Available online at www.sciencedirect.com
ScienceDirect
Journal of Computational and Applied Mathematics 210 (2007) 167–174
JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS
www.elsevier.com/locate/cam

Low complexity secant quasi-Newton minimization algorithms for nonconvex functions

Carmine Di Fiore*, Stefano Fanelli, Paolo Zellini

Dipartimento di Matematica, Università di Roma "Tor Vergata", Via della Ricerca Scientifica, 1 00133 Roma, Italy

Received 14 September 2005; received in revised form 11 July 2006

Abstract

In this work some interesting relations between results on basic optimization and algorithms for nonconvex functions (such as BFGS and secant methods) are pointed out. In particular, some innovative tools for improving our recent secant BFGS-type and $\mathscr{L}QN$ algorithms are described in detail.

MSC: 51M04; 65H20; 65F30; 90C53

Keywords: Gnomon; Secant methods; Global convergence; Global optimization

1. Introduction

In nonconvex optimization there is a strong need to find out global convergence theorems for quasi-Newton methods. On one hand, this work stresses the relationship between the classical one-dimensional secant algorithm and the recent n-dimensional BFGS-type secant methods [5]. On the other hand, it is shown that a *discrete weak convexity assumption* plays a fundamental role in the convergence of several methods. In Section 2 we first recall how the discrete character of some classical procedures has an original link with elementary constructions of ancient geometry, giving a historical interpretation of *regula falsi* method for a given function $g(x)$. Then, a general convergence theorem is shown for the classical one-dimensional secant method with weaker assumptions on $g(x)$. Section 3 deals with BFGS-type n-dimensional methods by proving a global convergence theorem for the secant case. Section 4 shows a representation theorem for the descent direction of any minimization method. Finally, a general theorem ensuring the convergence to the global minimum of functions with continuous second derivatives is illustrated in Section 5.

2. Gnomon, regula falsi, secant method

In order to solve an equation $g(x) = 0$, where g is a continuous function, one can use the *regula falsi* technique. Given two approximations x_0 and x_1 of x_*, $g(x_*) = 0$, such that $g(x_0)g(x_1) < 0$, the next approximation x_2 is defined

* Corresponding author.
E-mail address: difiore@axp.mat.uniroma2.it (C. Di Fiore).

0377-0427/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2006.10.060

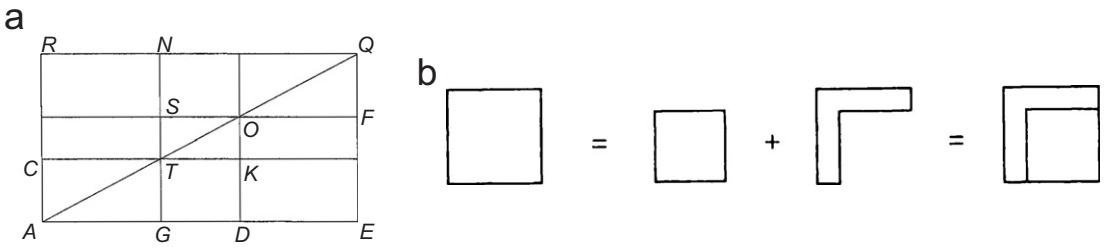

Fig. 1. (a) Regula falsi via gnomon; (b) square gnomon.

by the formula

$$x_2 = \frac{g(x_1)x_0 - g(x_0)x_1}{g(x_1) - g(x_0)},$$

i.e., x_2 is the intersection with the real axis of the line through $(x_0, g(x_0))$ and $(x_1, g(x_1))$. The following geometrical arguments, which use the gnomon theorem, give an interpretation of *regula falsi* applied to a *linear* equation $ax - b = 0$, $a \neq 0$, where we obviously have $x_2 = x_* = b/a$ [13]. The approximations x_0 and x_1 of x_* can be thought as the lengths of two aligned segments AG and AE (see Fig. 1(a)). The (computable) quantities ax_0 and ax_1 can be then thought as the values of the line ax in x_0 and x_1, i.e., the lengths of the segments GT and EQ, respectively. One has to find a segment AD=x_* such that the length of DO is equal to b. By the Gnomon Theorem (Euclid, I, 43) the areas of the rectangles NO and KF are equal. So,

$$\frac{FC + RS}{TS + FQ} = \frac{RK}{TS + FQ} = x_*.$$

But, in the same time, if $g(x) = ax - b$, then

$$\frac{FC + RS}{TS + FQ} = \frac{(b - ax_0)x_1 + (ax_1 - b)x_0}{(b - ax_0) + (ax_1 - b)} = \frac{g(x_1)x_0 - g(x_0)x_1}{g(x_1) - g(x_0)}.$$

Thus, $x_2^{\text{regula falsi}} \equiv (g(x_1)x_0 - g(x_0)x_1)/(g(x_1) - g(x_0)) = x_*$.

The Gnomon Theorem, together with the constructions of Euclidean *geometry* based on gnomon additions or subtractions like in Fig. 1(b), is the very basis of *algebraic*, iterative methods for solving systems of nonlinear equations. Historically, the methods due to Viète, Newton, Raphson, as well as secant algorithm and *regula falsi*, depend on the elementary increasing or reducing the area of a square (Fig. 1(b)) [13]. Problems (in modern terms: equations) of various degrees have been solved by elementary versions of Newton or secant-type methods by the gnomon scheme in old Babylonian, Chinese and Indian mathematics, in Greek and Arabian traditions [1,8,9,12,13], in Fibonacci's *Liber Abaci*, in Italian algebraic treatises of XVI century and, finally, in modern computer theory [13].

The *regula falsi* method converges by assuming convexity or concavity and has, in general, a linear rate of convergence. An improvement of *regula falsi* is the well known secant method. Let us consider the secant iterative formula when applied for computing the zeroes x_* of $g(x)$:

$$x_{k+1} = \frac{g(x_k)x_{k-1} - g(x_{k-1})x_k}{g(x_k) - g(x_{k-1})}, \quad g(x_k) \neq g(x_{k-1}). \tag{1}$$

Observe that if $g(x) = f'(x)$ for a differentiable function f, then (1) can be rewritten as $x_{k+1} = x_k - f'(x_k)/a_k$ where a_k is defined by the equation

$$a_k(x_k - x_{k-1}) = f'(x_k) - f'(x_{k-1}). \tag{2}$$

We will see the importance of the *secant* equation (2) in order to extend the application of the secant method to functions of several variables.

In the following Theorem 1 we give a general convergence result for the secant 1-d method with *no assumptions on the derivatives of the function* $g(x)$ *in* x_*. An essential key in the proof of Theorem 1 is the following representation of

the error $x_{k+1} - x_*$, in terms of the first and second order divided differences of g. Note that this representation extends to the nondifferentiable case the well known expression for the error involving g' and g''.

$$x_{k+1} - x_* = (x_k - x_*) - \frac{g(x_k)}{g[x_{k-1}, x_k]} = (x_k - x_*) - \frac{g[x_k, x_*]}{g[x_{k-1}, x_k]}\,(x_k - x_*)$$

$$= (x_k - x_*)(x_{k-1} - x_*)\,\frac{g[x_{k-1}, x_k, x_*]}{g[x_{k-1}, x_k]}.$$

Theorem 1 (*Secant 1-d convergence*). *Assume* $g(x) \in C^0(a, b)$: $g(a)g(b) < 0$, x_*: $g(x_*) = 0$. *Let* $g[\xi_0, \ldots, \xi_j]$ *denote Newton's divided differences of order j. Let* γ *be the positive root of the equation* $\gamma^2 = \gamma + 1$, *i.e.*, $\gamma \simeq 1.618\ldots$.

If $\exists k*$ *such that the settings*

$\overline{I}_{x_*,\varrho} = \{x : |x - x_*| \leqslant \rho\}$ *where* $\varrho = \max\{|x_{k^*} - x_*|, |x_{k^*+1} - x_*|\}$,
$M_1 = \sup_{x,y \in \overline{I}_{x_*,\varrho}\, x, y \neq x_*} |g[x, y, x_*]| < +\infty$,
$m_1 = \inf_{x,y \in \overline{I}_{x_*,\varrho}} |g[x, y]| > 0$,
$d_k = |x_k - x_*| M_1/m_1$

imply $d_{k^*} < 1$, $\quad d_{k^*+1} < d_{k^*}^{\gamma}$,

then the secant 1*-d sequence* (1) *is s.t.* $|x_k - x_*| < |x_{k^*} - x_*|$, $\forall k \geqslant k^*$, *and*

$$\lim_{k\to\infty} x_k = x_*.$$

Moreover, the order p of convergence is at least γ.

Proof. Let us prove by a *parallel inductive* procedure that $\forall i \geqslant 0$ s.t. $x_{k^*+i} \neq x_*$

$$d_{k^*+i} \leqslant d_{k^*}^{\gamma^i}, \quad x_{k^*+i} \in \overline{I}_{x_*,\varrho}.$$

For $i = 0, 1$ thesis is verified. By hypothesis we have $d_{k^*} \leqslant d_{k^*}^{\gamma^0}$, $d_{k^*+1} \leqslant d_{k^*}^{\gamma}$ and $x_{k^*} \in \overline{I}_{x_*,\varrho}$, $x_{k^*+1} \in \overline{I}_{x_*,\varrho}$. Assume the thesis is true for i and prove it for $i + 1$. By definition and the inductive assumption

$$d_{k^*+i+1} = \frac{M_1}{m_1}|x_{k^*+i+1} - x_*| = \frac{M_1}{m_1}\,|x_{k^*+i} - x_*||x_{k^*+i-1} - x_*|\,\frac{|g[x_{k^*+i-1}, x_{k^*+i}, x_*]|}{|g[x_{k^*+i-1}, x_{k^*+i}]|}$$

$$\leqslant \frac{M_1}{m_1}\,|x_{k^*+i} - x_*|\,\frac{M_1}{m_1}\,|x_{k^*+i-1} - x_*| = d_{k^*+i}d_{k^*+i-1} < d_{k^*}^{\gamma^i} d_{k^*}^{\gamma^{i-1}} = d_{k^*}^{\gamma^{i-1}\gamma^2} = d_{k^*}^{\gamma^{i+1}}$$

being $\gamma^2 = \gamma + 1$. Then

$$|x_{k^*+i+1} - x_*| = d_{k^*+i+1}\,\frac{m_1}{M_1} < d_{k^*}^{\gamma^{i+1}}\,\frac{m_1}{M_1} < d_{k^*}\,\frac{m_1}{M_1} = |x_{k^*} - x_*| < \varrho.$$

Thus, $x_{k^*+i+1} \in \overline{I}_{x_*,\varrho}$, $\{x_k\} \in \overline{I}_{x_*,\varrho}$, $\forall k \geqslant k^*$, and

$$\lim_{k\to\infty} |x_k - x_*| = \lim_{i\to\infty} |x_{k^*+i} - x_*| = \frac{m_1}{M_1} \lim_{i\to\infty} d_{k^*+i} \leqslant \frac{m_1}{M_1} \lim_{i\to\infty} d_{k^*}^{\gamma^i} = 0.$$

Let us prove now that the convergence order of the method is at least γ. By the assumptions we have $\forall k > k^*$ (if $x_k \neq x_*$)

$$|x_{k+1} - x_*| \leqslant |x_k - x_*||x_{k-1} - x_*|\,\frac{M_1}{m_1},$$

which proves that the order of convergence is superlinear $\Rightarrow$ $p > 1$. Assume $p = \gamma + \varepsilon$, ε real unknown, and set $y_k = |x_{k+1} - x_*|/|x_k - x_*|^p$. Then $\lim_{k\to\infty} y_k = l$, $0 < l < +\infty$, and

$$y_k \leqslant |x_k - x_*|^{1-p}|x_{k-1} - x_*|\,\frac{M_1}{m_1} = y_{k-1}^{1-p}|x_{k-1} - x_*|^{p(1-p)+1}\,\frac{M_1}{m_1}.$$

Since $y_{k-1}^{1-p} \to l^{1-p}$, $k \to \infty$, it must be $p(1-p)+1 \leqslant 0$, i.e. $(\gamma+\varepsilon)(1-\gamma-\varepsilon)+1 \leqslant 0$. From this inequality, being $\gamma^2 = \gamma + 1$, one deduces that $\varepsilon \leqslant 1-2\gamma$ or $\varepsilon \geqslant 0$. But $\varepsilon \leqslant 1-2\gamma \Rightarrow p \leqslant 1-\gamma < 0$, which is absurd. Thus, $\varepsilon \geqslant 0$. □

Remarks. *Convergence* is guaranteed *without convexity or concavity assumptions* on $g(x)$ and *under no asymptotic conditions on* d_k. The inequalities

$$d_{k^*} < 1, \quad d_{k^*+1} < d_{k^*}^{\gamma} \tag{3}$$

must be verified in fact *in two consecutive iterations only*. Secant method is therefore quite different from Newton's method, which requires the local convexity/concavity of the function $g(x)$. From an operational point of view, (3) *is automatically satisfied in the majority of cases after a suitable initial number of iterations*.

3. The n-dimensional case: BFGS-type methods

We are now interested in the minimization of a function f of n variables:

$$\text{find } \mathbf{x}_* \text{ such that } f(\mathbf{x}_*) = \min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}). \tag{4}$$

Recall that a quasi-Newton minimization method solving (4) is said to be *secant* if it exploits a search direction of type $\mathbf{d}_k = -A_k^{-1}\mathbf{g}_k$, where $\mathbf{g}_k$ is the gradient $\mathbf{g} = \nabla f$ evaluated in $\mathbf{x}_k$, and A_k solves the *secant equation*:

$$A_k(\mathbf{x}_k - \mathbf{x}_{k-1}) = \mathbf{g}_k - \mathbf{g}_{k-1}. \tag{5}$$

Observe that (5) is an obvious generalization of (2). However, here $n > 1$, thus A_k is not uniquely determined. In the following we consider the *BFGS-type* algorithms, introduced in [5] and including the classical BFGS method as well as the new $\mathscr{L}QN$ methods. The BFGS-type minimizing sequence $\{\mathbf{x}_k\}_{k=0}^{+\infty}$ can be defined either by a secant (S) or by a nonsecant (NS) iterative scheme, as follows:

$$\mathbf{x}_0 \in \mathbb{R}^n, \quad \tilde{B}_0 = I, \quad \mathbf{d}_0 = -\mathbf{g}_0. \textit{ For } k = 0, 1, \ldots$$

$$\begin{cases} \mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{d}_k, \\ B_{k+1} = \varphi(\tilde{B}_k, \mathbf{s}_k, \mathbf{y}_k), \quad \mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k, \quad \mathbf{y}_k = \mathbf{g}_{k+1} - \mathbf{g}_k, \\ \text{define } \tilde{B}_{k+1} \text{ positive definite (pd)} \\ \text{set } A_{k+1} = \begin{cases} B_{k+1} & (\text{pd since } \mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k > 0)\ S, \\ \tilde{B}_{k+1}, & NS, \end{cases} \\ \mathbf{d}_{k+1} = -A_{k+1}^{-1}\mathbf{g}_{k+1} \quad \leftarrow \text{descent direction}, \end{cases}$$

where $\lambda_k > 0$ is chosen such that $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$ and $\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k > 0$, and φ is the classical BFGS Hessian updating formula

$$\varphi(B, \mathbf{s}, \mathbf{y}) = B + \frac{1}{\mathbf{y}^{\mathrm{T}}\mathbf{s}}\mathbf{y}\mathbf{y}^{\mathrm{T}} - \frac{1}{\mathbf{s}^{\mathrm{T}}B\mathbf{s}}B\mathbf{s}\mathbf{s}^{\mathrm{T}}B. \tag{6}$$

Let us recall two main BFGS-type *properties*:

- *BFGS-types are well defined descent algorithms*: In fact, $f \in C^1$, bounded below and $\mathbf{g}_k^{\mathrm{T}}\mathbf{d}_k < 0 \Rightarrow$ the Armijo–Goldstein (or Wolfe) set

$$AG = \{\lambda \in \mathbb{R}^+ : f(\mathbf{x}_k + \lambda\mathbf{d}_k) - f(\mathbf{x}_k) \leqslant \lambda c_1 \mathbf{d}_k^{\mathrm{T}}\mathbf{g}_k \text{ and}$$

$$(\mathbf{g}(\mathbf{x}_k + \lambda\mathbf{d}_k) - \mathbf{g}_k)^{\mathrm{T}}\lambda\mathbf{d}_k \geqslant (1-c_2)(-\mathbf{g}_k^{\mathrm{T}}\mathbf{d}_k)\},$$

$0 < c_1 < c_2 < 1$, *is not empty*. Thus $\lambda_k \in AG \Rightarrow f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$ and $\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k > 0$, which, in turn, implies that $\varphi(\tilde{B}_k, \mathbf{s}_k, \mathbf{y}_k)$ is positive definite.

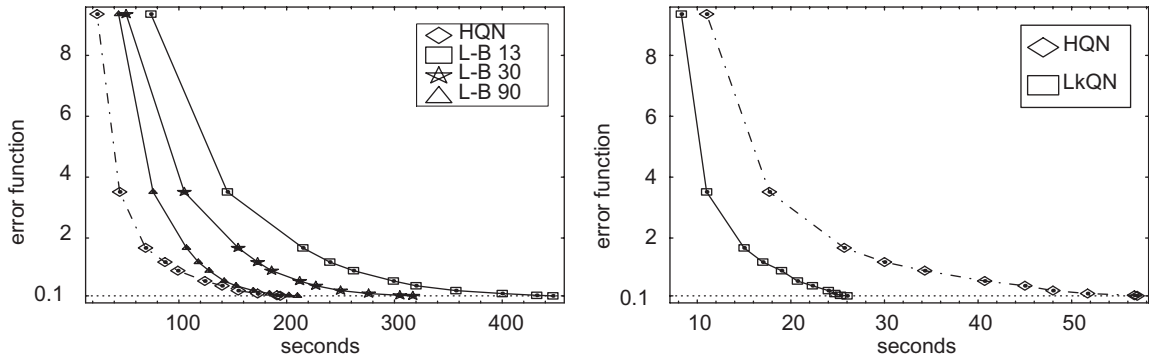

Fig. 2. $\mathscr{L}QN$ (HQN, LkQN) compared with limited memory BFGS (L-B).

- *S BFGS-types are secant algorithms for any choice of* $\tilde{B}_k$: i.e., for S BFGS-type, the matrix $A_{k+1} = \varphi(\tilde{B}_k, \mathbf{s}_k, \mathbf{y}_k)$ used in the definition of the new direction $\mathbf{d}_{k+1}$ solves the *secant equation* (5) with k, $k+1$ replacing $k-1$, k, respectively. Note that this is usually a necessary condition for superlinear convergence of quasi-Newton methods. Moreover, observe that the secant equation also implies $\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k = \mathbf{s}_k^{\mathrm{T}} A_{k+1}\mathbf{s}_k$, i.e., for any secant algorithm with A_{k+1} positive definite (pd), the condition $\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k > 0$ must be verified.

BFGS-type methods have been introduced to solve efficiently minimization problems with a big number n of unknowns. In particular, if f is the error function of a neural network, then n can be extremely large (see [2]). So, for these problems, classical BFGS methods [4,10], where $\tilde{B}_k = B_k$, requiring $\mathrm{O}(n^2)$ flops per step and $\mathrm{O}(n^2)$ memory allocations, cannot be efficiently implemented. A simple choice of $\tilde{B}_k$ with $\tilde{B}_k \neq B_k$ which yields a reduction of complexity to $\mathrm{O}(n)$ is $\tilde{B}_k = \alpha_k I$, $\alpha_k > 0$. In the generalized Battiti–Shanno algorithms considered in [6] new choices of α_k competitive with the best known are proposed.

A reduction of complexity to $\mathrm{O}(n \log n)$, $\mathrm{O}(n)$ can be obtained by the more significant $\mathscr{L}QN$ and adaptive $\mathscr{L}QN$ methods [5,2,6,7], where $\tilde{B}_k$ is chosen equal to the best fit to B_k in a matrix algebra $\mathscr{L}$, i.e., $\tilde{B}_k = \mathscr{L}_{B_k}$ with

$$\|\mathscr{L}_{B_k} - B_k\|_{\mathrm{F}} = \min_{X \in \mathscr{L}} \|X - B_k\|_{\mathrm{F}}, \quad \|\cdot\|_{\mathrm{F}} = \text{Frobenius norm}.$$

If $\mathscr{L} = \{UDU^* : D = \text{diagonal}\}$ where U is an $n \times n$ unitary matrix, then the complexity of $\mathscr{L}QN$ is given by the cost of the transform $U\mathbf{z}$ [5]. In particular, it is $\mathrm{O}(n \log n)$ if $U =$ Fourier, Hartley [5,2] and $\mathrm{O}(n)$ if $U =$ product of two Householder operators [7]. In adaptive $\mathscr{L}QN$ the algebra $\mathscr{L}$ is changed during the optimization procedure. In Fig. 2, we see that S $\mathscr{L}QN$, $\mathscr{L}$=Hartley algebra (HQN), and adaptive S $\mathscr{L}QN$ (LkQN) outperform the limited memory BFGS method (L-B) [10] in the learning process of a neural network associated to the ionosphere data set [2,7] ($n = 1408$). We point out that L-B is a well known adaptation of classical BFGS to large scale problems [10], whose complexity per step is $\mathrm{O}(mn)$ with $m =$ the number of vector pairs $(\mathbf{s}_j, \mathbf{y}_j)$ utilized for the Hessian approximation (in Fig. 2 $m = 13, 30, 90$).

NS BFGS-type methods have not such a good experimental behaviour, and in fact the secant equation (5) is not verified when $A_{k+1} = \tilde{B}_{k+1}$. However, we remark that

- *NS BFGS-type methods are globally convergent* provided that $\det B_k \leqslant \det \tilde{B}_k$, $\mathrm{tr}\, B_k \geqslant \mathrm{tr}\, \tilde{B}_k$, and

$$\frac{\|\mathbf{g}(\mathbf{x}_k + \lambda_k \mathbf{d}_k) - \mathbf{g}_k\|^2}{(\mathbf{g}(\mathbf{x}_k + \lambda_k \mathbf{d}_k) - \mathbf{g}_k)^{\mathrm{T}}(\lambda_k \mathbf{d}_k)} = \frac{\|\mathbf{y}_k\|^2}{\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k} \leqslant M \tag{8}$$

(see [5]). Note that the above conditions on $\tilde{B}_k$ allow to extend Powell's proof for BFGS [11] to BFGS-type and are satisfied for $\tilde{B}_k = \mathscr{L}_{B_k}$. We also underline that assumption (8), on the current guesses $\mathbf{x}_k$ and $\mathbf{x}_{k+1}$, is automatically satisfied if $f(x)$ is a convex function [11]. So, (8) can be seen as a sort of (discrete) *weak convexity assumption*.

In order to obtain *global convergence of secant BFGS-type* methods, we may first apply to our case two known results:

Lemma 1 (*Nocedal and Wright [10]*). *Let $f \in C^1$ be bounded below. Apply an S BFGS-type method with $\lambda_k \in AG$. Assume, for each k,*

$$\|\mathbf{g}_{k+1} - \mathbf{g}_k\| \leqslant M\|\mathbf{x}_{k+1} - \mathbf{x}_k\|. \tag{9}$$

Then $\sum_k (\cos \widehat{-\mathbf{g}_k, \mathbf{d}_k})^2 \|\mathbf{g}_k\|^2 < +\infty$.

Corollary 1 (*Nocedal and Wright [10]*). *If* $\mathbf{g}$ *is Lipschitz and* $\cos \widehat{-\mathbf{g}_{k_i}, \mathbf{d}_{k_i}} \geqslant c > 0$, *then* $\mathbf{g}_{k_i} \to \mathbf{0}$.

Observe that (8) implies the (discrete) Lipschitz condition (9). Thus we also have the following:

Theorem 2. (8) & $\cos \widehat{-\mathbf{g}_{k_i}, \mathbf{d}_{k_i}} \geqslant c > 0 \Rightarrow \mathbf{g}_{k_i} \to \mathbf{0}$.

In other words, a weak convexity assumption together with a lower boundness of $\cos \widehat{-\mathbf{g}_{k_i}, \mathbf{d}_{k_i}}$ guarantees a global convergence behaviour of S BFGS-type methods.

4. A characterization of descent directions

Given any (secant or not secant) minimization iterative scheme

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{d}_k, \quad \lambda_k > 0, \tag{10}$$

the search direction $\mathbf{d}_k$ is required (at least for most iterations k) to be a *descent direction* in the current guess $\mathbf{x}_k$:

$$\mathbf{d}_k^{\mathrm{T}} \mathbf{g}_k < 0. \tag{11}$$

In [10] it is claimed that $\mathbf{d}_k$ is *often* defined from $\mathbf{g}_k = \nabla f(\mathbf{x}_k)$ itself and by a symmetric nonsingular matrix A_k via the equation

$$A_k \mathbf{d}_k = -\mathbf{g}_k. \tag{12}$$

Now, in Theorem 3, we observe that

- *any descent direction* $\mathbf{d}_k$ *always solves Eq.* (12) *for some* pd A_k.

We also show that if the angle between $-\mathbf{g}_k$ and $\mathbf{d}_k$ is uniformly less than 90°, then the matrices A_k can be chosen with bounded condition number.

Intuitively, a direction $\mathbf{d}_k$ satisfying (11) implies a fast *local* convergence, provided that the family of A_k solving (12) *can* approximate the Hessian of f. When $\mathbf{x}_k$ is not in the neighbourhood of $\mathbf{x}_*$, the latter requirement on A_k is no longer necessary; the aim is, in fact, to obtain after a suitable small number of steps an approximation $\mathbf{x}_k$: f is convex in the set $\{\mathbf{x} : f(\mathbf{x}) \leqslant f(\mathbf{x}_k)\}$. Again, intuitively, a direction $\mathbf{d}_k$ satisfying (11) implies such a *global* convergence provided that the family of A_k solving (12) has some properties analogous to the ones of $\lambda_k \in AG$. We underline that if $\lambda_k \in AG$, the sequence $\{\mathbf{x}_k\}$ has already a global convergence property.

Theorem 3. *If* $\mathbf{d}_k$ *is a descent direction in* $\mathbf{x}_k$ *for a function f, i.e.,* $\mathbf{d}_k^{\mathrm{T}} \mathbf{g}_k < 0$, *then* $\mathbf{d}_k = -A_k^{-1} \mathbf{g}_k$ *for some positive definite matrix* A_k. *Moreover, if* $\cos \widehat{-\mathbf{g}_k, \mathbf{d}_k} \geqslant c > 0$, *then* A_k *can be chosen such that* $\mathrm{cond}\,(A_k) \leqslant M_c$.

Proof. The first assertion is proved by choosing

$$A_k = \varphi(\alpha_k I, \mathbf{d}_k, -\mathbf{g}_k), \quad \alpha_k > 0, \tag{13}$$

where φ is defined in (6). For the second assertion choose the same A_k with $\alpha_k = \|\mathbf{g}_k\|/\|\mathbf{d}_k\|$. In fact, the characteristic polynomial of $\varphi(\alpha_k I, \mathbf{d}_k, -\mathbf{g}_k)$ is $c(z) = (z-\alpha_k)^{n-2}((z-\alpha_k)^2 - c_1(z-\alpha_k) + c_2)$. Thus, the computation of c_1 and c_2 yields the n eigenvalues of $\varphi(\alpha_k I, \mathbf{d}_k, -\mathbf{g}_k)$, i.e., α_k with multiplicity $n-2$ and

$$z_{\mp} = \alpha_k + \tfrac{1}{2}[p_k - \alpha_k \mp \sqrt{(p_k-\alpha_k)^2 - 4\alpha_k(q_k - p_k)}],$$

where $p_k = \|\mathbf{g}_k\|^2/\mathbf{d}_k^{\mathrm{T}}(-\mathbf{g}_k)$, $q_k = \mathbf{d}_k^{\mathrm{T}}(-\mathbf{g}_k)/\|\mathbf{d}_k\|^2$. Moreover, we have $z_- \leqslant \alpha_k \leqslant z_+$, being $q_k - p_k \leqslant 0$. So, the condition number of A_k is $\mathrm{cond}(A_k) = ((p_k+\alpha_k)^2/(2\alpha_k q_k))(1 + \sqrt{1 - 4\alpha_k q_k/(p_k+\alpha_k)^2}) - 1$. Thus, $\mathrm{cond}\,(A_k) \leqslant M$ iff $(p_k+\alpha_k)^2/(\alpha_k q_k) \leqslant \hat{M}$ and hence when the following three inequalities hold:

$$\frac{1}{\cos \widehat{-\mathbf{g}_k, \mathbf{d}_k}} \leqslant M_1, \quad \alpha_k \frac{\|\mathbf{d}_k\|}{\|\mathbf{g}_k\|} \frac{1}{\cos \widehat{-\mathbf{g}_k, \mathbf{d}_k}} \leqslant M_2, \quad \frac{1}{\alpha_k} \frac{\|\mathbf{g}_k\|}{\|\mathbf{d}_k\|} \frac{1}{\cos \widehat{-\mathbf{g}_k, \mathbf{d}_k}} \leqslant M_3.$$

If $\cos \widehat{-\mathbf{g}_k, \mathbf{d}_k} \geqslant c > 0$, the latter inequalities are satisfied for $\alpha_k = \|\mathbf{g}_k\|/\|\mathbf{d}_k\|$. □

It is well known that if A_k is pd with condition number bounded by M, then $\cos \widehat{-\mathbf{g}_k, -A_k^{-1}\mathbf{g}_k} \geqslant 1/M > 0$ [10]. So, we have the following:

Corollary 2. *A sequence* $\mathbf{d}_k$ *is such that* $\cos \widehat{-\mathbf{g}_k, \mathbf{d}_k} \geqslant c > 0$ *iff* $\mathbf{d}_k = -A_k^{-1}\mathbf{g}_k$ *for some sequence* A_k *of* pd *matrices satisfying the inequality*

$$\|A_k\|\|A_k^{-1}\| \leqslant M. \tag{14}$$

The previous corollary suggests that a suitable choice of A_k, step by step, could give to the sequence $\{\mathbf{x}_k\}$ a local and/or global convergence property.

5. A global optimization theorem

Theorem 2 and Corollary 2 can have significant applications in global optimization. In particular, (14) allows to derive an important result, which is an operational extension of a theorem proved in [6] (see Theorem 2). We point out that condition (15) in the next Theorem 4 is trivially satisfied if f is convex, but is particularly meaningful for a general class of nonconvex functions (see [6, Definition 1]).

Theorem 4. *Given* $f \in C^2$, *let* $f_{\min}$ *indicate the value of its global minimum. Assume that for the quasi-Newton sequence* $\mathbf{d}_0 = -\mathbf{g}_0$, $\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{d}_k$, $\mathbf{d}_{k+1} = -A_{k+1}^{-1}\mathbf{g}_{k+1}$ *we have*

$$\forall \varepsilon_a \in \Re^+, \quad \exists \varepsilon_s \in \Re^+ : \|\mathbf{g}_k\| > \varepsilon_s \quad \textit{apart from } k : f(\mathbf{x}_k) - f_{\min} < \varepsilon_a. \tag{15}$$

Moreover, let conditions (8) *and* (14) *be satisfied.*

Then, $\forall \varepsilon_a \in \Re^+, \exists k^{**} : \forall k > k^{**}$:

$$f(\mathbf{x}_k) - f_{\min} < \varepsilon_a.$$

Proof. By applying Corollary 2 and Theorem 2, we have that (8) & (14) $\Rightarrow \mathbf{g}_{k_i} \to \mathbf{0}$. But, by (15) this implies the desired result. □

Operational applications: Consider problems in which the function $f(\mathbf{x})$ has a *finite number of local minima* and *the value of its global minimum* $f_{\min}$ *can be estimated in advance*. The latter hypotheses are often satisfied if $f(\mathbf{x})$ is the error function of a neural network and are in general assumed in the literature in many algorithms of probabilistic type (such as Simulated Annealing or Multistart) or as necessary conditions for the convergence of deterministic methods (f.i. TRUST [3]). Since the value $f_{\min}$ is known, in every local search one can perform *a finite number of iterations* until the computed stationary point can be "recognized" as a local minimum or a global one so that the conditions (14)

and (15) are locally satisfied for suitable M and ε_s, respectively. We underline that *the estimation of the upper bound M* in (14) *is not required* in every local search and *is not strictly related to the values* $\mathbf{g}_{k_i}$, $\mathbf{d}_{k_i}$ *computed by the algorithm*, differently from the assumption (theoretically equivalent!) $\cos \widehat{-\mathbf{g}_{k_i}, \mathbf{d}_{k_i}} \geqslant 1/M > 0$. Moreover, since the matrices A_k can be chosen arbitrarily, several "tunneling" criteria [3] may be utilized in order to escape from a local minimum. The choice of the most suitable structures for A_k is the object of our future research.

References

[1] A. Boeckh, Philolaos des Pythagoreers Lehren, Berlin, 1819.
[2] A. Bortoletti, C. Di Fiore, S. Fanelli, P. Zellini, A new class of quasi-Newtonian methods for optimal learning in MLP-networks, IEEE Trans. Neural Networks 14 (2003) 263–273.
[3] B.C. Cetin, J. Barhern, J.W. Burdick, Terminal repeller unconstrained subenergy tunnelling (TRUST) for fast global optimization, J. Optim. Theory Appl. 77 (1993) 97–126.
[4] J.E. Dennis Jr., R.B. Schnabel, Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Prentice-Hall, Englewood Cliffs, NJ, 1983.
[5] C. Di Fiore, S. Fanelli, F. Lepore, P. Zellini, Matrix algebras in quasi-Newton methods for unconstrained optimization, Numer. Math. 94 (2003) 479–500.
[6] C. Di Fiore, S. Fanelli, P. Zellini, An efficient generalization of Battiti–Shanno's quasi-Newton algorithm for learning in MLP-networks, ICONIP'04, Calcutta, 2004, pp. 483–488.
[7] C. Di Fiore, S. Fanelli, P. Zellini, Low complexity minimization algorithms, Numer. Linear Algebra with Appl. 12 (2005) 755–768.
[8] M.J. Gazalé, Gnomon, from Pharaohs to Fractals, Princeton University Press, Princeton, 1999.
[9] S. Kangshen, J.N. Crossley, A.W.C. Lun, The Nine Chapters on the Mathematical Art, Oxford University Press, New York, 1999.
[10] J. Nocedal, S.J. Wright, Numerical Optimization, Springer, Berlin, 1999.
[11] M.J.D. Powell, Some global convergence properties of a variable metric algorithm for minimization without line searches, in: R.W. Cottle, et al. (Eds.), Nonlinear Programming, SIAM-AMS Proceedings, vol. 9, Providence, 1976, pp. 53–72.
[12] A.P. Youschkevitch, Les mathématiques arabes VIII–XV siècles, Vrin, Paris, 1976.
[13] P. Zellini, Gnomon, Adelphi, Milano, 1999.