On the complexity of the steepest-descent method with exact linesearches

by C. Cartis, N. I. M. Gould and Ph. L. Toint

Report NAXYS-16-2012 9 September 2012

University of Edinburgh, Edinburgh, EH9 3JZ, Scotland (UK)

Rutherford Appleton Laboratory, Chilton, OX11 0QX, England (UK)

University of Namur, 61, rue de Bruxelles, B5000 Namur (Belgium)

http://www.fundp.ac.be/sciences/naxys

On the complexity of the steepest-descent with exact linesearches

C. Cartis, N. I. M. Gould and Ph. L. Toint

9 September 2012

Abstract

The worst-case complexity of the steepest-descent algorithm with exact linesearches for unconstrained smooth optimization is analyzed, and it is shown that the number of iterations of this algorithm which may be necessary to find an iterate at which the norm of the objective function's gradient is less that a prescribed ϵ is, essentially, a multiple of $1/\epsilon^2$, as is the case for variants of the same algorithms using inexact linesearches.

1 Introduction

The worst-case analysis of optimization algorithms for finding unconstrained stationary points of nonlinear non-convex functions has recently been considered in a number of contributions (see Nesterov, 2004, Nesterov and Polyak, 2006, Cartis, Gould and Toint, 2011*a*, 2011*b*, 2011*c*, 2012*a*, 2012*b*, 2012*c*, Vicente, 2010, Bian, Chen and Ye, 2012, Gratton, Sartenaer and Toint, 2008, or Jarre, 2011, to cite a few). In particular, the study of the steepest-descent method, the most archetypal method for unconstrained nonlinear optimization, was considered by several authors, whose analysis differ primarily by the particular technique used for (possibly approximately) minimizing the objective function along the steepest-descent direction. An upper bound on the number of iterations required to obtain an approximate stationary point was given by Nesterov (2004) using a variant of the algorithm where the step is computed using the knowledge of a global Lipschitz constant on the gradient of the objective function. He showed that at most $O(\epsilon^{-2})$ iterations might be needed to find an iterate at which the Euclidean norm of the gradient is below a generic tolerance $\epsilon > 0$. As it turns out, his result also applies to the "pure" steepest-descent algorithm, that is the variant using exact linesearches. A lower complexity bound was also obtained by Cartis, Gould and Toint (2010), where it was shown that the bound of $O(\epsilon^{-2})$ iterations is essentially tight for a version using a Goldstein type linesearch. However, this result depends on a one-dimensional counter-example where the objective function is monotonically decreasing, in which case an exact linesearch would obviously give much better results. The purpose of this short paper is to close the remaining conceptual gap, that is to show that the lower bound of $O(\epsilon^{-2})$ iterations also holds for the steepest-descent algorithm with exact linesearches when applied on functions with globally Lipschitz continuous gradient.

The next section recalls the algorithms and the assumptions required for our complexity analysis. Section 3 proposes an example of worst-case behaviour for the method, while Section 4 is devoted to verifying that the example does satisfy the assumptions stated. A few words of conclusion are presented in Section 5.

2 The steepest-descent method with exact linesearches

We consider the unconstrained minimization problem

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

where $f(x)$ is a smooth function from $\mathbb{R}^n$ into $\mathbb{R}$. One of the simplest and oldest algorithm for solving this problem is the steepest-descent method by Cauchy (1847), whose iterates are defined, for a given initial guess x_0, by the simple iteration

$$x_{k+1} = \arg\min_{t \geq 0} f(x_k - t g_k), \qquad (k \geq 0) \tag{2.2}$$

where $g_k = \nabla_x f(x_k)$ and where ties are broken by choosing the first minimizer of $f(x_k - t g_k)$ if there is more than one (say). This choice is of course most often numerically unrealistic, except for special functions $f(x)$ such as quadratics, where the minimizer can be determined analytically. But it remains an ideal that numerically sounder techniques attempt to imitate, justifying our curiosity.

The assumptions we make on problem (2.1) are as follows.

$\boxed{\textbf{AF.0}}$ $f(x)$ is bounded below on $\mathbb{R}^n$, that is there exists a constant $\kappa_{\rm lbf}$ such that, for all $x \in \mathbb{R}^n$,

$$f(x) \geq \kappa_{\rm lbf}.$$

$\boxed{\textbf{AF.1}}$ $f(x)$ is continuously differentiable on $\mathbb{R}^n$.

$\boxed{\textbf{AF.2}}$ $g(x) = \nabla_x f(x)$ is Lipschitz continuous on $\mathbb{R}^n$, that is there exists a constant $L_g \geq 0$ such that, for all $x, y \in \mathbb{R}^n$,

$$\|g(x) - g(y)\| \leq L_g \|x - y\|.$$

Here and below, $\|\cdot\|$ stands for the Euclidean norm.

We now briefly recall the upper complexity bound for algorithm (2.2) by suitably reformulating the result of Nesterov (2004).

Theorem 2.1 Suppose that AF.0–AF.2 hold. Then there exists a constant $\kappa_{\rm upp}$ depending on x_0 and possibly on n such that, for all $\epsilon \in (0,1)$ at most

$$\left\lceil \frac{\kappa_{\rm upp}}{\epsilon^2} \right\rceil \tag{2.3}$$

iterations of method (2.2) are needed to obtain an iterate x_k such that $\|g_k\| \leq \epsilon$.

Proof. We first note that AF.1, Taylor's expansion at x_k and AF.2 give that, for each $k \geq 0$,

$$f(x_k) - f(x_k - t g_k) \geq f(x_k) - f(x_k) + t\|g_k\|^2 - \tfrac{1}{2} t^2 L_g \|g_k\|^2$$

for any $t \geq 0$. Maximizing the right-hand side of this inequality with respect to t, we obtain that

$$f(x_k) - f(x_k - \frac{1}{L_g} g_k) \geq \frac{1}{2L_g}\|g_k\|^2 \geq \frac{\epsilon^2}{2L_g} \tag{2.4}$$

for each iteration k, as long as $\|g_k\| > \epsilon$.

But (2.2) ensures that the slope of $f(x_k - tg_k)$ must be zero at $x_{k+1} = x_k - t_k g_k$, giving that, for all k,

$$0 = \langle g_k, g(x_{k+1}) \rangle = \|g_k\|^2 + \langle g_k, g(x_k - t_k g_k) - g_k \rangle \geq \|g_k\|^2 (1 - L_g t_k),$$

where we used the Cauchy-Schwartz inequality and AF.2. This implies that t_k, the argument of the (first) minimum in (2.2), is such that $t_k \geq 1/L_g$ and therefore, because of (2.4), that, for each k,

$$f(x_k) - f(x_k - t_k g_k) \geq f(x_k) - f(x_k - \frac{1}{L_g} g_k) \geq \frac{\epsilon^2}{2L_g}$$

as long as $\|g_k\| > \epsilon$. Thus a maximum number of

$$\left\lceil \frac{2L_g(f(x_0) - \kappa_{\text{lbf}})}{\epsilon^2} \right\rceil \stackrel{\text{def}}{=} \left\lceil \frac{\kappa_{\text{upp}}}{\epsilon^2} \right\rceil$$

such iterations may take place before x_k is found such that $\|g_k\| \leq \epsilon$. □

The purpose of the present paper is to show that the bound (2.3) is essentially tight, which cannot be deduced from the one-dimensional example of Cartis et al. (2010). The next section describes how to build a new two-dimensional example where algorithm (2.2) essentially requires $\mathcal{O}(\epsilon^{-2})$ iterations to achieve $\|g_k\| \leq \epsilon$.

3 Constructing a counter-example

Because, as in Cartis et al. (2010), our example is based on polynomial Hermite interpolation, we first state and prove crucial properties of this type of interpolation.

Theorem 3.1 Assume that real values f_0, g_0, h_0, f_T, g_T, h_T and $T > 0$ are known. Then there exists a fifth order polynomial $p(t) \stackrel{\text{def}}{=} c_0 + c_1 t + c_2 t^2 + c_3 t^3 + c_4 t^4 + c_5 t^5$, $t \in [0, T]$, such that

$$p(0) = f_0, \quad p'(0) = g_0 \quad \text{and} \quad p''(0) = h_0,$$

$$p(T) = f_T, \quad p'(T) = g_T \quad \text{and} \quad p''(T) = h_T.$$

The coefficients of this polynomial are given by

$$\begin{aligned} c_0 = f_0, \quad c_1 = g_0, \quad c_2 = \tfrac{1}{2}h_0, \quad c_3 = \frac{1}{T}(10r_0 - 4r_1 + \tfrac{1}{2}r_2), \\ c_4 = \frac{1}{T^2}(-15r_0 + 7r_1 - r_2) \ \text{ and } \ c_5 = \frac{1}{T^3}(6r_0 - 3r_1 + \tfrac{1}{2}r_2), \end{aligned} \tag{3.5}$$

where

$$r_0 = \frac{1}{T^2}(f_T - f_0 - g_0 T - \tfrac{1}{2}h_0 T^2), \quad r_1 = \frac{1}{T}(g_T - g_0 - h_0 T) \ \text{ and } \ r_2 = h_T - h_0.$$

Moreover, if there are non-negative constants κ_0, κ_1 and κ_2 such that

$$|r_1| \leq \kappa_0, \quad |r_1| \leq \kappa_1 \quad \text{and} \quad |r_2| \leq \kappa_2, \tag{3.6}$$

Then there exists $\kappa_f \geq 0$, $\kappa_g \geq 0$ and $\kappa_h \geq 0$ only depending on κ_0, κ_1 and κ_2 such that, for all $t \in [0, T]$,

$$|p(t)| \leq |f_0| + |g_0|T + \tfrac{1}{2}|h_0|T^2 + \kappa_f T^2, \tag{3.7}$$

$$|p'(t)| \leq |g_0| + \tfrac{1}{2}|h_0|T + \kappa_g T, \quad \text{and} \quad |p''(t)| \leq |h_0| + \kappa_h. \tag{3.8}$$

Proof. (See Cartis et al., 2011*c*.) Using the form of $p(t)$, we write the desired interpolation conditions as

$$p(0) = c_0 = f_0, \quad p'(0) = c_1 = g_0, \quad p''(0) = 2c_2 = h_0 \tag{3.9}$$

(which immeditaley gives the desired values for c_0, c_1 and c_2) and

$$\begin{array}{lclcl} p(T) & = & c_0 + c_1T + c_2T^2 + c_3T^3 + c_4T^4 + c_5T^5 & = & f_T, \\ p'(T) & = & c_1 + 2c_2T + 3c_3T^2 + 4c_4T^3 + 5c_5T^4 & = & g_T, \\ p''(T) & = & 2c_2 + 6c_3T + 12c_4T^2 + 20c_5T^3 & = & h_T. \end{array}$$

These conditions can the be re-expressed as a linear system with unknowns c_3, c_4 an c_5, whose solution exists and turns out to be

$$\left(\begin{array}{c} c_3 \\ c_4 \\ c_5 \end{array}\right) = \left(\begin{array}{ccc} \frac{1}{T} & 0 & 0 \\ 0 & \frac{1}{T^2} & 0 \\ 0 & 0 & \frac{1}{T^3} \end{array}\right) \left(\begin{array}{rrr} 10 & -4 & \frac{1}{2} \\ -15 & 7 & -1 \\ 6 & -3 & \frac{1}{2} \end{array}\right) \left(\begin{array}{c} \frac{1}{T^2}[f_T - f_0 - g_0T - \frac{1}{2}h_0T^2] \\ \frac{1}{T}[g_T - g_0 - h_0T] \\ h_T - h_0 \end{array}\right),$$

completeing the proof of (3.5). Taking absolute values in this relation, we obtain that

$$\left(\begin{array}{c} |c_3| \\ |c_4| \\ |c_5| \end{array}\right) \le \left(\begin{array}{c} \frac{1}{T}[10\kappa_0 + 4\kappa_1 + \frac{1}{2}\kappa_2] \\ \frac{1}{T^2}[15\kappa_0 + 7\kappa_1 + \kappa_2] \\ \frac{1}{T^3}[6\kappa_0 + 3\kappa_1 + \frac{1}{2}\kappa_2] \end{array}\right) \stackrel{\rm def}{=} \left(\begin{array}{c} \kappa_{c3}/T \\ \kappa_{c4}/T^2 \\ \kappa_{c5}/T^3 \end{array}\right).$$

As a consequence, we have that, for all $t \in [0, T]$,

$$|p(t)| \le |f_0 + g_0T + \tfrac{1}{2}h_0T^2| + (\kappa_{c3} + \kappa_{c4} + \kappa_{c5})T^2, \tag{3.10}$$

which gives (3.7) with $\kappa_f \stackrel{\rm def}{=} \kappa_0 + \kappa_{c3} + \kappa_{c4} + \kappa_{c5}$. Similarly, we obtain that, for all $t \in [0, T]$,

$$|p'(t)| \le |g_0 + h_0T| + (3\kappa_{c3} + 4\kappa_{c4} + 5\kappa_{c5})T \tag{3.11}$$

yieldling the first part of (3.8) with $\kappa_g \stackrel{\rm def}{=} \kappa_1 + 3\kappa_{c3} + 4\kappa_{c4} + 5\kappa_{c5}$, and

$$|p''(t)| \le |h_0| + (6\kappa_{c3} + 12\kappa_{c4} + 20\kappa_{c5}), \tag{3.12}$$

from which the second part of (3.8) finally follows with $\kappa_h \stackrel{\rm def}{=} \kappa_2 + 6\kappa_{c3} + 12\kappa_{c4} + 20\kappa_{c5}$. □

We now turn to construction our worst-case example for the steepest-descent method (2.2). The idea is to fix an arbitrary $\tau \in (0, \frac{1}{3}]$ and then to define $f(x, y)$, the objective function in the example as the sum of $f_1(x)$ and $f_2(x, y)$. As in Cartis et al. (2010), $f_1(x)$ is defined by piecewise Hermite polynomial interpolation between the sequence of iterates

$$x_0 = 0, \quad x_{k+1} = x_k + \sigma_k \quad (k \ge 0) \tag{3.13}$$

of the values

$$f_1(x_0) = \zeta(1+2\eta), \quad f_1(x_{k+1}) = f_1(x_k) - \sigma_k^2, \quad f'(x_k) = -\sigma_k, \quad \text{and} \quad f_1''(x_k) = 0, \tag{3.14}$$

where $\zeta(\cdot)$ is the Riemann zeta function and

$$\eta = \eta(\tau) \stackrel{\rm def}{=} \frac{1}{2-\tau} - \frac{1}{2} = \frac{\tau}{4-2\tau} \in (0, \tfrac{1}{2}) \quad \text{and} \quad \sigma_k \stackrel{\rm def}{=} \left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta}. \tag{3.15}$$

From (3.5), we then find that, for $x \in [x_k, x_{k+1}]$ and $t = (x - x_k)/\sigma_k$,

$$f_1(x) = f_1(x_k) - \sigma_k^2 t + \sigma_k(\sigma_k - \sigma_{k+1})\big[-4t^3 + 7t^4 - 3t^5\big],$$

$$f_1'(x) = -\sigma_k + (\sigma_k - \sigma_{k+1})\big[-12t^2 + 28t^3 - 15t^4\big], \tag{3.16}$$

and

$$f_1''(x) = \frac{\sigma_k - \sigma_{k+1}}{\sigma_k}\big[-24t + 84t - 60t^3\big]. \tag{3.17}$$

It is easy to verify that, for $t \in [0,1]$

$$-12t^2 + 28t^3 - 15t^4 = -t^2[12 - 28t + 15t^2] \le 1$$

and thus, using (3.16), that

$$f_1'(x) \le -\sigma_{k+1} < 0 \quad \text{for all} \quad x \in [x_k, x_{k+1}]. \tag{3.18}$$

In addition, taking into account that

$$0 < \frac{\sigma_k - \sigma_{k+1}}{\sigma_k} = 1 - \left(\frac{k+1}{k+2}\right)^{\frac{1}{2}+\eta} < 1,$$

for $k \ge 0$ and that $t \in [0,1]$ if $x \in [x_k, x_{k+1}]$, we obtain by a straightforward majoration in (3.17) that

$$|f_1''(x)| < 168 \tag{3.19}$$

for $x \in [x_k, x_{k+1}]$, which in turn implies that $f_1''(x)$ is uniformly bounded for all $x \ge 0$. The behaviour of $f_1(x)$ and of its first and second derivatives are pictured in Figure 3.1 on this page.

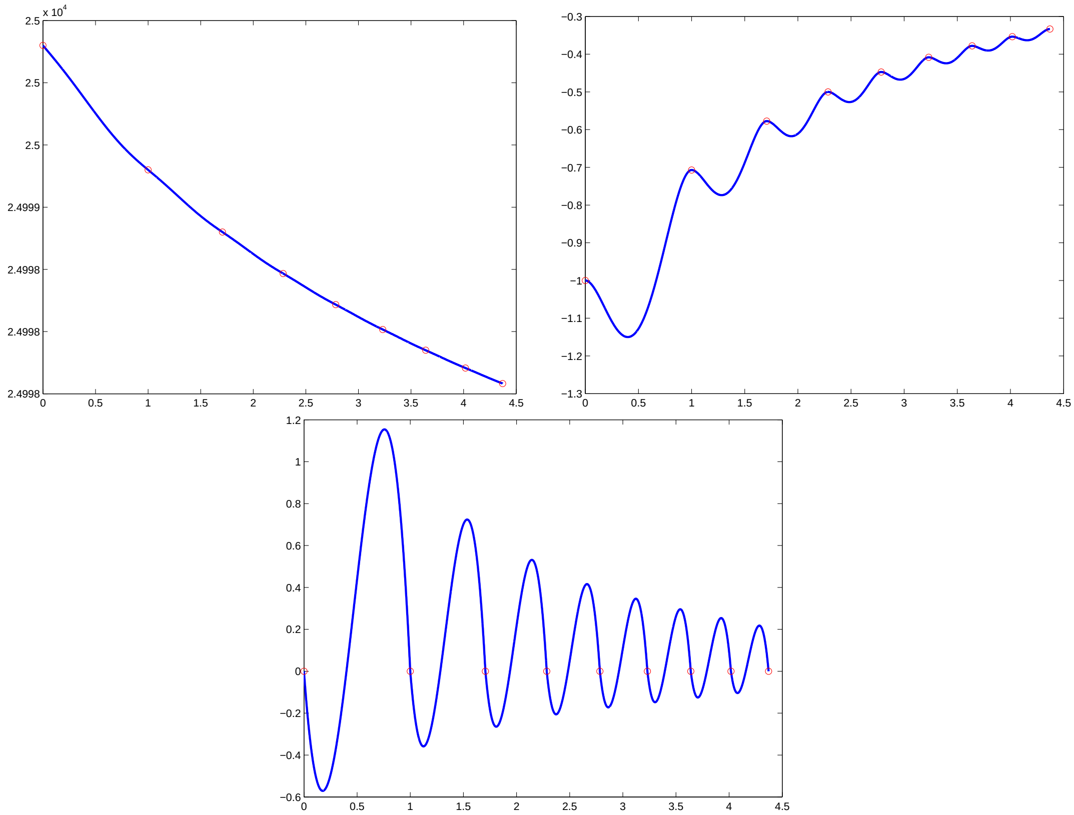

Figure 3.1: The function $f_1(x)$ and its first two derivatives (from top to bottom and left to right) on the first 8 intervals

We now turn to the specification of the function $f_2(x,y)$, whose role is to limit the iterates in the y-direction to a progressively narrower "corridor", thereby forcing the iteration path to oscillate between its lower and upper limits. We have already prescribed that the x-components of the successive iterates are given by (3.13), and $f_2(x, y)$ will be constructed in such a way that its gradient at the iterates is equal in norm to that of $f_1(x)$ but alternating in sign, thus generating the necessary orthogonality conditions and the oscillating iteration path.

More specifically (and in accordance with (3.13)), define, for all $k \geq 0$,

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = \begin{pmatrix} x_k \\ y_k \end{pmatrix} + \begin{pmatrix} \sigma_k \\ (-1)^k \sigma_k \end{pmatrix}$$

with

$$\sigma_k = -f_1'(x_k), \tag{3.20}$$

defining the zig-zaging piecewise linear iteration path $y(x)$ illustrated in Figure 3.2.

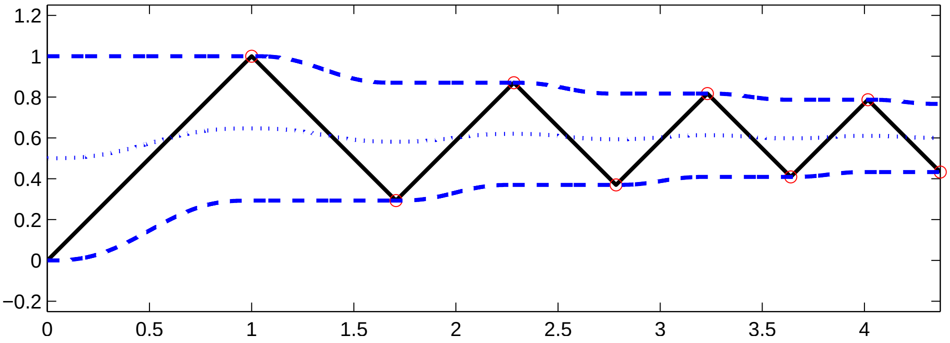

Figure 3.2: The iteration path $y(x)$ (plain), the upper and lower boundaries $y_{\text{low}}(x)$ and $y_{\text{up}}(x)$ (dashed) and $y_{\text{mid}}(x)$ (dotted) for $k = 1, \ldots, 8$ and $\eta = 10^{-5}$.

We now define the lower and upper boundaries of the "corridor" containing the iterates. This is achieved by defining the lower boundary $y_{\text{low}}(x)$ as a twice continuously differentiable curve that interpolates the y coordinates of the iterates of index $2k$ ($k \geq 0$) and is constant on the intervals $[x_{2k-1}, x_{2k}]$, yielding

$$y_{\text{low}}(x_{2k-1}) = y_{\text{low}}(x_{2k}) = y_{2k}.$$

Polynomial Hermite interpolation is used to twice continuously connect the constants parts. The upper boundary $y_{\text{up}}(x)$ is defined in the same way to interpolate the y coordinates of the iterates of index $2k+1$ ($k \geq 0$), being constant on $[x_{2k}, x_{2k+1}]$, yielding

$$y_{\text{up}}(x_{2k}) = y_{\text{up}}(x_{2k+1}) = y_{2k+1}.$$

Both $y_{\text{low}}(x)$ and $y_{\text{up}}(x)$ are shown on Figure 3.2, as well as their average $y_{\text{mid}}(x) = \frac{1}{2}(y_{\text{up}}(x) + y_{\text{low}}(x))$. If we define

$$\delta(x) \stackrel{\text{def}}{=} y_{\text{up}}(x) - y_{\text{low}}(x), \tag{3.21}$$

(the corridor width at x), we note that, by construction,

$$\delta(x_k) = \sigma_k \quad \text{for all} \quad k \geq 0. \tag{3.22}$$

Moreover, since the interpolation conditions defining $y_{\text{up}}(x)$ are given (for $x \in [x_{2k-1}, x_{2k}]$, say) by

$$y_{\text{up}}(x_{2k-1}) = \sum_{i=0}^{2k-2} (-1)^i \sigma_i \quad \text{and} \quad y_{\text{up}}(x_{2k}) = y_{\text{up}}(x_{2k+1}) = \sum_{i=0}^{2k} (-1)^i \sigma_i, \tag{3.23}$$

$$y_{\text{up}}'(x_{2k-1}) = y_{\text{up}}'(x_{2k}) = 0 \quad \text{and} \quad y_{\text{up}}''(x_{2k-1}) = y_{\text{up}}''(x_{2k}) = 0, \tag{3.24}$$

a closer inspection of the interpolating polynomial (see (3.5)) reveals that, for $x \in [x_{2k-1}, x_{2k}]$,

$$y_{\rm up}(x) = y_{\rm up}(x_{2k-1}) - (\sigma_{2k-1} - \sigma_{2k})\big[10t^3 - 15t^4 + 6t^5\big], \tag{3.25}$$

where $t = (x - x_{2k-1})/\sigma_{2k-1}$. Symmetrically, we have that, for $x \in [x_{2k}, x_{2k+1}]$,

$$y_{\rm low}(x) = y_{\rm low}(x_{2k}) + (\sigma_{2k} - \sigma_{2k+1})\big[10t^3 - 15t^4 + 6t^5\big], \tag{3.26}$$

where $t = (x - x_{2k})/\sigma_{2k}$. We thus obtain from (3.25) and (3.26), using (3.22) and defining $t = (x - x_k)/\sigma_k$, that, for $x \in [x_k, x_{k+1}]$

$$\delta(x) = \sigma_k - (\sigma_k - \sigma_{k+1})\big[10t^3 - 15t^4 + 6t^5\big] \tag{3.27}$$

and

$$y_{\rm mid}(x) = y_{\rm mid}(x_k) + \tfrac{1}{2}(-1)^k(\sigma_k - \sigma_{k+1})\big[10t^3 - 15t^4 + 6t^5\big]. \tag{3.28}$$

These two last relations yield that

$$\delta'(x) = 2(-1)^{k+1} y'_{\rm mid}(x) = -30\,\frac{\sigma_k - \sigma_{k+1}}{\sigma_k}\big[t^2 - 2t^3 + t^4\big] \le 0, \tag{3.29}$$

and also that

$$\delta''(x) = 2(-1)^{k+1} y''_{\rm mid}(x) = -60\,\frac{\sigma_k - \sigma_{k+1}}{\sigma_k^2}\big[t - 3t^2 + 2t^3\big]. \tag{3.30}$$

The last inequality in (3.29) results from the decreasing nature of σ_k and the fact that $1 - 2t + t^2 = (1-t)^2 \ge 0$ for $t \in [0,1]$. It immediately implies, with (3.22) and (3.29), that $\delta(x)$ is non-increasing and that

$$\sigma_k = \delta(x_k) \ge \delta(x) \ge \delta(x_{k+1}) = \sigma_{k+1} \quad \text{for} \quad x \in [x_k, x_{k+1}]. \tag{3.31}$$

The next step is to define, for each x, $f_2(x,y)$ as a twice continuously differentiable function of y whose value is small between $y_{\rm low}(x)$ and $y_{\rm up}(x)$ and first increases before levelling off when the distance of y to the corridor increases, thereby keeping the iterates within the corridor. The details of $f_2(x,y)$ are given by

$$f_2(x,y) = \left\{ \begin{array}{ll} 8\,\delta(x)^2 & \text{if } y \le y_{\rm low}(x) - 1 \\ (y - y_{\rm mid}(x))^2 & \text{if } y \in [y_{\rm low}(x), y_{\rm up}(x)] \\ 8\,\delta(x)^2 & \text{if } y \ge y_{\rm up}(x) + 1 \end{array} \right. \tag{3.32}$$

where Hermite interpolation is once more used to twice continuously connect the first and second interval, as well as the second and third. In the first of these intervals, $f_2(x,y)$ is thus defined by a fifth order polynomial translated to $[0,1]$, with boundary conditions on this latter interval given by

$$p(0) = 8\,\delta(x)^2, \quad p'(0) = 0, \quad p''(0) = 0$$

and

$$p(1) = (y_{\rm low}(x) - y_{\rm mid}(x))^2, \quad p'(1) = 2(y_{\rm low}(x) - y_{\rm mid}(x)) = -\delta(x) \quad \text{and} \quad p''(1) = 2.$$

The interpolation conditions on the second interval are symmetrically defined. Figure 3.3 shows the shape of $f_2(x,y)$ for fixed x.

Note that $f_2(x,y)$ is symmetric in y with respect to $y_{\rm mid}(x)$ by construction. Note also that, using (3.21), the definition of $y_{\rm mid}(x)$ and (3.22),

$$\frac{\partial f_2}{\partial y}(x_{2k}, y_{2k}) = \frac{\partial f_2}{\partial y}(x_{2k}, y_{\rm low}(x_{2k})) = 2(y_{\rm low}(x_{2k}) - y_{\rm mid}(x_{2k})) = -\delta(x_{2k}) = -\sigma_{2k} \tag{3.33}$$

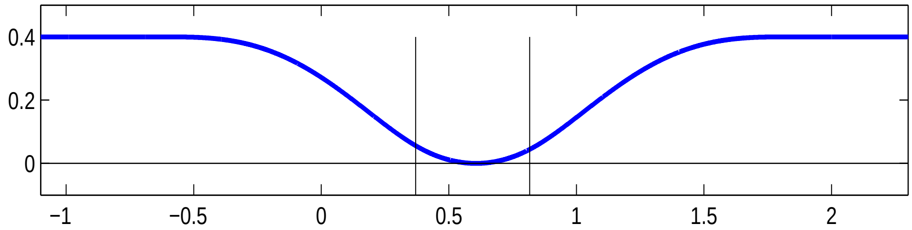

Figure 3.3: The shape of $f_2(x,y)$ for $x = x_2$ and $\eta = 10^{-5}$, the vertical lines indicating the values of $y_{\rm low}(x_2)$ and $y_{\rm up}(x_2)$.

and, similarly,

$$\begin{aligned} \frac{\partial f_2}{\partial y}(x_{2k+1}, y_{2k+1}) &= \frac{\partial f_2}{\partial y}(x_{2k+1}, y_{\rm up}(x_{2k+1})) \\ &= 2(y_{\rm up}(x_{2k+1}) - y_{\rm mid}(x_{2k+1})) \\ &= \delta(x_{2k+1}) \\ &= \sigma_{2k+1}. \end{aligned} \tag{3.34}$$

Note also that, because of (3.32) and (3.29) taken at $x = x_k$ (i.e. $t = 0$),

$$\frac{\partial f_2}{\partial x}(x_k, y_k) = -2(y_k - y_{\rm mid}(x_k))y'_{\rm mid}(x_k) = 0. \tag{3.35}$$

We finally define the objective function of our minimization problem (2.1) by

$$f(x,y) \stackrel{\rm def}{=} \left\{ \begin{array}{ll} \frac{1}{\sqrt{2}}\left[f_1(x) + f_2(x,y)\right] & \text{for } \ x \geq 0, y \in \mathbb{R}, \\ \frac{1}{\sqrt{2}}\left[f_1(0) + xf'_1(0) + f_2(0,y)\right] & \text{for } \ x < 0, y \in \mathbb{R}, \end{array} \right. \tag{3.36}$$

whose contour lines, superimposed on the path of iterates, are shown in Figure 3.4.

We thus obtain, using (3.14), (3.33)-(3.34) and (3.35), that

$$g_{SD2}(x_k, y_k) = -\frac{1}{\sqrt{2}} \left(\begin{array}{c} \sigma_k \\ (-1)^k \sigma_k \end{array} \right), \tag{3.37}$$

and therefore that

$$\|g_{SD2}(x_k)\| = \sigma_k.$$

Because of the definition of σ_k in (3.15), this implies that the algorithm will require, for any $\epsilon \in (0,1)$, at least

$$\left\lfloor \frac{1}{\epsilon^{2-\tau}} \right\rfloor \tag{3.38}$$

iterations to produce an iterate x_k such that $\|g_k\| \leq \epsilon$. This allows us to conclude, as desired, that the evaluation complexity bound of $\mathcal{O}(\epsilon^{-2})$ is essentially sharp, provided we can show that $f(x,y)$ is bounded below and has a globally Lipschitz continuous gradient, and that the slope of $f(x,y)$ is always non-positive along the trajectory. This is the object of the next section.

4 Verifying the example

We start by a useful auxiliary result.

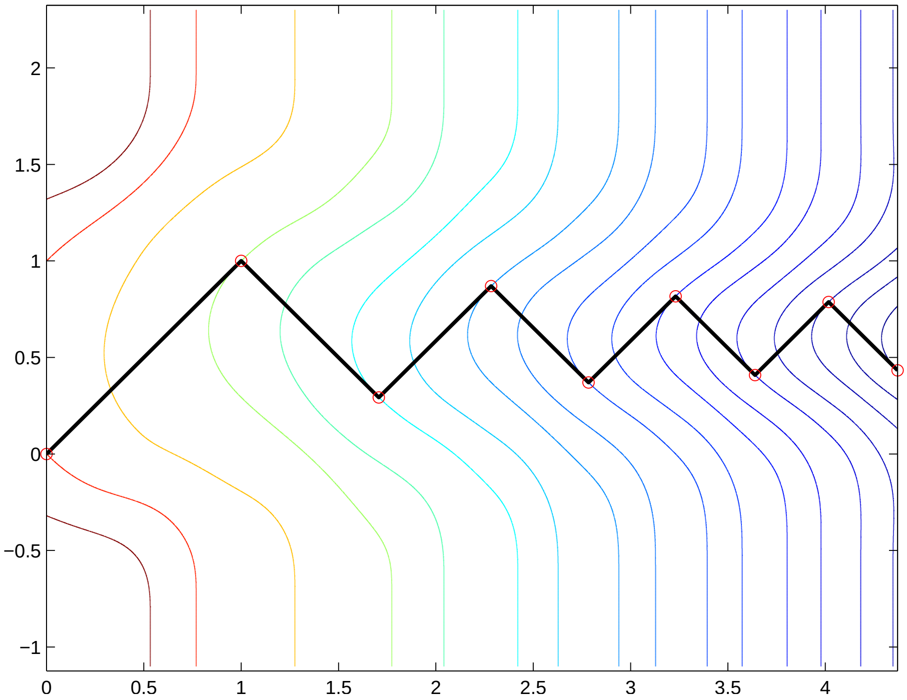

Figure 3.4: The countour lines of $f(x,y)$ and the path of iterates for $\eta = 10^{-5}$.

Lemma 4.1 The values of

$$f_2(x,y), \quad \frac{\partial f_2}{\partial x}(x,y), \quad \frac{\partial f_2}{\partial y}(x,y), \quad \frac{\partial^2 f_2}{\partial x^2}(x,y), \quad \frac{\partial^2 f_2}{\partial y^2}(x,y) \quad \text{and} \quad \frac{\partial^2 f_2}{\partial x \partial y}(x,y)$$

are uniformly bounded (in absolute value) for all $x \geq 0$ and $y \in [y_{\rm low}(x) - 1, y_{\rm low}(x)] \cup [y_{\rm up}(x), y_{\rm up}(x) + 1]$.

Proof. Because, for each x and $y \in [y_{\rm low}(x) - 1, y_{\rm low}(x)]$, $f_2(x,y)$ is a polynomial in y on an interval of length one, its values and that of its first and second derivatives with respect to y are uniformly bounded (in absolute value) provided its coefficients are uniformly bounded, which is the case (see (3.5) with $T = 1$ in Theorem 3.1, page 3) if the quantities

$$|8\,\delta(x)^2 - (y_{\rm up}(x) - y_{\rm mid}(x))^2 - 0 - \tfrac{1}{2}0| \quad \text{and} \quad |\delta(x) - 0 - 0| \tag{4.39}$$

are themselves uniformly bounded (the third component of the right-hand side of (3.5) being identically equal to 2). But this is the case for the first term in (4.39) since

$$|8\,\delta(x)^2 - (y_{\rm up}(x) - y_{\rm mid}(x))^2| = |8\,\delta(x)^2 - \tfrac{1}{4}\delta(x)^2| < 8\delta(x)^2 \leq 8,$$

and for the second because of (3.31) and the bound $\sigma_k \leq 1$. What about the derivatives with respect to x (for $y \in [y_{\rm low}(x) - 1, y_{\rm low}(x)]$)? Since $f_2(x,y)$ is defined, in this interval, as a polynomial in y shifted to $[0,1]$, the dependence in x is entirely captured by the coefficients $c_0, \ldots c_5$ of this polynomial, themselves depending on the boundary conditions

$$c_0 = 8\,\delta(x)^2, \quad c_1 = 0 \quad \text{and} \quad c_2 = 0 \tag{4.40}$$

and (3.5). The boundedness of the first and second derivatives of $c_0, \ldots, c_5$ (as functions of x) are then implied by (4.42) and the boundedness of the two terms in (4.39), which we already verified. Finally, the second derivative of $f_2(x, y)$ with respect to x and y (for $y \in [y_{\rm low}(x)-1, y_{\rm low}(x)]$) is also a polynomial on an (shifted) interval of length one, obtained by differentiating $c_1, \ldots, c_5$ with respect to x in the polynomial corresponding to the derivative of $f_2(x, y)$ with respect to y. Because we just verified that the first derivatives of $c_0, \ldots, c_5$ with respect to x are themselves uniformly bounded in x, this must also be the case of the cross-derivatives of $f_2(x, y)$. By symmetry, the conclusion of the lemma also holds for all $x \geq 0$ and $y \in [y_{\rm up}(x), y_{\rm up}(x)+1]$. □

Theorem 4.2 The function $f(x, y)$ is uniformly bounded below on $\mathbb{R}^2$.

Proof. Observe first that (3.18) implies that $f_1(x)$ is bounded below because

$$\sum_{i=0}^{\infty} \left[f_1(x_k) - f_1(x_{k+1}) \right] = \sum_{i=0}^{\infty} \sigma_k^2 = \zeta(1+2\eta) < \infty.$$

Moreover, it also results from this last observation that $f_1(x) \geq 0$ for all $x \geq 0$ (and thus also for all $x \in \mathbb{R}$). The fact that $f_2(x, y)$ is also uniformly bounded below results from its definition in (3.32) and Lemma 4.1. The desired conclusion then follows from (3.36). □

The verification that the gradient of $f(x, y)$ admits a uniform Lipschitz constant is a more lengthy calculation, which is the object of the next theorem. It depends on the observation that

$$\begin{aligned} 0 \leq \sigma_k - \sigma_{k+1} &= \left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta} - \left(\frac{1}{k+2}\right)^{\frac{1}{2}+\eta} \\ &\leq \left(\tfrac{1}{2}+\eta\right)\left(\frac{1}{k+2}\right)^{-\frac{1}{2}+\eta}\left(\frac{1}{k+1} - \frac{1}{k+2}\right) \\ &= \left(\tfrac{1}{2}+\eta\right)\left(\frac{1}{k+2}\right)^{\frac{1}{2}+\eta}\left(\frac{1}{k+1}\right) \\ &\leq \left(\tfrac{1}{2}+\eta\right)\sigma_k^2 \end{aligned} \tag{4.41}$$

where we used the bound $\eta < \frac{1}{10} < \frac{1}{2}$ and the resulting concavity of $t^{\frac{1}{2}+\eta}$.

Theorem 4.3 The gradient of function $f(x, y)$ is uniformly Lipschitz continuous on $\mathbb{R}^2$.

Proof. Let us consider the functions $\delta(x)$ and $y_{\rm mid}(x)$. Remembering (3.29), (3.30) and (4.41) and using the fact that $t \in [0, 1]$ when $x \in [x_k, x_{k+1}]$, we easily deduce that, for x in this interval,

$$\max\left[|\delta'(x)|, |\delta''(x)|, |y'_{\rm mid}(x)|, |y''_{\rm mid}(x)|\right] \leq 360\left(\tfrac{1}{2}+\eta\right) \stackrel{\rm def}{=} \kappa_{\rm dy}. \tag{4.42}$$

We now turn to the analysis of the second derivatives of $f_2(x, y)$.

- Consider first the case where $y \in [y_{\rm low}(x), y_{\rm up}(x)]$. In this interval, we obtain, for $x \in [x_k, x_{k+1}]$, that

$$\frac{\partial^2 f_2}{\partial y^2}(x, y) = 2. \tag{4.43}$$

Moreover

$$\frac{\partial f_2}{\partial x}(x,y) = -2(y - y_{\rm mid}(x))y'_{\rm mid}(x) \tag{4.44}$$

and thus

$$\frac{\partial^2 f_2}{\partial x^2}(x,y) = 2y'_{\rm mid}(x)^2 - 2(y - y_{\rm mid}(x))y''_{\rm mid}(x).$$

Taking absolute values and noting that, because of the definition of $y_{\rm mid}(x)$ and (4.42),

$$|y - y_{\rm mid}(x)| \le \tfrac{1}{2}\delta(x) \le \tfrac{1}{2}\sigma_k \tag{4.45}$$

for $x \in [x_k, x_{k+1}]$ and $y \in [y_{\rm low}(x), y_{\rm sup}(x)]$, we obtain, for x and y in these intervals, that

$$\left|\frac{\partial^2 f_2}{\partial x^2}(x,y)\right| \le 2\kappa_{\rm dy}^2 + \sigma_k\kappa_{\rm dy} \le 2\kappa_{\rm dy}^2 + \kappa_{\rm dy}, \tag{4.46}$$

where we also used (4.42) and the bound $\sigma_k \le 1$. Finally, for x and y in the same intervals, we have that

$$\left|\frac{\partial^2 f_2}{\partial x\,\partial y}(x,y)\right| = |-2y'_{\rm mid}(x)| \le 2\kappa_{\rm dy}$$

where we used (4.44) and (4.42). Considering this last relation together with (4.43) and (4.46), we thus conclude that the second derivatives of $f_2(x,y)$ are uniformly bounded for all $x \ge 0$ and all $y \in [y_{\rm low}(x), y_{\rm sup}(x)]$.

• The case where $y \in [y_{\rm low}(x)-1, y_{\rm low}(x)] \cup [y_{\rm up}(x), y_{\rm up}(x)+1]$ is covered by Lemma 4.1.

• To conclude our analysis, we are thus left with checking the boundedness of the second derivatives of $f_2(x,y)$ for $y \ge y_{\rm up} + 1$ and $y \le y_{\rm low} - 1$. In these intervals, $f_2(x,y) = 8\,\delta(x)^2$, whose second derivatives are bounded because of (4.42). We may therefore finally assess that the second derivatives of $f_2(x,y)$ are bounded for all $x \ge 0$ and all y. (Figure 4.5 shows the second derivative of $f_2(x,y)$ with respect to y for $x = x_2$.)

We may now combine this last conclusion with (3.19) and (3.36) to deduce that $f(x,y)$ has uniformly bounded second derivatives for all $(x,y) \in \mathbb{R}^2$. The desired Lipschitz continuity of its gradient then follows. □

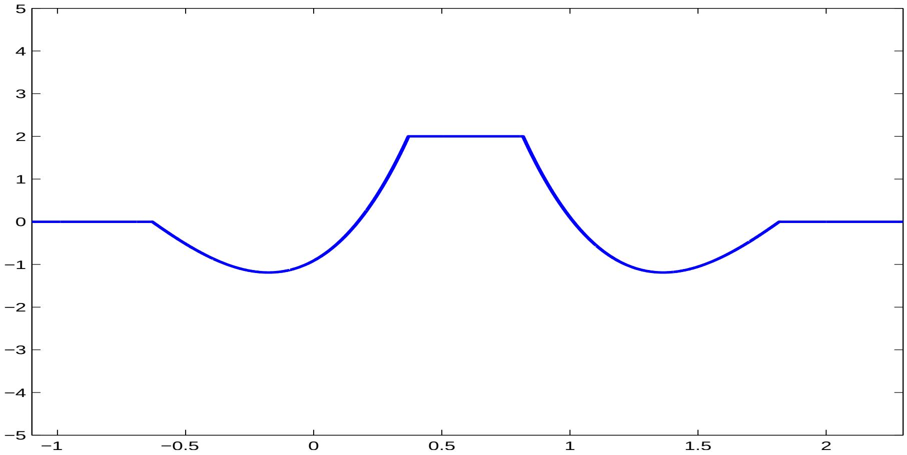

Figure 4.5: The second derivative of $f_2(x_2, y)$, for $\eta = 10^{-5}$.

We conclude the construction of our example by verifying that the sequence of iterates can indeed be obtained from the steepest-descent method with exact linesearches.

Theorem 4.4 The iterate (x_{k+1}, y_{k+1}) is the first minimizer along the steepest descent direction from (x_k, y_k).

Proof. The theorem statement is equivalent to verifying that the slope

$$\begin{aligned}\omega(x) &= \left\langle \nabla f(x, y(x)), \tfrac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ (-1^k) \end{pmatrix} \right\rangle \\ &= \tfrac{1}{2} \left[f_1'(x) + \frac{\partial f_2}{\partial x}(x, y) + (-1)^k \frac{\partial f_2}{\partial y}(x, y) \right]\end{aligned}$$

of $f(x, y)$ on $[x_k, x_{k+1}]$, which is given by

$$\omega(x) = \tfrac{1}{2} f_1'(x) + [y(x) - y_{\rm mid}(x)][(-1)^k - y'_{\rm mid}(x)], \tag{4.47}$$

is always non-positive and is zero only at the iterates (the corners of the trajectory). To prove this property, we first observe that, because of (3.29),

$$(-1)^k - y'_{\rm mid}(x) = (-1)^k \left[1 - |y'_{\rm mid}(x)|\right] \tag{4.48}$$

Observe now that (4.41) and the decreasing nature of σ_k together give that, for $k > 0$,

$$\left| \frac{\sigma_{2k} - \sigma_{2k-1}}{\sigma_{2k-1}} \right| \le (\tfrac{1}{2} + \eta)\sigma_{2k-1} < \tfrac{1}{2} + \eta \le 0.6.$$

where the last inequality follows from the bound $\eta \le \frac{1}{10}$. Hence, recalling (3.29) and using the fact that $\max_{t\in[0,1]} t^2(1-t)^2 = \frac{1}{16}$, we obtain that, for $x \in [x_{2k-1}, x_{2k}]$,

$$|y'_{\rm mid}(x)| < 15 \times 0.6 \max_{t\in[0,1]} t^2(1-t)^2 < 0.57. \tag{4.49}$$

Similarly, (4.41) and the decreasing nature of σ_k imply that, for $k > 0$,

$$\left| \frac{\sigma_{2k+1} - \sigma_{2k}}{\sigma_{2k}} \right| \le (\tfrac{1}{2} + \eta)\sigma_{2k} < \tfrac{1}{2} + \eta \le 0.6$$

while, for $k = 0$,

$$\left| \frac{\sigma_1 - \sigma_0}{\sigma_0} \right| = \left| \left(\frac{1}{2}\right)^{\frac{1}{2}+\eta} - 1 \right| < 0.6.$$

This thus gives that

$$|y'_{\rm mid}(x)| < 15 \times 0.6 \max_{t\in[0,1]} t^2(1-t)^2 < 0.57 \tag{4.50}$$

for $x \in [x_{2k}, x_{2k+1}]$. Combining (4.49) and (4.50), we obtain that $|y'_{\rm mid}(x)| < 1$ for all $x \ge 0$, and therefore, using (4.48), that

$$|(-1)^k - y'_{\rm mid}(x)| \le 1,$$

for all $x \in [x_k, x_{k+1}]$, where the inequality is strict except at x_k and x_{k+1} since $y'_{\rm mid}(x_k) = y'_{\rm mid}(x_{k+1}) = 0$. Hence we obtain, using (4.45), that, for $x \in [x_k, x_{k+1}]$,

$$\big| [y(x) - y_{\rm mid}(x)] \, [(-1)^k - y'_{\rm mid}(x)] \big| \le \tfrac{1}{2}\delta(x). \tag{4.51}$$

Moreover, since, at the leftmost boundary of $[x_{2k-1}, x_{2k}]$,

$$y(x_{2k-1}) - y_{\rm mid}(x_{2k-1}) = \tfrac{1}{2}\delta(x_{2k-1}) = \tfrac{1}{2}\sigma_{2k-1}$$

and, at the leftmost boundary of $[x_{2k}, x_{2k+1}]$,

$$y(x_{2k}) - y_{\rm mid}(x_{2k}) = -\tfrac{1}{2}\delta(x_{2k}) = -\tfrac{1}{2}\sigma_{2k}$$

(where we used (3.31)), we deduce from (4.48) that the inequality in (4.51) can only hold as an equality at x_{k+1}.

Our penultimate step to is note that (3.16) and (3.27) together give that, for $x \in [x_k, x_{k+1}]$ and $t = (x - x_k)/\sigma_k$,

$$f_1'(x) + \delta(x) = (\sigma_k - \sigma_{k+1})\big[-12\,t^2 + 18\,t^3 - 6\,t^5\big] \leq 0, \tag{4.52}$$

where, again, the inequality is strict in the interior of the interval (see Figure 4.6).

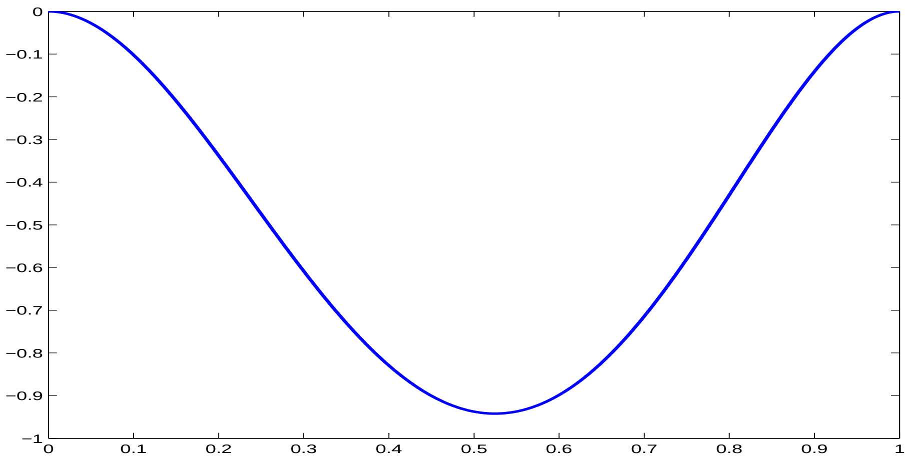

Figure 4.6: The polynomial $-12\,t^2 + 18\,t^3 - 6\,t^5$ on $[0, 1]$.

Combining finally (4.47), (4.51) and (4.52), we obtain that, for all $k \geq 0$,

$$\omega(x) < 0 \;\text{ for }\; x \in [x_k, x_{k+1}), \tag{4.53}$$

and (x_{k+1}, y_{k+1}) is indeed the first local minimizer of $f(x, y)$ along the steepest-descent direction at iterate (x_k, y_k). □

This last theorem is illustrated in Figure 4.7, and completes the construction of our example.

5 Conclusions

We have constructed an example where, for an arbitrary $\tau > 0$, the steepest-descent method with exact linesearches takes at least a multiple of $\epsilon^{-2+\tau}$ iterations to find an approximate stationary point at which $\|g_k\| \leq \epsilon$, for any $\epsilon \in (0, 1)$. This result closes the gap left by Cartis et al. (2010) who could not accomodate this type of linesearch corresponding to the archetypal, if very often impractical, definition of the method. Given that we have shown in this last paper that it is impossible to obtain an $O(\epsilon^{-2})$ worst-case complexity *for all* ϵ, this is probably the best result that can be obtained.

As was the case in this last paper, our example may furthermore be adapted to cover the case where the level sets of the objective are finite by extending $f(x, y)$ beyond the approximate minimizer. This is achieved by smoothly prolongating $f_1(x)$ beyond this point with a suitably increasing function and by, say, keeping the width of the "corridor" constant in this part of the plane. Such an example may therefore be constructed for every $\epsilon \in (0, 1)$.

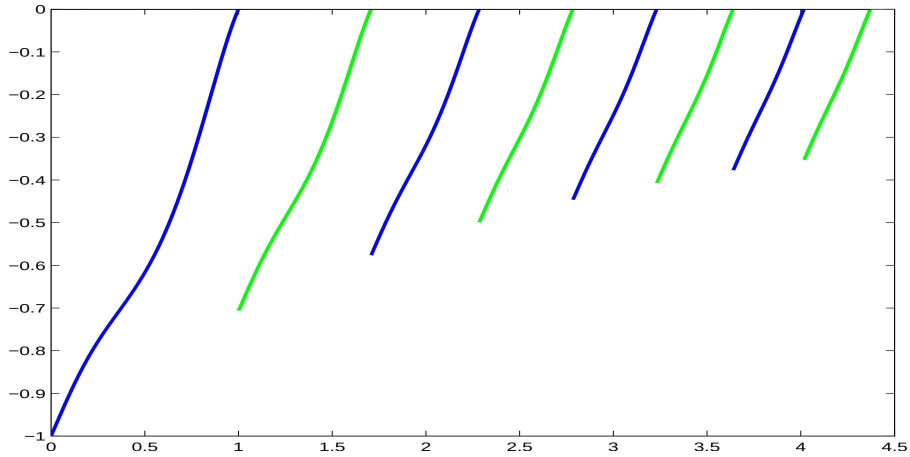

Figure 4.7: The behaviour of $\omega(x)$ for $\eta = 10^{-5}$.

References

W. Bian, X. Chen, and Y. Ye. Complexity analysis of interior point algorithms for non-Lipschitz and nonconvex minimization. *Mathematical Programming, Series A*, (submitted), 2012.

C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization. *SIAM Journal on Optimization*, **20**(6), 2833–2852, 2010.

C. Cartis, N. I. M. Gould, and Ph. L. Toint. Adaptive cubic overestimation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Mathematical Programming, Series A*, **127**(2), 245–295, 2011*a*.

C. Cartis, N. I. M. Gould, and Ph. L. Toint. An adaptive cubic regularisation algorithm for nonconvex optimization with convex constraints and its function-evaluation complexity. *IMA Journal of Numerical Analysis*, (to appear), 2011*b*.

C. Cartis, N. I. M. Gould, and Ph. L. Toint. Optimal Newton-type methods for nonconvex optimization. Technical Report naXys-17-2011, Namur Centre for Complex Systems (naXys), FUNDP-University of Namur, Namur, Belgium, 2011*c*.

C. Cartis, N. I. M. Gould, and Ph. L. Toint. Complexity bounds for second-order optimality in unconstrained optimization. *Journal of Complexity*, **28**, 93–108, 2012*a*.

C. Cartis, N. I. M. Gould, and Ph. L. Toint. A note about the complexity of minimizing Nesterov's smooth Chebyshev-Rosenbrock function. *Optimization Methods and Software*, (to appear), 2012*b*.

C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the oracle complexity of first-order and derivative-free algorithms for smooth nonconvex minimization. *SIAM Journal on Optimization*, **22**(1), 66–86, 2012*c*.

A. Cauchy. Méthode générale pour la résolution des systèmes d'équations simultannées. *Comptes Rendus de l'Académie des Sciences*, pp. 536–538, 1847.

S. Gratton, A. Sartenaer, and Ph. L. Toint. Recursive trust-region methods for multiscale nonlinear optimization. *SIAM Journal on Optimization*, **19**(1), 414–444, 2008.

F. Jarre. On Nesterovs smooth Chebyshev-Rosenbrock function. Technical report, University of Düsseldorf, Düsseldorf, Germany, May 2011.

Yu. Nesterov. *Introductory Lectures on Convex Optimization*. Applied Optimization. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.

Yu. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming, Series A*, **108**(1), 177–205, 2006.

L. N. Vicente. Worst case complexity of direct search. Technical report, Department of Mathematics, University of Coimbra, Coimbra, Portugal, May 2010. Preprint 10-17, revised 2011.