Solving L_p-Norm Problems and Applications

Yuying Li*

TR 93-1331
March 1993

Department of Computer Science
Cornell University
Ithaca, NY 14853-7501

*Computer Science Department, Cornell University, Ithaca, New York, 14853.
Research partially supported by the Applied Mathematical Sciences Research Program (KC-04-02) of the Office of Energy Research of the U.S. Department of Energy under grant DE-FG02-86ER25013.A000, and in part by NSF, AFOSR and ONR through grant DMS-8920550, and by the Cornell Theory Center, which receives major funding from the National Science Foundation and IBM Corporation, with additional support from New York State and members of its Corporate Research Institute.

SOLVING L_P-NORM PROBLEMS AND APPLICATIONS

YUYING LI*

Abstract. The l_p norm discrete estimation problem $\min_{x\in\Re^n} \|b - A^T x\|_p^p$ has been solved in many data analysis applications, e.g. geophysical modeling. Recently, a new globally convergent Newton method (called GNCS) has been proposed for solving l_p problems with $1 \le p \le 2$ [5]. This method is much faster than the widely used IRLS method when $1 \le p \le 1.5$ and comparable to it when $p > 1.5$. In this paper, modification is made to the line search procedure so that the GNCS method is applicable for l_p problems with $1 \le p < \infty$. The global convergence results for l_1 problems are obtained under weaker assumptions than required in [2]. In addition, the usefulness of l_p norm solution with $1 \le p \le 2$ is demonstrated by applying the GNCS algorithm to a synthetic geophysical tomographic inversion problem. Additional numerical results are included to support the efficiency of GNCS.

Key Words. linear regression, discrete estimation, tomographic inversion, IRLS, GNCS, linear programming, Newton method

Subject Classification. AMS/MOS: 65H10, 65K05, 65K10.

1. Introduction. In empirical science, the following discrete approximation problem is often solved

$$\min_{x\in\Re^n} \|A^T x - b\|_p^p. \tag{1.1}$$

The least squares solution ($p = 2$) has been mostly used because it is an application of the principle of maximum likelihood under the normal distribution assumption. The l_1 solution has shown to be an increasingly attractive alternative to l_2 since the l_1 solution has the maximum resistance property: a small number of isolated large errors do not usually change the solution (e.g.,[1], [10],[13]). However, the l_1 solution may not be unique while there exists exactly one l_p solution for any l_p problem with $p > 1$. Thus it may be desirable to obtain the l_p solution with $1 < p < 2$.

As an example, consider a synthetic geophysical tomographic inversion problem described in [10] (See §5 for details). The exact solution l_p solution with $1 \le p < \infty$ when no error is present is given in Figure 1. Now assuming both large spike errors and small Gaussian noise in the data, the least squares and l_1 solutions are shown in Figure 2 and Figure 3 respectively. Compared to Figure 1, it is clear that the l_1 solution is less affected by the large spike error and thus is preferable.

---

* Computer Science Department, Cornell University, Ithaca, New York, 14853. Research partially supported by the Applied Mathematical Sciences Research Program (KC-04-02) of the Office of Energy Research of the U.S. Department of Energy under grant DE-FG02-86ER25013.A000, and in part by NSF, AFOSR, and ONR through grant DMS-8920550, and by the Cornell Theory Center which receives major funding from the National Science Foundation and IBM Corporation, with additional support from New York State and members of its Corporate Research Institute.

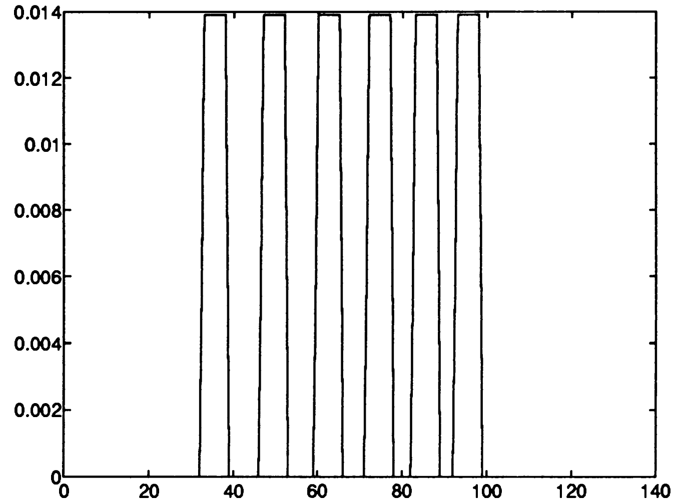

FIG. 1. *Exact Solution for a Synthetic Geophysical Tomographic Inversion Problem*

The l_p norm solution with $1 < p < 2$ has been used less often in practice. Nontheless, the l_p solutions, with $1 < p < 2$, can be useful when one seeks a solution with some degree of resistance property and the character of least squares solution simultaneously(e.g, [10], [14]). For the synthetic seismic tomographic inversion problem considered, the l_p solutions, with p close to unity, have similar error resistance properties (see Figure 4 and 5).

The IRLS method [9] has been widely used to solve (1.1), particularly in geophysical data processing. Since the choice of the parameter p may need to be determined experimentally [11], the fact that the IRLS method can solve l_p minimization problems for $1 \leq p < 2$ is certainly an advantage. Even for l_1 problems for which many LP-related methods exist, IRLS can still be more desirable because it works on the original problem directly. This is particularly true when a crude approximate solution is satisfactory. Moreover, the l_p norm, $1 \leq p < 2$, is decreased at each iteration and a readily available good starting point can be immediately exploited.

Although the IRLS method has been widely used, it can be extremely slow when p is close to unity (e.g., [9] and [5]). The slow convergence of IRLS, when p is close to unity, is not due to the lack of line search. Indeed, an IRLS method with a line search (IRLSL) implemented in [5] has similar slow convergence behavior, though small improvements over IRLS are observed. Moreover, numerical examples illustrate that the methods are slow globally ([5]. see also §4).

Based on the complementary slackness conditions, a new globally convergent Newton algorithm called GNCS is proposed in [5]. It uses the same line search procedure of IRLSL. Compared to IRLSL, the method has better convergence property and better practical performance. The method is computationally very similar to the IRLS method and retains all the advantages of the IRLS method over other approaches.

In this paper, we first briefly review the GNCS method (§2). A modification is made to the

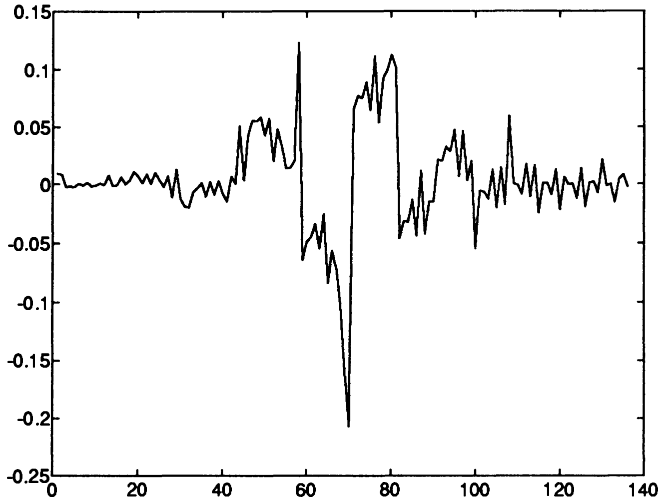

FIG. 2. l_2 *Solution Least Squares Solution with Spike and Gaussian Erros*

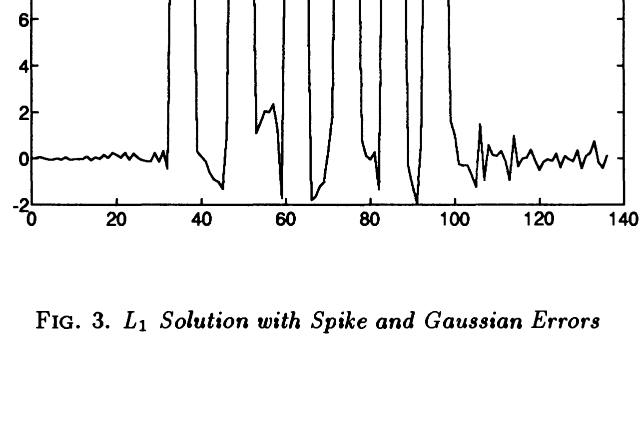

FIG. 3. L_1 *Solution with Spike and Gaussian Errors*

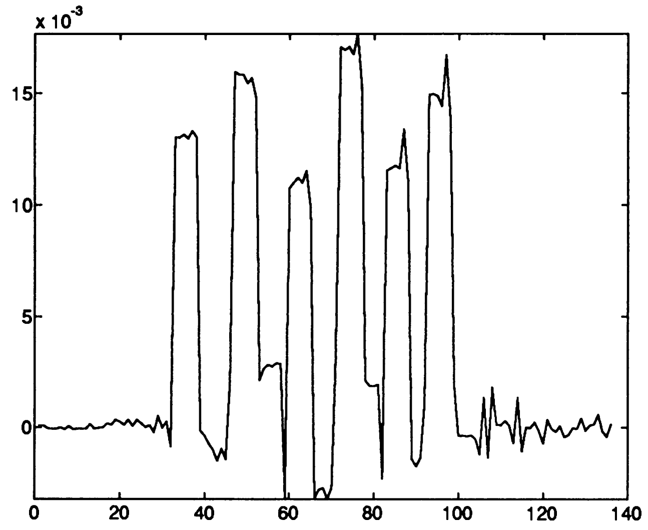

FIG. 4. l_p *Solution,* $p = 1.2$*, with Spike and Gaussian Errors*

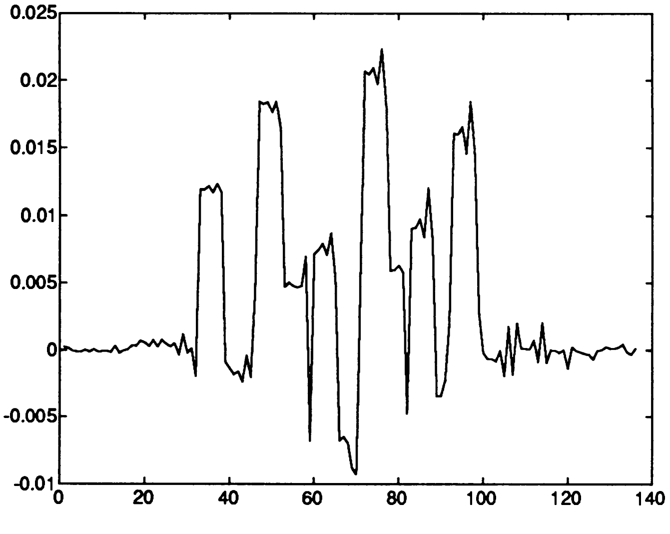

FIG. 5. l_p *Solution,* $p = 1.3$*, with Spike and Gaussian Errors*

simple line search procedure adopted in [5] to extend the GNCS method for l_p problems with $1 \le p < \infty$. In §3, a global convergence result of the GNCS method, in the case of $p = 1$, is obtained under a weaker condition than the one assumed in [2]. In §3, we report some of our numerical experience with the method for large sparse problems and demonstrate its advantages over the IRLS approach. Finally, in §5, the GNCS method is applied to a synthetic geophysical tomographic inversion problem to investigate the l_p norm solution in the presence of a few large erratic data and Gaussian noise and thus demonstrate the usefulness of the l_p norm solution when $1 \le p \le 2$.

2. The GNCS method. The motivation of the GNCS method has been described in [5] and [2]. In this section, we briefly summarize the basic components of the method. Modifications are made to the line search procedure so that GNCS solves l_p problems with $1 \le p < \infty$.

Assume that A has full rank. Let the rows of the matrix Z form a basis for the null space of A, i.e., $AZ^T = 0$. The problem (1.1) is equivalent to the following

$$\begin{aligned} &\min_{r \in \Re^m} \{\phi(r) \;=\; \|r\|_p^p\} \\ \text{(2.1)} \qquad &\text{subject to} \quad Zr = Zb. \end{aligned}$$

This formulation is used because of the following reason: the points of nondifferentiability (either first or second order) are simply expressed as $r_i = 0$ for some i. Notice that Z is not needed computationally.

The objective of GNCS is to derive a method which is efficient for all l_p problems with p ranging from 1 to ∞. However, the problems when $p = 1$ and $p > 1$ are very different mathematical problems. The objective function is piecewise differentiable when $p = 1$ while continuously differentiable for $p > 1$. This is reflected by the different optimality conditions in the two cases. For $p > 1$, the optimality conditions for (2.1) formulation are simply that there exists $w \in \Re^{m-n}$ such that

$$\nabla\phi(r) - Z^T w = 0 \quad \text{and} \quad Zr = Zb. \tag{2.2}$$

When $p = 1$, the optimality conditions are

$$\mathrm{diag}(r)(\nabla\phi(r) - Z^T w) = 0, \quad Zr = Zb \tag{2.3}$$

and $|\lambda = Z^T w| \le e$. We use e to denote the vector with each component equal to unity and $\mathrm{diag}(r)$ represents a diagonal matrix with the components of r on the diagonal. The condition is referred to as the complementarity condition. Let $g \stackrel{\text{def}}{=} \nabla\phi(r)$. If, for every $r_i = 0$, $\lambda_i \ne g_i$, strict complementarity is said to be satisfied. It is easy to see that the multipliers w (and thus λ) are uniquely defined at points satisfying the strict complementarity condition.

In order to develop a method which solves the l_p norm problems for all p efficiently, one system of equations is chosen as a template to determine suitable descent directions for all p. The IRLS method chooses the system (2.2) to determine such directions [5]. But (2.2) does not include the optimality condition for $p = 1$. This explains the inefficiency of the IRLS method for p close or equal to unity.

The key observation which motivates the development of GNCS is to use the system (2.3). The advantage of using (2.3) is the fact that, in addition to being part of the optimality condition for $p = 1$, (2.3) is the optimality condition for $p > 1$ if there are no zero residuals at the solution. In this case, a method based on the Newton directions defined by (2.3) will be quickly convergent if suitable stepsizes are taken. When there are zero residuals at the solution for $1 < p < \infty$, convergence is achieved through a globalization technique described below. When $1 < p < 2$ and there exist zero residuals at the solution, it is not clear how one can achieve fast convergence since the function $\phi(r)$ is nonlinear and the Hessian matrix is not defined at the solution. In this case, it is reasonable to achieve linear convergence. When $p > 2$, the globalization technique ensures fast convergence if the Hessian matrix of $\psi(x) \stackrel{\text{def}}{=} \|A^T x - b\|_p^p$ is nonsingular.

In this paper, we define the multiplication and division of two vectors are to be interpreted as componentwise operations. Let $D_r = \text{diag}(|r|)$, $\lambda = Z^T w$ and $D_\lambda = \text{diag}(\text{sgn}(r)(pg - \lambda))$. The Newton step for (2.3) is defined by

$$(2.4) \qquad [D_\lambda, -D_r Z^T] \begin{bmatrix} d \\ d_w \end{bmatrix} = -D_r(g - Z^T w) \quad \text{and} \quad Zr = Zb.$$

Alternatively, let $d = A^T d_x$ and we have

$$(2.5) \qquad [D_\lambda A^T, -D_r Z^T] \begin{bmatrix} d_x \\ d_w \end{bmatrix} = -D_r(g - Z^T w).$$

Hence

$$(2.6) \qquad d = -A^T (A D_\lambda D_r^{-1} A^T)^{-1} A g.$$

It is not difficult to show that the Newton direction defined by (2.6) is a descent direction for the objective function $\phi(r)$ when sufficiently close to the solution ([2],[5]). However, it may not be a descent direction far away from the solution. Thus, a globalization technique is required.

Since the IRLSL method is globally convergent [9], one way to achieve global convergence is to combine the Newton direction (2.6) with the IRLS direction.

It can be easily verified [5] that the IRLS direction is $d = A^T d_x$ where

$$(2.7) \qquad [\text{diag}(|g|) A^T, -D_r Z^T] \begin{bmatrix} d_x \\ d_w \end{bmatrix} = -D_r(g - Z^T w).$$

By observing the similarity between (2.7) and (2.6), a globalization technique is formed by introducing a diagonal matrix D_θ whose diagonal entries are componentwise convex combinations of the components of $\text{diag}(g)$ and D_λ:

$$(2.8) \qquad D_\theta = \text{diag}(|\theta g + (e - \theta)(pg - \lambda)|)$$

Here θ is a vector. Furthermore,

$$\theta = \frac{\eta e}{\gamma|g| + \eta e}, \quad \eta = \max(\max(\frac{|D_r(g-\lambda)|}{\phi(r_0)}), \max(\max(|\lambda| - |g|, 0))) \tag{2.9}$$

where $0 < \gamma < 1$ is a constant. We use the Matlab [7] definition for the function max: $\max(x)$ denotes the maximum component of a vector x and the value of $\max(x, y)$ is a vector whose components are the maximum of the corresponding components of x and y.

The following lemma can be easily established from the definition of D_θ and θ (e.g., [5]). It indicates that the diagonal matrix D_θ behaves like $\mathrm{diag}(|g|)$ globally.

LEMMA 1. *Suppose* $0 < \gamma < 1$ *and* $1 \leq p < \infty$. *Assume* D_θ *is defined by (2.8). Then* D_θ *satisfies*

$$(p - 1 + (1-\gamma)\min(\theta))\mathrm{diag}(|g|) \leq |D_\theta| \leq (p+1)\mathrm{diag}(|g|). \tag{2.10}$$

To complete the description for the algorithm, we only need to discuss our line search strategy. The line search technique used in [5] exploits the special convex structure of the objective function $\phi(r)$ and the property of the descent direction determined by the GNCS algorithm. It is a finite procedure when $1 \leq p < 2$.

Let $\mathcal{J}$ denote the ordered (ascent) steplengths which result in zero residuals moving along a descent direction d_k, i.e.,

$$\mathcal{J} = \{\beta_k^i : \beta_k^i = -\frac{r_{jk}}{d_{jk}}, \ r_{jk}d_{jk} < 0, \ \beta_k^0 \leq \beta_k^1 \leq \cdots, \beta_k^l\}. \tag{2.11}$$

We call $\{\beta_k^i\}$ breakpoints. For notational convenience, we define $\beta_k^0 \stackrel{\text{def}}{=} 0$.

Let $\breve{\alpha}_k$ denote the optimal stepsize along d_k for the following quadratic interpolation function $U_k(r)$

$$U_k(r) = \frac{1}{2}\sum_{i=1}^{m} \frac{|g_{ik}|}{|r_{ik}|} r_i^2 + \sum_{i=1}^{m}(|r_{ik}|^p - \frac{1}{2}\frac{|g_{ik}|}{|r_{ik}|}|r_{ik}|^2),$$

where $g_k = \nabla\phi(r_k)$. Hence

$$\breve{\alpha}_k = -\frac{{g_k}^T d_k}{{d_k}^T \mathrm{diag}(p|r_k|^{p-2}) d_k}.$$

From Lemma 1 and

$$(AD_{\theta k} D_{r_k}^{-1} A^T) d_{xk} = -Ag_k,$$

it is straightforward to verify that, for $1 \leq p < 2$,

$$(p - 1 + (1-\gamma)\min(\theta_k)) \leq \breve{\alpha}_k \leq (p+1). \tag{2.12}$$

In addition, it can be shown that [5], for $1 \le p < 2$,

$$\phi(r_k + \breve{\alpha}_k d_k) < \phi(r_k) + \frac{1}{2} g_k^T d_k$$

Hence, for $1 < p < 2$, the sequence $\{\breve{\alpha}_k\}$ is uniformly bounded away from zero and produces a sufficient decrease of the objective function. However, when $p > 2$, an increase of the objective function may occur with this stepsize because $U(r)$ is not a dominant quadratic function for $\phi(r)$ when $p > 2$.

We define $\breve{\alpha}_k$ for $p > 2$ as follows.

$$\breve{\alpha}_k = -\frac{1}{p} \times \frac{g_k{}^T d_k}{d_k{}^T \mathrm{diag}(p(|r_k|)^{p-2}) d_k}. \tag{2.13}$$

Given a small constant $\epsilon > 0$. Define β_k^* to be the first breakpoint, in the interval $(\min\{(p-1)\epsilon, \breve{\alpha}_k\}, \alpha_{\max})$, at which the direction d_k ceased to be descent:

$$\beta_k^* \stackrel{\text{def}}{=} \min\{\beta_k^i : \ \min\{(p-1)\epsilon, \breve{\alpha}_k\} < \beta_k^i < \alpha_{\max} \text{ and } g(r_k + \beta_k^i d_k)^T d_k \ge 0\}.$$

Here a large upper bound $\alpha_{\max} > \frac{p+1}{p}$ is imposed on β_k^* and β_k^* is defined to be $\alpha_{\max}$ if no such β_k^i exists. This definition of β_k^* is different from the one used in [5] where it was required that $\beta_k^* > \breve{\alpha}_k$. We find this new definition computationally more efficient because it allows α_k to be better approximation to the exact minimum.

Assume $0 < \rho < 1$. The stepsize is determined in the following way:

$$\alpha_k = \mathbf{first}(\alpha \in \{\beta_k^*, 1, \breve{\alpha}_k, \rho\breve{\alpha}_k, \rho^2\breve{\alpha}_k \cdots\} \ : \ \phi(r_k + \alpha d_k) < \phi(r_k) + \frac{p-1}{p}\beta_f \alpha g_k^T d_k), \tag{2.14}$$

i.e., α_k equals the first stepsize among the ordered set (by appearance) $\{\beta_k^*, 1, \breve{\alpha}_k, \rho\breve{\alpha}_k, \rho^2\breve{\alpha}_k \cdots\}$ such that the sufficient decrease condition is satisfied. Notice that, for $p = 1$, β_k^* is always taken. If $1 \le p < 2$, $\alpha_k \le \breve{\alpha}_k$ always gives sufficient decrease and $\phi(r)$ is evaluated at most three times. When $p > 2$, more evaluations may be required but our computation experience shows that $\breve{\alpha}_k$, defined above, often produces a sufficient decrease.

The points with exact zero residual should be avoided since the objective function $\phi(r)$ may be nondifferentiable at such points. We achieve this by applying a step-back technique, when necessary:

$$\alpha_k \leftarrow \alpha_k^- + \tau_k(\alpha_k - \alpha_k^-) \tag{2.15}$$

where $\alpha_k^- = \max(\beta_k^i < \alpha_k)$. The amount of backtracking is determined by τ_k which is given as follows

$$\tau_k = \max(\tau, 1 - \frac{\eta_k}{\gamma + \eta_k}). \tag{2.16}$$

Here τ is a constant (e.g., 0.975). It is clear the backtracking distance converges to zero as the iterates converge to the solution, i.e., when $\{\eta_k\}$ converges to zero. Moreover, it is easy to verify that the same sufficient decrease condition is satisfied at the perturbed α_k due to convexity of $\phi(r)$.

A summary of the GNCS algorithm is given in Figure 6. We want to emphasize that the only difference between IRLSL and GNCS is the definition of scaling matrix D_k. In fact, there is only one line difference in the actual codes. Moreover, the main computation is in **Step 2**: a weighted least squares solve. For $1 < p < 2$, the line search (**Step 3**) is a finite procedure and takes, in the worst case $O(m)$ operations. For $p > 2$, we observe that the stepsize $\breve{\alpha}_k$ usually produces sufficient decrease. The weighted least squares solutions d_{x_k} can be obtained either through a direct solve or iterative solve. Notice that, for l_1 problems, the above line search procedure is slightly different from the exact line search technique used in [2].

3. Convergence Properties. The convergence property of the GNCS algorithm has been studied in [5] for $1 < p < 2$ and [2] for $p = 1$. Although the definition of β_k^* is slightly different from the one used in [5], it is easy to see that the convergence result established in [5] holds because $\{\beta_k^*\}$ is still uniformly bounded away from zero. For $p > 2$, the function $\phi(r)$ is at least twice continuously differentiable and the convergence theory follows from the conventional unconstrained minimization convergence analysis [3]. Hence, for $1 < p < \infty$, $\{r_k\}$ converges to r^*. Moreover, if $\{a_i : r_i = 0\}$ are linearly independent at r^*, $\{\lambda_k\}$ converges and $\{r_k\}$ converges to the solution.

The global converge results for l_1 problems, however, are obtained in [5] under fairly restrictive nondegeneracy assumptions. Even convergence to a vertex is established under the condition that an l_1 problem is both primal and dual nondegenerate at every vertex. In this section, we establish the global convergence results of GNCS for l_1 under weaker assumption. Assume A is of full rank n. We prove that the sequence $\{\phi(r_k)\}$ converges and every limit point satisfies the complementary slackness condition, i.e., $\mathrm{diag}(r)(g - Z^T w) = 0$ for some w. If we further assume that $\{\lambda_k\}$ converges, every limit point of $\{r_k\}$ is a minimum of $\phi(r)$.

We assume that $p = 1$ for the rest of this section. The following recent result of Stewart [12] is used.

THEOREM 2. *Let $\mathcal{D}$ denote the set of all positive definite $m \times m$ real diagonal matrices. Let A be an $m \times n$ real matrix of rank n. Then there exist constants χ_A and $\bar{\chi}_A$ such that for any $D \in \mathcal{D}$,*

1. $\|(A^T D^{-1} A)^{-1} A^T D^{-1}\| \le \chi_A$, *and*
2. $\|A(A^T D^{-1} A)^{-1} A^T D^{-1}\| \le \bar{\chi}_A$.

Let P_k be the orthogonal projector onto the orthogonal space of ZD_k, i.e.,

$$P_k = I - D_k Z^T (Z(D_k)^2 Z^T)^{-1} Z D_k$$

Given an initial point $r_0 = A^T x_0 - b$ *with* $|r_0| > 0$, λ_0, ρ, γ, $0 < \beta_f < \frac{1}{2}$, *a small constant* ϵ *and a large constant* $\alpha_{\max} > \frac{1+p}{p}$
for $k = 1, 2, \cdots$ **do**

Step 1. Compute $g_k = p(|r_k|)^{p-1}\mathrm{sgn}(r_k)$ and

$$\theta_k = \frac{\eta_k e}{\gamma |g_k| + \eta_k e}, \quad \eta_k = \max(\max(\frac{|D_{rk}(g_k - \lambda_k)|}{\phi(r_0)}), \max(\max(|\lambda_k| - |g_k|, 0)))$$

Let $D_{rk} = \mathrm{diag}(|r_k|)$, $D_{\theta k} = \mathrm{diag}(|pg_k - (e - \theta_k)\lambda_k|)$ and define $D_k = (D_{rk} D_{\theta}{}_k^{-1})^{\frac{1}{2}}$;

Step 2. Compute the direction d_k by

$$\begin{cases} D_k^{-1} A^T dx_k \overset{\text{l.s.}}{=} -D_k g_k \\ d_k = A^T dx_k; \end{cases}$$

Update : $\lambda_{k+1} \leftarrow D_{r}{}_k^{-1} D_{\theta k} d_k + g_k$;

Step 3. Compute $\breve{\alpha}_k = -\frac{g_k{}^T d_k}{p d_k{}^T \mathrm{diag}(p(|r_k|)^{p-2}) d_k}$; Compute breakpoints $\{\beta_k^i\}$ and let $\beta_k^* = \min\{\beta_k^i : \min\{(p-1)\epsilon, \breve{\alpha}_k\} < \beta_k^i < \alpha_{\max}$ and $g(r_k + \beta_k^i d_k)^T d_k \geq 0\}$; Set $\alpha_k = \mathbf{first}(\alpha \in \{\beta_k^*, 1, \breve{\alpha}_k, \rho\breve{\alpha}_k, \rho^2\breve{\alpha}_k \cdots\} : \phi(r_k + \alpha d_k) < \phi(r_k) + \frac{p-1}{p}\beta_f \alpha g_k^T d_k)$;
Compute τ_k by

$$\tau_k = \max(\tau, 1 - \frac{\eta_k}{\gamma + \eta_k})$$

and perturb α_k according to $\alpha_k \leftarrow \beta_k^- + \tau_k(\alpha_k - \alpha_k^-)$ if necessary; Update: $x_{k+1} \leftarrow x_k + \alpha_k d_k$;

FIG. 6. *The GNCS Algorithm*

where D_k equals $(|D_{rk}D_{\theta k}{}^{-1}|)^{\frac{1}{2}}$. Then

$$
(3.1) \qquad \begin{aligned} d_k &= -A^T(A(D_k)^{-2}A^T)^{-1}Ag_k \\ &= -D_kP_kD_kg_k \\ &= -(D_k)^2(g_k - \lambda_{k+1}), \end{aligned}
$$

where $\lambda_{k+1} = Z^T w_{k+1}$ and w_{k+1} is the least squares solution to

$$
D_k Z^T w_{k+1} \stackrel{\text{l.s.}}{=} D_k g_k.
$$

Using Stewart's theorem we immediately have the following result.

LEMMA 3. *The sequences* $\{w_k\}$ *and* $\{\lambda_k\}$ *are bounded, i.e., there exists* $\chi_\lambda > 0$ *such that* $\|w_k\| \le \chi_\lambda$ *and* $\|\lambda_k\| \le \chi_\lambda$.

LEMMA 4. *The sequences* $\{d_k\}$ *and* $\{\alpha_k d_k\}$ *converge to zero if* $\lim_{k\to\infty} \inf\{\eta_k\} > 0$.

Proof. Since $p = 1$, the objective function $\phi(r)$ is piecewise linear. Therefore

$$
\begin{aligned} \phi(r_k) - \phi(r_0) &\le \sum_{j=0}^{k-1}(\phi(r_j + (1-\tau_k)\min\{\beta_j^1, \alpha_{\max}\}d_j) - \phi(r_j)) \\ &\le \sum_{j=0}^{k-1} \min(1-\tau, \eta_k)\min\{\beta_j^1, \alpha_{\max}\}{g_j}^T d_j. \end{aligned}
$$

By assumption, there exists $\chi_\eta > 0$ such that $\eta_k > \chi_\eta$ for sufficiently large k. Using definition (2.9) of θ_k and the boundedness of $\{\lambda_k\}$, there exists $\chi_\theta > 0$ such that $\theta_k > \chi_\theta$ for sufficiently large k.

From (2.10), we have $|D_{\theta k}| \ge ((1-\gamma)\chi_\theta)I$. Hence

$$
(3.2) \qquad |g_k - (e - \theta_k)\lambda_k| \ge (1-\gamma)\chi_\theta.
$$

Using (2.11) and (3.1),

$$
\beta_k^1 = \frac{|g_{j_k} - (1-\theta_{j_k})\lambda_{j_k}|}{|g_{j_k} - \lambda_{j_{k+1}}|}.
$$

Hence $\beta_k^1 \ge \frac{(1-\gamma)\chi_\theta}{(1+\chi_\lambda)}$, we conclude that $\lim_{k\to\infty} g_k^T d_k = 0$. This is equivalent to $\lim_{k\to\infty} \|P_kD_kg_k\| = 0$.

The boundedness of $\{D_k\}$ follows directly from the definition $D_k = (|D_{rk}D_{\theta k}^{-1}|)^{\frac{1}{2}}$, (3.2) and the boundedness of $\{r_k\}$. Since $d_k = -D_kP_kD_kg_k$ and $\lim_{k\to\infty} P_kD_kg_k = 0$, we conclude that $\lim_{k\to\infty} d_k = 0$. Moreover, from the line search assumption, we have $\alpha_k \le \alpha_{\max}$. Hence $\lim_{k\to\infty} \alpha_k d_k = 0$. ∎

Next we establish the two main results of this section. Theorem 5 establishes that any limit point is a vertex and Theorem 6 shows that any limit point is a minimum assuming the convergence of $\{\lambda_k\}$.

THEOREM 5. *The sequence $\{\phi(r_k)\}$ converges. Moreover, at any limit point $\bar{r}$ of $\{r_k\}$, there exists $\bar{w}$ such that* $\mathrm{diag}(\bar{r})(\bar{g} - Z^T\bar{w}) = 0$.

Proof. We prove the results by considering two separate cases.

First we consider the case that zero is a limit point of $\{\eta_k\}$. This means that there exists a subsequence r_{l_k} converging to the solution r^*. Since $\{\phi(r_k)\}$ is monotonically decreasing, this means that $\{\phi(r_k)\}$ converges to the minimum value $\{\phi(r^*)\}$. Therefore, any limit point $\bar{r}$ is a minimum. Hence there exists $\bar{w}$ such that $\mathrm{diag}(\bar{r})(\bar{g} - Z^T\bar{w}) = 0$.

Now let us assume that zero is not a limit point of $\{\eta_k\}$. Since $\{r_k\}$ is bounded and, using Lemma 4, $\{\alpha_k d_k\}$ converges to zero, the limit points of $\{r_k\}$ form a closed connected set (see [8]). Since $\{\phi(r_k)\}$ is monotonically decreasing, bounded below and $\phi(r)$ is continuous, there exists $\bar{r}$ such that $\lim_{k\to\infty}\phi(r_k) = \phi(\bar{r})$. Using Lemma 4, $\{d_k\}$ converges to zero. The equation (3.1) immediately implies that, at any limit point $\bar{r}$ of $\{r_k\}$, there exists $\bar{w}$ such that $\mathrm{diag}(\bar{r})(\bar{g} - Z^T\bar{w}) = 0$. ∎

THEOREM 6. *Assume that $\{\lambda_k = Z^T w_k\}$ converges to $Z^T\bar{w}$. Then any limit point of $\{r_k\}$ is a solution of (2.1).*

Proof. Assume that $\bar{r}$ is a limit point of $\{r_k\}$. From Theorem 5, there exists $\bar{\lambda} = Z^T\bar{w}$ such that $\mathrm{diag}(\bar{r})(\bar{g} - \bar{\lambda}) = 0$. This means, by assumption, that the sequence $\{\lambda_k\}$ converges to $\bar{\lambda}$. Therefore, there exists a constant K such that λ_{ik} remains the same sign for $k > K$ if $\bar{\lambda}_k \neq 0$.

Let $\mathcal{I}^0 \stackrel{\text{def}}{=} \{i \;:\; \bar{r}_i = 0\}$ and $\mathcal{I}^1 \stackrel{\text{def}}{=} \{i \;:\; \bar{r}_i \neq 0\}$. To prove that $\bar{r}$ is a solution, we need to show $|\bar{\lambda}_i| \leq 1, \forall i \in \mathcal{I}^0$. We prove this by contradiction.

Assume the existence of $i^0 \in \mathcal{I}^0$ such that $|\bar{\lambda}_{i^0}| > 1$. Then the r_{i^0} is zero at every limit point since the complementarity condition is satisfied. In other words, $\{r_{i^0 k}\}$ converges to 0. We consider two cases.

Assume that there exists a $k > K$ with $r_{i^0 k}\lambda_{i^0 k} > 0$. Then $g_{i^0 k}(g_{i^0 k} - \lambda_{i^0 k+1}) < 0$. Hence $d_{i^0 k} r_{i^0 k} > 0$ and $|r_{i^0 k+1}| > |r_{i^0 k}|$. This is impossible because $\bar{r}_{i^0} = 0$.

The only other possibility is that, for all $k > K$, $r_{i^0 k}\lambda_{i^0 k} < 0$. But we prove next that $r_{i^0 k+1}\lambda_{i^0 k+1} > 0$ for some sufficiently large $k > K$ which is a contradiction.

By definition (2.11) and (3.1), a breakpoint is given by the following formula:

$$\beta_k^i = \frac{|r_{jk}g_{jk} - (1-\theta_{jk})\lambda_{jk}|}{|r_{jk}(g_{jk} - \lambda_{jk+1})|} = \frac{|g_{jk} - (1-\theta_{jk})\lambda_{jk}|}{|g_{jk} - \lambda_{jk+1}|}.$$

Notice that, $\theta_{kj} = \theta_{ki}$ for $j \neq i$. For notation simplicity, we regard θ_k as a scalar. We partition the set $\mathcal{I}^0$ as follows

$$\mathcal{I}_k^{\text{small}} \stackrel{\text{def}}{=} \{i \in \mathcal{I}^0 \;:\; |\bar{\lambda}_i| > 1 \textbf{ or } (|\bar{\lambda}_i| = 1 \textbf{ and } \bar{\lambda}_i g_{ik} < 0)\}$$

$$\mathcal{I}_k^{\text{large}} \stackrel{\text{def}}{=} \{i \in \mathcal{I}^0 \ : \ |\bar\lambda_i| < 1 \textbf{ or } (|\bar\lambda_i| = 1 \textbf{ and } \bar\lambda_i g_{ik} > 0)\}.$$

From (3.1), $g_i^k d_i^k < 0$ for all $i \in \mathcal{I}_k^{\text{small}}$. (Recall that $\bar\lambda_i g_{ik} < 0$, for $i \in \mathcal{I}_k^{\text{small}}$.)

From the fact that $0 < \bar\theta < 1$, it can be easily proved that, for $|\bar\lambda_i| \geq 1$ and $|\bar\lambda_j| < 1$

$$\frac{1 + (1-\bar\theta)|\bar\lambda_i|}{1 + |\bar\lambda_i|} < \frac{1 \pm (1-\bar\theta)\bar\lambda_j}{1 \pm \bar\lambda_j}.$$

Moreover, for $|\bar\lambda_i| \geq 1$ and $|\bar\lambda_j| = 1$, then for k sufficiently large, if $g_{ik}\bar\lambda_i < 0$ and $g_{j_k}\bar\lambda_j > 0$,

$$\frac{1 + (1-\theta_k)|\lambda_{ik}|}{1 + |\lambda_{ik+1}|} < \frac{1 - (1-\theta_k) g_{j_k}\lambda_{j_k}}{1 - g_{j_k}\lambda_{j_{k+1}}}$$

and

$$\frac{1 + (1-\theta_k)|\lambda_{ik}|}{1 + |\lambda_{ik+1}|} < 1 < \alpha_{\max}.$$

Hence, when k is sufficiently large, the breakpoints in $\mathcal{I}_k^{\text{large}}$ will be larger than the breakpoints in $\mathcal{I}_k^{\text{small}}$.

In addition, from $A\bar\lambda = 0$, we obtain

$$A_{\mathcal{I}^1 k} g_{\mathcal{I}^1 k} = -A_{\mathcal{I}^0} \bar\lambda_{\mathcal{I}^0}. \tag{3.3}$$

By definition, $d_k = -A^T dx_k$. Thus

$$\begin{aligned} {g_{\mathcal{I}^1 k}}^T d_{\mathcal{I}^1 k} &= \bar g_{\mathcal{I}^1}^T A_{\mathcal{I}^1}^T dx_k && (\text{since } g_{\mathcal{I}^1 k} = \bar g_{\mathcal{I}^1}) \\ &= -\bar\lambda_{\mathcal{I}^0}^T A_{\mathcal{I}^0}^T dx_k && (\text{from (3.3)}). \end{aligned}$$

Hence

$$ {g_{\mathcal{I}^1 k}}^T d_{\mathcal{I}^1 k} = -\bar\lambda_{\mathcal{I}^0}^T d_{\mathcal{I}^0 k}. \tag{3.4}$$

Now let g_+ denote the new gradient after passing all the breakpoints in $\mathcal{I}_k^{\text{small}}$. Then

$$\begin{aligned} g_+^T d_k &= -\sum_{i \in \mathcal{I}_k^{\text{small}}} g_{ik} d_{ik} + \sum_{j \in \mathcal{I}_k^{\text{large}}} {g_{j_k}}^T d_{j_k} + {g_{\mathcal{I}^1 k}}^T d_{\mathcal{I}^1 k} \\ &= -\sum_{i \in \mathcal{I}_k^{\text{small}}} g_{ik} d_{ik} + \sum_{j \in \mathcal{I}_k^{\text{large}}} {g_{j_k}}^T d_{j_k} - \bar\lambda_{\mathcal{I}^0}^T d_{\mathcal{I}^0 k} \qquad (\text{from (3.4)}) \\ &= \sum_{i \in \mathcal{I}_k^{\text{small}}} (-\bar\lambda_i - g_{ik}) d_{ik} + \sum_{j \in \mathcal{I}_k^{\text{large}}} (g_{j_k} - \bar\lambda_j) d_{j_k}. \end{aligned}$$

But $-\bar\lambda_i d_{ik} \leq g_{ik} d_{ik} < 0$, for $i \in \mathcal{I}_k^{\text{small}}$. Moreover, since $g_{j_k} d_{j_k} < 0$ if $|\bar\lambda_j| < 1$, $-|\bar\lambda_j d_{j_k}| \geq g_{j_k} d_{j_k}$ for $j \in \mathcal{I}_k^{\text{large}}$ with $|\bar\lambda_j| < 1$. In addition $g_{j_k} - \bar\lambda_j = 0$ if $j \in \mathcal{I}^{\text{large}}$ and $g_{j_k}\bar\lambda_j = 1$. Hence

$$\sum_{i \in \mathcal{I}_k^{\text{small}}} (-\bar\lambda_i - g_{ik}) d_{ik} < 0 \quad \text{and} \quad \sum_{j \in \mathcal{I}_k^{\text{large}}} (g_{j_k} - \bar\lambda_j) d_{j_k} \leq 0.$$

We conclude $g_+{}^T d_k < 0$. Therefore, $r_{i^0 k+1}\lambda_{i^0 k+2} > 0$ which is contradiction. ∎

If we assume that an l_1 problem is primal and dual nondegenerate at every vertex as in [2], then $\{\lambda_k\}$ converges and the result that the limit point of $\{r_k\}$ is the solution follows directly from Theorem 6. Therefore, the assumption that $\{\lambda_k\}$ converges is weaker. In fact, from the linear programming theory, $\{\lambda_k\}$ converges if the strict complementarity condition is satisfied at every limit point of $\{r_k\}$.

Assume that a limit point $\bar{r}$ of $\{r_k\}$ is both primal and dual nondegenerate. The sequence $\{r_k\}$ converges to r^* superlinearly (see [2]). The local superlinear convergence property of the algorithm suggests efficiency of the method when a good initial guess of the solution is available. This will be demonstrated by a geophysical tomographic inversion problem later.

4. Comparison between GNCS and IRLSL. Before we examine the efficiency of the two methods, we emphasize that the computational cost of GNCS and IRLSL is equivalent: a weighted least squares solve (same structure) and the same line search procedure. The computational superiority of the GNCS method over the IRLS method, when $1 \leq p < 2$ has been indicated in [5]. In this section, we make a more detailed comparison between the two methods. Moreover, examples of the performance of the both methods when $p > 2$ are included. Some limited experience with large sparse problems is also reported.

For the results reported in this paper, the parameters required by the algorithms are set as follows:

$$\tau \leftarrow 0.975, \qquad \beta_f \leftarrow 10^{-5}, \qquad \gamma \leftarrow 0.99, \qquad \epsilon \leftarrow 0.01, \qquad \alpha_{\max} \leftarrow 10^4.$$

Moreover, we stop the computation when

$$\textbf{either } \frac{|\phi(r^{k+1}) - \phi(r^k)|}{\phi(r^{k+1})} < \tau_s \textbf{ or } \eta^k < \tau_s \textbf{ or } \text{itcount} > 50$$

where τ_s has been set to $\frac{1}{2}10^{-11}$ and *itcount* denotes the number of iterations.

All the experiments are done in Matlab [7]. Each entry of a table is generated from 10 random problem instances using Matlab random matrix generator. The maximum number of iterations and the average number of iterations for problem of each size are reported.

For completeness, we include Table 1 to compare the typical behavior of GNCS and IRLSL when $1 \leq p < 2$. When $p = 1$, IRLSL fails to compute a solution with required accuracy after 50 iterations which is the maximum number of iterations allowed. This suggests that the efficiency of GNCS is not solely due to the local superlinear convergence of GNCS.

In order to illustrate that the GNCS method is better than IRLSL even when one only seeks a crude approximation, we have run both algorithms for 5 iterations only. For $1 \leq p \leq 1.6$, the computed function values for GNCS are always strictly smaller than that of IRLSL at the end of the fifth iteration. Moreover, the closer p is to one, the better approximation the GNCS solutions are. Therefore, GNCS is better than IRLSL even in the early stage of computation. When p is

Number of Iterations ($m = 200$, $n = 100$)										
p	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9
IRLSL	50/50	27.9/30	18.2/21	15/17	12.2/15	7.1/8	8.5/12	6.4/8	6.1/8	4.5/5
GNCS	16.7/20	10.7/11	10/11	9.4/10	8.6/10	7.6/9	7.3/8	6.6/7	6.5/7	5.6/6

TABLE 1
Comparison between IRLSL and GNCS ($1 \leq p < 2$)

greater than 1.6, the computed solutions from both methods are fairly close to the solutions at the end of the fifth iteration.

Table 2 lists comparisons between IRLSL and GNCS for $p > 2$. The two methods are comparable in this case.

Number of Iterations ($m = 200$, $n = 100$)										
p	3	4	5	6	7	8	9	10	11	12
IRLSL	5.4/6	6.2/7	7.1/8	7.8/9	8.3/9	8.8/11	9.6/11	10.2/12	11.3/13	11.7/13
GNCS	6.4/8	7.1/8	7.7/9	8.4/10	8.5/9	9.1/11	10/11	10.7/13	11.5/13	11.6/13

TABLE 2
Comparison between IRLSL and GNCS ($p > 2$)

In Table 1, a least squares solution is chosen as a starting point. Theoretically, a starting point has to satisfy the requirement that there is no zero residual, i.e., $|r_0| > 0$. The GNCS method maintains this for subsequent iterations. Numerically, however, a positive constant about the size of machine precision is added to the diagonal scaling matrix D_r and D_θ to prevent numerical zero due to the inexactness of the representation of real numbers on a computer. Hence, in practical implementation, a starting point with zero residual may be allowed.

We experiment with the least squares solution and a vertex as starting points in Table 3. The vertex starting point is determined by the solving the first n equations exactly. Since the vertex starting point is chosen arbitrarily without attempting to minimize the objective function, it is not surprising that the least squares starting point is better. Indeed, the function value $\phi(r)$ of the least squares starting point is much lower than the vertex starting point we pick.

It is interesting to note, however, the reasonable good performance of the GNCS algorithm starting from an arbitrary vertex. The iterates often leave the starting vertex in one iteration and approach the solution. This will be further illustrated by the tomographic inversion problem described in the next section.

We want to emphasize that the GNCS method is a superlinearly convergent method and it can start from anywhere. Nonetheless, the GNCS algorithm benefits from a warm start, i.e., a point close to a solution. This is illustrated by the last row of Table 3 which provides the number

AVG/MAX Number of Iterations for l_1 Problems						
	(200,20)	(200,100)	(200,180)	(300,50)	(300,150)	(300,280)
ls	13.8/17	15.6/17	12.3/14	21.9/31	18.7/21	12.4/14
Vertex	18.8/23	26/33	19.2/25	28.5/36	32.7/40	20.1/25
NearOpt(0.001)	5.6/7	6.6/8	6.4/8	6.4/7	7.7/8	6.3/8

TABLE 3
Performance of GNCS with Different Starting Point

AVG/MAX Number of Iterations							
(m,n)	(p=1)	(p=1.1)	(p=1.2)	(p=1.3)	(p=1.4)	(p=1.5)	(p=1.8)
(1000,100)	7.4/8	11.6/13	10/11	8.8/10	8.2/10	7.1/8	5.7/6
(2000,100)	8/9	12/13	10/12	8.9/11	8/9	6.7/8	5.3/6
(3000,100)	8.4/10	11.8/12	9.9/12	8.6/10	7.7/8	6.6/8	5.1/6
(1000,300)	17.8/20	11.5/13	10.4/11	9.1/10	8.2/9	7.5/9	6.1/7

TABLE 4
Performance of GNCS for Larger Problems

of average and maximum iterations required by GNCS starting from a perturbed optimal point $x_0 = x_* + 0.001 * \text{rand}(n,1)$. Here $rand(n,1)$ is a Matlab function which returns a random vector in $\Re^n$.

We have also done experiments with some random large sparse problems. The results are reported in Table 4. The problems have about 1% of nonzero entries.

Our experience clearly indicates that the GNCS method is much better than the IRLS method when p is less than 1.5. In our implementation, a QR factorization is solved at each iteration to obtain a descent direction; however, it is also possible to apply an iterative procedure, e.g., conjugate gradients, as has been done for IRLS in [10].

5. An Application of l_p Norm Minimization. In this section, we compare the performance of IRLSL and GNCS for a synthetic geophysical tomographic inversion problem described in [10]. The parameter settings are the same as before.

The origin of the word "tomography" comes from the Greek word "tomos" meaning cut or slice and "graphos" meaning drawing. The objective of tomography inversion is to reconstruct the slices of an object based on wavefield observations which have passed through the object and are recorded on its surface [6]. The essential idea behind travel time tomography is based on the following fact: the travel time for a seismic ray is the integration of slowness along the ray path,

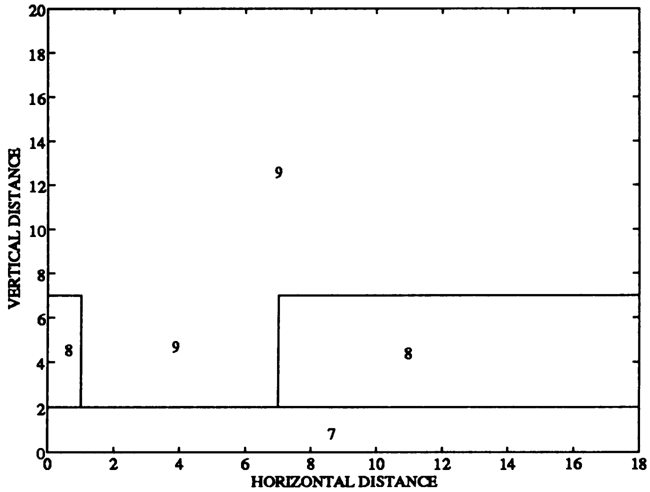

FIG. 7. *Geometry for Vertical Seismic Profile*

i.e.,

$$t(ray) = \int_{ray} s(x, y, z) d\ell$$

where $s(x, y, z)$ is slowness (reciprocal velocity) at the point (x, y, z) and $d\ell$ is the differential length along the ray. Here the ray path itself depends on slowness. The idea of linearized tomographic inversion is to assume an initial slowness model and use the difference between the computed travel times and the observed travel times to invert for the slowness perturbations. This results in an overdetermined system of equation

$$Ax = b,$$

where $b \stackrel{\text{def}}{=} \delta t$ are the differences between the travel times computed for the initial model and the observed travel times, $x \stackrel{\text{def}}{=} \delta s$ are the differences between the initial slowness model and the true slowness. The a_{ij} element of the rectangle matrix A is the distance the ith ray traveled in the jth cell. The linearization process can be repeated to produce the improved velocity models.

The overdetermined system of equations is solved by minimizing a norm of the residuals. The usual choice has been the l_2 norm. However, increasingly, the l_1 measure has shown increasingly to be advantageous (e.g., [10]). The l_p solutions, $1 < p < 2$, have been used but far less often. Our intention is to indicate, through a synthetic tomographic inversion example, that l_p norms with p between 1 and 1.5 may be eligible candidates as well.

We consider a synthetic vertical seismic profile, used in [10], in which a square velocity anomaly is to be reconstructed. Figure 7 shows the geometry of this synthetic vertical seismic profile: the square velocity anomaly appears in the middle layer. Figure 8 shows the ray illumination from 18 sources on the surface to 18 receivers on the left side of the model. Straight rays

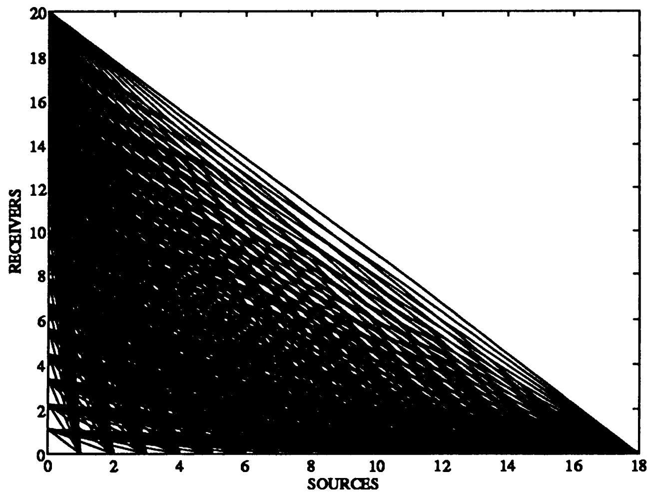

FIG. 8. *Ray illuminzation from 18 sources and 18 reveivers*

are assumed and the lower left triangular region was covered with 136 squares cells of constant slowness.

The initial guess consisted of the three layers with constant slowness, i.e., without the square anomaly in the middle layer. Figure 1 shows the graph of the exact model (see also [10] and [4]). It is not difficult to see that all the l_p solutions are equivalent to the exact model.

Next we introduce four noise spikes of the magnitude $0.9 \times \max(|b|)$. Figure 9 graphs the least squares solution in the presence of those spike errors. Figure 10 graphs the l_1 solution. The l_p solutions when 1.2, 1.3 are shown in Figure 11 and 12. The number of iterations taken are listed in Table 5. It is clear that the l_p solutions, $p = 1, 1.2, 1.3$, recover the exact solution well. The l_2 solution, on the other hand, totally distort the exact solution. Furthermore, GNCS computes a solution more quickly than IRLSL when $p = 1$.

p	1	1.1	1.2	1.3	1.4	1.5
GNCS	4	11	10	10	8	8
IRLSL	11	11	14	14	11	8

TABLE 5
Number of Iterations for Different p with Spike Errors

Now we consider the case where some Gaussian noise is present in travel times perturbation b. Assume that the noise has a normal distribution with deviation of $0.01 \times \max(|b|)$. The l_1 solution is computed by GNCS in 16 iterations while $IRLSL$ failed to compute a solution with the required accuracy after 50 iterations. The least squares solution and the l_1 solutions, computed by GNCS, are shown in Figure 14 and 13. The l_1 and l_2 solutions look very similar. Once again,

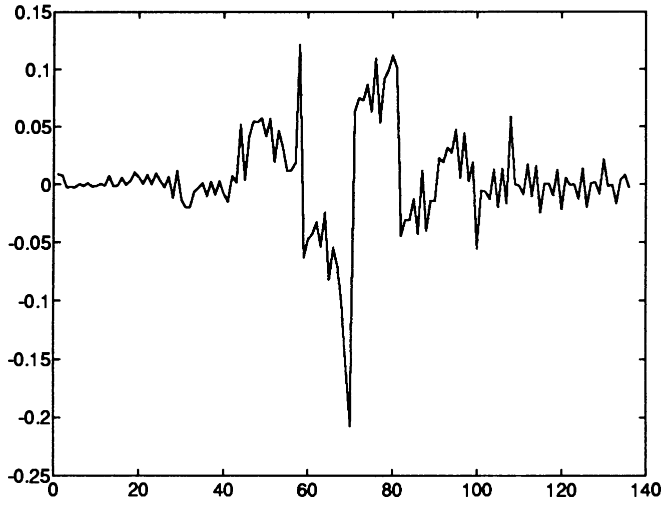

FIG. 9. *Least Squares Solution with Spike Errors*

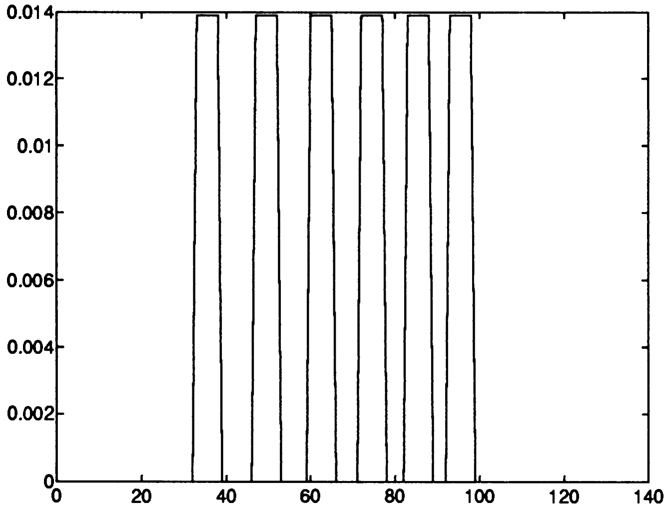

FIG. 10. L_1 *Solution with Spike Errors*

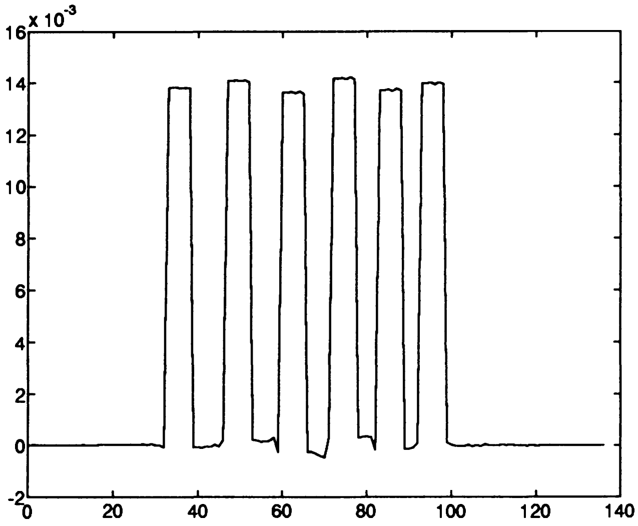

FIG. 11. *l_p Solution, $p = 1.2$, with Spike Errors*

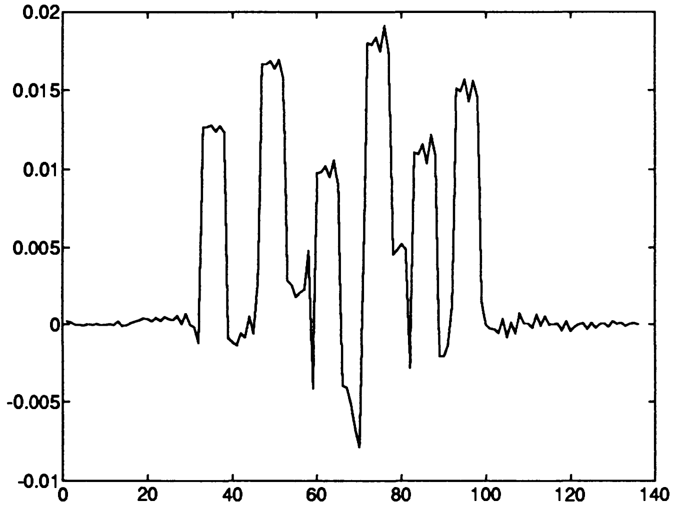

FIG. 12. *l_p Solution, $p = 1.3$, with Spike Errors*

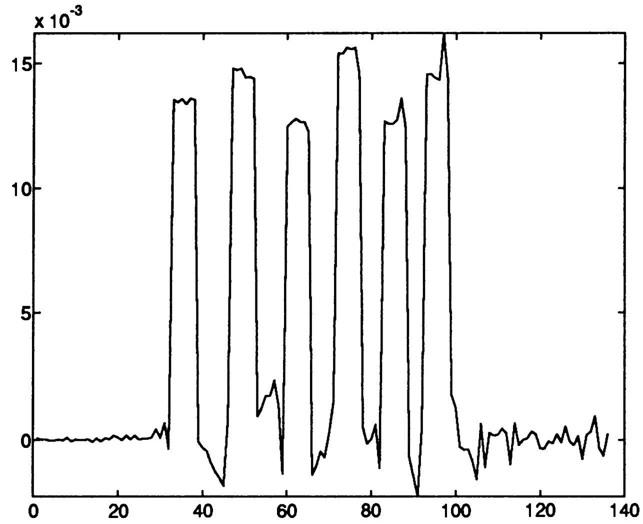

FIG. 13. L_1 *Solution with Gauss Errors*

p	1.1	1.2	1.3	1.4	1.5
GNCS	23	13	11	10	8
IRLSL	50	20	16	11	10

TABLE 6
Number of Iterations for Different p with Spike and Gaussian Errors

GNCS is more efficient than IRLSL for p close or equal to one.

Finally, we consider the more realistic situation when both the Gaussian noise and isolated spike errors described above are added simultaneously. Table 6 lists the number of iterations taken by each method. The l_1, l_2, $l_{1.3}$ solutions computed by GNCS are displayed in Figure 3, 2 and 5 respectively. The GNCS method is definitely preferable to IRLSL when $1 \leq p < 1.5$.

The synthetic geophysical tomographic inversion problem illustrates that l_p solutions, $1 \leq p \leq 1.5$, have similar error resistant quality and the closer p is to one, the more error resistant the solution is. GNCS is significantly more efficient than IRLSL for this range of p.

6. Conclusion. In this paper, we obtain the convergence of the GNCS method proposed in [5] under much less restrictive conditions. The performance comparisons between IRLSL and GNCS over a synthetic geophysical tomographic inversion problem are reported. The possible application of the l_p solutions, with $1 \leq p < 1.5$, is illustrated. Both methods are extended for l_p problems with $p > 2$.

Our computational experience for both random generated problem and synthetic geophysical tomographic inversion problem clearly indicate the superiority of GNCS over IRLSL. While the two methods are computationally cost-equivalent, GNCS computes a solution with required accuracy much faster for $1 \leq p \leq 1.5$. The performance of the two methods are comparable for

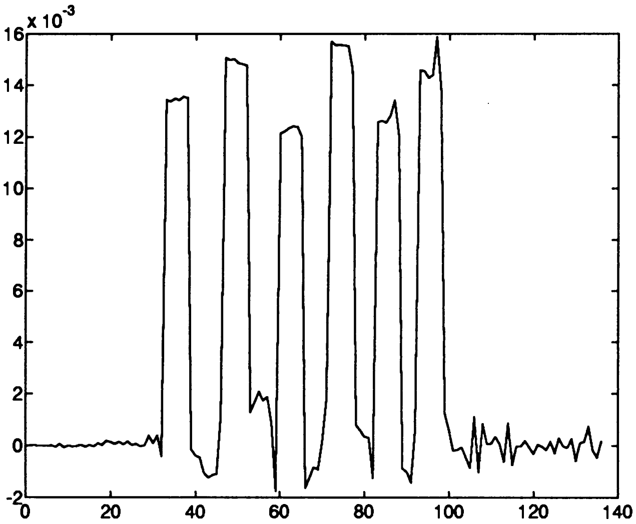

FIG. 14. *Least Squares Solution with Gauss Errors*

$p > 1.5$.

7. Acknowledgement. The author would like to thank Tom Coleman for his careful reading of the manuscript and suggestions.

REFERENCES

[1] J. F. Claerbout and F. Muir, *Robust modeling with erratic data*, Geophysics, 38 (1973), pp. 826–844.

[2] T. F. Coleman and Y. Li, *A globally and quadratically convergent affine scaling method for linear l_1 problems*, Mathematical Programming, 56 (1992), pp. 189–222.

[3] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization*, Prentice-Hall, 1983.

[4] A. G. John A. Scales and S. Treitel, *Fast l_p solution of large, sparse, linear systems application to seismic travel time tomography*, Journal of Computational Physics, 75 (1988), pp. 314–333.

[5] Y. Li, *A globally and quadratically convergent method for linear l_p problems*, SIAM Journal on Optimization, (to appear).

[6] L. R. Lines, *Applications of seismic traveltime tomograph–a review*, in Geophysical Inversion, 1992, pp. 333–337. SIAM.

[7] C. B. Moler, J. Little, S. Bangert, and S. Kleiman, *ProMatlab User's guide*, MathWorks, Sherborn, MA, 1987.

[8] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, 1970.

[9] M. R. Osborne, *Finite Algorithms in Optimization and Data Analysis*, Wiley Series in Probability and Mathematical Statistics, 1985.

[10] J. A. Scales, *Tomographic inversion via the conjugate gradient method*, Geophysics, 52 (1987), pp. 179–185.

[11] H. Sp'ath, *Mathematical Algorithms for Linear Regression*, Academic Press, 1992.

[12] G. W. Stewart, *On scaled projections and pseudoinverse*, Linear Algebra and its Applications, 112 (1989), pp. 189–193.

[13] H. L. Taylor, S. C. Banks, and J. F. McCoy, *Deconvolution with the l_1 norm*, Geophysics, (1979), pp. 39–52.

[14] R. Yarlagadda, J. B. Bednar, and T. L. Watt, *Fast algorithms for l_p deconvolution*, IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-33 (1985), pp. 174–182.