An example of slow convergence for Newton's method on a function with globally Lipschitz continuous Hessian

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

3 May 2013

Abstract

An example is presented where Newton's method for unconstrained minimization is applied to find an ϵ-approximate first-order critical point of a smooth function and takes a multiple of ϵ^{-2} iterations and function evaluations to terminate, which is as many as the steepest-descent method in its worst-case. The novel feature of the proposed example is that the objective function has a globally Lipschitz-continuous Hessian, while a previous example published by the same authors only ensured this critical property along the path of iterates, which is impossible to verify *a priori*.

1 Introduction

Evaluation complexity of nonlinear optimization leads to many surprises, one of the most notable ones being that, in the worst case, steepest-descent and Newton's method are equally slow for unconstrained optimization, thereby suggesting that the use of second-order information as implemented in the standard Newton's method is useless in the worst-case scenario. This counter-intuitive conclusion was presented in Cartis, Gould and Toint (2010), where an example of slow convergence of Newton's method was presented. This example shows that Newton's method, when applied on a sufficiently smooth objective function $f(x)$, may generate a sequence of iterates such that

$$\|\nabla_x f(x_k)\| \geq \left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta},$$

where x_k is the k-th iterate and η is an arbitrarily small positive number. This implies that the considered iteration will not terminate with

$$\|\nabla_x f(x_k)\| \leq \epsilon$$

and $\epsilon \in (0,1)$ for $k < \epsilon^{-2+\tau}$, where $\tau = \eta/(1+2\eta)$. This shows that the evaluation complexity of Newton's method for smooth unconstrained optimization is essentially $O(\epsilon^{-2})$, as is known to be the case for steepest descent (see Nesterov, 2004, pages 26-29). This example satisfies the assumption that the second derivatives of $f(x)$ are Lipschitz continuous *along the path of iterates*, that is, more specifically, that

$$\|\nabla_{xx} f(x_k + t(x_{k+1} - x_k)) - \nabla_{xx} f(x_k)\| \leq L_H \|x_{k+1} - x_k\|$$

for all $t \in (0,1)$ and some $L_H \geq 0$ independent of k. The gradient of this example is globally Lipschitz continuous.

While being formally adequate (and verified by the example if Cartis et al., 2010), this assumption has two significant drawbacks. The first is that it cannot be verified *a priori*, before the minimization algorithm is applied to the problem at hand. The second is that it has to be made for an infinite sequence

*School of Mathematics, University of Edinburgh, The King's Buildings, Edinburgh, EH9 3JZ, Scotland, UK. Email: coralia.cartis@ed.ac.uk. This author's work is supported by the EPSRC Grant EP/I028854/1.

†Numerical Analysis Group, Rutherford Appleton Laboratory, Chilton, OX11 0QX, England (UK). Email: nick.gould@stfc.ac.uk.

‡Namur Center for Complex Systems (naXys) and Department of Mathematics, University of Namur, 61, rue de Bruxelles, B-5000 Namur, Belgium. Email: philippe.toint@unamur.be.

of iterates, at least if a result is sought which is valid for all $\epsilon \in (0,1)$. It is therefore desirable to verify if the stronger but simpler assumption that $f(x)$ admits globally Lipschitz continuous second derivatives still allows the construction of an example with $O(\epsilon^{-2})$ evaluation complexity. It is the purpose of the present note to discuss such an example.

2 The example

Consider the unconstrained nonlinear minimization problem given by

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

where f is a twice continuously differentiable function from $\mathbb{R}^n$ into $\mathbb{R}$, which we assume is bounded below. The standard Newton' method for solving this problem is outlined as Algorithm 2.1, where we use the notation

$$g_k \stackrel{\rm def}{=} \nabla_x f(x_k) \quad \text{and} \quad H_k \stackrel{\rm def}{=} \nabla_{xx} f(x_k).$$

Algorithm 2.1: Newton's method for unconstrained minimization

Step 0: Initialization. A starting point $x_0 \in \mathbb{R}^n$ is given. Compute $f(x_0)$ and g_0. Set $k = 0$.

Step 1: Check for termination. If $\|g_k\| \leq \epsilon$, terminate with x_k as approximate first-order critical point.

Step 2: Step computation. Compute H_k and the step s_k as the solution of the linear system

$$H_k s_k = -g_k \tag{2.1}$$

Step 3: Accept the next iterate. Set $x_{k+1} = x_k + s_k$, increment k by one and go to Step 1.

Of course, the method as described in this outline makes strong (favourable) assumptions: it assumes that the matrix H_k is positive definite for each k and also that $f(x_k + s_k)$ sufficiently reduces $f(x_k)$ for being accepted without any specific globalization procedure, such as linesearch, trust-region or filter. However, since our example will ensure both these properties, the above description is sufficient for our purpose.

2.1 A putative iteration

Our example is two-dimensional. We therefore aim at building a function $f(x,y)$ from $\mathbb{R}^2$ into $\mathbb{R}$ such that, for any $\epsilon \in (0,1)$, Newton's method essentially takes ϵ^{-2} iterations to find an approximate first-order critical point (x_ϵ, y_ϵ) such that

$$\|\nabla_{(x,y)} f(x_\epsilon, y_\epsilon)\| \leq \epsilon \tag{2.2}$$

when started with $(x_0, y_0) = (0,0)$.

Consider $\tau \in (0,1)$ and an hypothetical sequence of iterates $\{(x_k, y_k)\}_{k=0}^{\infty}$ such that g_k, H_k and $f_k = f(x_k, y_k)$ are defined by the relations

$$g_k = -\begin{pmatrix} \left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta} \\ \left(\frac{1}{k+1}\right)^2 \end{pmatrix} \qquad H_k = \begin{pmatrix} 1 & 0 \\ 0 & \left(\frac{1}{k+1}\right)^2 \end{pmatrix}, \tag{2.3}$$

for $k \geq 0$ and

$$f_0 = \zeta(1+2\eta) + \frac{\pi^2}{6}, \qquad f_k = f_{k-1} - \frac{1}{2}\left[\left(\frac{1}{k+1}\right)^{1+2\eta} + \left(\frac{1}{k+1}\right)^2\right] \quad \text{for } k \geq 1, \tag{2.4}$$

where

$$\eta = \eta(\tau) \stackrel{\rm def}{=} \frac{\tau}{4-2\tau} = \frac{1}{2-\tau} - \frac{1}{2}.$$

If this sequence of iterates can be generated by Newton's method starting from the origin and applied on a twice-continuously differentiable function with globally Lipschitz continuous Hessian, then one may check that

$$\|g_k\| > \left|\frac{\partial f}{\partial x}(x_k, y_k)\right| = \left(\frac{1}{k+1}\right)^{\frac{1}{2-\tau}} > \epsilon \quad \text{for} \quad k \leq \epsilon^{-2+\tau},$$

and also that

$$\|g_k\| \leq \left(\frac{1}{k+1}\right)^{\frac{1}{2-\tau}} + \left(\frac{1}{k+1}\right)^2 \leq 2\left(\frac{1}{k+1}\right)^{\frac{1}{2-\tau}} \leq \epsilon \quad \text{for} \quad k \leq 4\epsilon^{-2+\tau}.$$

As a consequence, the algorithm will stop for k in a fixed multiple of ϵ^{-2} iterations.

2.2 A well-defined Newton scheme

The first step in our construction is to note that, since we consider Newton's method, the step s_k at iteration k is defined by the relation (2.1), which, together with (2.3), yields that

$$s_k = \begin{pmatrix} s_{k,x} \\ \\ s_{k,y} \end{pmatrix} = \begin{pmatrix} \left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta} \\ \\ 1 \end{pmatrix}, \tag{2.5}$$

and therefore that

$$x_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad x_k = \begin{pmatrix} \displaystyle\sum_{j=0}^{k-1} \left(\frac{1}{j+1}\right)^{\frac{1}{2}+\eta} \\ \\ k \end{pmatrix}. \tag{2.6}$$

The predicted value of the quadratic model at iteration k

$$q_k(x_k + s_x, y_k + s_y) \stackrel{\rm def}{=} f_k + \langle g_k, s\rangle + \tfrac{1}{2}\langle s, H_k s\rangle \tag{2.7}$$

at $(x_{k+1}, y_{k+1}) = (x_k + s_{k,x}, y_k + s_{k,y})$ is therefore given by

$$q_k(x_{k+1}, y_{k+1}) = f_k + \langle g_k, s_k\rangle + \tfrac{1}{2}\langle s_k, H_k s_k\rangle = f_k - \frac{1}{2}\left[\left(\frac{1}{k+1}\right)^{1+2\eta} + \left(\frac{1}{k+1}\right)^2\right] = f_{k+1}, \tag{2.8}$$

where the last equality results from (2.4). Thus the value of the k-th local quadratic at the trial point exactly corresponds to that planned for the objective function itself at the next iterate of our putative sequence. As a consequence, the sequence defined by (2.6) can be obtained from a well-defined Newton iteration where the local quadratic is minimized at every iteration yielding sufficient decrease, provided we can find a sufficiently smooth function f interpolating (2.3) at the points (2.6). Also note that the sequence $\{f_k\}_{k=0}^{\infty}$ is bounded below by zero due to the definition of the Riemann zeta function, which is finite since $1 + 2\eta > 1$. Figure 2.1 illustrates the sequence of quadratic models and the path of iterates for increasing k.

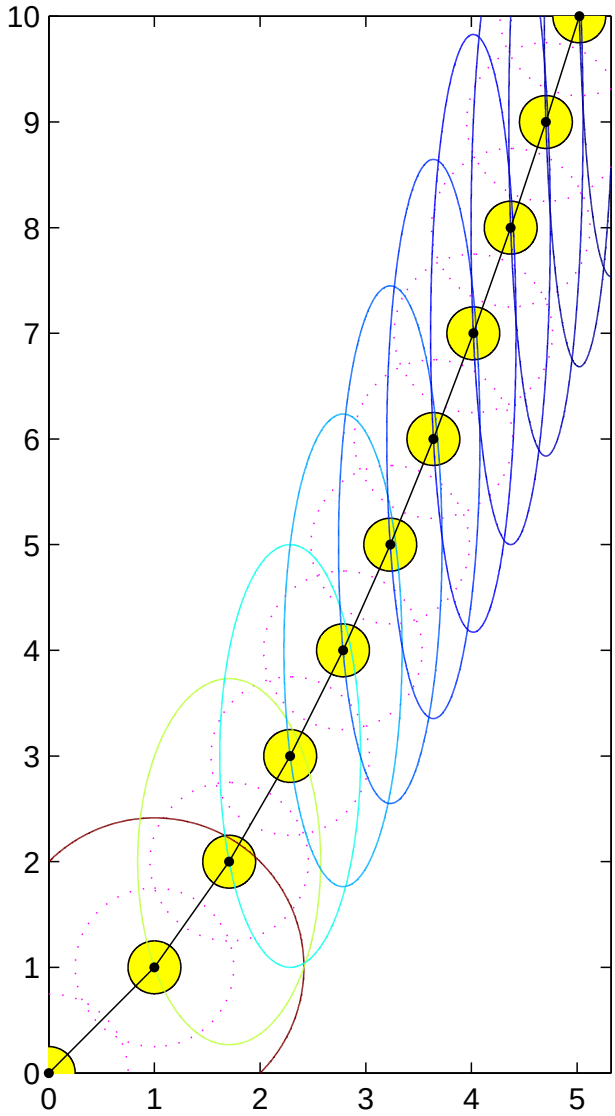

Figure 2.1: Contour line of $q_k(x,y) = f_k$ (full lines), the disks around the iterates of radius $\frac{1}{4}$ (shaded) and $\frac{3}{4}$ (dotted), and the path of iterates for $k = 0, \ldots 10$.

2.3 Building the objective function

The next step in our example construction is thus to build a smooth function with bounded second and third derivatives (which implies that its gradient and Hessian are Lipschitz continuous) interpolating (2.3) at (2.6). The idea is to exploit the large components of the step along the second dimension (see (2.5)) to define $f(x,y)$ to be identical to $q_k(x,y)$ in a domain around (x_k, y_k). More specifically, define, for $k \geq 1$,

$$\delta(\alpha) \stackrel{\rm def}{=} \begin{cases} 1 & \text{if } 0 \leq \alpha \leq \frac{1}{4}, \\ 16\alpha^3 \left[-5 + 15\alpha - 12\alpha^2\right] & \text{if } \frac{1}{4} \leq \alpha \leq \frac{3}{4}, \\ 0 & \text{if } \alpha \geq \frac{3}{4}. \end{cases} \tag{2.9}$$

This piecewise polynomial is defined[(1)] to be identically equal to 1 near the origin, and to smoothly decrease to zero (with bounded first, second and third derivatives) between $\frac{1}{4}$ and $\frac{3}{4}$. Using this function, we may then define, for each $k \geq 0$, a *local support* function

$$s_k(x,y) \stackrel{\rm def}{=} \delta\left(\left\| \begin{array}{c} x - x_k \\ y - y_k \end{array} \right\| \right).$$

which is identically equal to 1 in a (spherical) neighbourhood of (x_k, y_k) and decreases smoothly (with bounded first, second and third derivatives) to zero for all points whose distance to (x_k, y_k) exceeds $\frac{3}{4}$. The shapes of a support function at $(1, 1)$ is shown in Figure 2.2.

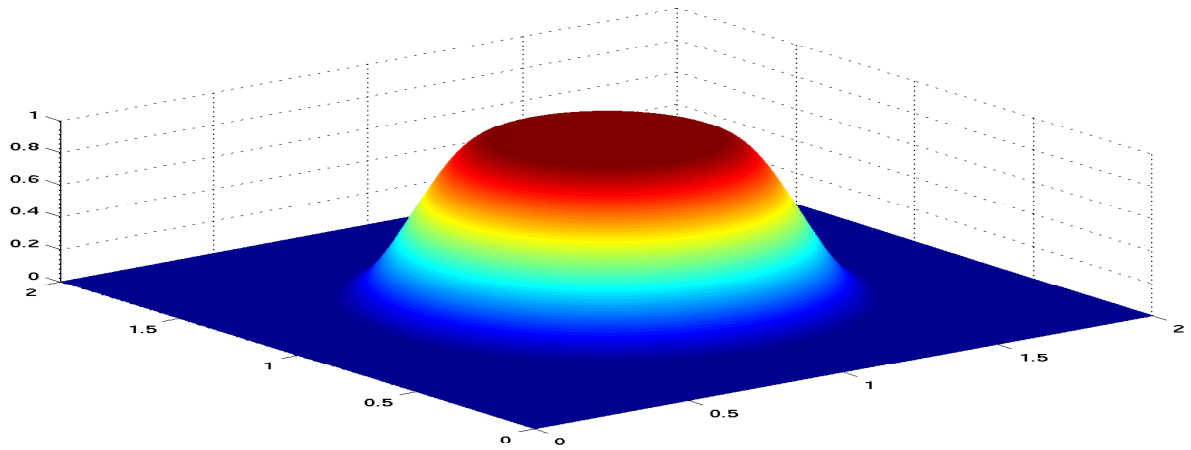

Figure 2.2: The shape of the support function centered at $(1, 1)$

We are now in position to define the 'useful part' of objective function as

$$f_{SN1}(x,y) = \sum_{k=0}^{\infty} s_k(x,y) q_k(x,y) \tag{2.10}$$

for all (x, y) in $\mathbb{R}^2$. Note that the infinite sum in (2.10) is obviously convergent everywhere as it involves at most two nonzero terms for each (x, y), because the distance between successive iterates exceeds 1. This function would already serve our purpose, as it obviously interpolates (2.3)-(2.4), is bounded below and has bounded first, second and third derivatives since the large values of $q_k(x, y)$ that occur in their expressions always occur far from (x_k, y_k) and are thus annihilated by the support function.

However, for the sake of illustration, we modify $f_{SN1}(x, y)$ to build

$$f_{SN}(x,y) = f_{SN1}(x,y) + \left[1 - \sum_{k=0}^{\infty} s_k(x,y) \right] f_{BCK}(x,y), \tag{2.11}$$

where the background function $f_{BCK}(x, y)$ only depends on y (i.e., $\nabla_x f_{BCK}(x, y) = 0$ for all x) and ensures that $f_{BCK}(x, y_k) = f_k$ for all x such that $|x - x_k| \geq \frac{3}{4}$. Its value is computed, for $y \geq 0$, by successive Hermite interpolation of the conditions (2.4) and

$$\nabla_y f_{BCK}(x,y) = \left(\frac{1}{k+1} \right)^{\frac{1}{2}+\eta}, \quad \nabla_{yy} f_{BCK}(x,y) = 0,$$

at $y_k = 0, 1, \ldots$. The shape of the resulting function as a function of y and of its third derivative are shown in Figures 2.3 and 2.4, respectively. It may clearly be extended to the complete real axis without altering its smoothness properties.

[(1)] Using Hermite interpolation, with boundary conditions

$$\delta(\tfrac{1}{4}) = 1, \quad \delta'(\tfrac{1}{4}) = 0, \quad \delta''(\tfrac{1}{4}) = 0, \quad \delta(\tfrac{3}{4}) = 0, \quad \delta'(\tfrac{3}{4}) = 0, \quad \delta''(\tfrac{3}{4}) = 0.$$

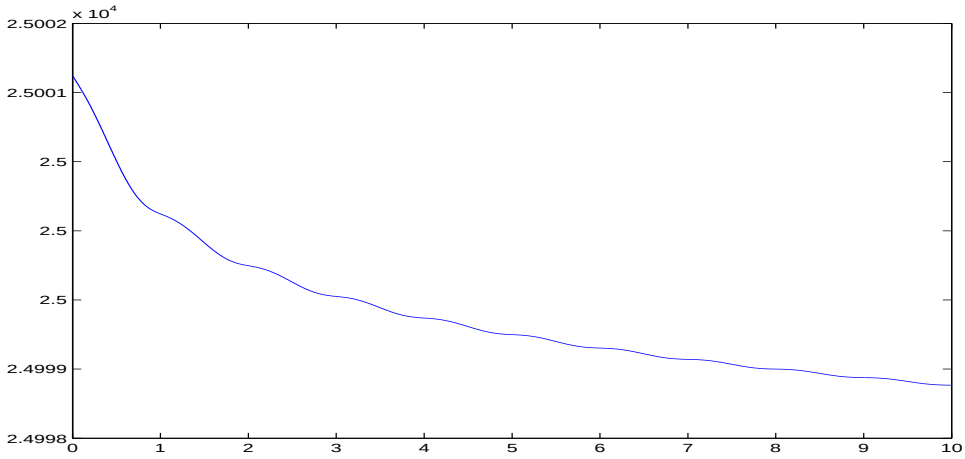

Figure 2.3: The shape of $f_{BCK}(x, y)$ as a function of y $(k = 0, \ldots, 10)$

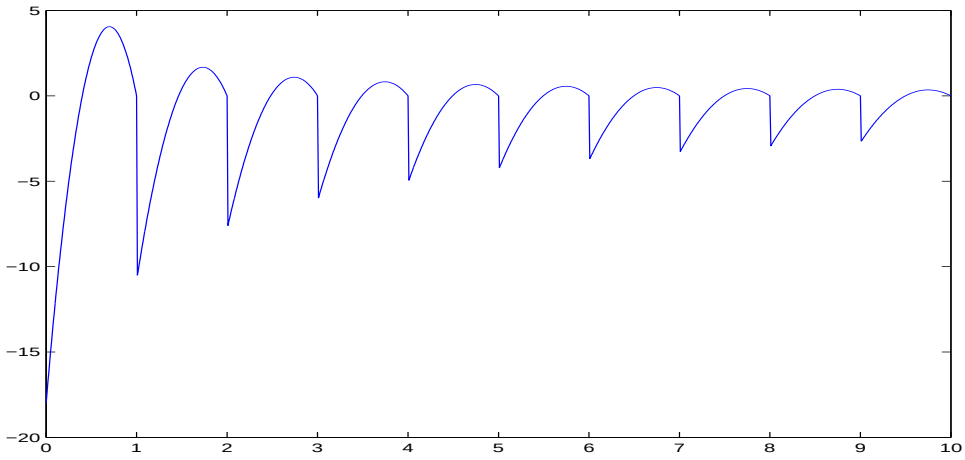

Figure 2.4: The shape of the third derivative of $f_{BCK}(x, y)$ as a function of y $(k = 0, \ldots, 10)$

The modification (2.11) has the effect of 'filling the lansdscape' around the path of iterates, considerably diminishing the variations of the function in the domain of interest. The countour lines of $f_{SN}(x, y)$ as given by (2.11) are shown in Figure 2.5 on the following page, together with the iteration path. A perspective view is provided in Figure 2.6 on page 8.

3 Conclusions and perspectives

We have produced a two dimensional example where the standard Newton's method is well-defined and converge slowly, in the sense that an ϵ-approximate first-order critical point is not found in less than ϵ^{-2} iterations, an evaluation complexity identical to that of the steepest-descent method. The objective function in this example has globally Lipschitz-continuous second derivatives, showing that this slowly convergent behaviour may still be observed under stronger but simpler assumptions than those used in Cartis et al. (2010). It is interesting to note that this example also applies if a trust-region globalization is used in conjunction with Newton's method, since, because of (2.8), all iterates are very successful and the iteration sequence is thus identical to that analyzed here whenever the initial trust-region radius Δ_0 is chosen larger than $\|s_0\| = \sqrt{2}$.

We also observe that the construction used above can be extended to produce examples with smooth functions interpolating general iteration conditions (such as (2.3)-(2.4)) provided the successive iterates remain uniformly bounded away from each other. Producing such a sequence of iterates from an original

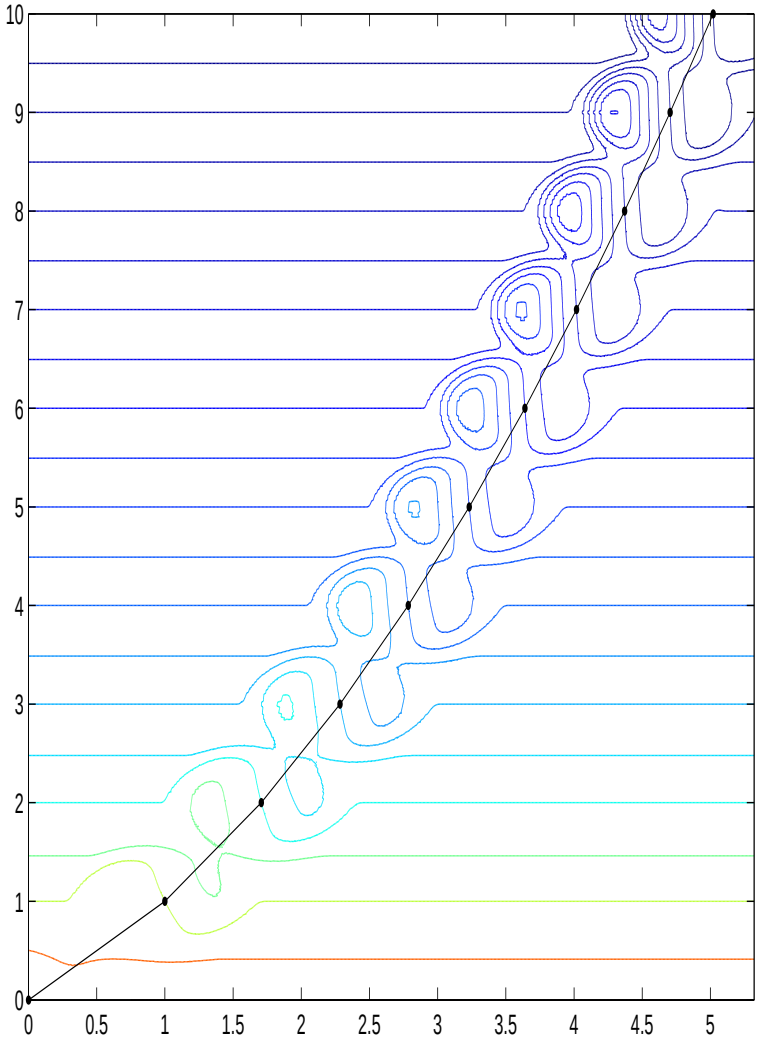

Figure 2.5: Contour lines of $f_{SN}(x, y)$ and the path of iterates for $k = 0, \ldots 10$.

slowly-converging one-dimensional sequence whose iterates asymptotically coalesce can be achieved, as is the case here, by extending the dimensionality of the example.

Acknowledgements

The third author is indebted to S. Bellavia; B. Morini and the University of Florence (Italy) for their kind support under the Azione 2 "Permanenza presso le unita' administrativa di studiosi stranieri di chiara fama" during a visit where this note was finalized.

References

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.

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

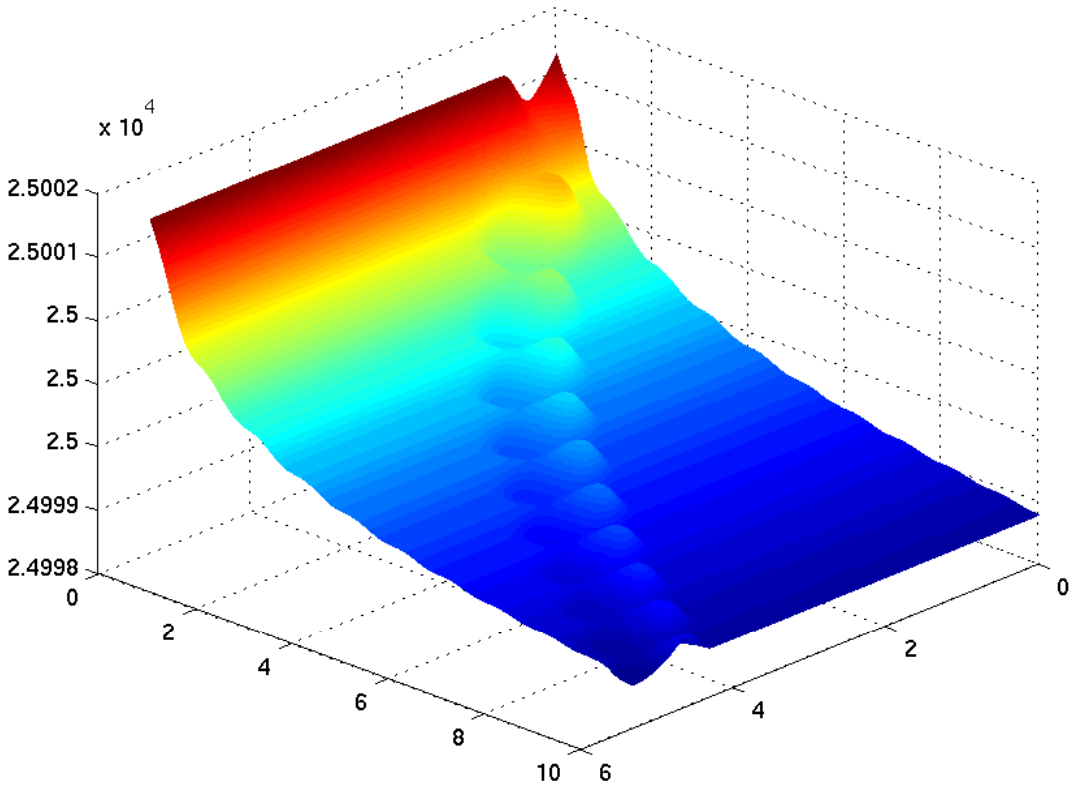

Figure 2.6: A view of $f_{SN}(x, y)$