

On the Convergence of Huber Approximation for the Nonlinear ℓ_1 Problem

Mustafa Ç. Pınar
Department of Industrial Engineering
Bilkent University
06533 Ankara, Turkey

Wolfgang M. Hartmann
SAS Institute
Cary, NC USA

December 4, 1998

Abstract

The smooth Huber approximation to the nonlinear ℓ_1 problem was proposed by Tishler and Zang (1982), and further developed in Yang (1995). In the present note, we use the ideas of Gould (1989) to give a new algorithm with rate of convergence results for the smooth Huber approximation.

1 Introduction

In this note we investigate a new algorithm for the nonlinear ℓ_1 estimation problem. Let $c_i : \Re^n \mapsto \Re$ be at least twice continuously differentiable functions for each $i = 1, \dots, m$. We want to find a minimizing point for the following function

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

This is an important problem in statistics, curve fitting and engineering design. In statistics, when measurement errors are not normally distributed (e.g., Cauchy distributed) the above problem may yield more reliable estimates than the nonlinear least squares problem; see Tishler and Zang (1982).

From a computational point of view, the nonlinear 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 nonlinear 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), Anderson and Osborne (1977b), 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 nonlinear 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), 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 the El-Attar et al. function to engineering problems in plasticity. The El-Attar et al. function is known as the hyperboloid approximation in location literature; see Andersen (1996).

The method given in the present note is akin to the algorithm of Tishler and Zang (1982) and to that of Yang (1995). It uses an approximation function known as the 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 the 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.

The proposed algorithm is essentially an adaptation of a quadratic penalty function algorithm proposed by Gould (1989) to solve nonlinear programming problems with equality constraints. We will use Gould's ideas in the context of an approximation algorithm for

the nonlinear 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.

In the next two sections we describe the proposed algorithm, and we summarize convergence and rate of convergence results. For ease of exposition we state the results without proofs as they can be obtained, mutatis mutandis, from the proofs of corresponding results in Gould (1989).

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

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

where

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

Before stating the algorithm we will give some definitions. Let $A(x, \mu) = \{i \mid |c_i(x)| \leq \mu\}$ represent the active set at x . $\nabla c_A(x)$ denotes a matrix with columns $\nabla c_i(x)$ where $i \in A(x, \mu)$. The Lagrange multiplier estimates λ_i are defined for all $i \in A(x, \mu)$ as:

$$\lambda_i = \frac{c_i(x)}{\mu}. \quad (4)$$

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

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

We define the quantity \bar{G}

$$\bar{G}(x, \lambda) = \sum_{i \in A^c(x, \mu)} \text{sgn}(c_i(x)) \nabla^2 c_i(x) + \sum_{i \in A(x, \mu)} \lambda_i \nabla^2 c_i(x) \quad (6)$$

which is nothing else than the Jacobian of \bar{g} , and, the $(n+t) \times (n+t)$ matrix

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

which is a perturbed version of the familiar augmented Karush-Kuhn-Tucker matrix. Finally, if there exist multipliers λ_i^* such that $-1 \leq \lambda_i^* \leq 1$ and

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

where $A(x^*) = \{i \mid c_i(x^*) = 0\}$, we say that x^* is a KKT point. 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 μ_{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 Compute function, gradient and Hessian values at $x^{(k,0)}$. Let $\bar{\lambda}^{(k,0)} = \bar{\lambda}(x^{(k,0)}, \mu^{(k)})$. Furthermore, compute $\bar{g}(x^{(k,0)}, \bar{\lambda}^{(k,0)})$, $\bar{G}(x^{(k,0)}, \bar{\lambda}^{(k,0)})$ and $K(x^{(k,0)}, \bar{\lambda}^{(k,0)}, \mu^{(k)})$. Let $\ell = 0$.

Step 1.1 If

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

then

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

and continue from Step 2.

Step 1.2 Find $p^{(k,\ell)}$ that satisfies:

$$-\bar{g}(x^{(k,\ell)}, \bar{\lambda}^{(k,\ell)})^T p^{(k,\ell)} \geq \epsilon \mu^{(k)} \|\bar{g}(x^{(k,\ell)}, \bar{\lambda}^{(k,\ell)})\|_2 \|p^{(k,\ell)}\|_2. \quad (10)$$

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 t negative eigenvalues, the rest of the eigenvalues are positive; see Gould (1986)) then, compute $p^{(k,\ell)}$ as a Newton direction from the system below:

$$\begin{bmatrix} \bar{G}(x^{(k,\ell)}, \bar{\lambda}^{(k,\ell)}) & \nabla c_A(x^{(k,\ell)})^T \\ \nabla c_A(x^{(k,\ell)}) & -\mu^{(k)} I \end{bmatrix} \begin{pmatrix} p^{(k,\ell)} \\ r^{(k,\ell)} \end{pmatrix} = - \begin{pmatrix} \bar{g}(x^{(k,\ell)}, \bar{\lambda}^{(k,\ell)}) \\ 0 \end{pmatrix} \quad (11)$$

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

$$\Phi(x^{(k,l)} + \alpha^{(k,l)} p^{(k,l)}, \mu^{(k)}) \leq \Phi(x^{(k,l)}, \mu^{(k)}) + \beta_1 \alpha^{(k,l)} \bar{g}(x^{(k,l)}, \bar{\lambda}^{(k,l)})^T p^{(k,l)} \quad (12)$$

$$\bar{g}(x^{(k,l)} + \alpha^{(k,l)} p^{(k,l)}, \bar{\lambda}(x^{(k,l)} + \alpha^{(k,l)} p^{(k,l)}))^T p^{(k,l)} \geq \beta_2 \bar{g}(x^{(k,l)}, \bar{\lambda}^{(k,l)})^T p^{(k,l)}. \quad (13)$$

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 $\ell \leftarrow \ell + 1$.

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., is invertible) compute $p^{(k)}$ from the linear system of equations below:

$$\begin{bmatrix} \bar{G}(x^{*(k)}, \lambda^{*(k)}) & \nabla c_A(x^{*(k)})^T \\ \nabla c_A(x^{*(k)}) & -\mu^{(k)} I \end{bmatrix} \begin{pmatrix} p^{(k)} \\ r^{(k)} \end{pmatrix} = - \begin{pmatrix} \bar{g}(x^{*(k)}, \lambda^{*(k)}) \\ c_A(x^{*(k)}) - \mu^{(k+1)} \lambda^{*(k)} \end{pmatrix} \quad (14)$$

and, let $x_a^{*(k)} = x^{*(k)} + p^{(k)}$. If

$$\|\bar{g}(x_a^{*(k)}, \bar{\lambda}(x_a^{*(k)}, \mu^{(k+1)}))\|_2 \leq \max\{\tau, \|\bar{g}(x^{*(k)}, \bar{\lambda}(x^{*(k)}, \mu^{(k+1)}))\|_2\} \quad (15)$$

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

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

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) satisfy 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 live in a bounded domain Ω .

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 (1983).

A2 The sequence $\{\mu^{(k)}\}$ goes to zero as k goes to infinity.

A3 At every limit point x^* of the sequence $\{x^{*(k)}\}$ x^* , and the corresponding limit point λ^* of the sequence $\{\lambda^{*(k)}\}$ λ^* , 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 $|A(x^*)| \leq n$ following Proposition 2.22 of Madsen (1985).

Theorem 1 *Let x^* be a limit point of the sequence $\{x^{*(k)}\}$.*

- Under A1, A2 and A3, x^* is a KKT point. The sequence $\{\lambda^{*(k)}\}$ converges to a vector of Lagrange multipliers.*
- For all indices k corresponding to the subsequence of $\{x^{*(k)}\}$ convergent to x^* the following error estimates hold when $\mu^{(k)} \rightarrow 0^+$:*

$$\lambda^{*(k)} = \lambda^* + o(1), \quad (16)$$

$$c_A(x^{*(k)}) = \mu^{(k)} \lambda^* + o(\mu^{(k)}). \quad (17)$$

The set of indices A uses in c_A above refer to the active set at x^* . That is, $A = \{i | c_i(x^*) = 0\}$. Notice that under assumption A3, the algorithm identifies the optimal active set in a finite number of iterations.

One needs two further assumptions before stating a sharper convergence result.

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, A2, A3, A4, and 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 $\mu^{(k)} \rightarrow 0^+$:*

$$x^{*(k)} = x^* + O(\mu^{(k)}), \quad (18)$$

$$\lambda^{*(k)} = \lambda^* + O(\mu^{(k)}), \quad (19)$$

$$c_A(x^{*(k)}) = \mu^{(k)} \lambda^* + O(\mu^{(k)2}). \quad (20)$$

Now, we can begin with the local convergence results.

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

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 $a_k = O_s(b_k)$ for two sequences a_k and b_k converging to zero if $c_2|b_k| \leq |a_k| \leq c_1|b_k|$ for all $k \geq k_0$ and some constants c_1 and c_2 .

Theorem 3 *Under A1, A2, A3, A4, A5, and A6 for all indices k corresponding to a convergent subsequence the following estimates hold:*

$$\bar{g}(x^{*(k)}, \bar{\lambda}(x^{*(k)}, \mu^{(k+1)})) = O_s(\mu^{(k)} / \mu^{(k+1)}), \quad (21)$$

$$\bar{g}(x_a^{*(k)}, \bar{\lambda}(x_a^{*(k)}, \mu^{(k+1)})) = O(\mu^{(k)2} / \mu^{(k+1)}). \quad (22)$$

The proof of this result follows from verbatim repetition of the proof of Theorem 5.1 of Gould (1989) with one minor exception. One needs to make sure that 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 (14) is $O(\mu^{(k)})$. This observation along with (14), (15) and (18) imply that

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

Then the active set identification property follows using A3.

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 (21) 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 \mathcal{K} to denote the set of indices corresponding to indices k associated with convergent subsequences.

Theorem 4 *Under A1–A7, for all $k \in \mathcal{K}$ the $k + 1$ st inner iteration begins from the alternative starting point $x_a^{*(k)}$ as defined in (14).*

The proof of this theorem follows directly from (15) which governs the use of $x_a^{*(k)}$, assumption A6 and the estimate (22) of the previous theorem.

The next step in the analysis is reached after some technical lemmata. First, a bound on the search direction vector is derived. It is easy to show that

$$p^{(k+1,0)} = O(\mu^{(k)2}). \quad (23)$$

Second, bounds on the eigenvalues of the Jacobian matrix of Φ are obtained using a result of Murray (1971). These results hold in our case, mutatis mutandis. An important intermediate result worth mentioning here is that for all large enough $k \in \mathcal{K}$ the matrix $K(x^{(k+1,0)}, \bar{\lambda}^{(k+1,0)}, \mu^{(k+1)})$ (the first matrix of $k + 1$ st inner iteration) satisfies the second-order sufficiency condition of assumption A4. Now, one can continue with the next theorem.

Theorem 5 *Under A1–A7, for all sufficiently large $k \in \mathcal{K}$ the following hold:*

- (a) *The Newton direction $p^{*(k+1,0)}$ obtained from (11) always satisfies (10).*
- (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 3.1 *Under A1–A7, for all sufficiently large $k \in \mathcal{K}$ 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 (11).

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 $\bar{g}(x^{(k+1,1)}, \bar{\lambda}^{(k+1,1)}) = O(\mu^{(k)4}/\mu^{(k+1)})$ for all sufficiently large $k \in \mathcal{K}$. This leads to the following theorem and its corollary.

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

Corollary 3.2 Under A1–A7, assume that the entire sequence $\{x^{*(k)}\}$ converges. Then, if $\{\mu^{(k)}\}$ converges Q -linearly the $\{x^{*(k)}\}$ converges R -linearly. If $\{\mu^{(k)}\}$ converges Q -superlinearly $\{x^{*(k)}\}$ converges R -superlinearly.

Gould (1989) shows that the assumptions are sufficient to establish that the whole sequence converges, and thus to strengthen the R rates to Q rates. This needs one more technical result. The interested reader is referred to this reference.

As a concluding note, a dense version of the algorithm was coded, and tested on 27 test problems with up to 15 variables and 100 equations. With the exception of four test problems, the algorithm displays the behavior predicted by the theoretical analysis outlined above. Clearly, the use of the augmented system opens up possibilities for the efficient exploitation of sparsity for the solution of large problems. This topic will be treated in the future.

References

- K. Andersen (1996) An efficient Newton barrier method for minimizing a sum of Euclidean norms, *SIAM J. Optim.* 6, 74-95.
- D.H. Anderson and M.R. Osborne (1977a) Discrete, nonlinear approximation problems in polyhedral norms, *Numer. Math.* 28, 143-156.
- D.H. Anderson and M.R. Osborne (1977b) Discrete, nonlinear approximation problems in polyhedral norms. A Levenberg-like algorithm, *Numer. Math.* 28, 157-170.
- R. H. Bartels and A.R. Conn (1982) An approach to nonlinear ℓ_1 data fitting. In J.P. Hennart, ed. *Numerical Analysis: Lecture Notes in Mathematics*, Springer Verlag, New York, 48-58.
- A. Ben-Tal and M. Teboulle (1989) A smoothing technique for non-differentiable optimization problems, *Lecture Notes in Mathematics*, 1405, 1-11.
- A. Ben-Tal, M. Teboulle and W.H. Yang (1991) A least-squares-based method for a class of nonsmooth minimization problems with applications in plasticity, *Appl. Math. Optim.* 24, 273-288.
- J.-P. Dussault (1995) Numerical stability and efficiency of penalty algorithms, *SIAM J. Numer. Anal.* 32, 296-317.
- R.A. El-Attar, M. Vidyasagar and S.R.K. Dutta (1979) An algorithm for ℓ_1 norm minimization with application to nonlinear ℓ_1 approximation, *SIAM J. Numer. Anal.* 16, 70-86.
- R. Gonin and A.H. Money (1989) *Nonlinear L_p Norm Estimation*, Marcel Dekker, New York.
- N.I.M. Gould (1989) On the convergence of a sequential penalty function method for constrained optimization, *SIAM J. Numer. Anal.* 26(1), 107-128.
- N.I.M. Gould (1986) On the accurate determination of search directions for simple differentiable penalty functions, *IMA J. Numer. Anal.* 6, 357-372.
- N.I.M. Gould (1985) On practical conditions for existence and uniqueness of solutions to the general equality constrained quadratic programming problems, *Math. Prog.* 32, 90-99.

- J. Hald and K. Madsen (1985) Combined LP and quasi-Newton methods for nonlinear ℓ_1 optimization, *SIAM J. Numer. Anal.* 22, 68-80.
- K. Madsen (1985) *Minimization of Nonlinear Approximation Functions*, Doctor Technics Thesis, Technical University of Denmark.
- K. Madsen and H.B. Nielsen (1993) A finite smoothing algorithm for linear ℓ_1 estimation, *SIAM J. Optim.* 3, 223-235.
- K. Madsen, H.B. Nielsen and M.Ç. Pinar (1996) A new finite continuation algorithm for linear programming, *SIAM J. Optim.* 6, 600-616.
- R.A. McLean and G.A. Watson (1980) Numerical methods for nonlinear discrete L_1 approximation problems. In: L. Collatz, G. Meinardus and H. Werner, eds. *Numerical Methods of Approximation Theory*, Birkhäuser Verlag, Basel.
- W. Murray (1971) Analytical expressions for eigenvalues and eigenvectors of the Hessian matrices of barrier and penalty functions, *J. Opt. Theory and Appl.* 7, 189-196.
- W. Murray and M. Overton (1981) A projected Lagrangian algorithm for nonlinear ℓ_1 optimization, *SIAM J. Sci. Stat. Comput.* 2, 207-224.
- M.R. Osborne and G.A. Watson (1971) On an algorithm for discrete nonlinear L_1 approximation, *Comput. J.* 14, 184-188.
- M. Overton (1982) Algorithms for nonlinear ℓ_1 and ℓ_∞ fitting, In: M.J.D. Powell, ed. *Nonlinear Optimization*, Academic Press, London, 91-101.
- A. Tishler and I. Zang (1982) An absolute deviations curve fitting algorithm for nonlinear models. In: S.H. Zanakis and J.S. Rustagi, eds. *Optimization in Statistics, TIMS Studies in Management Science* 19, North Holland.
- Z. Yang (1995) An algorithm for nonlinear L_1 curve-fitting based on the smooth approximation, *Comp. Stat. and Data Anal.* 19, 45-52.