Computational Optimization and Applications, 22, 5–35, 2002

On the Behavior of the Gradient Norm in the Steepest Descent Method*

JORGE NOCEDAL† nocedal@ece.nwu.edu, www.eecs.nwu.edu/~nocedal/PSfiles
ECE Department, Northwestern University, Evanston, Il 60208, USA

ANNICK SARTENAER** Annick.Sartenaer@fundp.ac.be
Department of Mathematics, Facultés Universitaires Notre-Dame de la Paix, Namur, Belgium

CIYOU ZHU
Department of Electrical and Computer Engineering, Northwestern University, Evanston, Il 60208-3118, USA

Abstract. It is well known that the norm of the gradient may be unreliable as a stopping test in unconstrained optimization, and that it often exhibits oscillations in the course of the optimization. In this paper we present results descibing the properties of the gradient norm for the steepest descent method applied to quadratic objective functions. We also make some general observations that apply to nonlinear problems, relating the gradient norm, the objective function value, and the path generated by the iterates.

Keywords: nonlinear optimization, unconstrained optimization, steepest descent method, behavior of gradient norm

1. Introduction

The sequence of gradient norms generated by algorithms for unconstrained optimization often exhibits oscillatory behavior, but it is not well understood whether the size of the oscillations is related to the conditioning of the problem and to the rate of convergence of the iteration. Since the norm of the gradient is often used in termination rules, it is also interesting to ask under what circumstances does it provide a good estimate of the accuracy in the optimal function value. In this paper we study the properties of the gradient norm for the steepest descent method applied to a quadratic objective function. We also present some results describing the path followed by the iterates, and the final accuracy in the function obtained in the presence of rounding errors.

We write the unconstrained optimization problem as

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

*Travel support for this research was provided by NATO grant CRG 960688.
†This author was supported by National Science Foundation grant CDA-9726385 and by Department of Energy grant DE-FG02-87ER25047-A004.
**Research Associate of the Belgian National Fund for Scientific Research.

Table 1. Final objective value and final square norm of the gradient obtained by two optimization methods on the PENALTY3 problem.

Algorithm	f	$\|g\|_2^2$
L-BFGS ($m = 5$)	$9.999458658 \times 10^{-4}$	6.66×10^{-5}
Inexact Newton	$9.999701976 \times 10^{-4}$	1.29×10^{-13}

where f is a twice continuously differentiable function whose gradient will be denoted by g.

The motivation for this work arose during the development of a limited memory code (L-BFGS-B) for bound constrained optimization [5, 14]. We observed that for some problems this code was unable to reduce the gradient norm $\|g(x)\|$ as much as we desired, but that LANCELOT [7] had no difficulties in doing so. Initially we reported this as a *failure* of the limited memory code to achieve high accuracy in the solution, but a closer examination of the results revealed that in some of these runs the limited memory code had actually produced a lower function value than LANCELOT. Several examples of this behavior are described in [14]. In Table 1 we present a striking example that was obtained when the inexact Newton method described in [6] and the limited memory code L-BFGS-B [14] (using $m = 5$ correction pairs) were applied to the unconstrained optimization problem PENALTY3 from the CUTE collection [4]. Both methods were run until no further progress could be made in reducing the objective function; we report the final function values and gradient square norms obtained by each method. (All the computations reported in this paper were performed in IEEE double precision arithmetic.)

This behavior of limited memory methods (and more generally of quasi-Newton methods) has been noted by other researchers [9, 12], and confirms the well-known fact that the gradient norm can be an unreliable measure of accuracy in the objective function f (see for example Chapter 8 in [10]).

Nevertheless there are good reasons for using the gradient norm to terminate optimization calculations. We know that it must be zero at a solution, its value is available at every iteration of a gradient-related method, and it requires no knowledge of the optimal function value f^* or the solution vector x^*. Because of this, it is used extensively in automatic stopping tests. For example, a variety of algorithms for constrained optimization, such as augmented Lagrangian and barrier methods, require the solution of unconstrained optimization subproblems, and the termination tests for these subproblems are usually based on the norm of the gradient.

The paper is organized as follows. In Section 2, we make some observations relating the size of the gradient and the accuracy in the objective function; they apply to general objective functions and are independent of the minimization algorithms used to solve the problem. The rest of the paper concentrates on the steepest descent method applied to quadratic functions. Section 3 summarizes the important results developed by Akaike [1] and extended by Forsythe [8]. In Section 4 we present an upper bound on the maximum oscillation in the gradient norm that can occur at any iteration, and in Section 5 we analyze the asymptotic behavior of the gradient norm in detail. The most important and relevant

results presented in Sections 3 to 5 are summarized in Theorem 5.2. In Section 6 we study the special case of a two-dimensional quadratic. We conclude in Section 7 by making some observations on the final accuracy in the objective function.

Notation. Machine accuracy (or unit roundoff) is denoted by **u**. We denote the condition number of a matrix A by $\gamma(A)$, or simply by γ when the argument is clear. Throughout the paper $\|\cdot\|$ denotes the ℓ_2 or Euclidean norm.

2. Accuracy in f vs gradient norm

Let us explore the relationship between the accuracy in the objective function, as measured by difference in function values

$$f(x) - f^*, \tag{2.1}$$

and the norm of the gradient,

$$\|g(x)\|, \tag{2.2}$$

which must be zero at a solution. Other norms can be used, but for the sake of concreteness we will focus our attention on the Euclidean norm of the gradient. Most of the results given in this section can be found in [10], but we derive them for clarity and completeness.

Using Taylor's theorem we have

$$f(x) = f^* + g(x^*)^T(x - x^*) + \frac{1}{2}(x - x^*)^T \hat{G}(x - x^*),$$

where $\hat{G} = \nabla^2 f(\zeta)$ for some ζ in the line segment connecting x and x^*. Noting that $g(x^*) = 0$ we obtain

$$f(x) - f^* = \frac{1}{2}\lambda(x)\|x - x^*\|^2, \tag{2.3}$$

where $\lambda(x)$ is the Rayleigh quotient of $\hat{G}$ in the direction $x - x^*$, and is defined by

$$\lambda(x) = \frac{(x - x^*)^T \hat{G}(x - x^*)}{\|x - x^*\|^2}. \tag{2.4}$$

Let us now consider the gradient. Taylor's theorem gives

$$g(x) = g(x^*) + \bar{G}(x - x^*),$$

where

$$\bar{G} = \int_0^1 \nabla^2 f(x + \tau(x^* - x))\, d\tau.$$

Thus

$$\|g(x)\|^2 = \bar{\lambda}(x)\|x - x^*\|^2, \tag{2.5}$$

where

$$\bar{\lambda}(x) = \frac{(x - x^*)^T \bar{G}^2 (x - x^*)}{\|x - x^*\|^2} \tag{2.6}$$

is the Rayleigh quotient of $\bar{G}^2$ in the direction $x - x^*$. Thus $f(x) - f^*$ and $\|g(x)\|^2$ are both proportional to $\|x - x^*\|^2$, and combining (2.3) and (2.5) we obtain

$$f(x) - f^* = \frac{1}{2}\left[\frac{(x - x^*)^T \hat{G}(x - x^*)}{(x - x^*)^T \bar{G}^2 (x - x^*)}\right]\|g(x)\|^2. \tag{2.7}$$

There is a simple geometrical interpretation of (2.7) in the case where the objective function is a strongly convex quadratic,

$$f(x) = \frac{1}{2}x^T G x,$$

where G is positive definite. In this case $\hat{G} = \bar{G} = G$ and (2.7) becomes

$$f(x) - f^* = \frac{1}{2}\left[\frac{\|z\|^2}{z^T G z}\right]\|g(x)\|^2, \tag{2.8}$$

where $z = G^{\frac{1}{2}}(x - x^*)$. In figure 1 we plot contours of f and $\|g\|^2$ for the case $f(x) = (x_1^2 + 5x_2^2)/2$. Note that since $\|g(x)\|^2 = x^T G^2 x$, the contours of $\|g\|^2$ are more elongated than those of f. Let us consider the points $\hat{x} = (0, 2/\sqrt{5})$ and $x = (2, 0)$, which have the same objective function value. It is clear from figure 1 that

$$\|g(\hat{x})\|^2 > \|g(x)\|^2,$$

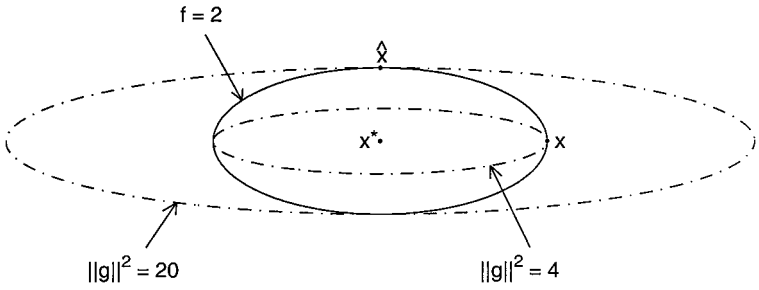

Figure 1. Contours of $f(x) = \frac{1}{2}(x_1^2 + 5x_2^2)$ and $\|g(x)\|^2 = x_1^2 + 25x_2^2$.

so that the gradient norm does not provide useful information about the accuracy in the objective function in this case. Indeed, we see from (2.8) that the relative magnitudes of $f(x) - f^*$ and $\|g(x)\|^2$ can vary as much as the condition number of the Hessian. This observation shall be related to the results of Table 1, since the Hessian matrix of problem PENALTY3 at the solution has a condition number of 1.1×10^{13}.

Figure 1 also suggests that the *path* followed by the iterates of an optimization algorithm may determine whether a small or large final gradient norm is obtained. Let us suppose that the region inside the solid line in figure 1 now denotes the set of points for which the function values cannot be distinguished in machine arithmetic. If an iterate falls inside this region the algorithm will stop as it will not be able to improve the objective function. An algorithm that approaches this region near $\hat{x}$ will give a higher gradient value than one approaching near x, but the quality of the solution, as measured by the objective function, will not be worse at $\hat{x}$.

We will show below that the steepest descent method will normally approach a solution along a point such as x in figure 1. As a result it will produce a final gradient norm that will be small, compared to other gradient norms corresponding to equal function values. Quasi-Newton methods are less predictable. An examination of numerical results reveals that the path generated by their iterates varies from problem to problem, and a description of the behavior of their gradient norms remains an open question.

3. Akaike's results and some extensions

In the rest of the paper we focus on the steepest descent method, with exact line searches, applied to the strongly convex quadratic function

$$f(x) = \frac{1}{2}(x - x^*)^T Q(x - x^*), \tag{3.1}$$

where $Q \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix and $x \in \mathbb{R}^n$. We begin by reviewing results of Akaike [1] that play an important role in our analysis of the asymptotic behavior of the gradient norm in the steepest descent method.

An iteration of the steepest descent method is given by

$$x^{(k+1)} = x^{(k)} - \theta^{(k)} g^{(k)}, \tag{3.2}$$

where

$$g^{(k)} = g\big(x^{(k)}\big) = Q\big(x^{(k)} - x^*\big), \tag{3.3}$$

and

$$\theta^{(k)} = \frac{\big(g^{(k)}\big)^T g^{(k)}}{\big(g^{(k)}\big)^T Q g^{(k)}}. \tag{3.4}$$

Let $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ denote the eigenvalues of Q, and $\xi_1, \xi_2, \ldots, \xi_n$ the corres ponding set of (orthonormal) eigenvectors. Let $x^{(0)}$ be the starting point and, with respect to this point, define

$$\lambda_- = \min\{\lambda_i : \xi_i^T g^{(0)} \neq 0\} \quad \text{and} \quad \lambda^- = \max\{\lambda_i : \xi_i^T g^{(0)} \neq 0\}. \tag{3.5}$$

In order to rule out the trivial case where the steepest descent method (3.2)–(3.4) finds the solution after one single iteration, we make the following assumption.

Assumption 1. The starting point $x^{(0)}$ and the matrix Q are such that $\lambda_- < \lambda^-$.

Indeed when Assumption 1 does not hold, the initial gradient $g^{(0)}$ is an eigenvector of Q. We will also make the following assumption whose significance to the analysis will be discussed later on.

Assumption 2. The matrix Q in (3.1) satisfies

$$0 < \lambda_1 < \cdots < \lambda_n, \tag{3.6}$$

and the starting point is such that

$$\lambda_- = \lambda_1 \quad \text{and} \quad \lambda^- = \lambda_n. \tag{3.7}$$

Under Assumptions 1, Akaike shows in [1, Theorem 4] that the error $\epsilon^{(k)} = x^{(k)} - x^*$ of the k-th approximate solution tends to be approximated by a linear combination of two fixed eigenvectors of Q corresponding to the eigenvalues λ_- and λ^-. In particular, if Assumption 2 holds, the steepest descent method is asymptotically reduced to a search in the 2-dimensional subspace generated by the two eigenvectors corresponding to the largest and the smallest eigenvalues of Q. Akaike also shows in [1, Theorem 4] that $\epsilon^{(k)}$ alternates asymptotically in two fixed directions. In Proposition 3.1, we summarize the main results on which the proof of Theorem 4 in [1] is based.

To state the results we define $\alpha_i^{(k)}$, $i = 1, \ldots, n$, to be the components of $g^{(k)}$ along the eigenvectors ξ_i of Q, that is,

$$g^{(k)} = \sum_{i=1}^{n} \alpha_i^{(k)} \xi_i. \tag{3.8}$$

Proposition 3.1. *Suppose that Assumptions* 1 *and* 2 *hold, and that we apply the steepest descent method* (3.2)–(3.4) *to a strongly convex quadratic function. Then*
(*i*) *the following limits hold,*

$$\lim_{k\to\infty} \frac{\left(\alpha_i^{(2k)}\right)^2}{\sum_{j=1}^{n} \left(\alpha_j^{(2k)}\right)^2} = \begin{cases} \dfrac{1}{1+c^2}, & \text{if } i = 1, \\ 0, & \text{if } i = 2, \ldots, n-1, \\ \dfrac{c^2}{1+c^2}, & \text{if } i = n, \end{cases} \tag{3.9}$$

and

$$\lim_{k\to\infty} \frac{\left(\alpha_i^{(2k+1)}\right)^2}{\sum_{j=1}^n \left(\alpha_j^{(2k+1)}\right)^2} = \begin{cases} \dfrac{c^2}{1+c^2}, & \text{if } i = 1, \\ 0, & \text{if } i = 2, \ldots, n-1, \\ \dfrac{1}{1+c^2}, & \text{if } i = n, \end{cases} \tag{3.10}$$

for some non-zero c, *and*
(*ii*) *the components* $\alpha_1^{(2k)}$, $\alpha_n^{(2k)}$, $\alpha_1^{(2k+1)}$ *and* $\alpha_n^{(2k+1)}$ *have fixed signs for large* k.

Proof: Item (i) is clearly established in the first part of the proof of Theorem 4 in [1]. Item (ii) is a consequence of the second part of the proof of Theorem 4 in [1]. A clearer proof is given by Forsythe in [8, Theorem 4.12] and the comment that follows. Indeed, Forsythe shows that the even and odd normalized gradients (y_{2k} and y_{2k+1}, respectively, in Forsythe's notation), converge to a single point. The sequences of first and last components of these vectors,

$$\left\{ \frac{\alpha_i^{(2k)}}{\sqrt{\sum_{j=1}^n \left(\alpha_j^{(2k)}\right)^2}} \right\} \quad \text{and} \quad \left\{ \frac{\alpha_i^{(2k+1)}}{\sqrt{\sum_{j=1}^n \left(\alpha_j^{(2k+1)}\right)^2}} \right\} \tag{3.11}$$

for $i = 1$ and $i = n$ in our notation, thus converge to non-zero values, proving (ii). □

Proposition 3.2 gives the asymptotic rate of convergence of $f^{(k)}\,(=f(x^{(k)}))$, as derived by Akaike in [1, Page 11].

Proposition 3.2. *Under the assumptions of Proposition* 3.1, *the sequence of function values satisfies*

$$\lim_{k\to\infty} \frac{f^{(k+1)}}{f^{(k)}} = \frac{c^2(\gamma-1)^2}{(c^2+\gamma)(1+c^2\gamma)}, \tag{3.12}$$

where c *is the same constant as in Proposition* 3.1, *and* $\gamma = \lambda_n/\lambda_1$.

Proof: Akaike shows in [1, Page 11] that

$$\lim_{k\to\infty} \frac{f^{(k+1)}}{f^{(k)}} = (\lambda_n - \lambda_1)^2\{(\lambda_n+\lambda_1)^2 + (c-c^{-1})^2\lambda_1\lambda_n\}^{-1}, \tag{3.13}$$

where c is the same constant as in Proposition 3.1. We can rewrite this limit as

$$\lim_{k\to\infty} \frac{f^{(k+1)}}{f^{(k)}} = \frac{c^2(\gamma-1)^2}{c^2(1+\gamma)^2 + (c^2-1)^2\gamma}, \tag{3.14}$$

which is equivalent to (3.12). □

A simple computation shows that the right hand side of (3.12) is maximized when $c^2 = 1$; this gives the worst rate of convergence in the objective function.

Next, we extend Akaike's results to provide an interpretation for the meaning of c, and in particular, show that it is related to the ratio of the components of the gradient $g^{(k)}$ in the coordinate system defined by the eigenvectors ξ_1 and ξ_n. Before establishing this result, we make the following observations: Assumptions 1 and 2 guarantee that

$$\alpha_1^{(k)} \neq 0 \quad \text{and} \quad \alpha_n^{(k)} \neq 0 \quad \text{for all } k \geq 0. \tag{3.15}$$

Indeed, since $\alpha_i^{(k)} = \xi_i^T g^{(k)}$, (3.15) is obviously true for $k = 0$, by definition of λ_- and λ^- and by Assumption 2. For $k > 0$, observe that, by multiplying (3.2) by Q and using (3.3) and (3.8),

$$\alpha_i^{(k)} = \alpha_i^{(k-1)}\big(1 - \theta^{(k-1)}\lambda_i\big), \quad i = 1, \ldots, n, \tag{3.16}$$

and

$$\theta^{(k-1)} = \frac{\sum_{i=1}^n \big(\alpha_i^{(k-1)}\big)^2}{\sum_{i=1}^n \big(\alpha_i^{(k-1)}\big)^2 \lambda_i}, \tag{3.17}$$

by (3.4) and (3.8). It follows from Assumption 1, (3.7) and (3.17) that

$$\lambda_1 < \frac{1}{\theta^{(k-1)}} < \lambda_n$$

for all $k > 0$. Hence (3.15) also holds for $k > 0$, by (3.16).

We next consider the asymptotic behavior of the sequence of steplengths $\{\theta^{(k)}\}$.

Lemma 3.3. *Under the assumptions of Proposition* 3.1, *the following limits hold*,

$$\lim_{k\to\infty} \theta^{(2k)} = \frac{1 + c^2}{\lambda_1(1 + c^2\gamma)} \tag{3.18}$$

and

$$\lim_{k\to\infty} \theta^{(2k+1)} = \frac{1 + c^2}{\lambda_1(c^2 + \lambda)}, \tag{3.19}$$

where c is the same constant as in Proposition 3.1.

Proof: From (3.17), (3.9) and (3.10) we have

$$\lim_{k\to\infty} \big(\theta^{(2k)}\big)^{-1} = \frac{\lambda_1(1 + c^2\gamma)}{1 + c^2} \tag{3.20}$$

and

$$\lim_{k\to\infty} \left(\theta^{(2k+1)}\right)^{-1} = \frac{\lambda_1(c^2+\gamma)}{1+c^2}. \tag{3.21}$$

□

We can now provide an interpretation for the constant c.

Lemma 3.4. *Under the assumptions of Proposition* 3.1, *the constant c satisfies*

$$c = \lim_{k\to\infty} \frac{\alpha_n^{(2k)}}{\alpha_1^{(2k)}}, \tag{3.22}$$

and

$$c = -\lim_{k\to\infty} \frac{\alpha_1^{(2k+1)}}{\alpha_n^{(2k+1)}}. \tag{3.23}$$

Moreover c is uniquely determined by the starting point $x^{(0)}$ and by the eigenvalues and the eigenvectors of Q.

Proof: From (3.9) and (3.10) we have that

$$\lim_{k\to\infty} \frac{\left(\alpha_n^{(2k)}\right)^2}{\left(\alpha_1^{(2k)}\right)^2} = \lim_{k\to\infty} \frac{\left(\alpha_1^{(2k+1)}\right)^2}{\left(\alpha_n^{(2k+1)}\right)^2} = c^2. \tag{3.24}$$

These limits together with item (ii) of Proposition 3.1 are sufficient to ensure the convergence of the sequences $\{\alpha_n^{(2k)}/\alpha_1^{(2k)}\}$ and $\{\alpha_1^{(2k+1)}/\alpha_n^{(2k+1)}\}$. Hence we can deduce (3.22) from (3.24), without loss of generality. Now (3.16), (3.18) and (3.22) imply that

$$\lim_{k\to\infty} \frac{\alpha_1^{(2k+1)}}{\alpha_n^{(2k+1)}} = \lim_{k\to\infty} \frac{\alpha_1^{(2k)}\left(1-\theta^{(2k)}\lambda_1\right)}{\alpha_n^{(2k)}\left(1-\theta^{(2k)}\lambda_n\right)} = -c, \tag{3.25}$$

which proves (3.23).

Finally note that equalities (3.16) and (3.17) together with (3.22) or (3.23) show that c is uniquely determined by the values of $\alpha_i^{(0)}$, $i = 1, \ldots, n$ (and hence by the starting point $x^{(0)}$), and by the eigenvalues and the eigenvectors of Q. □

We now determine the range of values that c can attain, for a given starting point $x^{(0)}$. An important quantity in this analysis is the *minimum deviation* of the eigenvalues of Q from the midrange, as measured by

$$\delta = \min_{i\in\mathcal{I}} \left| \frac{\lambda_i - \frac{\lambda_n+\lambda_1}{2}}{\frac{\lambda_n-\lambda_1}{2}} \right|, \tag{3.26}$$

where

$$\mathcal{I} = \left\{ i = 2, \ldots, n-1 : \lambda_1 < \lambda_i < \lambda_n,\ \xi_i^T g^{(0)} \neq 0 \text{ and } \lambda_i \neq \left(\theta^{(k)}\right)^{-1} \forall k \geq 0 \right\}. \tag{3.27}$$

Note that $\delta \in [0, 1)$, and its value depends on $x^{(0)}$ through the definition of the set $\mathcal{I}$. Moreover, δ can only be near one if *all* the eigenvalues whose index is in $\mathcal{I}$ cluster around λ_1 and λ_n. It is also important to observe that, by the identity $\alpha_i^{(0)} = \xi_i^T g^{(0)}$ and (3.16),

$$i \in \mathcal{I} \Rightarrow \lambda_1 < \lambda_i < \lambda_n \quad \text{and} \quad \alpha_i^{(k)} \neq 0 \quad \text{for all } k \geq 0. \tag{3.28}$$

In other words, for $i \in \mathcal{I}$, the gradient component along the eigenvector ξ_i whose corresponding eigenvalue is strictly between λ_1 and λ_n is not discarded in the course of the algorithm.

The restriction on the possible values for c given by the following lemma is an obvious consequence of a result of Akaike (see [1, Page 12]) from which the author deduces that "the rate of convergence of the steepest descent method for ill-conditioned problems tends near to its worst possible value (reached for $c^2 = 1$), especially when there is some λ_i close to the midpoint $(\lambda_n + \lambda_1)/2$".

Lemma 3.5. *Under the assumptions of Proposition* 3.1, *and assuming that the set $\mathcal{I}$ is nonempty*, *c is restricted to the interval*

$$\phi_\delta^{-1} \leq c^2 \leq \phi_\delta, \tag{3.29}$$

where

$$\phi_\delta = \frac{2 + \eta_\delta + \sqrt{\eta_\delta^2 + 4\eta_\delta}}{2}, \tag{3.30}$$

and

$$\eta_\delta = 4\left(\frac{1+\delta^2}{1-\delta^2}\right). \tag{3.31}$$

Proof: Using the following inequality that holds for all $i \in \mathcal{I}$ (see [1, Page 12]),

$$\left(\frac{\lambda_n - \lambda_1}{2}\right)^2 + \left(\lambda_i - \frac{\lambda_n + \lambda_1}{2}\right)^2 \geq \frac{(1-c^2)^2}{2(1+c^2)^2}(\lambda_n - \lambda_1)^2, \tag{3.32}$$

Akaike shows that

$$\frac{(c^2-1)^2}{c^2} \leq \eta_i \tag{3.33}$$

for all $i \in \mathcal{I}$, where

$$\eta_i = 4\left(\frac{1+\delta_i^2}{1-\delta_i^2}\right), \tag{3.34}$$

and

$$\delta_i = \frac{\lambda_i - \frac{\lambda_n+\lambda_1}{2}}{\frac{\lambda_n-\lambda_1}{2}}. \tag{3.35}$$

Since $|\delta_i| < 1$ for all $i \in \mathcal{I}$, using the definition (3.26) of the minimum deviation δ, we obtain

$$\frac{(c^2-1)^2}{c^2} \leq \eta_\delta, \tag{3.36}$$

where η_δ is defined in (3.31). This last inequality is equivalent to (3.29). □

Note that, by (3.28), the requirement that the set $\mathcal{I}$ be nonempty in the assumptions of Lemma 3.5 guarantees that at least one gradient component along an eigenvector ξ_i whose corresponding eigenvalue is strictly between λ_1 and λ_n is not discarded in the course of the algorithm. If $\mathcal{I}$ is empty, the steepest descent method will be reduced to a search in the 2-dimensional subspace generated by ξ_1 and ξ_n *after a finite number of iterations* rather than asymptotically. In that case, the behavior of the method is not typical: it coincides with that for the 2-dimensional case, which as we will see in Section 6, has some special properties.

Figure 2 illustrates the possible values of c^2 as a function of δ. It is clear that ϕ_δ increases very slowly with δ—except when δ approaches 1, when it diverges to ∞. Note also that the value $c^2 = 1$ giving the worst rate of convergence in f is always contained in the range of possible values of c. The definitions (3.30) and (3.31) imply that ϕ_δ (and hence the set of possible values of c^2) is exclusively determined by δ (for a fixed starting point), and thus by the distribution of the inner eigenvalues of Q—and is in general not directly dependent on the condition number γ, since we can vary γ while leaving δ unchanged.

Lemma 3.5 specifies the interval (3.29) of possible values of c^2, but it does not state that all values of c^2 in this interval can be attained for some starting point. We performed, however, some numerical experiments that suggest that this is indeed the case. For the two quadratic functions $f(x) = (x_1^2 + 4x_2^2 + 16x_3^2)/2$ and $f(x) = (x_1^2 + 75x_2^2 + 80x_3^2)/2$, we generated 1000 random starting points on the unit sphere, applied the steepest descent method (3.2)–(3.4), and recorded the values of c^2, as given by (3.22) and (3.23). Figure 3 shows the plots of the recorded values for the two examples (one "x" per recorded value of c^2). We note that for the first example, $\delta = 0.6$ and $[\phi_\delta^{-1},\ \phi_\delta] = [0.0961,\ 10.4039]$, while for the second example, $\delta = 0.8734$ and $[\phi_\delta^{-1},\ \phi_\delta] = [0.0315,\ 31.7036]$.

Assumption 2 has been made throughout this section to simplify the exposition. We note, however, that (3.6) can be relaxed without altering the results stated here, as discussed by Forsythe [8, Section 5]. On the other hand, (3.7) is assumed for convenience and without loss of generality.

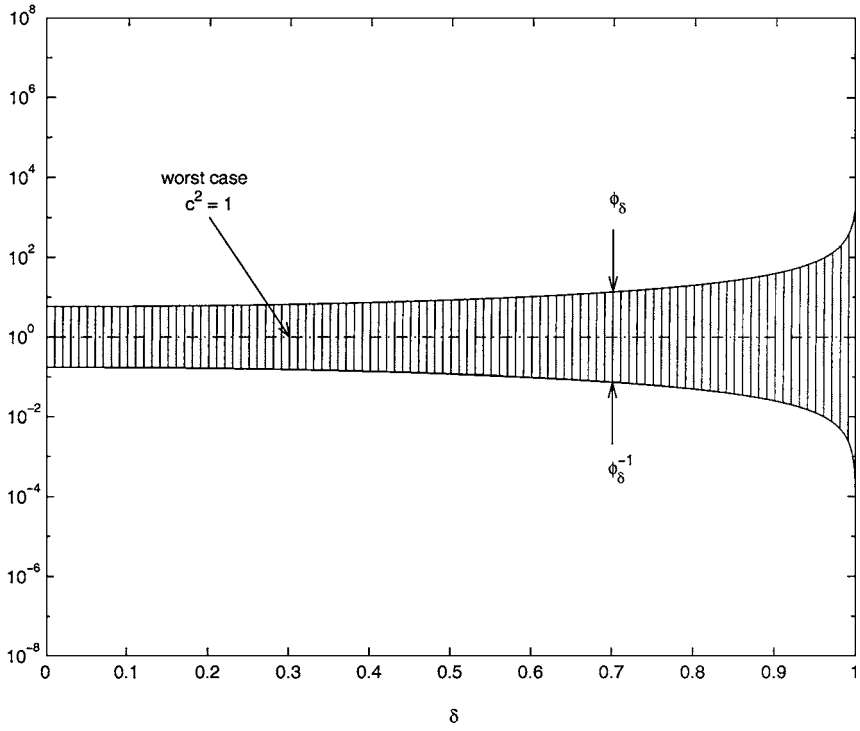

Figure 2. Intervals $[\phi_\delta^{-1},\ \phi_\delta]$ of possible values of c^2, as a function of $\delta \in [0, 1)$.

4. Maximum oscillation in the gradient norm

The following result provides an upper bound on the growth of the gradient norm as a function of the condition number γ. This bound holds, not only asymptotically, but at each iteration, and its derivation is independent from the results of Section 3.

Theorem 4.1. *At each iteration of the steepest descent method* (3.2)–(3.4) *applied to a strongly convex quadratic function*,

$$\frac{\left\|g^{(k+1)}\right\|^2}{\left\|g^{(k)}\right\|^2} \leq \frac{(\gamma-1)^2}{4\gamma}, \tag{4.1}$$

where $\gamma = \lambda_n/\lambda_1$.

Proof: The proof is similar to that used in [11] to establish the rate of convergence of the objective function for the steepest descent method. By (3.2) and (3.3), we have

$$g^{(k+1)} = g^{(k)} - \theta^{(k)} Q g^{(k)}.$$

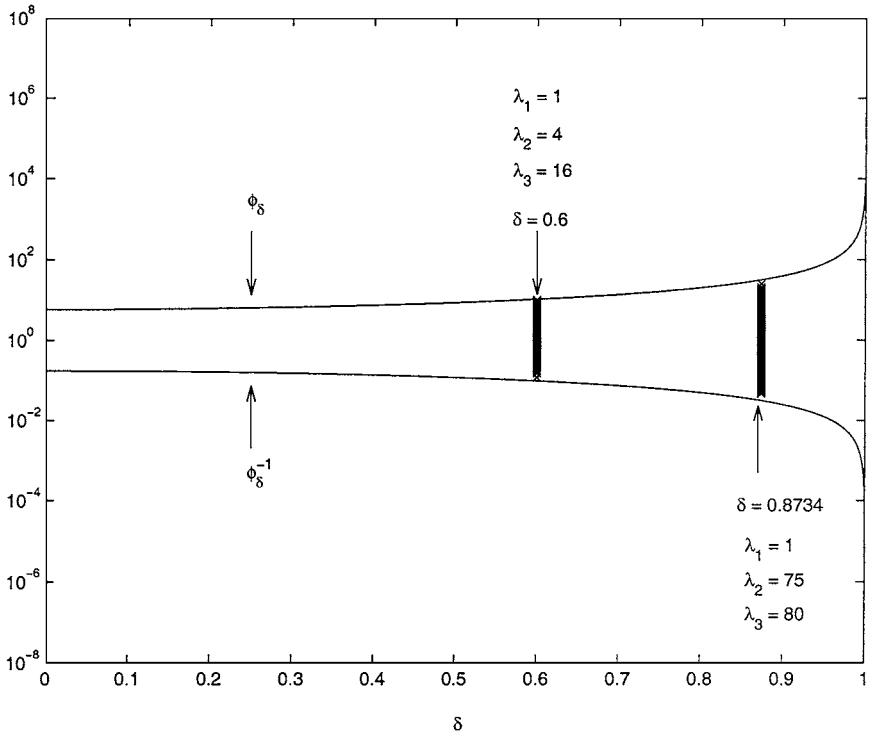

Figure 3. Values of c^2 (dark bars) attained from a sample of 1000 random starting points, for two quadratic functions in three variables.

Therefore

$$\left\|g^{(k+1)}\right\|^2 = \left\|g^{(k)}\right\|^2 - 2\theta^{(k)}\left(g^{(k)}\right)^T Qg^{(k)} + \left(\theta^{(k)}\right)^2\left(g^{(k)}\right)^T Q^2 g^{(k)}.$$

Substituting (3.4) in the above expression yields

$$\left\|g^{(k+1)}\right\|^2 = \left\{\frac{\left\|g^{(k)}\right\|^2\left\|Qg^{(k)}\right\|^2}{\left(\left(g^{(k)}\right)^T Qg^{(k)}\right)^2} - 1\right\}\left\|g^{(k)}\right\|^2. \tag{4.2}$$

By introducing $z^{(k)} = Q^{1/2}g^{(k)}$, we may rewrite this equation as

$$\left\|g^{(k+1)}\right\|^2 = \left\{\frac{\left(\left(z^{(k)}\right)^T Q^{-1}z^{(k)}\right)\left(\left(z^{(k)}\right)^T Qz^{(k)}\right)}{\left(\left(z^{(k)}\right)^T z^{(k)}\right)^2} - 1\right\}\left\|g^{(k)}\right\|^2. \tag{4.3}$$

Using the Kantorovich inequality (see [11]), we have

$$\frac{\left(\left(z^{(k)}\right)^T Q^{-1} z^{(k)}\right)\left(\left(z^{(k)}\right)^T Q z^{(k)}\right)}{\left(\left(z^{(k)}\right)^T z^{(k)}\right)^2} \leq \frac{(1+\gamma)^2}{4\gamma}.$$

Substituting this inequality in (4.3) yields the desired bound (4.1). □

This result implies that, for the gradient norm to increase, it is necessary that $(\gamma - 1)^2 > 4\gamma$, that is,

$$\gamma > 3 + 2\sqrt{2}. \tag{4.4}$$

Conversely, if the condition number of Q satisfies $\gamma \leq 3 + 2\sqrt{2}$, then the sequence of gradient norms $\{\|g^{(k)}\|\}$ generated by the steepest descent method (3.2)–(3.4) is monotonically decreasing. We can also deduce from this theorem that, if large oscillations in the gradient are observed, the problem must be ill-conditioned.

5. Asymptotic behavior of the gradient norm

Theorem 4.1 might suggest that, for ill-conditioned problems, the norm of the gradient can exhibit extreme growth at some iterations. Of course, since the gradient converges to zero (in exact arithmetic), there must exist iterations at which it decreases, and in general we can expect oscillatory behavior.

In the next theorem, we study the one-step and two-step ratios of gradient norms and establish their limiting values in terms of γ and the constant c from Section 3.

Theorem 5.1. *Suppose that Assumptions* 1 *and* 2 *hold. When applying the steepest descent method* (3.2)–(3.4) *to a strongly convex quadratic function, we have both*

$$\lim_{k\to\infty} \frac{\left\|g^{(2k+1)}\right\|^2}{\left\|g^{(2k)}\right\|^2} = \frac{c^2(\gamma-1)^2}{(1+c^2\gamma)^2}, \tag{5.1}$$

and

$$\lim_{k\to\infty} \frac{\left\|g^{(2k+2)}\right\|^2}{\left\|g^{(2k+1)}\right\|^2} = \frac{c^2(\gamma-1)^2}{(c^2+\gamma)^2}, \tag{5.2}$$

where c is the same constant as in Proposition 3.1. *Moreover, the two-step asymptotic rate of convergence of the gradient norm is equal to the one-step asymptotic rate in the function value, i.e.*

$$\lim_{k\to\infty} \frac{\left\|g^{(k+2)}\right\|}{\left\|g^{(k)}\right\|} = \lim_{k\to\infty} \frac{f^{(k+1)}}{f^{(k)}}. \tag{5.3}$$

Proof: Using (3.8), (3.15) and (3.16), we have that

$$\begin{aligned}\frac{\left\|g^{(2k+1)}\right\|^2}{\left\|g^{(2k)}\right\|^2} &= \frac{\sum_{i=1}^n \left(\alpha_i^{(2k+1)}\right)^2}{\sum_{i=1}^n \left(\alpha_i^{(2k)}\right)^2} \\ &= \frac{\left(\alpha_1^{(2k+1)}\right)^2 \sum_{i=1}^n \left(\left(\alpha_i^{(2k+1)}\right)^2 / \left(\alpha_1^{(2k+1)}\right)^2\right)}{\left(\alpha_1^{(2k)}\right)^2 \sum_{i=1}^n \left(\left(\alpha_i^{(2k)}\right)^2 / \left(\alpha_1^{(2k)}\right)^2\right)} \\ &= \frac{\left(1-\theta^{(2k)}\lambda_1\right)^2 \sum_{i=1}^n \left(\left(\alpha_i^{(2k+1)}\right)^2 / \left(\alpha_1^{(2k+1)}\right)^2\right)}{\sum_{i=1}^n \left(\left(\alpha_i^{(2k)}\right)^2 / \left(\alpha_1^{(2k)}\right)^2\right)}. \end{aligned} \tag{5.4}$$

As in the proof of Lemma 3.4, we observe that (3.9) and (3.10) yield

$$\lim_{k\to\infty} \frac{\left(\alpha_n^{(2k)}\right)^2}{\left(\alpha_1^{(2k)}\right)^2} = \lim_{k\to\infty} \frac{\left(\alpha_1^{(2k+1)}\right)^2}{\left(\alpha_n^{(2k+1)}\right)^2} = c^2 \tag{5.5}$$

and, for $i = 2, \ldots, n-1$,

$$\lim_{k\to\infty} \frac{\left(\alpha_i^{(k)}\right)^2}{\left(\alpha_1^{(k)}\right)^2} = \lim_{k\to\infty} \frac{\left(\alpha_i^{(k)}\right)^2}{\left(\alpha_n^{(k)}\right)^2} = 0. \tag{5.6}$$

We thus deduce (5.1) from (5.4) using these limits and (3.18) in Lemma 3.3. The proof of (5.2) is similar, but uses (3.19) rather than (3.18), and (5.3) is an obvious consequence of Proposition 3.2, (5.1) and (5.2). □

It is interesting to note that the two limits (5.1) and (5.2) coincide if and only if $c^2 = 1$, which as we recall gives the worst rate of convergence in the objective function. Indeed, for this value of c^2 the three limits (5.1), (5.2) and (3.12) are the same. Thus, if $c^2 = 1$, the *one-step* rates of convergence of $\|g^{(k)}\|^2$ and $f^{(k)}$ are the same, and the sequence of gradient norms will be monotonically decreasing for all sufficiently large k. These observations indicate that we cannot use the amplitude of the oscillations in the gradient norm as a sign that the starting point has caused the worst rate of convergence in f to take place; nor does the lack of oscillations in the gradient norm imply that the condition number of the Hessian Q is moderate. But, as noted earlier, since (4.1) is of order $O(\gamma)$, it is correct to state that if the oscillations in the gradient norm are large, then the condition number of Q must be large.

In the next section, we will make use of the results of Theorems 4.1 and 5.1 to make further observations about the asymptotic oscillatory behavior of the gradient norm.

5.1. Oscillations in the gradient norms

For a given problem, the choice of initial point determines both whether oscillations in the gradient norm will take place and the magnitude of the oscillations. Unlike the 2-dimensional

case (see Section 6) we will not be able to directly characterize the regions of initial points in $\mathbb{R}^n$ for which oscillations in the gradient norm take place. Instead we follow an indirect approach, using the results established so far, to make some observations about the largest possible oscillation and about the relationship between the rate of convergence in f and the oscillatory behavior of the gradient norm. These observations apply to most, but not all, problems.

We assume throughout this section that $x^{(0)}$ is fixed and γ is large enough that (4.4) holds. We first ask whether the upper bound given in (4.1)—which gives the maximum increase in the gradient norm, at one iteration—can be attained, asymptotically. Using (4.1), (5.1) and (5.2), we set up the equations

$$\frac{c^2(\gamma-1)^2}{(1+c^2\gamma)^2} = \frac{(\gamma-1)^2}{4\gamma} \quad \text{and} \quad \frac{c^2(\gamma-1)^2}{(c^2+\gamma)^2} = \frac{(\gamma-1)^2}{4\gamma},$$

whose solutions are

$$c^2 = 1/\gamma \quad \text{and} \quad c^2 = \gamma, \tag{5.7}$$

respectively. If c takes one of these values, then the maximum possible oscillation in $\|g\|$ (for that γ) will occur asymptotically.

From the one-step asymptotic behavior (5.1) and (5.2), we can also deduce that the gradient norm will grow (and thus oscillate) for sufficiently large k if one of the following conditions is satisfied:

$$\frac{c^2(\gamma-1)^2}{(1+c^2\gamma)^2} > 1 \quad \text{or} \quad \frac{c^2(\gamma-1)^2}{(c^2+\gamma)^2} > 1.$$

These two inequalities yield

$$\frac{l_\gamma}{\gamma^2} < c^2 < \frac{u_\gamma}{\gamma^2} \quad \text{and} \quad l_\gamma < c^2 < u_\gamma, \tag{5.8}$$

where

$$l_\gamma = \frac{(\gamma-1)^2 - 2\gamma - (\gamma-1)\sqrt{(\gamma-1)^2-4\gamma}}{2}, \tag{5.9}$$

and

$$u_\gamma = \frac{(\gamma-1)^2 - 2\gamma + (\gamma-1)\sqrt{(\gamma-1)^2-4\gamma}}{2}. \tag{5.10}$$

Since the bounds in (5.8) depend only on γ, we have found a simple relationship between c and γ that ensures oscillations in the gradient.

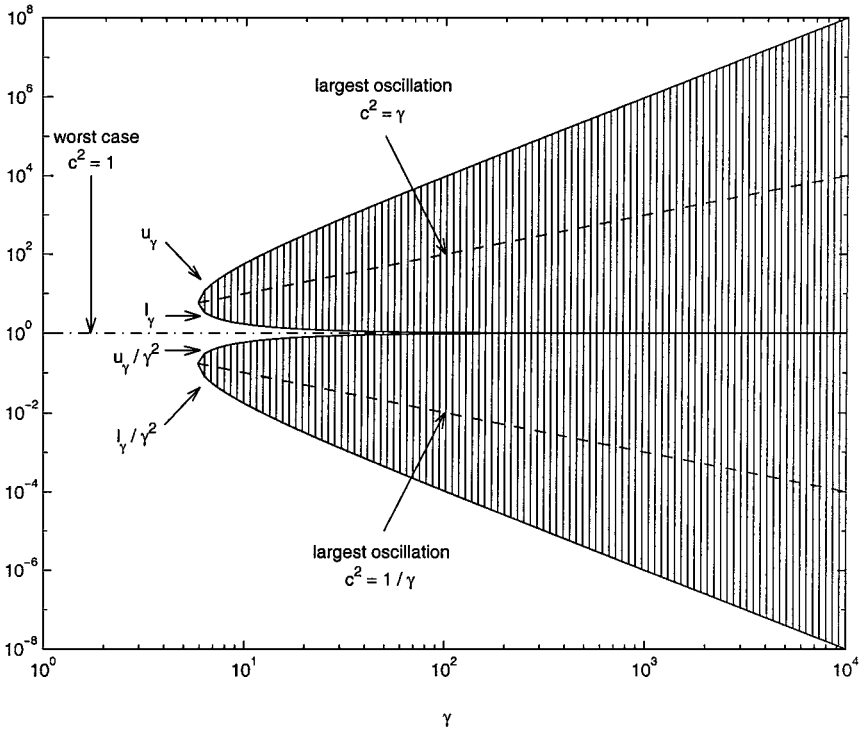

Figure 4. Intervals $(l_\gamma/\gamma^2,\ u_\gamma/\gamma^2)$ and $(l_\gamma,\ u_\gamma)$ of c^2 values for which the gradient norm experiences oscillation asymptotically, for $\gamma \in (3 + 2\sqrt{2},\ 10^4]$.

Figure 4 illustrates the values of c^2 and γ satisfying (5.8). The two dashed lines represent the values $c^2 = 1/\gamma$ and $c^2 = \gamma$ corresponding to the largest possible growth in $\|g\|$ (see (5.7)). Since l_γ and u_γ satisfy

$$1 < l_\gamma < \gamma < u_\gamma < \gamma^2 \quad \text{and} \quad l_\gamma u_\gamma = \gamma^2 \tag{5.11}$$

for all γ satisfying (4.4), and since

$$\lim_{\gamma\to\infty} l_\gamma = 1 \quad \text{and} \quad \lim_{\gamma\to\infty} u_\gamma = \infty, \tag{5.12}$$

whe see that as γ tends to infinity, the intervals (5.8) expand to cover (0, 1) and (1, ∞), respectively, but never overlap. Thus the value $c^2 = 1$, which gives rise to the worst rate of convergence in f, is not contained in the shaded area of figure 4. This is consistent with our previous observation that oscillations in the gradient norm do not occur in this case.

We have seen in Section 3, however, that the values of c^2 must be restricted to the interval (3.29). In figure 5 we superimpose over figure 4 the set of possible values of c^2 (shaded

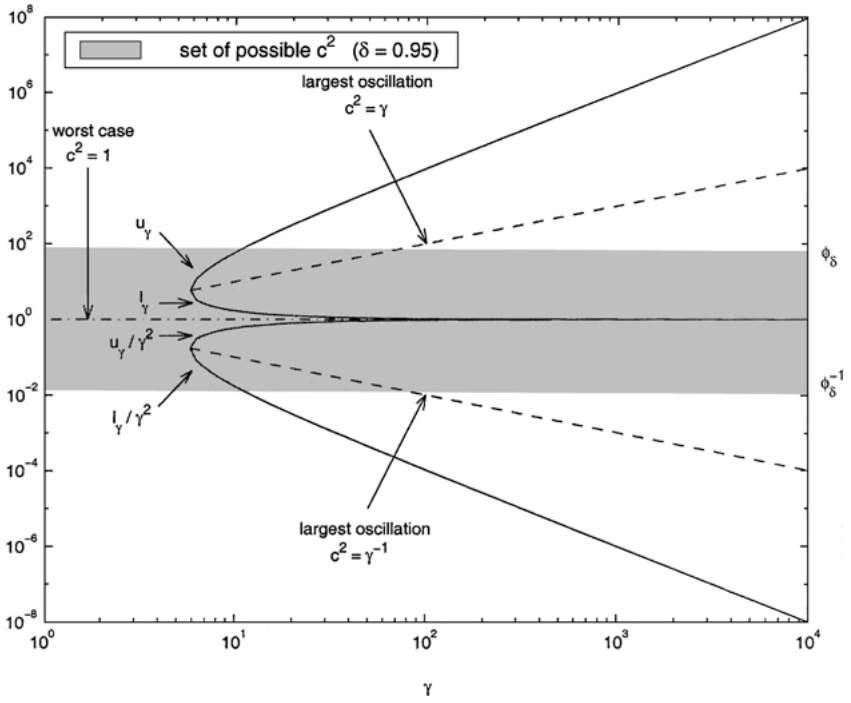

Figure 5. Possible c^2 values as a function of γ, for $\delta = 0.95$, superimposed on the set of values of c^2 and γ for which oscillation in the gradient norm takes place.

region) for $\delta = 0.95$. (Note that $\delta = 0.95$ yields a rather large set of possible values of c^2 and corresponds to a spectrum of Q whose eigenvalues are relatively far from $(\lambda_1 + \lambda_n)/2$.) Let us now consider how large can we expect the oscillations in $\|g\|$ to be. It is immediately apparent from figure 5 that the shaded region of possible values of c^2 considerably limits the size of the oscillations in $\|g\|$—compared to the maximum value which occurs when the values of c^2 and γ lie on the dashed lines. More specifically, if

$$\phi_\delta < \gamma, \tag{5.13}$$

the one-step growth in the gradient norm will not approach the upper bound given by (4.1), regardless of the starting point. *Moreover*, *as* γ *increases*, *the gap between the maximum actual oscillation in* $\|g\|$ *and the upper bound* (4.1) *will widen*. Condition (5.13) will be satisfied for most ill-conditioned problems and for most starting points since we have observed in Section 3 that ϕ_δ is small except when δ is close to one. For example, even in the mildly ill-conditioned case when $\gamma = 200$ we find that δ has to be greater than 0.98 for (5.13) to be violated.

We conclude this section by making an interesting remark relating the rate of convergence in f and the behavior of the gradient norm. Consider the right hand side of (3.12) as a function

of c^2, when $\gamma > 1$ is held fixed. This function is monotonically increasing for $c^2 \in (0, 1)$ and monotonically decreasing in $(1, \infty)$. Therefore:

(i) the rate of convergence in f decreases for $c^2 \in (0, 1)$;
(ii) the rate of convergence in f increases for $c^2 \in (1, \infty)$.

In terms of figure 5, as we move away vertically from both sides of the dash-dot line corresponding to $c^2 = 1$, the rate of convergence in f improves monotonically.

Let us now consider the oscillations in the gradient norm. If we vary c^2 for a fixed value of γ, it is easy to see that:

(iii) the right hand side in (5.1) is monotonically increasing for $c^2 \leq 1/\gamma$ and monotonically decreasing otherwise;
(iv) the right hand side in (5.2) is monotonically increasing for $c^2 \leq \gamma$ and monotonically decreasing otherwise.

We must, however, focus only on the possible values of c^2. For the current case where condition (5.13) holds, c^2 must satisfy

$$c^2 > 1/\gamma \quad \text{or} \quad c^2 < \gamma,$$

by (3.29) in Lemma 3.5. From this and (iii) and (iv), we deduce (see figure 5) that when increasing or decreasing c^2 vertically (i.e. for fixed γ) away from the value 1 until it reaches the border of the shaded area of possible values of c^2, the oscillations in the gradient increase (for either the odd or even iterates). More precisely by moving c^2 away from the value 1, we first obtain values of c^2 for which oscillations in the gradient will not occur (since the curves in figure 5 do not touch along the line $c^2 = 1$), while varying c^2 further generates values for which oscillations of increasing magnitude take place. Combining these observations with (i) and (ii) we deduce that *if* (5.13) *holds* (*which should be often the case*) *the asymptotic behavior of the steepest descent method is such that the larger the oscillation in* $\|g\|$, *the faster the convergence rate in* f. This observation was contrary to our initial expectations, as we had speculated that the largest oscillations in the gradient would characterize the most unfavorable starting points.

5.2. *Path followed by the iterates*

As we mentioned in Section 3, Akaike has shown (see [1, Theorem 4]) that if Assumptions 1 and 2 hold, the steepest descent method is asymptotically reduced to a search in the 2-dimensional subspace generated by the eigenvectors ξ_1 and ξ_n. Let us therefore consider the restriction of f to this subspace, and observe the values of the ratio

$$\frac{\alpha_n^{(k)}}{\alpha_1^{(k)}}. \tag{5.14}$$

Due to the definition of $\alpha_i^{(k)}$, this ratio is the slope of the gradient $g^{(k)}$ restricted to the space spanned by ξ_1 and ξ_n. We deduce from (3.22), (3.23) and (3.29) that, for a given value of δ,

$$\left[\frac{\alpha_n^{(k)}}{\alpha_1^{(k)}}\right]^2 \in \left[\phi_\delta^{-1}, \phi_\delta\right], \tag{5.15}$$

asymptotically. Since these intervals are generally narrow, the possible values for the slope of the gradient are greatly restricted, and imply that the iterates approach the solution along a path that is close to the eigenvector corresponding to the smallest eigenvalue of Q. This is associated with relatively small gradient norms, as we discussed in Section 2.

To illustrate this, we plot in figure 6 the contours of $f = (x_1^2 + 49x_n^2)/2$, which can be considered as the restriction of some quadratic function to $\mathbb{R}^2$. Let us assume that $[\phi_\delta^{-1}, \phi_\delta] = [0.1, 10]$, which corresponds to $\delta \simeq 0.58$. The sets of points for which the slope of the gradient satisfies the restriction (5.15) has been highlighted in figure 6. (Note that the highlighted areas in figure 6 do not overlap at the left and right extreme points of the contours, because $\phi_\delta^{-1} > 0$.) As γ grows and the contours become more elongated, the highlighted areas shrink and move closer and closer to the horizontal axis.

Let us now consider an example in three dimensions and observe the path in $\mathbb{R}^3$ followed in by the iterates, for a given choice of the starting point. Figures 7 to 10 illustrate this path in the case when $f(x) = (x_1^2 + 4x_2^2 + 16x_3^2)/2$ and $x^{(0)} = (3.1, 1, 0.39)$. For this example, $\gamma = 16$, $\delta = 0.6$ and $c = 1.2$. Figures 7, 9 and 10 show the rather fast speed at which the

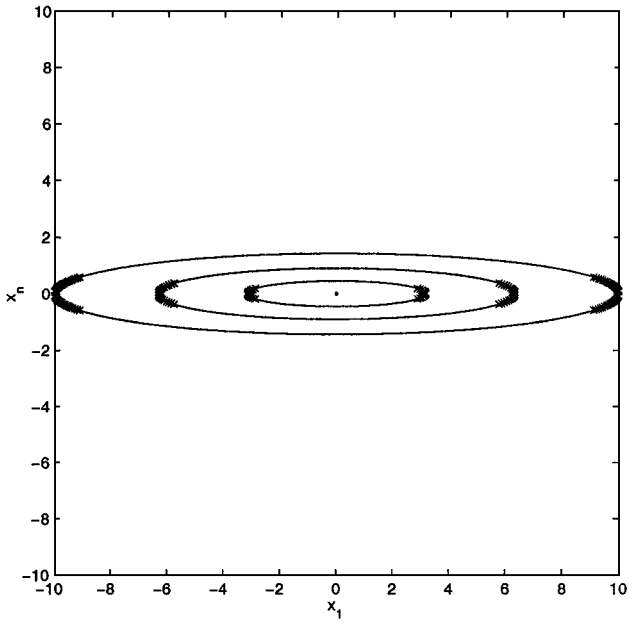

Figure 6. Sets of possible iterates, restricted to the 2-dimensional subspace spanned by ξ_1 and ξ_n, in the case when $[\phi_\delta^{-1}, \phi_\delta] = [0.1, 10]$.

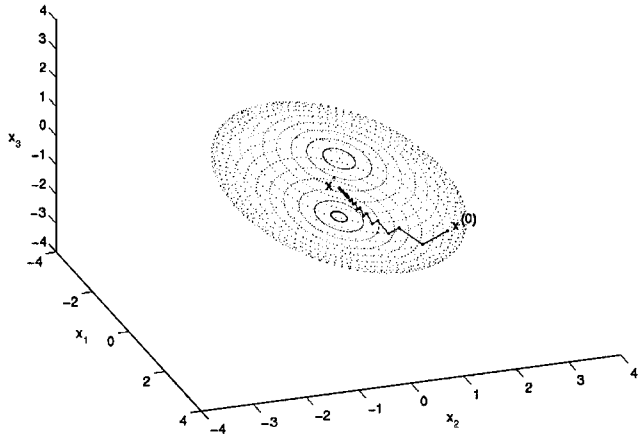

Figure 7. Example of path generated by the steepest descent method.

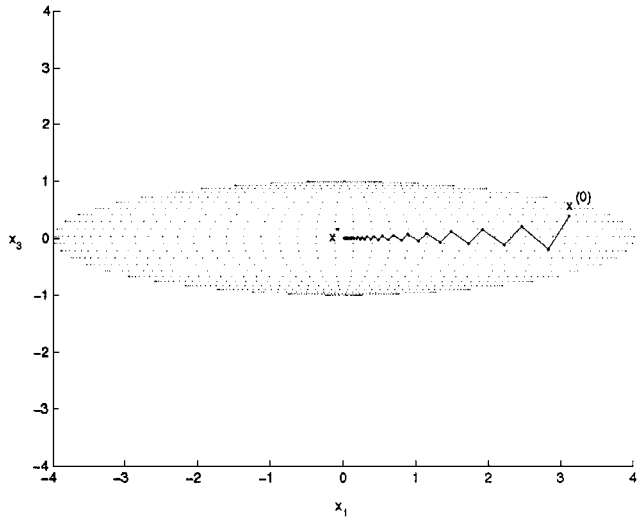

Figure 8. Viewpoint perpendicular to the x_1x_3-plane.

method is reduced asymptotically to a search in the x_1x_3-plane (that is, at which the second component becomes very small and converges to zero). Figure 8 shows that the iterates alternate asymptotically in two fixed directions. Figures 8 and 9 illustrate the fact that the path followed by the iterates is closely aligned with the eigenvector corresponding to the smallest eigenvalue.

In summary, by combining the results of Sections 2 and 3, we conclude that the steepest descent iterates will normally approach the solution along a path that will give a small final gradient norm, compared to the set of all gradient norm values corresponding to the same final function value.

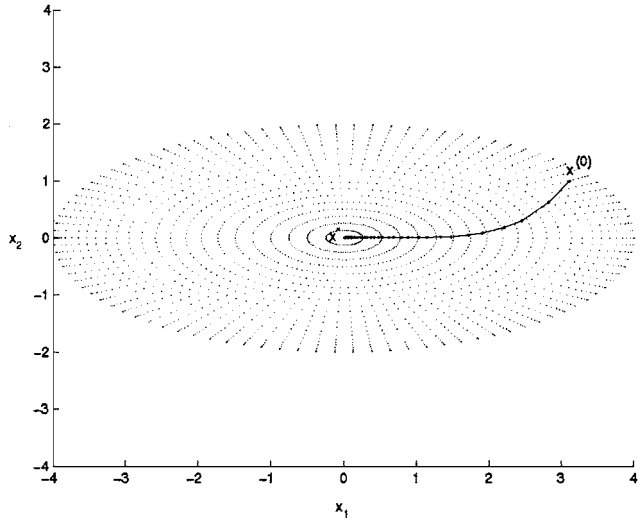

Figure 9. Viewpoint perpendicular to the x_1x_2-plane.

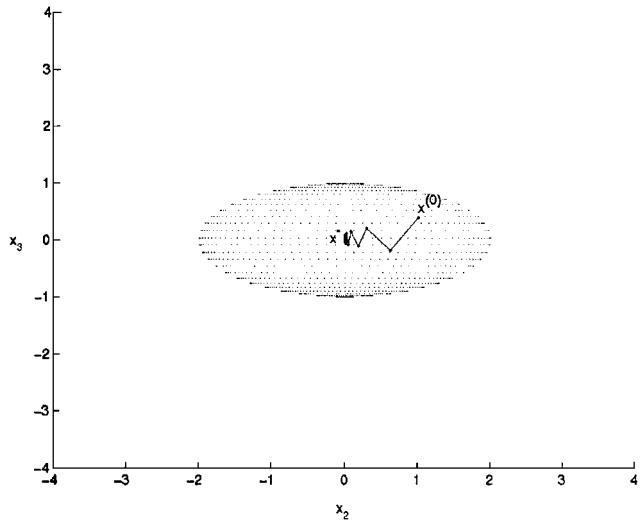

Figure 10. Viewpoint perpendicular to the x_2x_3-plane.

5.3. Summary of the results

For clarity, we now summarize the main results that have been presented in Sections 4 and 5. Since several of these results refer to the constant c, we also review its main properties presented in Section 3.

Theorem 5.2. *Suppose that we apply the steepest descent method* (3.2)–(3.4) *with exact line searches, starting from* $x^{(0)}$, *to the strongly convex quadratic function*

$$f(x) = \frac{1}{2}(x - x^*)^T Q(x - x^*), \tag{5.16}$$

where $Q \in \mathbb{R}^{n\times n}$ *is symmetric positive definite with eigenvalues* $0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ *and corresponding* (*orthonormal*) *eigenvectors* $\xi_1, \xi_2, \ldots, \xi_n$. *Then*

(i) *for all* $k \ge 0$,

$$\frac{\left\|g^{(k+1)}\right\|^2}{\left\|g^{(k)}\right\|^2} \le \frac{(\gamma-1)^2}{4\gamma}, \tag{5.17}$$

where $\gamma = \lambda_n/\lambda_1$;

(ii) *if Assumptions* 1 *and* 2 *in Section* 3 *hold, then*

$$\lim_{k\to\infty} \frac{\left\|g^{(2k+1)}\right\|^2}{\left\|g^{(2k)}\right\|^2} = \frac{c^2(\gamma-1)^2}{(1+c^2\gamma)^2}, \tag{5.18}$$

$$\lim_{k\to\infty} \frac{\left\|g^{(2k+2)}\right\|^2}{\left\|g^{(2k+1)}\right\|^2} = \frac{c^2(\gamma-1)^2}{(c^2+\gamma)^2}, \tag{5.19}$$

and

$$\lim_{k\to\infty} \frac{\left\|g^{(k+2)}\right\|}{\left\|g^{(k)}\right\|} = \lim_{k\to\infty} \frac{f^{(k+1)}}{f^{(k)}} = \frac{c^2(\gamma-1)^2}{(c^2+\gamma)(1+c^2\gamma)}, \tag{5.20}$$

for some constant c, *which satisfies the following three properties*:

(a) c *is given by the limits*

$$c = \lim_{k\to\infty} \frac{\alpha_n^{(2k)}}{\alpha_1^{(2k)}} = -\lim_{k\to\infty} \frac{\alpha_1^{(2k+1)}}{\alpha_n^{(2k+1)}}, \tag{5.21}$$

where $\alpha_i^{(k)}$, $i = 1, \ldots, n$, *are the components of* $g^{(k)}$ *along the eigenvectors* ξ_i *of* Q, *that is*,

$$g^{(k)} = \sum_{i=1}^{n} \alpha_i^{(k)} \xi_i; \tag{5.22}$$

(b) c *is uniquely determined by the starting point* $x^{(0)}$ *and by the eigenvalues and the eigenvectors of* Q;

(c) *if the set*

$$\mathcal{I} = \Big\{ i = 2, \ldots, n-1 : \lambda_1 < \lambda_i < \lambda_n,\ \xi_i^T g^{(0)} \neq 0 \\ \quad \textit{and} \quad \lambda_i \neq \left(\theta^{(k)}\right)^{-1}\ \forall k\ \geq 0 \Big\} \tag{5.23}$$

is nonempty, c is restricted to the interval

$$\phi_\delta^{-1} \leq c^2 \leq \phi_\delta, \tag{5.24}$$

where

$$\phi_\delta = \frac{2 + \eta_\delta + \sqrt{\eta_\delta^2 + 4\eta_\delta}}{2}, \tag{5.25}$$

with

$$\eta_\delta = 4\left(\frac{1+\delta^2}{1-\delta^2}\right) \quad \textit{and} \quad \delta = \min_{i \in \mathcal{I}} \left| \frac{\lambda_i - \frac{\lambda_n + \lambda_1}{2}}{\frac{\lambda_n - \lambda_1}{2}} \right|. \tag{5.26}$$

In Section 3 we described numerical experiments that suggest that the bounds (5.24) are tight. This observation, and the results summarized in Theorem 5.2 allowed us to make the series of observations about the behavior of the gradient norm presented in this section. More precisely, the characterization of the oscillatory behavior of the gradient norm derived in Section 5.1 is a direct consequence of results (5.17), (5.18) and (5.19) when combined with (5.24), while the characterization of the path followed by the iterates described in Section 5.2 is a result of the combination of (5.21), (5.22) and (5.24).

6. The 2-dimensional case

Since the set $\mathcal{I}$ in (3.27) is always empty in the 2-dimensional case, the assumptions of Lemma 3.5 are never satisfied and the values of c^2 will not be restricted to the interval (3.29). Therefore, we can expect a different behavior of the steepest descent method in the 2-dimensional case. In particular, we will be able to describe the behavior of the gradient norm at every iteration in terms of the starting point and the condition number γ. The rate of convergence in f is also easily characterized.

As the steepest descent method is invariant under the rotations and translations of the coordinates, let us assume, without losing generality, that

$$Q = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \quad \text{and} \quad x^* = (0, 0) \tag{6.27}$$

in (3.1), and that $0 < \lambda_1 < \lambda_2$.

Writing $x^{(k)} = (x_1^{(k)}, x_2^{(k)})$, relation (3.3) implies that

$$g^{(k)} = \big(\lambda_1 x_1^{(k)}, \lambda_2 x_2^{(k)}\big). \tag{6.28}$$

Let us define

$$\rho^{(k)} = \frac{x_1^{(k)}}{x_2^{(k)}}.$$

Using (3.2) and (3.4) it is easy to verify that $\rho^{(k+1)} = -\gamma^2/\rho^{(k)}$ for all $k \geq 0$, with $\gamma = \lambda_2/\lambda_1$, as pointed out in [3]. This implies that

$$\rho^{(2k)} = \rho^{(0)} \quad \text{and} \quad \rho^{(2k+1)} = -\frac{\gamma^2}{\rho^{(0)}}, \tag{6.29}$$

for all $k \geq 0$. Hence, the sequence of iterates $\{x^{(k)}\}$ zigzags between the pair of straight lines $x_2 = (1/\rho^{(0)})x_1$ and $x_2 = -(\rho^{(0)}/\gamma^2)x_1$, as is the case *asymptotically* in the n-dimensional case (see figures 7 and 8).

Observe now that (3.8), (6.28) and (6.29) imply that

$$\frac{\alpha_2^{(2k)}}{\alpha_1^{(2k)}} = \frac{\gamma}{\rho^{(2k)}} = \frac{\gamma}{\rho^{(0)}} \tag{6.30}$$

and

$$\frac{\alpha_2^{(2k+1)}}{\alpha_1^{(2k+1)}} = \frac{\gamma}{\rho^{(2k+1)}} = -\frac{\rho^{(0)}}{\gamma}, \tag{6.31}$$

for all $k \geq 0$. Hence the two subsequences $\{\alpha_2^{(2k)}/\alpha_1^{(2k)}\}$ and $\{\alpha_2^{(2k+1)}/\alpha_1^{(2k+1)}\}$ are both *constant* in the 2-dimensional case, and we can deduce from the definition of c in (3.22) that

$$c = \frac{\gamma}{\rho^{(0)}}. \tag{6.32}$$

In other words, c represents the *constant* slope $\gamma/\rho^{(0)}$ of the even subsequence of gradients $\{g^{(2k)}\}$ at each iteration (the *constant* slope of the odd subsequence $\{g^{(2k+1)}\}$ is equal to $-\rho^{(0)}/\gamma$).

As a consequence of this, the asymptotic analysis of the previous sections can now be replaced by an exact analysis based on the ratio $\rho^{(0)}$ (or equivalently the starting point $x^{(0)}$), whose choice in the 2-dimensional plane is obviously free. Indeed, it is easy to verify that (5.18), (5.19) and (5.20) hold for all $k \geq 0$, i.e.,

$$\frac{\|g^{(2k+1)}\|^2}{\|g^{(2k)}\|^2} = \frac{\gamma^2\big(\rho^{(0)}\big)^2(1-\gamma)^2}{\big(\big(\rho^{(0)}\big)^2 + \gamma^3\big)^2}, \tag{6.33}$$

$$\frac{\|g^{(2k+2)}\|^2}{\|g^{(2k+1)}\|^2} = \frac{\left(\rho^{(0)}\right)^2(\gamma-1)^2}{\left(\gamma+\left(\rho^{(0)}\right)^2\right)^2}, \tag{6.34}$$

$$\frac{f^{(k+1)}}{f^{(k)}} = \frac{\left(\rho^{(0)}\right)^2\gamma(\gamma-1)^2}{\left(\left(\rho^{(0)}\right)^2+\gamma^3\right)\left(\left(\rho^{(0)}\right)^2+\gamma\right)}, \tag{6.35}$$

and

$$\frac{\|g^{(k+2)}\|}{\|g^{(k)}\|} = \frac{f^{(k+1)}}{f^{(k)}}. \tag{6.36}$$

Let us now study under what conditions will the gradient norm oscillate. From (5.8) and (5.11) we see that oscillations will take place if the starting point satisfies

$$\gamma l_\gamma^{1/2} < \left|\rho^{(0)}\right| < \gamma u_\gamma^{1/2}, \tag{6.37}$$

or

$$l_\gamma^{1/2} < \left|\rho^{(0)}\right| < u_\gamma^{1/2}. \tag{6.38}$$

Moreover, since (6.33) and (6.34) are equalities the amplitude of the oscillations of the odd and even iterates is constant. Figure 11 gives a characterization of the oscillatory behavior

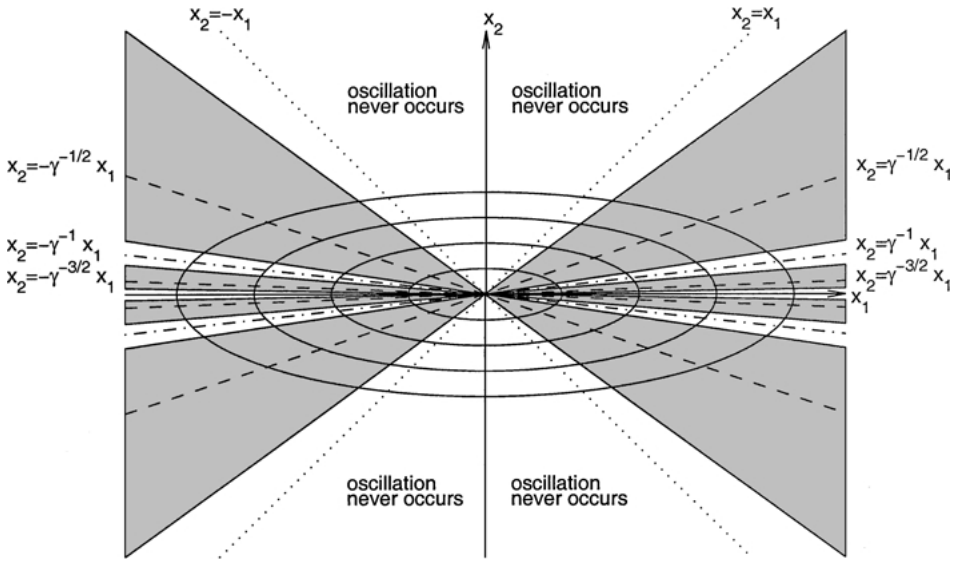

Figure 11. Characterization of the starting points for which the gradient norm will exhibit oscillations, in the 2-dimensional case. Here $\gamma = 9$.

of the gradient norm according to the choice of the starting point in the 2-dimensional plane, for the case $\lambda_1 = 1$ and $\lambda_2 = 9$. Conditions (6.37) and (6.38) determine two regions in each quadrant (see the shaded areas) for which the starting point will give rise to oscillations in the gradient. Observe that both conditions (6.37) and (6.38) together with (5.11) imply that oscillation will never occur when $|\rho^{(0)}| \leq 1$. For the first quadrant for instance, this corresponds to the region above the dotted line $x_2 = x_1$. Furthermore, because of (5.12), when γ increases and tends to infinity, the smaller shaded cone in each quadrant will tend to the horizontal axis, while the larger cone will expand to cover all the region $|\rho^{(0)}| > 1$, but without intersecting the smaller cone. Indeed, between these two cones lie the dash-dot lines corresponding to the worst case for the rate of convergence in f, which occurs when $|\rho^{(0)}| = \gamma$, and for which oscillations in the gradient norm will never occur. Finally, the largest oscillation in the gradient norm is obtained either when $|\rho^{(0)}| = \gamma^{3/2}$ or when $|\rho^{(0)}| = \gamma^{1/2}$ (see the dashed lines).

Let us now consider the rate of convergence in f. It can easily be verified that

$$\lim_{\gamma\to\infty} \frac{f^{(k+1)}}{f^{(k)}} = \begin{cases} 0, & \text{if } \left|\rho^{(0)}\right| < \gamma^{1/2}, \\ \dfrac{1}{2}, & \text{if } \left|\rho^{(0)}\right| = \gamma^{1/2}, \\ 1, & \text{if } \gamma^{1/2} < \left|\rho^{(0)}\right| < \gamma^{3/2}, \\ \dfrac{1}{2}, & \text{if } \left|\rho^{(0)}\right| = \gamma^{3/2}, \\ 0, & \text{if } \left|\rho^{(0)}\right| > \gamma^{3/2}. \end{cases} \tag{6.39}$$

Hence again, the rate of convergence may be characterized according to the region of the 2-dimensional plane in which the starting point $x^{(0)}$ lies. Three kinds of regions can be distinguished in each quadrant, as illustrated by figure 12 for the case $\lambda_1 = 1$ and $\lambda_2 = 9$. If $x^{(0)}$ is chosen outside the shaded areas (i.e. $|\rho^{(0)}| < \gamma^{1/2}$ or $|\rho^{(0)}| > \gamma^{3/2}$), the rate of convergence in f will be fast. If $x^{(0)}$ is selected on the boundary of the shaded areas (i.e. $|\rho^{(0)}| = \gamma^{1/2}$ or $|\rho^{(0)}| = \gamma^{3/2}$), the rate of convergence will be moderate. A starting point within the shaded

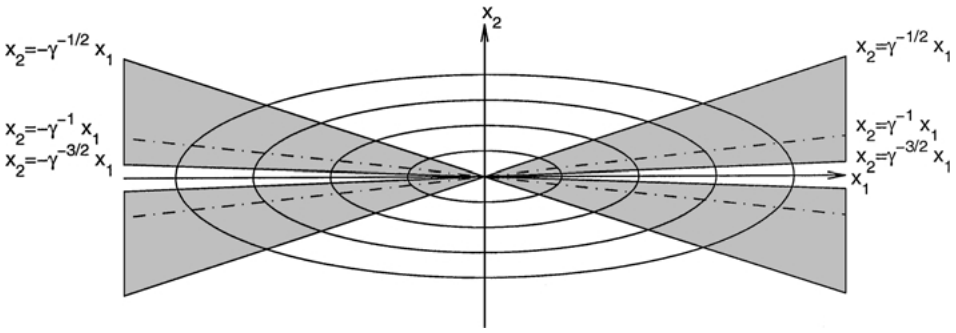

Figure 12. Characterization of the convergence rate in f in the 2-dimensional case according to the starting point (for $\gamma = 9$).

areas (i.e. $\gamma^{1/2} < |\rho^{(0)}| < \gamma^{3/2}$) will produce a slow rate of convergence—the slowest rate being reached for a starting point satisfying $|\rho^{(0)}| = \gamma$ (see the two dash-dot lines).

We note also that as the condition number γ grows and tends to infinity, the shaded areas in figure 12 shrink and tend towards the horizontal axis—which is the eigenvector ξ_1. Thus in the 2-dimensional case, if the starting point is chosen at random from, say, the uniform distribution, the chance of selecting a starting point that produces a fast rate of convergence *increases* with the condition number, a statement that cannot be made in the n-dimensional case. Indeed, we have seen in Section 5.2 that in the n-dimensional case, as the algorithm is progressively reduced to a search in the 2-dimensional subspace generated by ξ_1 and ξ_n, the iterates are generally attracted to the region near ξ_1—which is precisely the area where slow convergence in f prevails. This remark complements Akaike's analysis and illustrates some of the similarities and differences between the 2-dimensional and n-dimensional cases.

To conclude this section, we note from the fact that the shaded areas in figure 12 shrink and tend toward the horizontal axis as $\gamma \to \infty$, that for a fixed initial point $x^{(0)}$ (or equivalently $\rho^{(0)}$), the rate of convergence may even improve when γ increases (see figure 13). Indeed, it can be shown that the derivative with respect to γ of the right hand side term in (6.35) is negative if γ satisfies condition (4.4) and $\gamma^{1/2} \geq |\rho^{(0)}|$.

Given this, we should comment on the concluding remarks made in [2]. In that paper, the authors propose a two-point step size steepest descent method, and report numerical

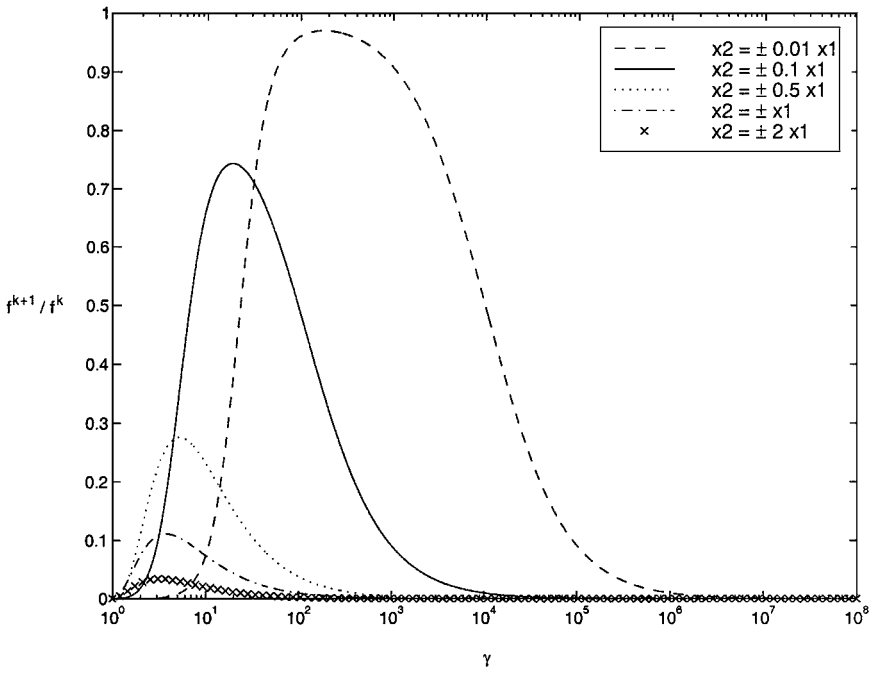

Figure 13. Rate of convergence of $f^{(k)}$ as a function of γ, in the 2-dimensional case, and for different choices of $\rho^{(0)}$.

experiments on a n-dimensional quadratic function for which the proposed method is faster than the classical steepest descent algorithm. To strengthen the numerical study, the authors analyze the convergence rate of their algorithm in the 2-dimensional case, and are surprised by the fact that the rate of convergence increases with the condition number of the Hessian matrix. They speculate that this could contribute to explain the numerical advantage of their method in the n-dimensional case. However, in the light of the analysis we have given above, that line of reasoning seems questionable. Even though in the 2-dimensional case the convergence rate of the steepest descent method may improve as the condition number increases, this is not true in the n-dimensional case. It is therefore necessary to show that the 2-dimensional case of the algorithm described in [2] is in fact representative of the n-dimensional case.

7. Final accuracy in f

Now that we have studied the behavior of the gradient norm, we conclude by making some observations on the final accuracy in the objective function, taking into account the effect of rounding errors.

For many optimization algorithms, the final accuracy in f, as measured by the difference $f(x) - f^*$, is intimately related to their speed of convergence. To illustrate this let us suppose that for all sufficiently large k there is a constant $0 < a < 1$ such that

$$f^{(k+1)} - f^* \leq a\big(f^{(k)} - f^*\big), \tag{7.1}$$

or equivalently, $f^{(k)} - f^{(k+1)} \geq (1 - a)(f^{(k)} - f^*)$. We let the algorithm iterate until the steps are so small that function values can no longer be distinguished in finite precision, i.e.

$$\frac{f^{(k)} - f^{(k+1)}}{f^{(k)}} \approx \mathbf{u}, \tag{7.2}$$

where we have assumed for convenience that $f^{(k)} \neq 0$ for all k sufficiently large. Thus $f^{(k)}$ is our best estimate of f^*. Assuming that the inequality (7.1) is tight we have

$$\frac{f^{(k)} - f^*}{f^{(k)}} \approx \frac{\mathbf{u}}{1 - a}. \tag{7.3}$$

Thus the slower the algorithm (the closer a is to 1) the fewer the correct digits in the final function value. We should note that this argument ignores the effects of roundoff errors in the computation of the iterates, which will prevent (7.1) from being sustained indefinitely.

For the steepest descent method with exact line searches, applied to a strongly convex quadratic function whose Hessian has a condition number γ, it is well known [11] that

$$a = \left(\frac{\gamma - 1}{\gamma + 1}\right)^2. \tag{7.4}$$

In addition, as argued by Akaike, we can expect (7.1) to be tight (see Section 3). Thus for this method the final accuracy in f is determined by the condition number of the Hessian. For large γ's, (7.3) can be approximated by

$$\frac{f^{(k)} - f^*}{f^{(k)}} \approx \frac{\gamma \mathbf{u}}{4}, \tag{7.5}$$

showing that the inaccuracy in f grows linearly with γ.

To test whether the behavior predicted by these relations can be observed in practice, even for non-quadratic objective functions, we performed numerical experiments using the quartic objective function in 100 variables,

$$\frac{1}{2}(x-1)^T D(x-1) + \frac{\sigma}{4}((x-1)^T B(x-1))^2 + 1, \tag{7.6}$$

where D was chosen as

$$D = \text{diag}[(1+\epsilon)^{-50}, (1+\epsilon)^{-49}, \ldots, (1+\epsilon)^{49}],$$

with $\epsilon = 0.18$, $\sigma = 0.18$, and

$$B = U^T U, \quad \text{with} \quad U = \begin{bmatrix} 1 & \ldots & 1 \\ & \ddots & \vdots \\ & & 1 \end{bmatrix}.$$

The starting point was chosen as $(-1)^i \times 50$ for $i = 1, \ldots, 100$. The Hessian matrix of this quartic function at the solution has a condition number of 1.3×10^7. We used double precision so that $\mathbf{u} \approx 2^{-16}$.

We used the steepest descent method, using the inexact line search of Moré and Thuente [13] that enforces the standard Wolfe conditions, and terminated it when no further decrease in the objectives function was possible. We obtained

$$f - f^* = 9.3D - 11 \quad \text{and} \quad \|g\|^2 = 4.4D - 13.$$

Note that there is a good agreement between our estimate (7.5) for the steepest descent method, which predicts approximately 10 correct digits in f, and these results—in spite of the fact that the problem was not quadratic.

References

1. H. Akaike, "On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method," Ann. Inst. Stat. Math. Tokyo, vol. 11, pp. 1–16, 1959.
2. J. Barzilai and J.M. Borwein, "Two-point step size gradient methods," IMA Journal of Numerical Analysis, vol. 8, pp. 141–148, 1988.

3. M.S. Bazaraa, H.D. Sherali, and C.M. Shetty, Nonlinear Programming, 2nd edn., John Wiley & Sons: New York, 1993.
4. I. Bongartz, A.R. Conn, N.I.M. Gould, and Ph.L. Toint, "CUTE: Constrained and unconstrained testing environment," ACM Transactions on Mathematical Software, vol. 21, no. 1, pp. 123–160, 1995.
5. R.H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," SIAM Journal on Scientific Computing, vol. 16, no. 5, pp. 1190–1208, 1995.
6. R. Byrd, J. Nocedal, and C. Zhu, "Towards a discrete Newton method with memory for large-scale optimization," in Nonlinear Optimization and Applications. G. Di Pillo and F. Giannessi (Eds.), Plenum: New York, 1996.
7. A.R. Conn, N.I.M. Gould, and Ph.L. Toint, "LANCELOT: A FORTRAN package for large-scale nonlinear optimization (Release A)," Number 17 in Springer Series in Computational Mathematics, Springer-Verlag: New York, 1992.
8. G.E. Forsythe, "On the asymptotic directions of the s-dimensional optimum gradient method," Numerische Mathematik, vol. 11, pp. 57–76, 1968.
9. J.Ch. Gilbert, Private communication, 1994.
10. P.E. Gill, W. Murray, and M.H. Wright, Practical Optimization, Academic Press: London, 1981.
11. D.G. Luenberger, Linear and Nonlinear Programming, 2nd edn., Addison-Wesley: Reading, MA, 1984.