Math. Program., Ser. B 87: 281–301 (2000)

Digital Object Identifier (DOI) 10.1007/s101070000147

M.J.D. Powell

# On the convergence of the DFP algorithm for unconstrained optimization when there are only two variables

This paper is dedicated to William C. Davidon, and commemorates his 70th birthday.

Received: June 16, 1999 / Accepted: December 24, 1999
Published online March 15, 2000 – © Springer-Verlag 2000

**Abstract**. Let the DFP algorithm for unconstrained optimization be applied to an objective function that has continuous second derivatives and bounded level sets, where each line search finds the first local minimum. It is proved that the calculated gradients are not bounded away from zero if there are only two variables. The new feature of this work is that there is no need for the objective function to be convex.

**Key words.** convergence theory – unconstrained optimization – variable metric algorithms

## 1. Introduction

It is a pleasure to write a paper that commemorates the contributions of Bill Davidon to variable metric methods for unconstrained optimization, because his brilliant original work on achieving quadratic termination (Davidon, 1959) provided the DFP algorithm that is also described in Fletcher and Powell (1963). Thus my career was helped greatly. That algorithm achieves wonderful efficiency in comparison with the steepest descent method, but convergence theorems for general smooth functions did not begin to appear until about 1970, and then the objective function was assumed to be convex. I am now particularly interested in convergence theorems or counter-examples for the algorithm when the objective function $F(\underline{x})$, $\underline{x} \in \mathbb{R}^n$, has the two properties

$$\left.\begin{array}{l}\text{The set } \mathcal{S} = \{\underline{x} : F(\underline{x}) \leq F(\underline{x}_1)\ \} \text{ is bounded, and} \\ \text{The function } F(\underline{x}),\ \underline{x} \in \mathcal{S}, \text{ has continuous second derivatives}\end{array}\right\}, \qquad (1)$$

where $\underline{x}_1$ is a given initial vector of variables. These properties allow some major departures from the convex case.

The existence of a convergence theorem or a counter-example depends on the line search conditions of the iterations of the algorithm. The analysis is interesting, and is more likely to be possible, if one restricts attention to "exact" line searches, which means that each step-length is calculated to give a local minimum of the one-dimensional line search objective function. Then the theorem of Dixon (1972) applies, stating the

M.J.D. Powell: Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Silver Street, Cambridge CB3 9EW, England

*Mathematics Subject Classification (1991):* 65K10, 49M07, 49M37

equivalence of other variable metric methods in the Broyden linear family to the DFP algorithm.

We are going to address the following version of the DFP algorithm, when the number of variables, namely $n$, is only two. In the description, $\underline{g}_k$ is the gradient $\underline{\nabla} F(\underline{x}_k)$, and $\underline{d}_k$ is the search direction of the $k$-th iteration. The conditions (1) ensure that the operations of each iteration are well-defined.

**Step 0:** Pick the starting point $\underline{x}_1 \in \mathbb{R}^n$, an $n \times n$ symmetric positive definite matrix $B_1$, and a positive tolerance $\varepsilon$. Set $k$ to 1.

**Step 1:** Terminate the calculation if the condition

$$\|\underline{g}_k\| \leq \varepsilon \tag{2}$$

is achieved.

**Step 2:** Otherwise, generate the search direction $\underline{d}_k$ by satisfying $B_k \underline{d}_k = -\underline{g}_k$.

**Step 3:** Set the step-length $\alpha_k$ to the largest positive number such that the line search function $F(\underline{x}_k + \alpha \underline{d}_k)$, $\alpha \geq 0$, decreases monotonically for $0 \leq \alpha \leq \alpha_k$. Then let the initial vector of variables for the next iteration be $\underline{x}_{k+1} = \underline{x}_k + \alpha_k \underline{d}_k$.

**Step 4:** Calculate the symmetric matrix $B_{k+1}$ by the DFP formula. Thus the quasi-Newton equation

$$B_{k+1} (\underline{x}_{k+1} - \underline{x}_k) = \underline{g}_{k+1} - \underline{g}_k \tag{3}$$

is obeyed in a way that ensures that $B_{k+1}$ is positive definite.

**Step 5:** Increase $k$ by one, and then go back to Step 1.

This method is not suitable for practical computation when $F$ is a general smooth objective function, because the calculation of $\alpha_k$ in Step 3 would require an infinite amount of work. Therefore we do not expect a convergence proof for the given algorithm to yield immediate improvements to existing software. On the other hand, the DFP algorithm has become of fundamental importance within the subject of nonlinear programming, so we take the view that it is worthwhile to study some theoretical questions that may help to explain its success.

We are going to prove that, if $n = 2$ and if the conditions (1) hold, then the termination condition (2) of the given algorithm is satisfied for a finite value of $k$. The details of the DFP formula for $B_{k+1}$ are irrelevant when there are only two variables. Indeed, Step 3 implies the property

$$\underline{g}_{k+1}^T \underline{d}_k = 0, \qquad k = 1, 2, 3, \ldots, \tag{4}$$

which is equivalent to the orthogonality of $B_{k+1}^{-1} \underline{g}_{k+1}$ to $B_{k+1} \underline{d}_k$. It follows from $\underline{d}_{k+1} = -B_{k+1}^{-1} \underline{g}_{k+1}$ and $\underline{x}_{k+1} - \underline{x}_k = \alpha_k \underline{d}_k$ that $\underline{d}_{k+1}$ is orthogonal to $B_{k+1} (\underline{x}_{k+1} - \underline{x}_k)$. Thus equation (3) provides the first of the conditions

$$\underline{d}_{k+1}^T (\underline{g}_{k+1} - \underline{g}_k) = 0 \qquad \text{and} \qquad \underline{d}_{k+1}^T \underline{g}_{k+1} < 0, \tag{5}$$

the other one being the descent property of the DFP algorithm when $\underline{d}_{k+1}$ is calculated. Expression (5) defines the direction of $\underline{d}_{k+1}$ uniquely for $n = 2$, the length of $\underline{d}_{k+1}$ being unimportant to the theoretical analysis because of the choice of $\alpha_{k+1}$. These remarks allow the matrices $B_k$, $k=1,2,3,\ldots$, to be removed from the given version of the DFP algorithm. Instead, we add to Step 0 that $\underline{d}_1$ is any vector that satisfies $\underline{d}_1^T\underline{g}_1 < 0$, we abolish Step 2, and we replace Step 4 by the statement that $\underline{d}_{k+1}$ is any vector in $\mathbb{R}^2$ that has the properties (5), except that there is no need to pick $\underline{d}_{k+1}$ if $\underline{g}_{k+1}$ is zero.

The search directions of the conjugate gradient algorithm (Polak and Ribière, 1969) also satisfy the conditions (5). Therefore, because $n=2$, our analysis applies to that method too, but some counter-examples to its termination are presented by Powell (1984). They include a two variable case when the step-length of every iteration gives the relations

$$\underline{g}_{k+1}^T\,\underline{d}_k = 0 \qquad \text{and} \qquad F(\underline{x}_{k+1}) < F(\underline{x}_k), \qquad k=1,2,3,\ldots, \tag{6}$$

but the line search function $F(\underline{x}_k + \alpha\,\underline{d}_k)$, $0 \le \alpha \le \alpha_k$, is not required to decrease monotonically. Therefore the monotonicity condition in Step 3 of the given algorithm is important to our proof of termination.

The proof is divided into three sections, that lead to a contradiction under the assumption that the inequality

$$\|\underline{g}_k\| > \varepsilon, \qquad k=1,2,3,\ldots, \tag{7}$$

holds for every positive integer $k$, where $\varepsilon$ is the positive tolerance that is set in Step 0. Now Theorem 2 of Powell (1972) states that, if the sequence $\underline{x}_k$, $k=1,2,3,\ldots$, converged to $\underline{x}_\star$, say, then $\underline{\nabla}F(\underline{x}_\star)$ would be zero. It follows from expression (7) that the sequence has more than one limit point. The purpose of Sect. 2 is to deduce that all the limit points of $\underline{x}_k$, $k=1,2,3,\ldots$, are collinear, and that the directions $\underline{d}_k$, $k=1,2,3,\ldots$, tend to be parallel to the straight line that contains the limit points. Therefore we assume in Sects. 3 and 4, without loss of generality, that the convex hull of the limit points is the straight line segment in $\mathbb{R}^2$ that joins $(-1,0)$ to $(1,0)$, the segment being finite because of the first part of expression (1).

Further, we introduce the notation

$$\gamma(x) = \Big[\, dF(x,y)/dy \,\Big]_{(x,0)}, \qquad -1 \le x \le 1, \tag{8}$$

for the derivative of the objective function in the $y$-direction on the line segment that has just been mentioned, where $x$ and $y$ are the components of $\underline{x} \in \mathbb{R}^2$. One of the lemmas of Sect. 2 establishes that $\gamma(x)$, $-1 \le x \le 1$, is bounded away from zero, and the final result of Sect. 3 is the property

$$\left| x + \frac{\gamma(x)}{\gamma'(x)} \right| \ge 1, \qquad -1 \le x \le 1, \tag{9}$$

which is trivial when $\gamma'(x)$ is zero, due to $\gamma(x) \ne 0$. The justification of this inequality requires much work. Therefore the analysis is presented in a way that allows Sect. 4 to be studied before the intricate part of Sect. 3.

The reader will find in Sect. 4 that the inequalities (7) and (9) lead to a contradiction, which completes the proof of termination of the given algorithm when $n=2$. Finally, there are some remarks in Sect. 5 on whether or not the conditions (1) imply termination for larger values of $n$.

The relevance of the analysis to algorithms that employ the PSB updating formula (Powell, 1970), instead of a variable metric one, is questionable. The PSB update achieves the quasi-Newton condition (3), but, because $B_{k+1}$ may not be positive definite, the next trial step $\underline{d}_{k+1}$ is usually generated by a trust region method instead of being given the value $\underline{d}_{k+1} = -B_{k+1}^{-1}\,\underline{g}_{k+1}$. If $B_{k+1}$ were nonsingular, then this value and the line search of the previous iteration would provide the first of the conditions (5), as mentioned already. The second of the conditions, however, may fail. For example, if $B_{k+1}$ is calculated from the data

$$\underline{x}_{k+1} - \underline{x}_k = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \underline{g}_k = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad \underline{g}_{k+1} = \begin{pmatrix} 0 \\ \sigma \end{pmatrix}, \quad B_k = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{10}$$

then PSB provides the matrix with diagonal elements of one and off-diagonal elements of $\sigma$. Thus singularity or loss of positive definiteness occurs if $|\sigma| = 1$ or $|\sigma| > 1$, respectively. Further, $\underline{d}_{k+1}^T\,\underline{g}_{k+1}$ is positive in the case $|\sigma| > 1$.

## 2. Proof of collinearity of the limit points

The assumption (7) implies that the number of iterations of the given algorithm is infinite, and already we have noted that the sequence $\underline{x}_k$, $k = 1, 2, 3, \ldots$, has more than one limit point. We consider the piecewise linear path in $\mathbb{R}^2$ that is constructed by drawing the straight line from $\underline{x}_k$ to $\underline{x}_{k+1}$ for every positive integer $k$, the results of this section being derived from the asymptotic form of the path as $k \to \infty$. We let $\mathcal{T} \subset \mathbb{R}^2$ denote the set of points of the asymptotic form, which are defined as follows. Because Step 3 of the given algorithm ensures that the objective function $F(\underline{x})$, $\underline{x} \in \mathbb{R}^2$, decreases monotonically on the path, the asymptotic form is contained in the set $\{\underline{x} : F(\underline{x}) = F_\star\}$, where $F_\star$ is the limit of the monotonic sequence $F(\underline{x}_k)$, $k = 1, 2, 3, \ldots$. Therefore $\underline{t}$ is an element of $\mathcal{T}$ if and only if $F(\underline{t})$ is equal to $F_\star$, and there is an infinite sequence of points on the path that converges to $\underline{t}$. In particular, $\mathcal{T}$ includes all the limit points of the vectors of variables $\underline{x}_k$, $k = 1, 2, 3, \ldots$. The required properties of $\mathcal{T}$ are presented as lemmas in order to give some structure to the details of the analysis.

**Lemma 1.** *$\mathcal{T}$ is closed.*

*Proof.* Let $\underline{t}_\star$ be in the closure of $\mathcal{T}$ and let $\eta$ be any positive number. We let $\hat{\underline{t}}(\eta)$ be an element of $\mathcal{T}$ that satisfies $\|\hat{\underline{t}}(\eta) - \underline{t}_\star\| \leq \frac{1}{2}\eta$, and then we let $\underline{t}(\eta)$ be a point on the piecewise linear path that satisfies $\|\underline{t}(\eta) - \hat{\underline{t}}(\eta)\| \leq \frac{1}{2}\eta$, which gives the condition $\|\underline{t}(\eta) - \underline{t}_\star\| \leq \eta$. Therefore, if $\eta$ runs through the values $(1/2)^j$, $j = 1, 2, 3, \ldots$, then the resultant sequence of points $\underline{t}(\eta)$ converges to $\underline{t}_\star$. Further, by combining the continuity of $F$ with $\underline{t}_\star$ in the closure of $\mathcal{T}$, we find $F(\underline{t}_\star) = F_\star$. It follows that $\underline{t}_\star$ is an element of $\mathcal{T}$ as required.

□

**Lemma 2.** *$\mathcal{T}$ is connected.*

*Proof.* If $\mathcal{T}$ were not connected, we could divide it into two parts, $\mathcal{T}_1$ and $\mathcal{T}_2$ say, such that $\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2$, and such that $\underline{t}_1 \in \mathcal{T}_1$ and $\underline{t}_2 \in \mathcal{T}_2$ imply $\|\underline{t}_1 - \underline{t}_2\| \geq \delta$, where $\delta$ is a positive constant. Further, we let $\mathcal{S}_1$ and $\mathcal{S}_2$ be the sets

$$\mathcal{S}_1 = \left\{ \underline{x} : \min_{\underline{t} \in \mathcal{T}_1} \|\underline{t} - \underline{x}\| \leq \tfrac{1}{4}\delta \right\} \quad \text{and} \quad \mathcal{S}_2 = \left\{ \underline{x} : \min_{\underline{t} \in \mathcal{T}_2} \|\underline{t} - \underline{x}\| \leq \tfrac{1}{4}\delta \right\}, \tag{11}$$

so $\mathcal{S}_1$ and $\mathcal{S}_2$ are also disjoint. Let $\mathcal{K}$ be the set of positive integers such that, for each $k$ in $\mathcal{K}$, the straight line segment between $\underline{x}_k$ and $\underline{x}_{k+1}$ reaches both $\mathcal{S}_1$ and $\mathcal{S}_2$. The number of elements of $\mathcal{K}$ is infinite, because otherwise the piecewise linear path would not have a limit point in $\mathcal{S}_1$ and a limit point in $\mathcal{S}_2$, which are required by the choices of $\mathcal{T}_1$ and $\mathcal{T}_2$. Moreover, for each $k \in \mathcal{K}$, we can let $\underline{t}_k$ be a point on the straight line from $\underline{x}_k$ to $\underline{x}_{k+1}$ that lies in the gap between $\mathcal{S}_1$ and $\mathcal{S}_2$, which gives the property $\|\underline{t}_k - \underline{t}\| > \frac{1}{4}\delta$, $\underline{t} \in \mathcal{T}$. On the other hand, the limit points of the sequence $\underline{t}_k$, $k \in \mathcal{K}$, are in $\mathcal{T}$. This contradiction completes the proof.

□

**Lemma 3.** *For every $\underline{t} \in \mathcal{T}$, the gradient $\underline{\nabla} F(\underline{t})$ is nonzero.*

*Proof.* We assume that $\underline{t}_\star \in \mathcal{T}$ satisfies $\underline{\nabla} F(\underline{t}_\star) = 0$, and we deduce a contradiction. Let $\underline{t}_j$, $j = 1, 2, 3, \ldots$, be a sequence of points on the piecewise linear path that has been mentioned that converges to $\underline{t}_\star$. Further, for each $j$, we let $k(j)$ be a positive integer such that $\underline{t}_j$ is on the line segment that joins $\underline{x}_{k(j)}$ to $\underline{x}_{k(j)+1}$. The condition $F(\underline{t}_\star) = F_\star$ implies that the sequence of integers $k(j)$, $j = 1, 2, 3, \ldots$, is divergent. Therefore, by choosing a subsequence of $\underline{t}_j$, $j = 1, 2, 3, \ldots$, if necessary, we assume without loss of generality that the integers $k(j)$, $j = 1, 2, 3, \ldots$, increase strictly monotonically. Let $\mathcal{K}$ be the set $\{k(j) : j = 1, 2, 3, \ldots\}$. Then, also without loss of generality, we replace $\mathcal{K}$ by a subset if necessary, so that the sequences $\underline{x}_k$, $k \in \mathcal{K}$, and $\underline{x}_{k+1}$, $k \in \mathcal{K}$, both converge, to $\hat{\underline{x}}_\star$ and $\check{\underline{x}}_\star$ say, respectively. It follows that $\hat{\underline{x}}_\star$, $\underline{t}_\star$ and $\check{\underline{x}}_\star$ are collinear, and that $\underline{t}_\star$ is strictly between $\hat{\underline{x}}_\star$ and $\check{\underline{x}}_\star$, due to the conditions

$$\|\underline{\nabla} F(\hat{\underline{x}}_\star)\| \geq \varepsilon, \qquad \|\underline{\nabla} F(\underline{t}_\star)\| = 0 \qquad \text{and} \qquad \|\underline{\nabla} F(\check{\underline{x}}_\star)\| \geq \varepsilon. \tag{12}$$

Further, the line segment from $\hat{\underline{x}}_\star$ to $\check{\underline{x}}_\star$ is a subset of $\mathcal{T}$, so the objective function takes the value $F_\star$ throughout the line segment. We also assume without loss of generality that the coordinates of $\underline{t}_\star$ and $\check{\underline{x}}_\star$ are $(0,0)$ and $(1,0)$, respectively, and that the second component of $\underline{\nabla} F(\check{\underline{x}}_\star)$ is positive, the first component of $\underline{\nabla} F(\underline{x})$ being zero for every $\underline{x}$ on the line segment. It follows from expression (12) that we can let $\underline{x}_\star$ be a point between $\underline{t}_\star$ and $\check{\underline{x}}_\star$ such that $\underline{\nabla} F(\underline{x}_\star)$ has the components $(0, \frac{1}{2}\varepsilon)$. Thus $\underline{x}_\star$ is the point $(c, 0)$, for some number $c$ that satisfies the strict inequalities $0 < c < 1$.

It also follows from the continuity of $\underline{\nabla} F$ that we can let $\delta$ be a positive constant such that the conditions

$$\left[ \frac{dF(x,y)}{dy} \right]_{(0,\theta)} \leq \tfrac{1}{4}\varepsilon \, \frac{1-c}{1+c} \qquad \text{and} \qquad \left[ \frac{dF(x,y)}{dy} \right]_{(c,\theta)} \geq \tfrac{1}{4}\varepsilon \tag{13}$$

hold for $0 \leq \theta \leq \delta$. Now, for every sufficiently large $k$ in $\mathcal{K}$, the line segment from $\underline{x}_k$ to $\underline{x}_{k+1}$ cuts both the line segment from $\underline{t}_\star = (0,0)$ to $(0,\delta)$ and the line segment from $\underline{x}_\star = (c, 0)$ to $(c, \delta)$. We let $\underline{a}_k$ and $\underline{b}_k$ be the points of intersection, and we let the coordinates of these points be $(0, \alpha_k)$ and $(c, \beta_k)$, respectively, so we are changing the meaning of $\alpha_k$ temporarily. Therefore the conditions (13) give the relations

$$F(\underline{a}_k) \le F_\star + \tfrac{1}{4}\varepsilon \, \frac{1-c}{1+c} \, \alpha_k \qquad \text{and} \qquad F(\underline{b}_k) \ge F_\star + \tfrac{1}{4}\varepsilon\beta_k. \tag{14}$$

Further, Step 3 of the algorithm of Sect. 1 provides $F(\underline{a}_k) \ge F(\underline{b}_k) > F_\star$. These remarks imply the bounds

$$0 < \beta_k \le \alpha_k \, (1-c)/(1+c), \tag{15}$$

for sufficiently large $k$ in $\mathcal{K}$. Moreover, because the straight line through $\underline{x}_k$ and $\underline{x}_{k+1}$ is also the straight line through $\underline{a}_k$ and $\underline{b}_k$, it has the equation

$$y = \alpha_k + (x/c)\,(\beta_k - \alpha_k), \qquad (x, y) \in \mathbb{R}^2, \tag{16}$$

so it intersects the $x$-axis at $(\xi_k, 0)$, where $\xi_k = \alpha_k \, c \,/(\alpha_k - \beta_k)$. It follows from expression (15) that $\xi_k$ is in the interval $c < \xi_k \le \frac{1}{2}(1+c)$.

When $k \in \mathcal{K}$ tends to infinity, however, the $x$-coordinate of $\underline{x}_{k+1}$ converges to one, so it becomes larger than $\frac{1}{2}(1+c)$. Thus the line segment from $\underline{x}_k$ to $\underline{x}_{k+1}$ cuts the $x$-axis at a point where the objective function takes the value $F_\star$, which is a contradiction. Therefore the lemma is true.

□

The final lemma of this section requires a well-known result that is included in Sect. 3 of Wolfe (1970), for instance. It is that the conditions (1) and (7) imply the property

$$\textstyle\sum_{k=1}^{\infty} \cos^2 \theta_k < \infty, \tag{17}$$

where $\theta_k$ is the angle between $\underline{d}_k$ and $-\underline{g}_k$ in the algorithm of Sect. 1.

**Lemma 4.** *The points of $\mathcal{T}$ are collinear.*

*Proof.* We assume that the lemma is false. Therefore we can let $\mathcal{C}$ be a circle of finite radius, $r$ say, that contains $\mathcal{T}$, and that includes at least three points of $\mathcal{T}$ on its circumference. Further, because the lemmas so far imply that $\mathcal{T}$ is a continuous curve that has two or no end-points, we can let $\underline{t}_\star$ be an interior point of $\mathcal{T}$ that is on the circumference of $\mathcal{C}$. We assume without loss of generality that $\underline{t}_\star$ and $\underline{\nabla} F(\underline{t}_\star)$ have the components $(0, 0)$ and $(0, 1)$, respectively. The reason for mentioning the circle is to deduce that $d^2F(x, y)/dx^2$ is nonzero at $\underline{x} = \underline{t}_\star$, where $x$ and $y$ are still the components of $\underline{x}$. We let $F_{xx}(\underline{t}_\star)$ denote this second derivative.

When $|\delta|$ is very small, the value of $F$ at $(\delta, 0)$ is $F_\star + \frac{1}{2}\delta^2 F_{xx}(\underline{t}_\star) + o(\delta^2)$. It follows from $F_y(\underline{t}_\star) = 1$ that the distance from the point $(\delta, -\frac{1}{2}\delta^2 F_{xx}(\underline{t}_\star))$ to $\mathcal{T}$ is $o(\delta^2)$. Hence the radius of curvature of $\mathcal{T}$ at $\underline{t}_\star$ is $1/|F_{xx}(\underline{t}_\star)|$. Therefore, because $\mathcal{T}$ is enclosed by the circle $\mathcal{C}$, we find the bound

$$|F_{xx}(\underline{t}_\star)| \ge 1\,/\,r. \tag{18}$$

We write this inequality in the form

$$|\underline{d}^T \nabla^2 F(\underline{t})\, \underline{d}| \geq \|\underline{d}\|^2 / r, \tag{19}$$

where $\underline{t}$ and $\underline{d}$ are $\underline{t}_\star$ and any vector that is orthogonal to $\underline{\nabla} F(\underline{t}_\star)$, respectively.

It follows from the continuity of first and second derivatives of $F$, and from $\underline{\nabla} F(\underline{t}_\star) \neq 0$, that $|\underline{d}^T \nabla^2 F(\underline{t})\, \underline{d}|$ is bounded below by a positive multiple of $\|\underline{d}\|^2$, if $\underline{d}$ is orthogonal to $\underline{\nabla} F(\underline{x})$, where $\underline{x}$ and $\underline{t}$ are any points that are sufficiently close to $\underline{t}_\star$. In other words, there exists a neighbourhood of $\underline{t}_\star$, $\mathcal{N}$ say, such that, if $\underline{t}$ and $\underline{x}$ are in $\mathcal{N}$, and if $\underline{d}^T \underline{\nabla} F(\underline{x})$ is zero, then the condition

$$|\underline{d}^T \nabla^2 F(\underline{t})\, \underline{d}| \geq \eta\, \|\underline{d}\|^2 \tag{20}$$

holds, where $\eta$ is a positive constant.

Now the nonzero curvature of $\mathcal{T}$ at $\underline{t}_\star$ implies that there exists a strictly increasing sequence of positive integers $\mathcal{K}$ such that $\underline{x}_k$, $k \in \mathcal{K}$, tends to $\underline{t}_\star$. Further, because nonzero curvature occurs at $\underline{t} \in \mathcal{T}$ if $\underline{t}$ is close to $\underline{t}_\star$, the distances $\|\underline{x}_{k+1} - \underline{x}_k\|$, $k \in \mathcal{K}$, converge to zero. Thus the properties $\underline{x}_k \in \mathcal{N}$ and $\underline{x}_{k+1} \in \mathcal{N}$ are achieved for an infinite number of values of $k$. We are going to deduce from condition (20) that $\cos\theta_{k+1}$ is bounded away from zero for all of them, which implies that the left hand side of expression (17) is infinite. Thus we will obtain a contradiction that completes the proof.

Let $\underline{x}_k$ and $\underline{x}_{k+1}$ be in $\mathcal{N}$. We pick $\underline{x} = \underline{x}_{k+1}$ and $\underline{d} = \underline{d}_k$, because then equation (4) shows the required orthogonality of $\underline{d}$ to $\underline{\nabla} F(\underline{x})$. We let $\underline{t}$ be on the line segment from $\underline{x}_k$ to $\underline{x}_{k+1}$, beginning with $\underline{t} = \underline{x}_{k+1}$. Now Step 3 of the given algorithm implies that the line search function $F(\underline{x}_k + \alpha\, \underline{d}_k)$, $\alpha \geq 0$, has a nonnegative second derivative at $\alpha = \alpha_k$, which is the condition $\underline{d}_k^T \nabla^2 F(\underline{x}_{k+1})\, \underline{d}_k \geq 0$. Therefore the modulus signs can be removed from the left hand side of expression (20), which is valid for every $\underline{t}$ in $\mathcal{N}$ as $\nabla^2 F$ is continuous.

We combine condition (20) for the range of values of $\underline{t}$ with the elementary identity

$$\underline{g}_{k+1} - \underline{g}_k = \int_{\theta=0}^{1} \nabla^2 F(\underline{x}_k + \theta\, [\underline{x}_{k+1} - \underline{x}_k]\,)\, (\underline{x}_{k+1} - \underline{x}_k)\, d\theta, \tag{21}$$

noting that the identity gives the bound

$$\|\underline{g}_{k+1} - \underline{g}_k\| \leq c\, \|\underline{x}_{k+1} - \underline{x}_k\|, \tag{22}$$

where $c$ is a positive constant. Specifically, we form the scalar product of both sides of equation (21) with $\underline{d}_k$, keeping the scalar product on the right hand side under the integral sign. It follows from condition (20) that the new integrand is bounded below by $\eta\, \|\underline{d}_k\|\, \|\underline{x}_{k+1} - \underline{x}_k\|$, as $\underline{x}_{k+1} - \underline{x}_k$ is a positive multiple of $\underline{d}_k$. Thus we deduce the relation

$$\underline{d}_k^T (\underline{g}_{k+1} - \underline{g}_k) \geq \eta\, \|\underline{d}_k\|\, \|\underline{x}_{k+1} - \underline{x}_k\| \geq (\eta/c)\, \|\underline{d}_k\|\, \|\underline{g}_{k+1} - \underline{g}_k\|, \tag{23}$$

the last assertion being due to the bound (22).

Inequality (23) shows that the cosine of the angle between $\underline{d}_k$ and $\underline{g}_{k+1} - \underline{g}_k$ is at least $\eta/c$. Moreover, equations (4) and (5) state that $\underline{g}_{k+1}$ and $\underline{d}_{k+1}$ are orthogonal to $\underline{d}_k$ and $\underline{g}_{k+1} - \underline{g}_k$, respectively, and there are only two variables. Therefore the modulus of the cosine of the angle between $\underline{d}_{k+1}$ and $\underline{g}_{k+1}$ is the same as the modulus of the cosine of the angle between $\underline{d}_k$ and $\underline{g}_{k+1}-\underline{g}_k$. These remarks imply the inequality $|\cos\theta_{k+1}|\geq\eta/c$ for an infinite number of values of $k$, which gives the required contradiction to expression (17).

□

As mentioned already, the analysis of this section allows us to assume without loss of generality that $\mathcal{T}$ is the straight line segment in $\mathbb{R}^2$ that joins $(-1,0)$ to $(1,0)$. Then the condition $F(\underline{t})=F_\star$, $\underline{t}\in\mathcal{T}$, and Lemma 3, imply that $\underline{\nabla}F(\underline{t})$ is a nonzero multiple of the second coordinate direction for every $\underline{t}$ in $\mathcal{T}$. We assume, also without loss of generality, that the multiple is positive. Thus $F(\underline{x}_k)>F_\star$, $k=1,2,3,\ldots$, and the definition of $\mathcal{T}$, cause the second component of $\underline{x}_k$ to be positive for all sufficiently large $k$, which allows us to assume this property for every $k$. Therefore, regarding the $x$-axis as horizontal in $\mathbb{R}^2$, the sequence $\underline{x}_k$, $k=1,2,3,\ldots$, approaches $\mathcal{T}$ from above. Further, because $\underline{g}_k$ tends to be vertical, it follows from the bound (17) that the search directions $\underline{d}_k$, $k=1,2,3,\ldots$, tend to be horizontal. In other words, the search directions become parallel to $\mathcal{T}$ in the limit $k\rightarrow\infty$, which is one of the assertions of Sect. 1.

## 3. Further analysis

Throughout the remainder of the paper, we let the scalings of the search directions have the property

$$\underline{d}_k^T\underline{g}_k=-\|\underline{g}_k\|^2,\qquad k=1,2,3,\ldots,\tag{24}$$

which does not lose generality, and which agrees with the second part of expression (5). It follows from $n=2$ and equation (4) that, for $k\geq 2$, $\underline{d}_k$ has the form $-\underline{g}_k+\beta_k\underline{d}_{k-1}$, where $\beta_k\in\mathbb{R}$ is determined by the first part of expression (5). Thus we derive the formula

$$\underline{d}_k=-\underline{g}_k+\frac{\underline{g}_k^T(\underline{g}_k-\underline{g}_{k-1})}{\underline{d}_{k-1}^T(\underline{g}_k-\underline{g}_{k-1})}\,\underline{d}_{k-1}=-\underline{g}_k+\frac{\underline{g}_k^T(\underline{g}_k-\underline{g}_{k-1})}{\|\underline{g}_{k-1}\|^2}\,\underline{d}_{k-1},\quad k\geq 2,\tag{25}$$

the last identity being a consequence of equations (4) and (24).

Moreover, the scaling (24) implies that the $\cos\theta_k$ term of inequality (17) has the value

$$\cos\theta_k=\|\underline{g}_k\|\,/\,\|\underline{d}_k\|,\qquad k=1,2,3,\ldots\,.\tag{26}$$

Thus inequality (17) would contradict the assumption (7) if an infinite subsequence of the norms $\|\underline{d}_k\|$, $k=1,2,3,\ldots$, were bounded. Therefore we may add the property

$$\|\underline{d}_k\|\rightarrow\infty\quad\text{as}\quad k\rightarrow\infty\tag{27}$$

to the conditions that have been noted already.

The limit (27) and equation (25) provide some useful relations. Firstly, because the assumptions (1) imply that the gradients $\underline{g}_k$, $k=1,2,3,\ldots$, are bounded, every $\underline{x}_k$ being in $\mathcal{S}$, they show that $\underline{d}_k$ tends to be a multiple of $\underline{d}_{k-1}$ as $k \to \infty$, which confirms the last remark of Sect. 2. They also give the condition

$$\|\underline{d}_k\| = \bigl(1+o(1)\bigr) \left| \frac{\underline{g}_k^T (\underline{g}_k - \underline{g}_{k-1})}{\|\underline{g}_{k-1}\|^2} \right| \|\underline{d}_{k-1}\|, \qquad k=2,3,4,\ldots, \tag{28}$$

where $1+o(1)$ denotes a factor that tends to one as $k \to \infty$. The sign of the term inside the modulus signs of condition (28) is going to be important. Therefore we introduce the disjoint sets

$$\mathcal{K}_{\text{same}} = \bigl\{k : \underline{g}_k^T (\underline{g}_k - \underline{g}_{k-1}) > 0\bigr\} \quad \text{and} \quad \mathcal{K}_{\text{opp}} = \bigl\{k : \underline{g}_k^T (\underline{g}_k - \underline{g}_{k-1}) < 0\bigr\}. \tag{29}$$

Thus $k \in \mathcal{K}_{\text{same}}$ or $k \in \mathcal{K}_{\text{opp}}$ correspond to the cases when the direction of $\underline{d}_k$ tends to be the same as or opposite to the direction of $\underline{d}_{k-1}$, respectively. If $\underline{g}_k^T (\underline{g}_k - \underline{g}_{k-1})$ were zero, then formula (25) would reduce to $\underline{d}_k = -\underline{g}_k$, which is not allowed by the limit (27) for sufficiently large $k$. Therefore, by deleting a finite number of iterations from the beginning of the calculation if necessary, we ensure that every iteration number is in one of the sets (29). Moreover, the analysis of the previous section implies that $\mathcal{K}_{\text{opp}}$ has an infinite number of elements.

Equations (28) and (29) and the Cauchy–Schwarz inequality imply the bound

$$\begin{aligned} \|\underline{d}_k\| &\le \bigl(1+o(1)\bigr) \left( \frac{\|\underline{g}_k\|}{\|\underline{g}_{k-1}\|} - \frac{\|\underline{g}_k\|^2}{\|\underline{g}_{k-1}\|^2} \right) \|\underline{d}_{k-1}\| \\ &= \|\underline{g}_k\| \bigl(1+o(1)\bigr) \left( 1 - \frac{\|\underline{g}_k\|}{\|\underline{g}_{k-1}\|} \right) \frac{\|\underline{d}_{k-1}\|}{\|\underline{g}_{k-1}\|}, \qquad k \in \mathcal{K}_{\text{opp}}. \end{aligned} \tag{30}$$

Moreover, the factor $(1 - \|\underline{g}_k\| / \|\underline{g}_{k-1}\|)$ is no greater than a constant that is strictly less than one, due to assumption (7) and the boundedness of $\|\underline{g}_{k-1}\|$. It follows from the meaning of $1+o(1)$ that there exists a constant integer $k_0$ such that the condition

$$\|\underline{d}_k\| \,/\, \|\underline{g}_k\| \le \|\underline{d}_{k-1}\| \,/\, \|\underline{g}_{k-1}\|, \qquad k \in \mathcal{K}_{\text{opp}}, \quad k \ge k_0, \tag{31}$$

is achieved.

The contradiction that will complete our work will come from an extension of the property (31). Specifically, letting $k$ be any sufficiently large integer in $\mathcal{K}_{\text{opp}}$, and letting $\ell(k)$ be the greatest element of $\mathcal{K}_{\text{opp}}$ that is less than $k$, it will be proved that $\|\underline{d}_k\| / \|\underline{g}_k\| \le \|\underline{d}_j\| / \|\underline{g}_j\|$ is satisfied for every integer $j$ in the interval $[\ell(k), k-1]$. Therefore, assuming that the elements of $\mathcal{K}_{\text{opp}}$ are arranged in ascending order, and choosing $j = \ell(k)$, the sequence $\|\underline{d}_k\| / \|\underline{g}_k\|$, $k \in \mathcal{K}_{\text{opp}}$, is monotonically decreasing for sufficiently large $k$. Thus the elements of the sequence are uniformly bounded, which implies that the norms $\|\underline{d}_k\|$, $k \in \mathcal{K}_{\text{opp}}$, are uniformly bounded too. On the other hand, our assumptions have provided the limit (27), which is the required contradiction.

The proof of inequality (9) occupies the remainder of this section, because it is needed by the method that gives the relation $\|\underline{d}_k\| / \|\underline{g}_k\| \le / \|\underline{d}_j\| / \|\underline{g}_j\|$, mentioned in the previous paragraph. The reader is advised to study Sect. 4 first, however, assuming that condition (9) is true. Thus a major interruption to the main argument is avoided, and the motivation for the following analysis is strengthened.

Again the analysis is divided into pieces by the use of lemmas. We employ the notation

$$\phi(x) = x + \gamma(x)/\gamma'(x), \qquad -1 \le x \le 1, \tag{32}$$

for the expression inside the modulus signs of inequality (9). We let $\phi(x)$ be $+\infty$ if $\gamma'(x)$ is zero, because we know from Sect. 2 that $\gamma(x)$, $-1 \le x \le 1$, is positive. We will establish the assertion (9) by supposing that it fails, and deducing a contradiction.

**Lemma 5.** *If inequality (9) does not hold, then we can assume without loss of generality that there exist numbers $a$ and $b$, satisfying $-1 < a < b < 1$, and having the properties*

$$-1 < \phi(a) < 1 \qquad \textit{and} \qquad \gamma'(x) > 0, \quad a \le x \le b. \tag{33}$$

*Further, there exists $x_\star$ in the set*

$$\mathcal{X} = \{x : \; b \le x \le 1, \; \gamma'(x) \ge 0\} \tag{34}$$

*such that $\phi(x_\star) = \inf\{\phi(x) : x \in \mathcal{X}\}$ is achieved, and the choices of $a$ and $b$ can provide the strict inequality $\phi(a) < \phi(x_\star)$.*

*Proof.* If condition (9) fails when $x = \hat{a}$, say, then it fails for every $x$ in $[-1, 1]$ that is sufficiently close to $\hat{a}$, because $\gamma$ and $\gamma'$ are continuous, and because $\min\{\gamma(x) : -1 \le x \le 1\} = \gamma_{\min}$, say, is positive. Therefore we can let $\hat{a}$ be an interior point of the interval $[-1, 1]$. Further, the possibility of replacing $x$ by $-x$ throughout the paper allows $\gamma'(\hat{a}) > 0$ to be assumed without loss of generality. Hence, by putting $\phi(\hat{a}) < 1$ and $\hat{a} > -1$ in the definition (32), we find the bound $\gamma'(\hat{a}) > \frac{1}{2}\gamma_{\min}$. Thus the conditions $\hat{a} \le x \le 1$ and $\gamma'(x) \ge \frac{1}{2}\gamma_{\min}$ are satisfied when $x = \hat{a}$.

We let $\mathcal{A}$ be the set of values of $x$ that minimize $\phi(x)$ subject to these conditions, and then we let $a$ be the greatest element of $\mathcal{A}$. The set $\mathcal{A}$ is well-defined and compact, due to the continuity of $\gamma$ and $\gamma'$, so $a$ is well-defined too. This choice and the definition (32) give $\phi(a) \le \phi(\hat{a}) < 1$ and $\phi(a) > a \ge \hat{a} > -1$, as required. Further, we let $b$ be any number in the open interval $(a, 1)$ such that the second part of expression (33) is also achieved, which is easy because $\gamma'(a)$ is positive and $\gamma'$ is continuous.

When considering the existence of $x_\star$, the condition $\gamma(x) \ge \gamma_{\min} > 0$, $-1 \le x \le 1$, allows us to restrict attention to values of $x \in \mathcal{X}$ such that $\gamma'(x)$ is bounded away from zero. Thus $\phi$ is continuous, so the existence of $x_\star$ follows from the compactness of the set (34). If $\phi(a) < \phi(x_\star)$ failed, then $x_\star$ would be a point in $[b, 1]$ satisfying $\gamma'(x_\star) > 0$ and $\phi(x_\star) \le \phi(a)$. Further, the last condition would give $\gamma(x_\star)/\gamma'(x_\star) \le 1 + \phi(a) < 2$, which would imply $\gamma'(x_\star) > \frac{1}{2}\gamma_{\min}$. It follows from $x_\star \ge b > a$ that the properties of $x_\star$ would contradict our choice of $a$. Therefore the proof is complete.

□

There are four more lemmas in this section, and we continue to assume the failure of condition (9). Therefore we let the numbers $a$, $b$ and $x_\star$ be as in Lemma 5. Further, we let $\mathcal{K}_\star$ be the set of integers $k$ such that $x_k \ge b$ and $x_{m(k)} \le a$ are satisfied, where $x_k$ and $x_{m(k)}$ are the first components of $\underline{x}_k$ and $\underline{x}_{m(k)}$, respectively, $m(k)$ being the greatest integer less than $k$ such that $x_{m(k)}$ is not in the open interval $(a, b)$. In other words, the strip $\{(x, y) \in \mathbb{R}^2 : a \le x \le b\}$ is crossed by the piecewise linear path that joins the points $\underline{x}_j$, $m(k) \le j \le k$, and $a < x_j < b$ holds if $j$ is any integer between $m(k)$ and $k$. It follows from the work of Sect. 2 that $\mathcal{K}_\star$ has an infinite number of elements. We will find, however, that the statements of Lemma 5 exclude $k$ from $\mathcal{K}_\star$ when $k$ is sufficiently large, which is the contradiction that will establish inequality (9).

Our argument addresses the point where the straight line through $\underline{x}_{k-1}$ and $\underline{x}_k$ cuts the $x$-axis in $\mathbb{R}^2$ for certain integers $k$. We let this point be $(\xi_k, 0)$ throughout the remainder of this section. Lemma 7 will provide a useful relation between $\xi_k$ and $\phi(x_k)$. The proof of that lemma and other parts of our analysis require the following expression for $\underline{\nabla} F(\underline{x})$.

**Lemma 6.** *Let $\underline{x} = (x, y)$ be any point of the set $\mathcal{S}$ of the conditions (1) such that $-1 \le x \le 1$ holds. Then $\underline{\nabla} F(\underline{x})$ has the form*

$$\underline{\nabla} F(\underline{x}) = \begin{pmatrix} y\,\gamma'(x) + o(y) \\ \gamma(x) + \mathcal{O}(y) \end{pmatrix}, \tag{35}$$

*where $\gamma$ is still the derivative (8), where $o(y)$ is a term whose ratio to $y$ tends to zero as $y \to 0$, and where $\mathcal{O}(y)$ is a term whose modulus is bounded above by a constant multiple of $|y|$.*

*Proof.* We are given $\underline{x} = (x, y)$, and we let $\underline{\hat{x}}$ be the point $(x, 0)$. The first component of $\underline{\nabla} F(\underline{\hat{x}})$ is zero because $F$ is constant on the straight line segment from $(-1, 0)$ to $(1, 0)$, and the second component is $\gamma(x)$ by the definition (8). Therefore it is sufficient to show that the first and second components of the difference $\underline{\nabla} F(\underline{x}) - \underline{\nabla} F(\underline{\hat{x}})$ are $y\,\gamma'(x) + o(y)$ and $\mathcal{O}(y)$, respectively. We infer the latter condition from the boundedness of second derivatives on compact sets and from the identity $\|\underline{x} - \underline{\hat{x}}\| = |y|$. Furthermore, the first condition can be deduced from the elementary relation

$$\underline{\nabla} F(\underline{x}) - \underline{\nabla} F(\underline{\hat{x}}) = \nabla^2 F(\underline{\hat{x}})\,(\underline{x} - \underline{\hat{x}}) + o(\|\underline{x} - \underline{\hat{x}}\|). \tag{36}$$

Specifically, because $\underline{x} - \underline{\hat{x}}$ is the vector $(0, y) \in \mathbb{R}^2$, the first component of the product $\nabla^2 F(\underline{\hat{x}})\,(\underline{x} - \underline{\hat{x}})$ is exactly $y\,\partial^2 F(\underline{\hat{x}})/\partial x\,\partial y = y\,\gamma'(x)$, so the identity $\|\underline{x} - \underline{\hat{x}}\| = |y|$ gives the required result.

□

**Lemma 7.** *Let $k$ be any integer such that the first and second components of the search direction $\underline{d}_{k-1}$ are positive and negative, respectively, and let $(\xi_k, 0)$ be the coordinates of the point where the straight line through $\underline{x}_{k-1}$ and $\underline{x}_k$ intersects the $x$-axis. If $\xi_k \le \rho$ holds for any $\rho \in \mathbb{R}$ that is independent of $k$, then $\xi_k$ has the property*

$$\xi_k = \phi(x_k) + o(1), \tag{37}$$

*where $o(1)$ is still a term that tends to zero as $k \to \infty$. Further, $\xi_k \le \rho$ implies $\gamma'(x_k) > 0$ for sufficiently large $k$.*

*Proof.* Straightforward algebra gives the formula

$$\xi_k = x_k - y_k\,(x_k - x_{k-1})/(y_k - y_{k-1}), \tag{38}$$

where $\underline{x}_k = (x_k, y_k)$ and $\underline{x}_{k-1} = (x_{k-1}, y_{k-1})$. Furthermore, because equation (4) shows that $\underline{d}_{k-1}$ is orthogonal to $\underline{\nabla} F(\underline{x}_k)$, Lemma 6 provides the relation

$$\frac{x_k - x_{k-1}}{y_k - y_{k-1}} = -\frac{\gamma(x_k) + \mathcal{O}(y_k)}{y_k\,\gamma'(x_k) + o(y_k)} = -\frac{\gamma(x_k) + o(1)}{y_k\,[\gamma'(x_k) + o(1)]}, \tag{39}$$

the last equation being due to $y_k \to 0$ as $k \to \infty$, which is one of the conclusions of Sect. 2. Therefore $\xi_k$ has the form

$$\xi_k = x_k + [\gamma(x_k) + o(1)] \,/\, [\gamma'(x_k) + o(1)]. \tag{40}$$

Now the conditions of the lemma with $y_k > 0$ imply $x_k < \xi_k \le \rho$, so, using $\gamma(x_k) \ge \gamma_{\min} > 0$, we deduce from expression (40) that $\gamma'(x_k)$ is bounded below by a positive constant for sufficiently large $k$. It follows that the right hand side of equation (40) has the form $x_k + \gamma(x_k)/\gamma'(x_k) + o(1)$. Thus the definition (32) gives $\xi_k = \phi(x_k) + o(1)$, which completes the proof of condition (37). This proof includes the assertion $\gamma'(x_k) > 0$ when $k$ is sufficiently large. Therefore the last statement of the lemma is also true.

□

For each $k \in \mathcal{K}_\star$, we consider the numbers $\xi_j$, $m(k)+1 \le j \le k$, where $\mathcal{K}_\star$ and $m(k)$ are defined after the proof of Lemma 5. These definitions provide $x_{m(k)} \le a < x_{m(k)+1}$ and $x_{k-1} < b \le x_k$, so the first components of $\underline{d}_{m(k)}$ and $\underline{d}_{k-1}$ are positive. It will be shown next that we may assume without loss of generality that their second components are negative.

We let $(a, \bar{y})$ be the point where the line segment from $\underline{x}_{m(k)}$ to $\underline{x}_{m(k)+1}$ cuts $x = a$. The line search of the algorithm of Sect. 1 satisfies $\underline{d}_{m(k)}^T \underline{\nabla} F(a, \bar{y}) \le 0$, which is combined with another application of Lemma 6. Specifically, letting $d_x$ and $d_y$ be the components of $\underline{d}_{m(k)}$, and using the form (35) of $\underline{\nabla} F(a, \bar{y})$, we find the inequality

$$d_x\,[\bar{y}\,\gamma'(a) + o(\bar{y})] + d_y\,[\gamma(a) + \mathcal{O}(\bar{y})] \le 0. \tag{41}$$

Now $\bar{y}$ is positive, and $\gamma(a)$ and $\gamma'(a)$ are positive constants. It follows from $d_x > 0$ that $d_y < 0$ occurs for sufficiently large $k$. Therefore, by deleting some early iterations of the algorithm if necessary, we obtain $d_y < 0$, $k \in \mathcal{K}_\star$, as claimed. Further, by analogy with equations (38) and (39), we deduce the bounds

$$a < \xi_{m(k)+1} = a - \bar{y}\,(d_x/d_y) \le a + \bar{y}\,[\gamma(a) + \mathcal{O}(\bar{y})] \,/\, [\bar{y}\,\gamma'(a) + o(\bar{y})], \tag{42}$$

which can be written in the form

$$a < \xi_{m(k)+1} \le \phi(a) + o(1), \qquad k \in \mathcal{K}_\star. \tag{43}$$

We have begun to prove that the second component of $\underline{d}_{k-1}$ is negative for $k \in \mathcal{K}_\star$. Indeed, the previous paragraph treats the possibility $m(k) = k-1$. Otherwise, when $m(k) \le k-2$ occurs, we have $a \le x_{k-1} \le b$. Therefore $\gamma'(x_{k-1}) \ge \gamma'_{\min}$ is satisfied, where $\gamma'_{\min}$ is the constant $\min\{\gamma'(x) : a \le x \le b\}$, which is positive due to Lemma 5.

Hence, remembering $y_{k-1} > 0$ and $\gamma(x_{k-1}) \geq \gamma_{\min} > 0$, we deduce from equation (35) that both components of $\underline{\nabla} F(\underline{x}_{k-1})$ are positive for sufficiently large $k$. We avoid the last proviso by deleting some early iterations of the algorithm if necessary. It follows from the descent condition $\underline{d}_{k-1}^T \underline{\nabla} F(\underline{x}_{k-1}) < 0$, and from the positivity of the first component of $\underline{d}_{k-1}$, that the second component of $\underline{d}_{k-1}$ is negative for every $k$ in $\mathcal{K}_\star$.

Therefore Lemma 7 is applicable for $k \in \mathcal{K}_\star$. Hence, letting $\rho$ be a constant such that $\rho > \phi(x_\star)$, we find the bound

$$\xi_k \geq \min[\phi(x_k) + o(1), \rho], \qquad k \in \mathcal{K}_\star. \tag{44}$$

Now, when $\xi_k \leq \rho$ occurs, then the last statement of Lemma 7 provides $\gamma'(x_k) > 0$ for sufficiently large $k$. It follows from $x_k \geq b$ that $x_k$ is in the set (34), so the choice of $x_\star$ gives $\phi(x_\star) \leq \phi(x_k)$. Hence condition (44) and $\rho > \phi(x_\star)$ imply that $\xi_k$ has the property

$$\xi_k \geq \phi(x_\star) + o(1), \qquad k \in \mathcal{K}_\star. \tag{45}$$

The contradiction that will complete the work of this section is suggested by the relations (43) and (45) when $m(k)$ is $k-1$. We see that in this case $\xi_k$ is bounded above by $\phi(a) + o(1)$ and is bounded below by $\phi(x_\star) + o(1)$. On the other hand, Lemma 5 establishes the strict inequality $\phi(a) < \phi(x_\star)$. Therefore the value $m(k) = k$–1 is excluded for sufficiently large $k$ in $\mathcal{K}_\star$. The analysis of the remaining situation $m(k) \leq k-2$ will be assisted by the following lemma.

**Lemma 8.** *Let $j$ be an integer such that $a \leq x_j \leq b$ and $x_{j-1} < x_j$ are satisfied. If $j$ is sufficiently large, and if $j$ is in the set $\mathcal{K}_{same}$ of expression (29), then the strict inequality $\xi_{j+1} < \xi_j$ is achieved. Moreover, if $j$ is sufficiently large, then the conditions $a \leq x_{j-1} < x_j \leq b$ are sufficient for $j$ to be in the set $\mathcal{K}_{same}$.*

*Proof.* By applying an argument in the paragraph after expression (43), we deduce from $a \leq x_j \leq b$ that both components of $\underline{g}_j = \underline{\nabla} F(\underline{x}_j)$ are positive for sufficiently large $j$. Therefore the line search condition $\underline{g}_j^T \underline{d}_{j-1} = 0$ implies that the two components of $\underline{d}_{j-1}$ have opposite signs, the first component being positive due to $x_{j-1} < x_j$. Thus $\xi_j$ is well-defined and satisfies $\xi_j > x_j$.

We also know from the work of Sect. 2 that, for large $j$, the directions of $\underline{d}_{j-1}$ and $\underline{d}_j$ tend to be parallel to the $x$-axis in $\mathbb{R}^2$. Hence $x_{j-1} < x_j$ and $j \in \mathcal{K}_{\text{same}}$ cause both directions to be near the positive coordinate direction $(1, 0)$. Further, the conditions $\underline{g}_j^T \underline{d}_{j-1} = 0$ and $\underline{g}_j^T \underline{d}_j < 0$ hold for every $j$, and $\underline{g}_j / \|\underline{g}_j\|$ tends to $(0, 1)$ as $j \to \infty$. Therefore, if $L_{j-1}$ and $L_j$ are the half-lines in $\mathbb{R}^2$ that begin at $\underline{x}_j$ and that have the directions $\underline{d}_{j-1}$ and $\underline{d}_j$, respectively, then $L_{j-1}$ can be mapped into $L_j$ by a small clockwise rotation about $\underline{x}_j$. Now $y_j > 0$ implies that a clockwise rotation of $L_{j-1}$ would decrease the first coordinate of the point where $L_{j-1}$ cuts the $x$-axis. Thus, because $(\xi_j, 0)$ and $(\xi_{j+1}, 0)$ are the points of intersection of $L_{j-1}$ and $L_j$ with the $x$-axis, the required inequality $\xi_{j+1} < \xi_j$ is achieved.

In order to prove the other statement of the lemma, we assume $a \leq x_{j-1} < x_j \leq b$, and we seek the sign of the scalar product $\underline{g}_j^T (\underline{g}_j - \underline{g}_{j-1})$. We recall the elementary identity

$$\underline{g}_j - \underline{g}_{j-1} = \int_0^1 \nabla^2 F\big(\underline{x}_{j-1} + \theta\,[\underline{x}_j - \underline{x}_{j-1}]\big)(\underline{x}_j - \underline{x}_{j-1})\,d\theta. \qquad (46)$$

Because we will find that the scalar product is of the same magnitude as $|x_j - x_{j-1}|$, we employ the notation $o(|x_j - x_{j-1}|)$ for terms whose ratio to $|x_j - x_{j-1}|$ tends to zero as $j \to \infty$. In particular, the direction of $\underline{d}_{j-1}$ provides the condition

$$\underline{x}_j - \underline{x}_{j-1} = \underline{\hat{x}}_j - \underline{\hat{x}}_{j-1} + o(|x_j - x_{j-1}|), \qquad (47)$$

where $\underline{\hat{x}}_{j-1}$ and $\underline{\hat{x}}_j$ are the points $(x_{j-1}, 0)$ and $(x_j, 0)$, respectively. Further, because $y_{j-1}$ and $y_j$ tend to zero as $j \to \infty$, we write expression (46) in the form

$$\begin{aligned} \underline{g}_j - \underline{g}_{j-1} &= \int_0^1 \nabla^2 F\big(\underline{\hat{x}}_{j-1} + \theta\,[\underline{\hat{x}}_j - \underline{\hat{x}}_{j-1}]\big)(\underline{\hat{x}}_j - \underline{\hat{x}}_{j-1})\,d\theta + o(|x_j - x_{j-1}|) \\ &= \underline{\nabla} F(\underline{\hat{x}}_j) - \underline{\nabla} F(\underline{\hat{x}}_{j-1}) + o(|x_j - x_{j-1}|), \end{aligned} \qquad (48)$$

where the last line is elementary. Moreover, $\underline{g}_j$ tends to have the components 0 and $\gamma(x_j)$ as $j \to \infty$. Thus condition (48) gives the equation

$$\underline{g}_j^T (\underline{g}_j - \underline{g}_{j-1}) = \gamma(x_j)\,[\gamma(x_j) - \gamma(x_{j-1})] + o(|x_j - x_{j-1}|), \qquad (49)$$

which is valid for all magnitudes of $|x_j - x_{j-1}|$ that are allowed by the assumptions $a \le x_{j-1} < x_j \le b$.

Now these assumptions imply that $\gamma(x_j) - \gamma(x_{j-1})$ is bounded below by the product $(x_j - x_{j-1}) \min\{\gamma'(x) : a \le x \le b\}$. Therefore equation (49) provides the inequality

$$\underline{g}_j^T (\underline{g}_j - \underline{g}_{j-1}) \ge (x_j - x_{j-1})\,[\gamma_{\min}\gamma'_{\min} + o(1)], \qquad (50)$$

where $\gamma_{\min}$ and $\gamma'_{\min}$ are positive constants that have been defined already. It follows that $j$ is in the set $\mathcal{K}_{\text{same}}$ for sufficiently large $j$, which completes the proof of the lemma.

□

Next we apply the lemma to the case when, in addition to $k \in \mathcal{K}_\star$ and $m(k) \le k-2$, the iteration number $m(k)+1$ is in the set $\mathcal{K}_{\text{same}}$. We recall that, if $j$ satisfies $x_{j-1} < x_j$ and $j \in \mathcal{K}_{\text{same}}$, then $x_j < x_{j+1}$ occurs. Thus the definition of $m(k)$ and the choice $j = m(k)+1$ provide $x_{m(k)} \le a < x_j < x_{j+1}$. Further, the last part of Lemma 8 shows that, if $x_{j+1} < b$ holds and if $j$ is increased by one, then the new $j$ is also in $\mathcal{K}_{\text{same}}$, which gives $a < x_j < x_{j+1}$ for the new $j$. By employing these properties recursively for additions of one to $j$ until $x_{j+1} \ge b$ is obtained, we deduce the conditions

$$x_{m(k)} \le a < x_{m(k)+1} < \cdots < x_{k-1} < b \le x_k. \qquad (51)$$

It follows from the first part of Lemma 8 that $\xi_{j+1} < \xi_j$ holds for every integer $j$ in the interval $[m(k)+1, k-1]$, so expression (43) implies $\xi_k \le \phi(a) + o(1)$. Hence the inequalities (45) and $\phi(a) < \phi(x_\star)$ exclude all large values of $k$ as before. It remains to exclude large values of $k$ in $\mathcal{K}_\star$, satisfying $m(k) \le k-2$, such that $m(k)+1$ is an element of $\mathcal{K}_{\text{opp}}$.

In this case the definitions of $m(k)$ and $\mathcal{K}_{\text{opp}}$ provide $x_{m(k)} \le a < x_{m(k)+2} < x_{m(k)+1} < b$. Most of our analysis of this situation will be presented in Lemma 9, where $j$ corresponds to $m(k)+2$. Specifically, assuming that $k$ is sufficiently large, it will be proved that $m(k)+2$ is in $\mathcal{K}_{\text{opp}}$, which gives $x_{m(k)+2} < x_{m(k)+3} < \xi_{m(k)+3}$. It will also be proved that $\xi_{m(k)+3}$ has the property $\xi_{m(k)+3} < \xi_{m(k)+1}$. If $x_{m(k)+3} \ge b$ occurs, then $k$ is equal to $m(k)+3$. Otherwise we have $a < x_{m(k)+2} < x_{m(k)+3} < b$, so the recursive argument of the previous paragraph allows expression (51) to be replaced by the conditions

$$x_{m(k)} \le a < x_{m(k)+2} < \cdots < x_{k-1} < b \le x_k. \tag{52}$$

Further, using the first part of Lemma 8 again, we find $\xi_k < \xi_{m(k)+3}$. Therefore $\xi_k \le \xi_{m(k)+3}$ occurs for all the integers $k$ that are being considered. Thus the assertion $\xi_{m(k)+3} < \xi_{m(k)+1}$ of the next lemma and expression (43) imply $\xi_k \le \phi(a) + o(1)$. It follows yet again that the bound (45) prevents $k$ from becoming large, which completes the contradiction of the hypothesis that the number of elements in $\mathcal{K}_\star$ is infinite. We have now established the required condition (9), assuming that Lemma 9 is true. Its proof is the last task of this section.

**Lemma 9.** *Let $j$ be an integer that satisfies $x_{j-2} < x_{j-1}$ and $a \le x_j < x_{j-1} \le b$. If $j$ is sufficiently large, then the inequalities $x_j < x_{j+1} < \xi_{j+1} < \xi_{j-1}$ are achieved.*

*Proof.* The conditions $a \le x_{j-1} < x_j \le b$ are assumed in the paragraph that includes equations (46) to (49), but it is elementary that the derivation of these equations is also valid for the current conditions $a \le x_j < x_{j-1} \le b$, because the search directions still tend to be parallel to the $x$-axis. Further, the analogue of inequality (50) is the property

$$\underline{g}_j^T (\underline{g}_j - \underline{g}_{j-1}) \le (x_j - x_{j-1}) \, [\gamma_{\min} \gamma'_{\min} + o(1)], \tag{53}$$

because now the sign of $x_j - x_{j-1}$ is negative. Thus $j$ is in the set $\mathcal{K}_{\text{opp}}$ when $j$ is large. It follows from $x_j < x_{j-1}$ that $x_j < x_{j+1}$ occurs as required.

Therefore we assume $j \in \mathcal{K}_{\text{opp}}$ for the remainder of the proof. The argument at the beginning of the proof of Lemma 8, with $j$ replaced by $j-1$, shows that we may also assume that $\xi_{j-1}$ is well-defined and satisfies $\xi_{j-1} > x_{j-1}$. Moreover, we let $j$ be at least the constant $k_0$ of condition (31), in order to have the inequality

$$\|\underline{g}_j\| \,/\, \|\underline{d}_j\| \ge \|\underline{g}_{j-1}\| \,/\, \|\underline{d}_{j-1}\|. \tag{54}$$

Thus the identity (26) implies $\theta_j \le \theta_{j-1}$, where $\theta_k$ is still the acute angle between the directions $\underline{d}_k$ and $-\underline{g}_k$ for every positive integer $k$. The angles $\theta_{j-1}$ and $\theta_j$ are shown in Fig. 1.

Let $L_k$ be the straight line through $\underline{x}_k$ and $\underline{x}_{k+1}$ for $k = j-2, \, j-1, \, j$. Then equation (4) states that $-\underline{g}_{j-1}$ and $-\underline{g}_j$ are orthogonal to $L_{j-2}$ and $L_{j-1}$, respectively. Thus the figure provides the identities

$$\psi_{j-1} + \theta_{j-1} = \psi_j + \theta_j = \pi/2, \tag{55}$$

where $\psi_k$ is the small acute angle between $L_{k-1}$ and $L_k$ for $k = j-1, \, j$. Therefore the conclusion $\theta_j \le \theta_{j-1}$ of the previous paragraph gives $\psi_j \ge \psi_{j-1}$. Hence we deduce

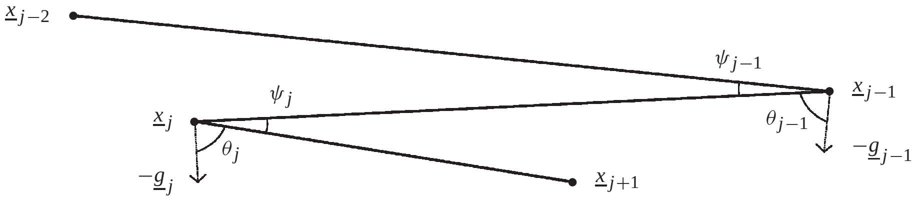

**Fig. 1.** The angles $\theta_{j-1}$ and $\theta_j$

from the figure that either $L_{j-2}$ is parallel to $L_j$ or that these lines meet at a point whose first coordinate is less than $x_j$. In other words, the straight line extension from $\underline{x}_j$ through $\underline{x}_{j+1}$ does not meet $L_{j-2}$. This extension, however, has to leave the region of $\mathbb{R}^2$ that is below $L_{j-2}$, above the $x$-axis, and to the right of the line $x = x_{j+1}$. It follows that $L_j$ cuts the $x$-axis at a point that is to the right of $(x_{j+1}, 0)$ and to the left of $(\xi_{j-1}, 0)$. Further, this point is $(\xi_{j+1}, 0)$ according to the definition of $\xi_{j+1}$. Therefore $\xi_{j+1}$ satisfies $x_{j+1} < \xi_{j+1} < \xi_{j-1}$. The analysis of this section is complete.

□

## 4. The convergence theorem

The method that will be used to complete the analysis is indicated in the paragraph that follows expression (31). Therefore we let $k$ be a sufficiently large element of $\mathcal{K}_{\text{opp}}$, and we let $\ell(k)$ be the greatest element of $\mathcal{K}_{\text{opp}}$ that is less than $k$, as before. If $j$ is an integer in the interval $[\ell(k), k-2]$, then our justification of the condition $\|\underline{d}_k\|/\|\underline{g}_k\| \leq \|\underline{d}_j\|/\|\underline{g}_j\|$ that has been mentioned requires the following lemma. The only application of the bound (9) occurs in its proof.

**Lemma 10.** *Let $j$ and $k$ be chosen in the way just described, and let $\rho$ be the ratio*

$$\rho = \min\{\gamma(x) : -1 \leq x \leq 1\} \,/\, \max\{\gamma(x) : -1 \leq x \leq 1\}, \tag{56}$$

*where $\gamma$ is the positive, continuous function (8). Then the inequality*

$$\left(\frac{\|\underline{g}_{k-1}\|}{\|\underline{g}_j\|} - 1\right)\left(1 - \frac{\|\underline{g}_k\|}{\|\underline{g}_{k-1}\|}\right) \leq \frac{1+o(1)}{1+\rho} \tag{57}$$

*holds, where $1+o(1)$ still denotes a number that tends to one as $k \to \infty$.*

*Proof.* The definition (29) of $\mathcal{K}_{\text{opp}}$ and $k \in \mathcal{K}_{\text{opp}}$ imply $\|\underline{g}_k\| < \|\underline{g}_{k-1}\|$. Thus inequality (57) is achieved when $\underline{g}_j$ satisfies $\|\underline{g}_{k-1}\|/\|\underline{g}_j\| \leq 1 + (1+\rho)^{-1}$. Therefore for the remainder of the proof we can assume that this property fails, so we have the condition

$$\|\underline{g}_{k-1}\| > \|\underline{g}_j\| + (1+\rho)^{-1}\,\|\underline{g}_j\| > \|\underline{g}_j\| + (1+\rho)^{-1}\varepsilon, \tag{58}$$

the last inequality being due to assumption (7). Moreover, the ratio $\|\underline{g}_{k-1}\|/\|\underline{g}_j\|$ is bounded above by $M/\varepsilon$, where $M$ is the constant $\sup\{\|\underline{g}_i\| \,:\, i = 1,2,3,\ldots\}$. Thus inequality (57) is also achieved if $\underline{g}_k$ satisfies $1-\|\underline{g}_k\|/\|\underline{g}_{k-1}\| \le (1+\rho)^{-1}$ $(\varepsilon/M)$. Therefore we can assume that this property fails too, which gives the condition

$$\|\underline{g}_{k-1}\| > \|\underline{g}_k\| + (1+\rho)^{-1}(\varepsilon/M)\,\|\underline{g}_{k-1}\| > \|\underline{g}_k\| + (1+\rho)^{-1}(\varepsilon^2/M). \tag{59}$$

We relate inequality (57) to the derivative (8) by letting $(x_j,0)$, $(x_{k-1},0)$ and $(x_k,0)$ be the points on the line segment from $(-1,0)$ to $(1,0)$ in $\mathbb{R}^2$ that are closest to $\underline{x}_j$, $\underline{x}_{k-1}$ and $\underline{x}_k$, respectively. Then, because the calculated vectors of variables tend to the line segment, and because the first and second components of $\underline{\nabla} F$ are zero and positive there, we have the formulae

$$\|\underline{g}_i\| = \big(1+o(1)\big)\,\gamma(x_i), \qquad i\in\{j,\;k-1,\;k\}. \tag{60}$$

Moreover, expressions (58) and (59) show that $\|\underline{g}_{k-1}\|/\|\underline{g}_j\|$ and $\|\underline{g}_k\|/\|\underline{g}_{k-1}\|$ are bounded away from one. It follows that assertion (57) is true if we establish the condition

$$\left(\frac{\gamma(x_{k-1})}{\gamma(x_j)}-1\right)\left(1-\frac{\gamma(x_k)}{\gamma(x_{k-1})}\right) \le \frac{1+o(1)}{1+\rho}. \tag{61}$$

Another advantage of expressions (58) and (59) is that they ensure that $|x_j - x_{k-1}|$ and $|x_{k-1}-x_k|$ are bounded away from zero as $k\to\infty$. Further, because the choice of $j$ causes every integer in the interval $[j+1,k-1]$ to be in the set $\mathcal{K}_{\text{same}}$, the component $x_{k-1}$ is strictly between $x_j$ and $x_k$ for sufficiently large $k$. Therefore symmetry allows the assumption

$$-1 \le x_j < x_{k-1} < x_k \le 1. \tag{62}$$

The proof will be completed by deducing inequality (61) from this assumption and the property (9).

It is elementary that we can write expression (9) in the form

$$-1\,/\,(1+x) \le \gamma'(x)\,/\,\gamma(x) \le 1\,/\,(1-x), \qquad -1\le x\le 1. \tag{63}$$

Hence, if $a$ and $b$ are any numbers that satisfy $-1\le a<b\le 1$, integration over the interval $a\le x\le b$ gives the bounds

$$-\log(1+b)+\log(1+a) \le \log(\gamma(b))-\log(\gamma(a)) \le -\log(1-b)+\log(1-a), \tag{64}$$

which are equivalent to the conditions

$$(1+a)\,/\,(1+b) \le \gamma(b)\,/\,\gamma(a) \le (1-a)\,/\,(1-b), \qquad -1\le a<b\le 1. \tag{65}$$

It follows from the ordering (62) that the left hand inequalities of the expression

$$\left.\begin{array}{l} \gamma(x_{k-1}) \;\le\; [(1-x_j)/(1-x_{k-1})]\,\gamma(x_j) \;\le\; 2\,\gamma(x_j)\,/\,(1-x_{k-1}) \\ \gamma(x_{k-1}) \;\le\; [(1+x_k)/(1+x_{k-1})]\,\gamma(x_k) \;\le\; 2\,\gamma(x_k)\,/\,(1+x_{k-1}) \end{array}\right\} \tag{66}$$

are achieved, and the right hand inequalities are due to $x_j \geq -1$ and $x_k \leq 1$. Thus we deduce the property

$$\gamma(x_{k-1}) \leq \min\left[\frac{2\,\gamma(x_j)}{1-x_{k-1}},\ \frac{2\,\gamma(x_k)}{1+x_{k-1}}\right] \leq \max_{-1\leq\theta\leq 1}\,\min\left[\frac{2\,\gamma(x_j)}{1-\theta},\ \frac{2\,\gamma(x_k)}{1+\theta}\right]. \qquad (67)$$

Further, $\theta$ maximizes the right hand side of this bound when the two terms inside the square brackets are equal, so $\theta$ takes the value $[\gamma(x_k)-\gamma(x_j)]\,/\,[\gamma(x_k)+\gamma(x_j)]$. Thus condition (67) provides the second of the inequalities

$$\max[\gamma(x_j),\,\gamma(x_k)] \leq \gamma(x_{k-1}) \leq \gamma(x_j) + \gamma(x_k), \qquad (68)$$

the first of them being due to the assumptions (58) and (59) for large enough $k$.

Now the left hand side of the required bound (61) increases monotonically if $\gamma(x_{k-1})$ runs through the interval (68), so we find the relation

$$\left(\frac{\gamma(x_{k-1})}{\gamma(x_j)} - 1\right)\left(1 - \frac{\gamma(x_k)}{\gamma(x_{k-1})}\right) \leq \frac{\gamma(x_k)}{\gamma(x_j)}\,\frac{\gamma(x_j)}{\gamma(x_j)+\gamma(x_k)} = \frac{\gamma(x_k)}{\gamma(x_j)+\gamma(x_k)}. \qquad (69)$$

Further, the definition (56) implies that the right hand side of this expression is at most $(1+\rho)^{-1}$. Therefore the lemma is true.

□

Next we establish the condition $\|\underline{d}_k\|/\|\underline{g}_k\| \leq \|\underline{d}_j\|/\|\underline{g}_j\|$, $\ell(k) \leq j \leq k-2$, that is the subject of the opening paragraph of this section.

**Lemma 11.** *There exists a constant integer $k_1$ such that, if $j$ and $k$ are chosen as in Lemma 10, then $k \geq k_1$ implies the inequality*

$$\|\underline{d}_k\|\,/\,\|\underline{g}_k\| \leq \|\underline{d}_j\|\,/\,\|\underline{g}_j\|. \qquad (70)$$

*Proof.* The choices of $j$ and $k$ are such that the integers $j+1,\ j+2,\ \ldots,\ k-1$ are all in the set $\mathcal{K}_{\text{same}}$. Therefore equation (28) gives the relation

$$\frac{\|\underline{d}_i\|}{\|\underline{d}_{i-1}\|} = \frac{\|\underline{g}_i\|^2}{\|\underline{g}_{i-1}\|^2}\left\{\big(1+o(1)\big)\left(1 - \frac{\underline{g}_i^T\,\underline{g}_{i-1}}{\|\underline{g}_i\|^2}\right)\right\} < \frac{\|\underline{g}_i\|^2}{\|\underline{g}_{i-1}\|^2}, \quad j+1 \leq i \leq k-1, \qquad (71)$$

for sufficiently large $k$, where the last part depends on the remark that the gradients tend to be multiples of the second coordinate vector, each multiplier being bounded above and bounded away from zero. We form the product over $i$ of the left and right hand sides of expression (71), except that we retain the middle term of the expression instead of the right hand term for one value of $i$. Thus we find the condition

$$\frac{\|\underline{d}_{k-1}\|}{\|\underline{d}_j\|} \leq \frac{\|\underline{g}_{k-1}\|^2}{\|\underline{g}_j\|^2}\,\min_{j+1\leq i\leq k-1}\left\{\big(1+o(1)\big)\left(1 - \frac{\underline{g}_i^T\,\underline{g}_{i-1}}{\|\underline{g}_i\|^2}\right)\right\}. \qquad (72)$$

If the last term in braces were at most $\varepsilon/M$, where $M$ is still the constant $\sup\{\|\underline{g}_i\| :\ i=1,2,3,\ldots\}$, then condition (72) would imply $\|\underline{d}_{k-1}\|/\|\underline{d}_j\| \leq \|\underline{g}_{k-1}\|/\|\underline{g}_j\|$, which is the same as $\|\underline{d}_{k-1}\|/\|\underline{g}_{k-1}\| \leq \|\underline{d}_j\|/\|\underline{g}_j\|$. Further, the property (31) would give the required result (70). Therefore, for the remainder of the proof, we may make the assumption

$$\big(1+o(1)\big)\Big(1-\underline{g}_i^T\underline{g}_{i-1}\,/\,\|\underline{g}_i\|^2\Big) \geq \varepsilon/M, \qquad j+1 \leq i \leq k-1. \tag{73}$$

This assumption and $\underline{g}_i^T\underline{g}_{i-1} \to \|\underline{g}_i\|\,\|\underline{g}_{i-1}\|$, $i\to\infty$, provide the relations

$$\left.\begin{array}{c}\big(1+o(1)\big)\Big(1-\underline{g}_i^T\underline{g}_{i-1}\,/\,\|\underline{g}_i\|^2\Big) = \big(1+o(1)\big)\Big(1-\|\underline{g}_{i-1}\|\,/\,\|\underline{g}_i\|\Big)\\ \big(1+o(1)\big)\Big(1-\|\underline{g}_{i-1}\|/\,\|\underline{g}_i\|\Big) \geq \varepsilon/M\\ \|\underline{g}_{i-1}\|\,/\,\|\underline{g}_i\| \leq \big(1+o(1)\big)\big(1-\varepsilon/M\big) < \Big(1-\tfrac{1}{2}\varepsilon/M\Big)\end{array}\right\}, \tag{74}$$

when $i$ is any integer in $[j+1,k-1]$ and $k$ is sufficiently large, the first part of the third property being a reformulation of the second one. By combining the first parts of expressions (71) and (74), we deduce the equation

$$\frac{\|\underline{d}_i\|}{\|\underline{d}_{i-1}\|} = \frac{\|\underline{g}_i\|}{\|\underline{g}_{i-1}\|}\left\{\big(1+o(1)\big)\left(\frac{\|\underline{g}_i\|}{\|\underline{g}_{i-1}\|}-1\right)\right\}, \qquad j+1 \leq i \leq k-1. \tag{75}$$

Then we form the product over all values of $i$ again. The product of the $1+o(1)$ terms is another $1+o(1)$ term, because $k-j$ is bounded above for large $k$ due to the following three remarks. Firstly, the step-lengths $\|\underline{x}_i-\underline{x}_{i-1}\|$, $j+1 \leq i \leq k-1$, are bounded away from zero due to the third line of expression (74), secondly the directions of these steps tend to be the same due to the choice of $j$, and thirdly the length of the line segment $\mathcal{T}$ is finite. Thus equation (75) provides the relation

$$\frac{\|\underline{d}_{k-1}\|}{\|\underline{d}_j\|} = \big(1+o(1)\big)\frac{\|\underline{g}_{k-1}\|}{\|\underline{g}_j\|}\prod_{i=j+1}^{k-1}\left(\frac{\|\underline{g}_i\|}{\|\underline{g}_{i-1}\|}-1\right). \tag{76}$$

Now it is elementary that, if $a\in\mathbb{R}$ and $b\in\mathbb{R}$ satisfy $a>1$ and $b>1$, then $(a-1)(b-1)<ab-1$ holds, which is helpful because, according to expression (74), all of the ratios $\|\underline{g}_i\|/\|\underline{g}_{i-1}\|$ exceed one. By applying this remark recursively to the terms of the product on the right of equation (76) when there is more than one term, we find the condition

$$\frac{\|\underline{d}_{k-1}\|}{\|\underline{d}_j\|} \leq \big(1+o(1)\big)\frac{\|\underline{g}_{k-1}\|}{\|\underline{g}_j\|}\left(\frac{\|\underline{g}_{k-1}\|}{\|\underline{g}_j\|}-1\right). \tag{77}$$

Moreover, inequality (30) gives the bound

$$\frac{\|\underline{d}_k\|}{\|\underline{g}_k\|} \leq \big(1+o(1)\big)\frac{\|\underline{d}_{k-1}\|}{\|\underline{g}_{k-1}\|}\left(1-\frac{\|\underline{g}_k\|}{\|\underline{g}_{k-1}\|}\right). \tag{78}$$

By combining expressions (77) and (78), we obtain the property

$$\frac{\|\underline{d}_k\|}{\|\underline{g}_k\|} \leq \bigl(1+o(1)\bigr)\,\frac{\|\underline{d}_j\|}{\|\underline{g}_j\|}\left(\frac{\|\underline{g}_{k-1}\|}{\|\underline{g}_j\|}-1\right)\left(1-\frac{\|\underline{g}_k\|}{\|\underline{g}_{k-1}\|}\right). \tag{79}$$

It follows from Lemma 10 that inequality (70) is achieved for sufficiently large $k$, because $\rho$ is a positive constant. The proof is complete.

□

It remains to apply the argument in the paragraph after inequality (31). Specifically, the bound (31) when $\ell(k)=k{-}1$, and Lemma 11 when $\ell(k)\leq k{-}2$, provide the condition

$$\|\underline{d}_k\| \,/\, \|\underline{g}_k\| \leq \|\underline{d}_{\ell(k)}\| \,/\, \|\underline{g}_{\ell(k)}\|, \qquad k\in\mathcal{K}_{\text{opp}}, \quad k\geq\max[k_0,k_1], \tag{80}$$

where $\ell(k)$ is still the greatest element of $\mathcal{K}_{\text{opp}}$ that is less than $k$. Thus the ratios $\|\underline{d}_k\|/\|\underline{g}_k\|$, $k\in\mathcal{K}_{\text{opp}}$, are uniformly bounded, which contradicts the limit (27), because the number of elements in $\mathcal{K}_{\text{opp}}$ is infinite. Therefore we have established the following result.

**Theorem 1.** *Let the DFP algorithm with exact line searches, as described in Sect. 1, be applied to an objective function of only two variables that has the properties (1). Then the termination condition (2) is achieved for some finite integer k.*

□

## 5. More than two variables

The work of this paper has made a small contribution to knowledge about properties of the DFP algorithm with exact line searches. We know after Sect. 2 that, if the gradients are bounded away from zero, then $\mathcal{T}$, which is the limit of the piecewise linear path that joins the calculated vectors of variables, is a straight line segment. Further, it is important to the subsequent analysis that the path is confined to one side of $\mathcal{T}$. This remark, however, depends on the assumption that the number of variables of the calculation is only two.

For more variables, the limiting path may still be a straight line segment, but infinite subsequences of $\underline{x}_k$, $k=1,2,3,\ldots$, can occur on all sides of $\mathcal{T}$, because in three or more dimensions there is room for a piecewise linear path to move from one side to the opposite side of a straight line without intersecting it. Thus it may be possible for $\|\underline{g}_k\|>\varepsilon$, $k=1,2,3,\ldots$, to hold for $n\geq 3$, where $\varepsilon$ is a positive constant.

Therefore Yu-hong Dai (private communications) and the author have put much effort into trying to construct an $n=3$ example, where the conditions (1) are satisfied, and where no iteration of the algorithm of Sect. 1 achieves the termination condition $\|\underline{g}_k\|\leq\varepsilon$ for some prescribed $\varepsilon>0$. They restrict attention to the case when the distance from $\underline{x}_k$ to the first coordinate axis tends to zero as $k\rightarrow\infty$. Further, letting $(\underline{x}_k)_j$ denote the $j$-th component of $\underline{x}_k$, they relate $\underline{x}_{k+\ell}$ to $\underline{x}_k$ by the equations

$$(\underline{x}_{k+\ell})_1=(\underline{x}_k)_1 \qquad \text{and} \qquad (\underline{x}_{k+\ell})_j=c\,(\underline{x}_k)_j, \qquad j=2,3,\ldots,n, \tag{81}$$

for every iteration number $k$, where $\ell$ and $c$ are a small positive integer and a constant from the interval $0 < c < 1$, respectively. Thus the sequence of calculated vectors of variables is defined by the choices of $\ell$, $c$ and $\underline{x}_k$, $k=1, 2, \ldots, \ell$. Then, using $n=3$, the gradients $\underline{g}_k$, $k=1, 2, 3, \ldots$, are deduced from the properties

$$\underline{g}_k^T(\underline{x}_k-\underline{x}_{k-1}) = 0 \qquad \text{and} \qquad (\underline{x}_{k+1}-\underline{x}_k)^T(\underline{g}_k-\underline{g}_{k-1}) = 0 \tag{82}$$

of the DFP algorithm with exact line searches, except that all the gradients can be scaled by a single constant. Examples of this kind are presented by Powell (1984) for the conjugate gradient method.

When $n = 3$, however, the need for the example to be consistent with the use of the DFP formula provides a substantial difference from the conjugate gradient method. This condition can be expressed as $\ell$ nonlinear equality constraints on the parameters of the example. Moreover, the line search of the algorithm of Sect. 1 imposes some nonlinear inequality constraints on the parameters that are analogous to ones that occur in Powell (1984). Thus the construction of the required example is expressed as a search for a feasible point of a nonlinear programming problem.

Several months of work on this problem have been unsuccessful, although Dai and Powell are very close to achieving a solution. Specifically, a feasible point can be found if an arbitrarily small relative tolerance is allowed in one of the line search conditions, and if all the other constraints have to hold. A suitable example does not occur in the limit as the tolerance tends to zero, because then two of the first components $(\underline{x}_k)_1$, $k=1, 2, \ldots, \ell$, tend to coincide.

A paper will not be written yet on that work, because there are still many unexplored ways of choosing the parameters that may give a feasible point. Indeed, it is hoped that further research will expose an $n=3$ optimization calculation, where the algorithm of Sect. 1 does not terminate, and where the conditions (81) are satisfied. The discovery of such a calculation would provide a nice contrast to the conclusions of our lengthy analysis of a DFP algorithm with exact line searches, when there are only two variables.

## References

1. Davidon, W.C. (1959): Variable metric method for minimization. Report ANL 5990 (revised), Argonne National Laboratory, Illinois, USA
2. Dixon, L.C.W. (1972): Quasi-Newton algorithms generate identical points. Math. Program. **2**, 383–387
3. Fletcher, R., Powell, M.J.D. (1963): A rapidly convergent descent method for minimization. Comput. J. **6**, 163–168
4. Polak, E., Ribière, G. (1969): Note sur la convergence de methodes de directions conjuguées. Rev. Française Informat. Recherche Opérationelle **16**, 35–43
5. Powell, M.J.D. (1970): A new algorithm for unconstrained optimization. In: Rosen, J.B., Mangasarian, O.L., Ritter, K., eds., Nonlinear Programming, 1970. Academic Press, New York, pp. 31–65
6. Powell, M.J.D. (1972): Some properties of the variable metric algorithm. In: Lootsma, F.A., ed., Numerical Methods for Nonlinear Optimization, 1971. Academic Press, London, pp. 1–17
7. Powell, M.J.D. (1984): Nonconvex minimization calculations and the conjugate gradient method. In: Griffiths, D.F., ed., Numerical Analysis Proceedings, 1983 (Lecture Notes in Mathematics 1066). Springer-Verlag, Berlin, pp. 122–141
8. Wolfe, P. (1970): Convergence theory in nonlinear programming. In: Abadie, J., ed., Integer and Nonlinear Programming, 1969. North-Holland, Amsterdam, pp. 1–36