=1}^{\infty} \frac{\|g_k\|^4}{\|d_k\|^2} < \infty. \tag{2.143}$$

In order to conclude the proof we have to show that $\|d_k\| \leqslant c\|g_k\|$ for some positive constant c. To this end notice that

$$\|y_k\| \leqslant L\alpha_k \|d_k\|$$

and thus we can write

$$\begin{aligned}
\left|\beta_{k+1}^{HZ}\right| &= \left| \frac{y_k^T g_{k+1}}{d_k^T y_k} - 2\frac{\|y_k\|^2 d_k^T g_{k+1}}{\left(d_k^T y_k\right)^2} \right| \\
&\leqslant \frac{\|y_k\| \|g_{k+1}\|}{\nu \alpha_k \|d_k\|^2} + 2\frac{\|y_k\|^2 \|d_k\| \|g_{k+1}\|}{\nu^2 \alpha_k^2 \|d_k\|^4} \\
&\leqslant \frac{L\alpha_k \|d_k\| \|g_{k+1}\|}{\nu \alpha_k \|d_k\|^2} + 2\frac{L^2 \alpha_k^2 \|d_k\|^3 \|g_{k+1}\|}{\nu^2 \alpha_k^2 \|d_k\|^4} \\
&\leqslant \left(\frac{L}{\nu} + \frac{2L^2}{\nu^2}\right) \frac{\|g_{k+1}\|}{\|d_k\|}.
\end{aligned} \tag{2.144}$$

Eventually, we have

$$\|d_{k+1}\| \leqslant \|g_{k+1}\| + \left|\beta_{k+1}^{HZ}\right| \|d_k\| \leqslant c\|g_{k+1}\|$$

with $c = 1 + L/\nu + 2L^2/\nu^2$ and from (2.143) also

$$\sum_{k=1}^{\infty} \|g_k\|^2 < \infty$$

which concludes the proof. □

Notice that Theorem 2.15 concerns the scheme defined by (2.129)–(2.130). It does not hold when we use (2.130), (2.133)–(2.134), but in that case we can refer to Theorem 2.14. Furthermore, $\hat{\beta}_{k+1}^{HZ}$ is defined in such a way that when x_k is close to a solution then η_k tends to $-\infty$ and thus $\hat{\beta}_{k+1}^{HZ} = \beta_{k+1}^{HZ}$. From the analysis given in the proof it follows also that the scheme with (2.133), (2.130) and (2.135) would be convergent for strongly convex functions, if d_{k+1} were a direction of descent.

Now we are in the position to show that if the cosine of the angle between d_k and g_{k+1} is small then the second term in the brackets in formula (2.132) can be neglected. Hager and Zhang estimate the norm of the second term in the brackets and the norm of d_{k+1} with respect to the norm of g_{k+1} They evaluate

$$\frac{|d_k^T g_{k+1}|}{|d_k^T y_k|} \leqslant c_1 \varepsilon \|g_{k+1}\| \tag{2.145}$$

$$\|d_{k+1}\| \geqslant \sqrt{1 - 2c_2\varepsilon}\|g_{k+1}\| \tag{2.146}$$

where c_1, c_2 are some positive numbers and ε is the cosine of the angle between d_k and g_{k+1}. As a result we have

$$\frac{|d_k^T g_{k+1}|}{|d_k^T y_k|} \frac{\|y_k\|}{\|d_{k+1}\|} \leqslant \frac{c_1 \varepsilon}{\sqrt{1 - 2c_2\varepsilon}} \tag{2.147}$$

and if ε is small then the left-hand side in (2.147) is also small. To show (2.145)–(2.146) we refer to condition (2.139) (and to (2.141)). We have

$$\frac{|d_k^T g_{k+1}|}{|d_k^T y_k|} \leqslant \frac{L}{\nu} \left|t_k^T g_{k+1}\right| \leqslant c_1 \varepsilon \|g_{k+1}\|,$$

where $t_k = d_k / \|d_k\|$. Furthermore, from (2.129),

$$\begin{aligned} \|d_{k+1}\|^2 &= \|g_{k+1}\|^2 - 2\beta_{k+1}^{HZ} d_k^T g_{k+1} + \left(\beta_k^{HZ}\right)^2 \|d_k\|^2 \\ &\geqslant \|g_{k+1}\|^2 - 2\beta_k^{HZ} d_k^T g_{k+1}, \end{aligned} \tag{2.148}$$

and from (2.144) we also have

$$\left|\beta_k^{HZ} d_k^T g_{k+1}\right| \leqslant c_2 \left|t_k^T g_{k+1}\right| \|g_{k+1}\| = c_2 \varepsilon \|g_{k+1}\|^2. \tag{2.149}$$

Equation (2.149) combined with (2.148) leads to (2.146).

On the basis of this analysis we can conclude that the Hager–Zhang algorithm achieves stronger convergence by impairing conjugacy.

2.11 Notes

The chapter is the introduction to nonlinear conjugate gradient algorithms. Our main interests is to establish the global convergence properties of several versions of the standard conjugate gradient algorithm such as Fletcher–Reeves, Polak–Ribière, or Hestenes–Stiefel.

Although it was quickly recognized that the Polak–Ribière is the most efficient conjugate gradient algorithm much efforts have been devoted to the analysis of the Fletcher–Reeves versions. This was mostly due to the very interesting result by Al-Baali [2] which opened the possibility of analyzing the conjugate gradient algorithms with the common feature of having $\|g_{k+1}\|^2$ in the formula for β_{k+1}. We provide original prove of the Al-Baali convergence theorem. The proofs of the related convergence results in Sects. 8–9 follow the same pattern and their ingenuities lie in constructing the proper sequences which, under the assumption that a method is not convergent, lead to the contradiction with the thesis of Theorem 2.1. While reporting these results we refer to ingenuities of the original proofs.

Some other schemes for updating β_{k+1} along the Fletcher–Reeves scheme have been proposed, for example Fletcher considered the method, which he called *conjugate descent* which is based on the formula [71]

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

In fact, if we look for a conjugate gradient algorithm which is derived directly from its quadratic counterpart we have two choices for the numerator in β_{k+1}: $\|g_{k+1}\|^2$ and $g_{k+1}^T y_k$, and three choices for the denominator: $\|g_k\|^2$, $-d_k^T g_k$ and $d_k^T y_k$ [92].

Consider now the literature on the convergence of the Polak–Ribière versions of a conjugate gradient algorithm. In that case we have the negative result by Powell [169] who showed, using 3 dimensional example, that even with exact line search, the Polak–Ribière algorithm can cycle indefinitely without converging to a stationary point (see also [40]). As it was shown earlier in [168] the crucial condition guaranteeing the convergence (under the exact line search assumption) is

$$\lim_{k\to\infty} \|x_{k+1} - x_k\| = 0.$$

This condition (often erroneously stated that step-sizes go to zero), which is difficult to ensure if the Wolfe conditions are applied in the line search, hampered attempts to construct a globally convergent conjugate gradient algorithm with the Polak–Ribière scheme. In Sect. 9 we show in fact a hybrid method which uses the Fletcher–Reeves update scheme to ensure its convergence. Another way of improving the global convergence properties of the Polak–Ribière versions is to employ different line search rules. Such an approach is followed in [88] (see also [89]) where a new Armijo rule is proposed – α_k is defined as

$$\alpha_k = \max\left[\beta^j \frac{\tau\,|d_k^T g_k|}{\|d_k\|^2}\right]$$

with $j \geqslant 0$ being the smallest integer number which leads to

$$f(x_{k+1}) - f(x_k) \leqslant -\mu\alpha_k\|d_k\|^2$$
$$-\eta_1\|g_{k+1}\|^2 \leqslant g_{k+1}^T d_k \leqslant \eta_2\|g_{k+1}\|^2$$

where $\beta,\ \mu \in (0,1)$ and $0 < \eta_2 < 1 < \eta_1$. In another venue of the same spirit they analyzed using trust region approach combined with the Polak–Ribière version of a conjugate gradient method [81].

As far as the other possibilities are concerned among all 6 combinations of numerators and denominators in coefficients β_{k+1} we have to refer to the method analyzed by Liu and Storey in [124] and based on the updating rule

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

Furthermore, we need to stress that the method discussed in Sect. 10 is in fact the Hestenes–Stiefel conjugate gradient algorithm with

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

provided that directional minimization is exact. The analysis of the method is provided closely following [91].

We end up the review of the standard nonlinear conjugate gradient algorithms by noting that the presented versions can be combined. Nazareth [143] introduces a two-parameter family of conjugate gradient methods:

$$\beta_{k+1} = \frac{\lambda_k \|g_{k+1}\|^2 + (1-\lambda_k) g_{k+1}^T y_k}{\tau_k \|g_k\|^2 + (1-\tau_k) d_k^T y_k}$$

where $\lambda_k, \tau_k \in [0,1]$. In [45] an even wider, a three parameter family is considered.

All versions of the conjugate gradient algorithm considered in this chapter we call standard ones. They are derived from Algorithm 1.4 presented in Chap. 1. It seems natural to look for extensions of the general conjugate gradient algorithm formulated as Algorithm 1.5. Surprisingly, that approach was first applied in nondifferentiable optimization so we postpone the presentation of the extension of the method of shortest residuals to Chap. 5 where the basic concepts of nondifferentiable optimization are also introduced.

Conjugate Gradient Algorithms in Nonconvex Optimization
Pytlak, R.
2009, XXVI, 478 p., Hardcover
ISBN: 978-3-540-85633-7