Available online at www.sciencedirect.com
SCIENCE DIRECT®
European Journal of Operational Research 169 (2006) 1096–1107
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Huber approximation for the non-linear ℓ_1 problem

Mustafa Ç. Pınar [a,*], Wolfgang M. Hartmann [b]

[a] *Department of Industrial Engineering, Bilkent University, 06533 Ankara, Turkey*
[b] *SAS Institute, Heidelberg, Germany*

Received 1 September 2003; accepted 18 October 2004
Available online 13 May 2005

Abstract

The smooth Huber approximation to the non-linear ℓ_1 problem was proposed by Tishler and Zang (1982), and further developed in Yang (1995). In the present paper, we use the ideas of Gould (1989) to give a new algorithm with rate of convergence results for the smooth Huber approximation. Results of computational tests are reported.
© 2005 Elsevier B.V. All rights reserved.

Keywords: Nonlinear programming; Non-differentiable optimization; Smoothing algorithms; Huber M-estimator

1. Introduction

In this paper we investigate a new algorithm for the non-linear ℓ_1 estimation problem, also known as the absolute deviations curve fitting problem in statistics. Let $c_i : \Re^n \mapsto \Re$ be at least twice continuously differentiable functions for each $i = 1, \ldots, m$. We want to find a minimizing point for the following function:

$$f(x) \equiv \sum_{i=1}^{m} |c_i(x)|. \tag{1}$$

From a statistical point of view, it is well known that the properties of the estimated parameters, i.e., optimal values of x, highly depend upon the underlying distribution of the error terms in the model. Basset and Koenker (1978) proved that the estimator based on the ℓ_1 problem above (a minimizing point of f) is a consistent and asymptotically normal estimator. They also discussed conditions under which the ℓ_1 estimator is superior to the least squares estimator. Since the ℓ_1 estimator does not square the contribution of

* Corresponding author. Tel.: +90 312 290 1514; fax: +90 312 266 4054.
E-mail address: mustafap@bilkent.edu.tr (M.Ç. Pınar).

0377-2217/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2004.10.029

errors, it may be less influenced by the presence of outliers in the data as opposed to the least squares estimator. Tishler and Zang (1982) observed that when measurement errors are Cauchy distributed the ℓ_1 solution yields more reliable estimates than the non-linear least squares problem.

From a computational point of view, the non-linear ℓ_1 estimation problem presents a major difficulty: its objective function is not continuously differentiable. Several algorithms have been proposed for solving the problem over the past three decades. Gonin and Money (1989) offer a classification of these algorithms into four categories:

1. *Gauss–Newton or Levenberg–Marquardt type algorithms.* These algorithms use first derivative information only and reduce the non-linear problem into a sequence of linear ℓ_1 estimation problems. Examples of this class of algorithms can be found in Osborne and Watson (1971), Anderson and Osborne (1977a,b), and McLean and Watson (1980).
2. *SQP type methods.* These algorithms utilize a sequence of quadratic programming (QP) subproblems along with an active set strategy. They incorporate second order information into the objective function of QP subproblems. Examples of this class are algorithms proposed by Murray and Overton (1981), Bartels and Conn (1982), and Overton (1982).
3. *Two phase or hybrid methods.* These algorithms aim at identifying the optimal active set in the first phase of the algorithm. With the active set identified the algorithm proceeds to the second phase where a system of non-linear equations is solved using a method with fast local convergence properties, e.g., Newton's method or a quasi-Newton method. Representatives of this type of algorithms are given by McLean and Watson (1980) and Hald and Madsen (1985).
4. *Smoothing or approximation algorithms.* These methods approximate the non-differentiable objective function by a differentiable function amenable to minimization by first- or second-order methods depending on the approximation. These methods, although not presented as such in the original sources, have a path-following flavor as well; see El-Attar et al. (1979), and Tishler and Zang (1982) for two different algorithmic contributions to this area. Ben-Tal and Teboulle (1989) derive smoothing functions for non-differentiable optimization problems including the ℓ_1 problems. Ben-Tal et al. (1991) applied El-Attar et al. function to engineering problems in plasticity. El-Attar et al. function is known as the hyperboloid approximation in location literature; see Andersen (1996).

The method given in the present paper is akin to the algorithm of Tishler and Zang (1982) and to that of Yang (1995). It uses an approximation function known as Huber's M-estimator function in the field of robust statistics. The method is similar to the successful method for the linear ℓ_1 problem developed by Madsen and Nielsen (1993) and Madsen et al. (1996). However, the proposed algorithm presents many theoretical and computational departures from the Tishler–Zang, Yang, and Madsen et al. cases:

- Unlike Tishler–Zang, Yang, and Madsen et al. it uses a sequence of inexactly minimized subproblems which are solved more and more accurately as the approximation becomes more accurate.
- Unlike Tishler–Zang and Yang method, it uses an extrapolation procedure which enables the two-step superlinear convergence property under a strict complementarity assumption.
- It uses second-order information effectively in that Newton's method coupled with a line search is employed to solve the Huber subproblems.
- Although it is the third contribution on the Huber approximation of the non-linear ℓ_1 function, our paper is the first to give rate of convergence results for the resulting algorithm.

The proposed algorithm is essentially an adaptation of a quadratic penalty function algorithm proposed by Gould (1989) to solve non-linear programming problems with equality constraints. The main contribution of the present paper is to use Gould's ideas in the context of an approximation algorithm for the

non-linear ℓ_1 estimation problem. We note that Dussault (1995) proposed a similar algorithm for variational inequality problems. Dussault (1998) extends these results to augmented Lagrangian-like penalty methods. However, he does not give computational results in his papers.

In the next two sections (Sections 2 and 3) we describe the proposed algorithm, and we give convergence and rate of convergence results. Section 4 is devoted to a summary of the numerical results. Unlike the previous contribution by Yang (1995) which does not give numerical results, we report the results of a careful implementation, and comparison with competing software.

2. The proposed algorithm

As the problem is non-differentiable at points where the functions c_i have zero value (although c_i's are smooth themselves) we propose an approximation technique which will replace the original problem by

$$\Phi(x) = \sum_{i=1}^{m} \phi(c_i(x)), \tag{2}$$

where

$$\phi(c_i(x)) = \begin{cases} \frac{c_i(x)^2}{2\mu}, & \text{if } |c_i(x)| \leqslant \mu, \\ |c_i(x)| - \mu/2, & \text{if } |c_i(x)| > \mu \end{cases} \tag{3}$$

for a positive scalar μ. The above function was proposed by Huber (1981) as a robust estimator when the measurement error distribution deviated from normality. We use the function as a smoothing approximation to the ℓ_1 function as in Madsen and Nielsen (1993). It is easy to verify that ϕ is a once continuously differentiable function of its argument, and that the following properties hold:

$$\lim_{\mu \to 0} \phi(t) = |t|$$

for scalar t, with

$$\lim_{\mu \to 0} \Phi(x) = f(x).$$

Therefore, when μ approaches zero, we get arbitrarily close to the true non-differentiable ℓ_1 function.

Before stating the algorithm we will give some definitions. Let $A(x,\mu) = \{i \mid |c_i(x)| \leqslant \mu\}$ represent the active set at (x,μ) and $A^c(x,\mu)$ its complement with respect to the index set $\{1,\ldots,m\}$. $\nabla c_A(x)$ denotes a matrix with columns $\nabla c_i(x)$ where $i \in A(x,\mu)$. The Lagrange multiplier estimates $\bar{\lambda}_i$, so-called as they are reminiscent of Lagrange multipliers in the Karush–Kuhn–Tucker (KKT) optimality conditions (8) below, are defined for all $i \in A(x,\mu)$ as

$$\bar{\lambda}_i = \bar{\lambda}_i(x,\mu) = \frac{c_i(x)}{\mu}. \tag{4}$$

Let $\bar{g}$ given below represent the gradient of the function $\Phi(x)$. The expression for $\bar{g}$ is given as

$$\bar{g}(x,\bar{\lambda}) = \sum_{i \in A^c(x,\mu)} \operatorname{sgn}(c_i(x)) \nabla c_i(x) + \sum_{i \in A(x,\mu)} \bar{\lambda}_i \nabla c_i(x). \tag{5}$$

We define the quantity $\overline{G}$ (derivative of $\bar{g}$ with respect to x while keeping $\bar{\lambda}$ fixed) as

$$\overline{G}(x,\bar{\lambda}) = \sum_{i \in A^c(x,\mu)} \operatorname{sgn}(c_i(x)) \nabla^2 c_i(x) + \sum_{i \in A(x,\mu)} \bar{\lambda}_i \nabla^2 c_i(x) \tag{6}$$

and the $(n+m) \times (n+m)$ matrix

$$K(x,\bar{\lambda},\mu)=\begin{bmatrix}\overline{G}(x,\bar{\lambda}) & \nabla c_A(x)^{\mathrm{T}}\\ \nabla c_A(x) & -\mu I\end{bmatrix}. \tag{7}$$

We say that x^* is a KKT point (first-order stationary point; see p. 43 of Madsen, 1985) if there exist multipliers λ_i^* such that $-1\leqslant\lambda_i^*\leqslant 1$ and

$$\sum_{i\in A^c(x^*)}\operatorname{sgn}(c_i(x^*))\nabla c_i(x^*)+\sum_{i\in A(x^*)}\lambda_i^*\nabla c_i(x^*)=0, \tag{8}$$

where $A(x^*)=\{i\mid c_i(x^*)=0\}$.

Now, the algorithm is the following:

Algorithm

Step 0. Let an initial point $x^{(0)}$ be given. Set the positive constants $\gamma,\tau,\beta_1,\beta_2,\epsilon,\mu^{(0)}$ and $\mu_{\min}$ as $\beta_1<0.5$, $\beta_1<\beta_2<1$, $\epsilon\ll 1$ and $\mu_{\min}\ll 1$. Let $k=0$ and $x^{(0,0)}=x^{(0)}$.

Step 1. Inner Iteration:

Step 1.0. Let $\bar{\lambda}^{(k,0)}=\bar{\lambda}(x^{(k,0)},\mu^{(k)})$. Compute $\bar{g}(x^{(k,0)},\bar{\lambda}^{(k,0)})$, $\overline{G}(x^{(k,0)},\bar{\lambda}^{(k,0)})$ and $K(x^{(k,0)},\bar{\lambda}^{(k,0)},\mu^{(k)})$. Let $\ell=0$.

Step 1.1. If

$$\|\bar{g}(x^{(k,\ell)},\bar{\lambda}^{(k,\ell)})\|_2\leqslant\gamma\mu^{(k)} \tag{9}$$

then

$$x^{*(k)}=x^{(k,\ell)}\quad\text{and}\quad\lambda^{*(k)}=\bar{\lambda}^{(k,\ell)} \tag{10}$$

and continue from Step 2.

Step 1.2. Find $p^{(k,\ell)}$ that satisfies the *descent condition*:

$$-\bar{g}(x^{(k,\ell)},\bar{\lambda}^{(k,\ell)})^{\mathrm{T}}p^{(k,\ell)}\geqslant\epsilon\mu^{(k)}\|\bar{g}(x^{(k,\ell)},\bar{\lambda}^{(k,\ell)})\|_2\|p^{(k,\ell)}\|_2, \tag{11}$$

i.e., if $K(x^{(k,\ell)},\bar{\lambda}^{(k,\ell)},\mu^{(k)})$ satisfies the second-order conditions (i.e., it is non-singular and it has precisely m negative eigenvalues, the rest of the eigenvalues are positive; see Gould, 1986) then, compute $p^{(k,\ell)}$ for the descent condition (11) as a Newton direction from the system below:

$$\begin{bmatrix}\overline{G}(x^{(k,\ell)},\bar{\lambda}^{(k,\ell)}) & \nabla c_A(x^{(k,\ell)})^{\mathrm{T}}\\ \nabla c_A(x^{(k,\ell)}) & -\mu^{(k)}I\end{bmatrix}\begin{pmatrix}p^{(k,\ell)}\\ r^{(k,\ell)}\end{pmatrix}=-\begin{pmatrix}\bar{g}(x^{(k,\ell)},\bar{\lambda}^{(k,\ell)})\\ 0\end{pmatrix}. \tag{12}$$

Otherwise, use Remark 2.

Step 1.3. Find a stepsize $\alpha^{(k,\ell)}$ that satisfies Armijo–Goldstein sufficient descent and curvature conditions

$$\Phi(x^{(k,l)}+\alpha^{(k,l)},\mu^{(k)})\leqslant\Phi(x^{(k,l)},\mu^{(k)})+\beta_1\alpha^{(k,l)}\bar{g}(x^{(k,l)},\bar{\lambda}^{(k,l)})^{\mathrm{T}}p^{(k,l)}, \tag{13}$$

$$\bar{g}(x^{(k,l)}+\alpha^{(k,l)}p^{(k,l)},\bar{\lambda}(x^{(k,l)}\alpha^{(k,l)}p^{(k,l)}))^{\mathrm{T}}p^{(k,l)}\geqslant\beta_2\bar{g}(x^{(k,l)},\bar{\lambda}^{(k,l)})^{\mathrm{T}}p^{(k,l)}. \tag{14}$$

If $p^{(k,\ell)}$ is indeed a Newton direction then always try first $\alpha^{(k,\ell)}=1$, i.e., try a full Newton step first.

Step 1.4. Move:

$$x^{(k,\ell+1)}=x^{(k,\ell)}+\alpha^{(k,\ell)}p^{(k,\ell)}$$

and let $\ell\leftarrow\ell+1$. Go to Step 1.1.

Step 2. If $\mu^{(k)} < \mu_{\min}$ then stop with the iterate $x^{*(k)}$ as an approximate solution. Otherwise, $\mu^{(k+1)}$ is set according to $0 < \mu^{(k+1)} < \mu^{(k)}$.

Step 3. If $K(x^{*(k)}, \lambda^{*(k)}, \mu^{(k)})$ satisfies the second-order condition (i.e., it is invertible and has precisely m negative eigenvalues) compute $p^{(k)}$ from the linear system of equations below:

$$\begin{bmatrix} \overline{G}(x^{*(k)}, \lambda^{*(k)}) & \nabla c_A(x^{*(k)})^{\mathrm{T}} \\ \nabla c_A(x^{*(k)}) & -\mu^{(k)} I \end{bmatrix} \begin{pmatrix} p^{(k)} \\ r^{(k)} \end{pmatrix} = -\begin{pmatrix} \bar{g}(x^{*(k)}, \lambda^{*(k)}) \\ c_A(x^{*(k)}) - \mu^{(k+1)} \lambda^{*(k)} \end{pmatrix} \tag{15}$$

and let

$$x_a^{*(k)} = x^{*(k)} + p^{(k)}. \tag{16}$$

If

$$\|\bar{g}(x_a^{*(k)}, \bar{\lambda}(x_a^{*(k)}, \mu^{(k+1)}))\|_2 \leqslant \max\{\tau, \|\bar{g}(x^{*(k)}, \bar{\lambda}(x^{*(k)}, \mu^{(k+1)}))\|_2\} \tag{17}$$

then

$$x^{(k+1,0)} = x_a^{*(k)}. \tag{18}$$

Otherwise, set $x^{(k+1,0)} = x^{*(k)}$; $k \leftarrow k+1$ go back to Step 1.

Some remarks concerning the algorithm are in order here.

Remark 1. In Step 1.1 we require only an inexact stationary point of the Huber approximation function. However, as γ becomes smaller, the accuracy becomes more stringent.

Remark 2. In Step 1.2 when the matrix K does not satisfy the second-order condition (i.e., is not invertible or fails to have precisely m negative eigenvalues) then we may use a direction of negative curvature (donc) or a direction of linear infinite descent (dolit), depending on which is applicable, (see Gould, 1986), as long as (11) is satisfied.

Remark 3. Note that Step 3 is an extrapolation procedure which applies a Newton step at the stationary point conditions of the Huber function using the reduced value of μ. However, it uses the previous value of μ so that the matrix K is available from Step 1.4 of the previous inner iteration.

3. Convergence and rate of convergence

In this section we give convergence and rate of convergence results for the algorithm of the previous section. The results follow along the lines of Gould (1989). Therefore, we omit the proofs whenever they are obtained, mutatis mutandis, by verbatim repetition of Gould's results. We point out the corresponding result of Gould (1989) for the interested reader's convenience.

Under a strict complementarity assumption, the algorithm is shown to converge in a locally two-step superlinearly convergent manner. The two-step superlinear convergence hinges on Step 3 in the following way:

- First, we can show using Gould's results that the sequence $\{\mu^{(k)}\}$ can be set as a superlinearly convergent sequence. This follows from the observation that eventually, the starting point of an inner iteration is always obtained from the linear system at Step 3.

- Second, eventually either this starting point of Step 3 or the first inner iterate obtained from it at Step 1.4 (which is ultimately a full Newton iterate with a step size of unity) satisfies the inner stopping criteria. Therefore, the iterates inherit the superlinear behavior of μ eventually but in a two-step fashion.

For the analysis, we will assume that $\mu_{\min} = 0$. The first global convergence result is stated under the following assumptions:

A1 All iterates x generated by the algorithm stay in a bounded domain Ω.
A2 The sequence $\{\mu^{(k)}\}$ goes to zero as k goes to infinity.
A3 At every limit point x^* of the sequence $\{x^{*(k)}\}$, and the corresponding limit point λ^* of the sequence $\{\lambda^{*(k)}\}$ (it is proved below in Theorem 1 that whenever $\{x^{*(k)}\}$ has a limit point, the sequence $\{\lambda^{*(k)}\}$ has a limit point), strict complementarity holds. That is, for $c_i(x^*) = 0$ one has $|\lambda_i^*| < 1$.

Assumption A3 implies that $\nabla c_A(x^*)$ is of full rank and that $|A(x^*)| \leqslant n$ following Proposition 2.22 of Madsen (1985).

The set of indices A used in c_A refers to the active set at x^*, unless otherwise stated. That is, $A = \{i \mid c_i(x^*) = 0\}$.

Theorem 1. *Let x^* be a limit point of the sequence $\{x^{*(k)}\}$.*
(a) *Under A1–A3, x^* is a KKT point. The sequence $\{\lambda^{*(k)}\}$ converges to a vector of Lagrange multipliers.*
(b) *For all indices k corresponding to the subsequence of $\{x^{*(k)}\}$ convergent to x^* the following error estimates hold when $\mu^{(k)} \to 0^+$:*

$$\lambda^{*(k)} = \lambda^* + o(1), \tag{19}$$

$$c_A(x^{*(k)}) = \mu^{(k)}\lambda^* + o(\mu^{(k)}). \tag{20}$$

Proof. First, we define for the purposes of the proof the quantity

$$g(x) = \sum_{i \in A^c(x,\mu)} \operatorname{sgn}(c_i(x))\nabla c_i(x).$$

Now, consider only those indices k for which a particular subsequence $\{x^{*(k)}\}$ converges to x^*. As $\nabla c_A(x^*)$ is of full rank, we may define

$$\lambda^* = -\nabla c_A(x^*)^{+\top} g(x^*).$$

Furthermore, for k sufficiently large, $\nabla c_A(x^{*(k)})^+$ exists, is bounded, and converges to $\nabla c_A(x^*)^+$. From (9) and (10), we have that

$$\|g(x^{*(k)}) + \nabla c_A(x^{*(k)})^\top \lambda^{*(k)}\|_2 = \|\bar{g}(x^{*(k)}, \lambda^{*(k)})\|_2 \leqslant \gamma\mu^{(k)}. \tag{21}$$

Thus, we deduce that

$$\|\nabla c_A(x^{*(k)})^{+\top} g(x^{*(k)}) + \lambda^{*(k)}\|_2 = \|\nabla c_A(x^{*(k)})^{+\top}(g(x^{*(k)}) + \nabla c_A(x^{*(k)})^\top \lambda^{*(k)})\|_2 \leqslant \gamma\mu^{(k)}\|\nabla c_A(x^{*(k)})^{+\top}\|_2. \tag{22}$$

Combine the identity

$$\lambda^{*(k)} - \lambda^* = (\nabla c_A(x^{*(k)})^{+\top} g(x^{*(k)}) + \lambda^{*(k)}) + (\nabla c_A(x^*)^{+\top} g(x^*) - \nabla c_A(x^{*(k)})^{+\top} g(x^{*(k)}))$$

with (22) to obtain the bound

$$\|\lambda^{*(k)} - \lambda^*\|_2 = \gamma\mu^{(k)} \|\nabla c_A(x^{*(k)})^{+\top}\|_2 + \|\nabla c_A(x^*)^{+\top} g(x^*) - \nabla c_A(x^{*(k)})^{+\top} g(x^{*(k)})\|_2. \tag{23}$$

Thus, as the right-hand side of (23) can be made arbitrarily close to zero by picking k large enough, $\lambda^{*(k)}$ is bounded for k sufficiently large and converges to λ^*. Furthermore, since $\|\lambda^{*(k)}\|_\infty \leqslant 1$ we have that $\|\lambda^*\|_\infty \leqslant 1$. Then, taking the limit of (21) as k approaches infinity, we deduce that

$$g(x^*) + \nabla c_A^\top(x^*)\lambda^* = 0. \tag{24}$$

Furthermore, multiplying (23) by $\mu^{(k)}$, we obtain the additional bound

$$\|c_A(x^{*(k)}) - \mu^{(k)}\lambda^*\|_2 \leqslant \gamma\mu^{(k)2} \|\nabla c_A(x^{*(k)})^{+\top}\|_2 + \mu^{(k)} \|\nabla c_A(x^*)^{+\top} g(x^*) - \nabla c_A(x^{*(k)})^{+\top} g(x^{*(k)})\|_2. \tag{25}$$

Taking the limit of (25) as k approaches infinity, we have that

$$c_A(x^*) = 0. \tag{26}$$

Hence, (24) and (26) imply that x^* is a Kuhn–Tucker point, and the (sub)sequence $\{\lambda^{*(k)}\}$ converges to the relevant vector of Lagrange multipliers. The asymptotic estimates (19) and (20) may be deduced from (23) and (25), respectively. □

Notice that under assumption A3, the algorithm identifies the optimal active set in a finite number of iterations. Under assumption A1, one can show that the inner iteration is finitely convergent under the condition that $\mu_{\min} > 0$ using the standard analysis of Dennis and Schnabel (1996).

One needs two further assumptions before stating a sharper convergence result identical, after the necessary changes, to Theorem 4.2 of Gould (1989).

A4 At every limit point x^* of the sequence $\{x^{*(k)}\}$ the matrix $K(x^*, \lambda^*, 0)$ has exactly $|A|$ negative eigenvalues, the remaining eigenvalues are positive.

The assumption above along with A3 can be shown to be a second-order sufficiency condition for x^* to be a local minimum; see Gould (1985).

A5 All functions c_i possess third derivatives, and assume bounded values within Ω.

Theorem 2. *Under A1–A5 the results of Theorem 1 are valid. Furthermore, for all convergent subsequences of the sequence $\{x^{*(k)}\}$ one has the following error estimates when $\mu^{(k)} \to 0^+$:*

$$x^{*(k)} = x^* + \mathrm{O}(\mu^{(k)}), \tag{27}$$

$$\lambda^{*(k)} = \lambda^* + \mathrm{O}(\mu^{(k)}), \tag{28}$$

$$c_A(x^{*(k)}) = \mu^{(k)}\lambda^* + \mathrm{O}(\mu^{(k)2}). \tag{29}$$

Now, we begin with the local convergence results.

A6 The sequence $\{\mu^{(k)}\}$ is adjusted so as to have $\mu^{(k+1)} \leqslant \sigma^{(k)}\mu^{(k)}$ with $\lim_{k\to\infty}\sigma^{(k)} = \sigma < 1$.

The assumption A6 ensures that the sequence $\{\mu^{(k)}\}$ is at least linearly convergent. The following is the most important intermediate result. For the purposes of this theorem, we say that $a_k = \mathrm{O}_s(b_k)$ for two se-

quences a_k and b_k converging to zero if $c_2|b_k| \leqslant |a_k| \leqslant c_1|b_k|$ for all $k \geqslant k_0$ and some constants c_1 and c_2. Although this theorem corresponds to Theorem 5.1 of Gould (1989), it requires a slight addition in our case. We therefore give the proof in its entirety for the sake of completeness.

Theorem 3. *Under A1–A6 for all indices k corresponding to a convergent subsequence the following estimates hold*:

$$\bar{g}(x^{*(k)}, \bar{\lambda}(x^{*(k)}, \mu^{(k+1)})) = \mathrm{O}_s(\mu^{(k)}/\mu^{(k+1)}), \tag{30}$$

$$\bar{g}(x_a^{*(k)}, \bar{\lambda}(x_a^{*(k)}, \mu^{(k+1)})) = \mathrm{O}(\mu^{(k)2}/\mu^{(k+1)}). \tag{31}$$

Proof. To verify (30), first we have that the estimate (20) yields

$$\overline{\lambda}(x^{*(k)}, \mu^{(k+1)}) - \lambda^{*(k)} = c_A(x^{*(k)})(1/\mu^{(k+1)} - 1/\mu^{(k)}) = (\mu^{(k)}/\mu^{(k+1)} - 1)\lambda^* + \mathrm{o}(\mu^{(k)}/\mu^{(k+1)}) \tag{32}$$

as k tends to infinity. From A6, we have that

$$1/2(1-\sigma)\mu^{(k)}/\mu^{(k+1)} \leqslant |\mu^{(k)}/\mu^{(k+1)} - 1| \leqslant \mu^{(k)}/\mu^{(k+1)} \tag{33}$$

for all large k. Therefore, combining (32) and (33), we have

$$(1/2(1-\sigma)(1-\varepsilon_1)\|\lambda^*\|_2)\mu^{(k)}/\mu^{(k+1)} \leqslant \|\overline{\lambda}(x^{*(k)}, \mu^{(k+1)}) - \lambda^{*(k)}\|_2 \leqslant ((1+\varepsilon_1)\|\lambda^*\|_2)\mu^{(k)}/\mu^{(k+1)} \tag{34}$$

for all k sufficiently large, where the terms $(1-\varepsilon_1)$ and $(1+\varepsilon_1)$ $(0 < \varepsilon_1 \ll 1)$ account for the asymptotically smaller terms in (34). Now, from (21) we obtain

$$\begin{aligned}
\bar{g}(x^{*(k)}, \overline{\lambda}(x^{*(k)}, \mu^{(k+1)})) &= \bar{g}(x^{*(k)}, \lambda^{*(k)}) + \nabla c_A^{\top}(x^{*(k)})(\overline{\lambda}(x^{*(k)}, \mu^{(k+1)}) - \lambda^{*(k)}) \\
&= \nabla c_A^{\top}(x^{*(k)})(\overline{\lambda}(x^{*(k)}, \mu^{(k+1)}) - \lambda^{*(k)}) + \mathrm{O}(\mu^{(k)}) \\
&= \nabla c_A^{\top}(x^{*(k)})(\overline{\lambda}(x^{*(k)}, \mu^{(k+1)}) - \lambda^{*(k)}) + \mathrm{o}(\mu^{(k)}/\mu^{(k+1)}).
\end{aligned} \tag{35}$$

Then, (34), (35), and the continuity of $\nabla c_A(x)$ give the bound

$$\|\bar{g}(x^{*(k)}, \overline{\lambda}(x^{*(k)}, \mu^{(k+1)}))\|_2 \leqslant (2(1+\varepsilon_1)(1+\varepsilon_2)\|\nabla c_A^{\top}(x^*)\|_2\|\lambda^*\|_2)\mu^{(k)}/\mu^{(k+1)} \tag{36}$$

for all k sufficiently large, where the term $(1+\varepsilon_2)$ $(0 < \varepsilon_2 \ll 1)$ accounts for the asymptotically smaller terms in (35) and the constant two occurs because of the bound $\|\nabla c_A^{\top}(x^{*(k)})\|_2 \leqslant 2\|\nabla c_A^{\top}(x^*)\|_2$. Premultiplying (35) by $\nabla c_A(x^{*(k)})^{+\top}$ gives

$$\overline{\lambda}(x^{*(k)}, \mu^{(k+1)}) - \lambda^{*(k)} = \nabla c_A(x^{*(k)})^{+\top}\bar{g}(x^{*(k)}, \overline{\lambda}(x^{*(k)}, \mu^{(k+1)})) + \mathrm{o}(\mu^{(k)}/\mu^{(k+1)}). \tag{37}$$

Using the continuity of $\nabla c_A(x)^{+\top}$ in some neighborhood of x^* this leads to

$$\|\overline{\lambda}(x^{*(k)}, \mu^{(k+1)}) - \lambda^{*(k)}\|_2 \leqslant 2(1+\varepsilon_2)\|\nabla c_A(x^*)^{+\top}\|_2\|\bar{g}(x^{*(k)}, \overline{\lambda}(x^{*(k)}, \mu^{(k+1)}))\|_2 \tag{38}$$

for all k sufficiently large, where the term $(1+\varepsilon_2)$ once again accounts for the asymptotically smaller term in (37). Inequalities (34) and (38) combine to give the bound

$$(1/4(1-\sigma)(1-\varepsilon_1)\|\lambda^*\|_2/(1+\varepsilon_2)\|\nabla c_A(x^*)^{+\top}\|_2)\mu^{(k)}/\mu^{(k+1)} \leqslant \|\bar{g}(x^{*(k)}, \overline{\lambda}(x^{*(k)}, \mu^{(k+1)}))\|_2 \tag{39}$$

for large k. The bounds (36) and (39) then imply (30).

For the estimate (31), observe that the coefficient matrix $K(x^{*(k)}, \lambda^{*(k)}, \mu^{*(k)})$ of (15) satisfies the second-order condition (and hence is non-singular) for large enough k from assumption A4 and Theorem 2. Hence $x_a^{*(k)}$ is defined by (16). The active set at a limit point of x^* of $\{x^{*(k)}\}$ is correctly identified for sufficiently

large k at $x_a^{*(k)}$. To see this, note first that the right-hand side of (15) is $O(\mu^{(k)})$. This observation along with (15), (17) and (27) implies that

$$x_a^{*(k)} = x^* + O(\mu^{(k)}).$$

Then the active set identification property follows using A3.

Now define

$$\lambda_a^{*(k)} = \lambda^{*(k)} + r^{(k)}, \tag{40}$$

where $r^{(k)}$ is given by (15). Then, by Taylor's expansion and (15) one has

$$\begin{bmatrix} \bar{g}(x_a^{*(k)}, \lambda_a^{*(k)}) \\ c_A(x_a^{*(k)}) - \mu^{(k+1)}\lambda_a^{*(k)} \end{bmatrix} = \begin{bmatrix} \overline{G}(x^{*(k)}, \lambda^{*(k)}) & \nabla c_A^\top(x^{*(k)}) \\ \nabla c_A(x^{*(k)}) & -\mu^{(k+1)} I \end{bmatrix} \begin{bmatrix} p^{(k)} \\ r^{(k)} \end{bmatrix} \tag{41}$$

$$= \begin{bmatrix} \bar{g}(x^{*(k)}, \lambda^{*(k)}) \\ c(x^{*(k)}) - \mu^{(k+1)}\lambda^{*(k)} \end{bmatrix} + O(\|p^{(k)}\|_2^2) + O(\|r^{(k)}\|_2^2) \tag{42}$$

$$= \begin{bmatrix} 0 \\ (\mu^{(k)} - \mu^{(k+1)}) r^{(k)} \end{bmatrix} + O(\|p^{(k)}\|_2^2) + O(\|r^{(k)}\|_2^2)$$

$$= O(\|p^{(k)}\|_2^2) + O(\|r^{(k)}\|_2^2) + O(\mu^{(k)}\|r^{(k)}\|_2).$$

Moreover, Eqs. (9), (19), and (20) ensure that the right-hand side of (15) is $O(\mu^{(k)})$.

Thus $\|p^{(k)}\|_2 = O(\mu^{(k)}) = \|r^{(k)}\|_2$ and (41) gives

$$\bar{g}(x_a^{*(k)}, \lambda_a^{*(k)}) = O(\mu^{(k)2}) \tag{43}$$

and

$$c_A(x^{*(k)}) - \mu^{(k+1)}\lambda_a^{*(k)} = O(\mu^{(k)2}). \tag{44}$$

But then, (44) and the definition of $\overline{\lambda}(x_a^{*(k)}, \mu^{(k+1)})$ give

$$\mu^{(k+1)}(\overline{\lambda}(x_a^{*(k)}, \mu^{(k+1)}) - \lambda_a^{*(k)}) = c_A(x_a^{*(k)}) - \mu^{(k+1)}\lambda_a^{*(k)} = O(\mu^{(k)2})$$

and hence

$$\overline{\lambda}(x_a^{*(k)}, \mu^{(k+1)}) - \lambda_a^{*(k)} = O(\mu^{(k)2}/\mu^{(k+1)}). \tag{45}$$

Now, Eqs. (43) and (45) combine to give

$$\bar{g}(x_a^{*(k)}, \overline{\lambda}(x_a^{*(k)}, \mu^{(k+1)})) = \bar{g}(x_a^{*(k)}, \lambda_a^{*(k)}) + \nabla c_A^\top(x_a^{*(k)})(\overline{\lambda}(x_a^{*(k)}, \mu^{(k+1)}) - \lambda_a^{*(k)}) = O(\mu^{(k)2}/\mu^{(k+1)}),$$

which establishes (31). □

Notice that under A6 the gradient at $x^{*(k)}$ is asymptotically larger than the gradient at the alternative starting point $x_a^{*(k)}$. This indicates that the alternative starting point $x_a^{*(k)}$ should be asymptotically preferable to $x^{*(k)}$. On the other hand, Theorem 3 gives a clue as to the choice of the sequence $\{\mu^{(k)}\}$. The value $\mu^{(k+1)}$ should be smaller than $\mu^{(k)}$, but larger than $\mu^{(k)2}$. This choice ensures that the sequence $\{\mu^{(k)}\}$ approaches zero in a Q-superlinearly convergent manner. This leads to the final assumption.

A7 As k goes to infinity the sequence $\{\mu^{(k)}\}$ is adjusted as $\mu^{(k)2}/\mu^{(k+1)} = o(1)$.

Notice here that under assumption A7 the gradient at $x^{*(k)}$ in the estimate (30) can get arbitrarily large whereas the gradient at $x_a^{*(k)}$ vanishes to zero. The next step is to show that the sequence $\{x^{*(k)}\}$ follows the Q-superlinearly convergent sequence $\{\mu^{(k)}\}$. In order to show this one needs to show (1) that asymptotically, the point $x_a^{*(k)}$ is always chosen as the starting point of the inner iterations, and (2) that this point or the first

Newton iterate obtained from this point satisfies the inner iteration stopping criterion (9). For convenience we use $\mathscr{K}$ to denote the set of indices corresponding to indices k associated with convergent subsequences.

Theorem 4. *Under A1–A7, for all $k \in \mathscr{K}$ the $k+1$st inner iteration begins from the alternative starting point $x_a^{*(k)}$ as defined in (15).*

The proof of this theorem follows directly from (17) which governs the use of $x_a^{*(k)}$, assumption A6 and the estimate (31) of the previous theorem.

Now, one can give the next theorem the proof of which is identical to that of Theorem 5.8 of Gould (1989). This result is a consequence of two technical intermediate results, namely Lemmas 5.5 and 5.8 of Gould (1989).

Theorem 5. *Under A1–A7, for all sufficiently large $k \in \mathscr{K}$ the following hold:*
(a) *The Newton direction $p^{*(k+1,0)}$ obtained from (12) always satisfies (11).*
(b) *The step length $\alpha^{(k+1,0)}$ used with the Newton direction is equal to one.*

Now, using the above theorem and the aforementioned second-order sufficiency property (c.f. assumption A4) of the matrix $K(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}, \mu^{(k+1)})$ the following corollary is obtained.

Corollary 1. *Under A1–A7, for all sufficiently large $k \in \mathscr{K}$ the following holds:*

$$x^{(k+1,1)} = x^{(k+1,0)} + p^{(k+1,0)},$$

where $p^{(k+1,0)}$ is the Newton direction obtained from (12).

The next step is to show that at the point $x^{(k+1,1)}$ of the previous corollary the gradient can be bounded. It is easy to show using Taylor series expansion that $\bar{g}(x^{(k+1,1)}, \bar{\lambda}^{(k+1,1)}) = \mathrm{O}(\mu^{(k)4}/\mu^{(k+1)})$ for all sufficiently large $k \in \mathscr{K}$. This leads to the following theorem and its corollary.

Theorem 6. *Under A1–A7, for all sufficiently large $k \in \mathscr{K}$, for $\ell \leqslant 1$ (9) holds.*

Corollary 2. *Under A1–A7, assume that the entire sequence $\{x^{*(k)}\}$ converges. Then,*
(a) *if $\{\mu^{(k)}\}$ converges Q-linearly the $\{x^{*(k)}\}$ converges R-linearly,*
(b) *if $\{\mu^{(k)}\}$ converges Q-superlinearly $\{x^{*(k)}\}$ converges R-superlinearly.*

4. Numerical results

In this section we summarize our computational experience with a preliminary version of the algorithm of the previous section. We believe more research effort will be necessary in future to reach a definite conclusion about the performance of the algorithm.

A version of the algorithm for dense matrix algebra was coded in C, and tested on 25 test problems with up to 15 variables and 100 equations. For the numerical linear algebraic tasks the algorithm uses a version of the symmetric indefinite matrix factorization techniques of Bunch and Parlett (1971). Using this factorization, the calculations can be arranged in such a way that computation of the eigenvalues of the matrix K are not necessary. For details, the reader is referred to Conn and Gould (1984). As in Gould (1989) we used $\tau = 0.1$, and $\gamma = 1$ although other choices should also be investigated in future work.

The results of our experiments with the algorithm of this paper, and two competing algorithms, the Hald and Madsen (1985) two-stage non-linear ℓ_1 algorithm, and the general purpose Nelder and Mead (1965) simplex algorithm are summarized below. The Hald–Madsen code is recognized to be the most efficient non-linear ℓ_1 code to date.

We report results with two different degrees of accuracy, 10^{-8} and 10^{-6}, in Table 1. The test problems are available in Hock and Schitkowski (1981) when no source is indicated. They can also be obtained from the author of the present paper upon request.

With the exception of five test problems, the algorithm displays the behavior predicted by the theoretical analysis outlined above. In problems Tishler–Zang (40×5), Hald and Madsen 1 LC, and Biggs I, the algorithm ran into numerical difficulties. In the problems Powell badly scaled function and Osborne I function, only a single value of μ was used with a large number of Newton iterations.

In the remaining 20 problems, superlinear μ sequences were used successfully. On the other hand, it is observed that the Hald–Madsen algorithm is the fastest in a larger number of test problems while our algorithm is fastest in some test cases. The reason for the larger number of function and Jacobian evaluations in our case is that in some test cases the algorithm takes many Newton steps for the initial value of μ. This indicates that the choice of initial μ along with a suitable starting point deserves further research. Another point that deserves further research is the choice of the search direction when the Newton system of Step 1.2 does not have any solution, or when it does have multiple solutions. The use of doncs results in poor directions of descent in the algorithm. In fact, we observed that the algorithm was competitive with the Hald–Madsen algorithm whenever doncs were not used. A stable and efficient alternative to doncs has to be carefully researched in the future. A trust region type algorithm may be investigated as an alternative here.

Table 1
Computational results

Problem			PH(6)		PH(8)		HM		NM
Description	m	n	F	Jac	F	Jac	F	Jac	F
Tishler and Zang (1982)	40	6	146	40	180	44	10	10	716
Tishler and Zang (1982)	40	3	192	115	236	119	22	22	701
Tishler and Zang (1982)	40	5	–	–	–	–	27	27	1202
El-Attar et al. (1979) (Gonin and Money, 1989, p. 49)	3	2	72	26	91	28	11	11	153
Madsen (1975)[a] (Gonin and Money, 1989, p. 51)	3	2	37	25	57	33	49	49	78
Hald and Madsen (1985): 0 LC	3	2	35	21	32	24	12	12	106
Hald and Madsen (1985): 1 LC	3	2	–	–	–	–	11	11	77
Jennrich and Sampson (1968)[a]	10	2	122	61	133	63	33	33	125
Rosenbrock function	2	2	57	46	61	47	31	31	428
Freudenstein and Roth function	2	2	18	17	19	18	28	28	58
Powell (1970)[a] badly scaled function	2	2	230	103	238	103	126	126	878
Brown badly scaled function	3	2	30	23	31	24	63	63	303
Beale (1958)[a] function	3	2	25	21	32	24	12	12	106
Helical Valley	3	3	45	36	49	38	14	14	305
Bard (1970)[a] function	15	3	52	33	147	51	10	10	165
Gauss function	15	3	67	33	227	127	11	11	176
Gulf Research and Development	100	3	63	37	63	37	21	21	293
Box (1966)[a] three dimensional function	10	3	124	75	137	75	20	20	437
Powell (1962)[a] singular function	4	4	31	23	53	28	90	90	405
Wood (Cox, 1969)[a] function	6	4	77	61	78	62	12	12	368
Kowalik and Osborne (1968)[a] function	11	4	108	57	186	71	10	10	279
Brown and Dennis (1971)[a] function	20	4	16	16	17	17	41	41	302
Osborne I (1972)[a] function	33	5	414	209	542	235	10	10	1218
Biggs (1971)[a] function	13	6	–	–	–	–	150	150	789
Osborne II (1972)[a] function	65	11	146	72	244	88	16	16	1508

PH(6): Pinar and Hartmann Algorithm with $\mu_{\min} = 10^{-6}$; PH(8): Pinar and Hartmann Algorithm with $\mu_{\min} = 10^{-8}$; HM: Hald and Madsen (1985) Algorithm; NM: Nelder and Mead (1965) Algorithm; F: number of function evaluations; Jac: number of Jacobian evaluations.

[a] See Hock and Schitkowski (1981).

References

Andersen, K., 1996. An efficient Newton barrier method for minimizing a sum of Euclidean norms. SIAM Journal on Optimization 6, 74–95.

Anderson, D.H., Osborne, M.R., 1977a. Discrete, nonlinear approximation problems in polyhedral norms. Numerische Mathematik 28, 143–156.

Anderson, D.H., Osborne, M.R., 1977b. Discrete, nonlinear approximation problems in polyhedral norms. A Levenberg-like algorithm. Numerische Mathematik 28, 157–170.

Bartels, R.H., Conn, A.R., 1982. An approach to nonlinear ℓ_1 data fitting. In: Hennart, J.P. (Ed.), Numerical analysis: Lecture notes in mathematics. Springer Verlag, New York, pp. 48–58.

Basset Jr., G., Koenker, R., 1978. Asymptotic theory of least absolute error regression. Journal of the American Statistical Association 73, 618–622.

Ben-Tal, A., Teboulle, M., 1989. A smoothing technique for non-differentiable optimization problems. Lecture Notes in Mathematics, Vol. 1405, 1–11.

Ben-Tal, A., Teboulle, M., Yang, W.H., 1991. A least-squares-based method for a class of nonsmooth minimization problems with applications in plasticity. Applied Mathematics and Optimization 24, 273–288.

Bunch, J.R., Parlett, B.N., 1971. Direct methods for solving symmetric indefinite systems of linear equations. SIAM Journal on Numerical Analysis 8, 639–655.

Conn, A.R., Gould, N.I.M., 1984. On the location of directions of infinite descent for nonlinear programming algorithms. SIAM Journal on Numerical Analysis 21, 1162–1179.

Dennis, J.E., Schnabel, R.B., 1996. Numerical Methods for Unconstrained Optimization and Nonlinear Equations. SIAM, Philadelphia.

Dussault, J.-P., 1995. Numerical stability and efficiency of penalty algorithms. SIAM Journal on Numerical Analysis 32, 296–317.

Dussault, J.-P., 1998. Augmented penalty algorithms. IMA Journal of Numerical Analysis 18, 355–372.

El-Attar, R.A., Vidyasagar, M., Dutta, S.R.K., 1979. An algorithm for ℓ_1 norm minimization with application to nonlinear ℓ_1 approximation. SIAM Journal on Numerical Analysis 16, 70–86.

Gonin, R., Money, A.H., 1989. Nonlinear L_p norm estimation. Marcel Dekker, New York.

Gould, N.I.M., 1985. On practical conditions for existence and uniqueness of solutions to the general equality constrained quadratic programming problems. Mathematical Programming 32, 90–99.

Gould, N.I.M., 1986. On the accurate determination of search directions for simple differentiable penalty functions. IMA Journal of Numerical Analysis 6, 357–372.

Gould, N.I.M., 1989. On the convergence of a sequential penalty function method for constrained optimization. SIAM Journal on Numerical Analysis 26 (1), 107–128.

Hald, J., Madsen, K., 1985. Combined LP and quasi-Newton methods for nonlinear ℓ_1 optimization. SIAM Journal on Numerical Analysis 22, 68–80.

Hock, W., Schitkowski, K., 1981. Test Examples for Nonlinear Programming Codes. In: Lecture Notes in Economics and Mathematical Systems, vol. 187, Springer-Verlag, Berlin.

Huber, P.J., 1981. Robust statistics. Wiley and Sons, New York.

Madsen, K., 1985. Minimization of nonlinear approximation functions, Doctor Technices Thesis, Technical University of Denmark.

Madsen, K., Nielsen, H.B., 1993. A finite smoothing algorithm for linear ℓ_1 estimation. SIAM Journal on Optimization 3, 223–235.

Madsen, K., Nielsen, H.B., Pınar, M.Ç., 1996. A new finite continuation algorithm for linear programming. SIAM Journal on Optimization 6, 600–616.

McLean, R.A., Watson, G.A., 1980. Numerical methods for nonlinear discrete L_1 approximation problems. In: Collatz, L., Meinardus, H., Werner, H. (Eds.), Numerical methods of approximation theory. Birkhäuser Verlag, Basel.

Murray, W., Overton, M., 1981. A projected Lagrangian algorithm for nonlinear ℓ_1 optimization. SIAM Journal on Scientific Computing 2, 207–224.

Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. The Computer Journal 7, 308–313.

Osborne, M.R., Watson, G.A., 1971. On an algorithm for discrete nonlinear L_1 approximation. The Computer Journal 14, 184–188.

Overton, M., 1982. Algorithms for nonlinear ℓ_1 and ℓ_∞ fitting. In: Powell, M.J.D. (Ed.), Nonlinear optimization. Academic Press, London, pp. 91–101.

Tishler, A., Zang, I., 1982. An absolute deviations curve fitting algorithm for nonlinear models. In: Zanakis, S.H., Rustagi, J.S. (Eds.), Optimization in statistics, TIMS studies in management science, 19. North Holland, Amsterdam.

Yang, Z., 1995. An algorithm for nonlinear L_1 curve-fitting based on the smooth approximation. Computational Statistics & Data Analysis 19, 45–52.