A Global and Quadratic Affine Scaling Method for Linear L_1 Problems*

Thomas F. Coleman**
Yuying Li ***

TR 89-1026
July 1989

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

---

*Written June 1989. Presented at the Third SIAM Conference on Optimization, in Boston, April 1989.
**Research partially supported by the Applied Mathematical Science Research Program (KC-04-02) of the Office of Energy Research of the U.S Dept. of Energy under grant DE-FG02-86ER25013.A000, by the U.S. Army Research Office through the Mathematical Sciences Institute, Cornell Univ., and by the Computational Mathematics Program of the National Science Foundation under grant DMS-8706133.
***Research partially supported by the U.S. Army Research Office through the Mathematical Sciences Institute, Cornell Univ., and by the Computational Mathematics Program of the National Science Foundation under grant DMS-8706133.

A GLOBAL AND QUADRATIC AFFINE SCALING METHOD FOR LINEAR L_1 PROBLEMS*

THOMAS F. COLEMAN† AND YUYING LI‡

Abstract. Recently, various interior point algorithms - related to the Karmarkar algorithm - have been developed for linear programming. In this paper, we first show how this "interior point" philosophy can be adapted to the linear l_1 problem (in which there are no feasibility constraints) to yield a globally convergent algorithm. We then show that the linear algorithm can be modified to provide a globally and ultimately quadratically convergent algorithm. This modified algorithm is significantly more efficient in practise: we present numerical results to support this claim.

Key Words. Linear l_1 estimation, linear programming, interior-point algorithm, simplex method, least absolute value regression, affine scaling method, Karmarkar

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

Abbreviated Title. An affine scaling method for linear l_1 problems

* Written June 1989. Presented at the Third SIAM Conference on Optimization, in Boston, April 1989.

† Computer Science Department and Center for Applied Mathematics, Cornell University, Ithaca, NY 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, by the U.S. Army Research Office through the Mathematical Sciences Institute, Cornell University, and by the Computational Mathematics Program of the National Science Foundation under grant DMS-8706133.

‡ Computer Science Department, Cornell University, Ithaca,NY 14853. Research partially supported by the U.S. Army Research Office through the Mathematical Sciences Institute, Cornell University and by the Computational Mathematics Program of the National Science Foundation under grant DMS-8706133.

1. Introduction. Let A be an n-by-m matrix with $m > n$ and consider the overdetermined system

$$A^T x \approx b \tag{1}$$

where row i of A^T is denoted by a_i^T, for $i = 1 : m$. The linear l_1 problem is to find a vector x which is a solution to

$$\min_{x \in \Re^n} \{\psi(x) \stackrel{def}{=} \sum_{i=1}^{m} |a_i^T x - b_i|\}. \tag{2}$$

Note that the function $\psi(x)$ is piecewise linear; ψ is not differentiable at any point x such that $a_i^T x = b_i$, for any index i. Moreover, it is well known that problem (2) is equivalent to the following linear program (LP):

$$\begin{aligned} & \min_{u,v,x} \sum_{j=1}^{m} (u_i + v_i) \\ \text{subject to} \quad & a_i^T x - u_i + v_i = b_i \\ & u_i \geq 0,\ v_i \geq 0,\ i = 1, \cdots, m. \end{aligned} \tag{3}$$

Thus, the l_1 problem can be solved by a general LP-solver. However, usually it is more efficient to use techniques especially tailored to problem (2), e.g., [1] and [2]. Most such methods are related to the simplex method in that they involve notions of an active set and pivoting or exchanging columns. These are finite algorithms.

The advent of affine scaling methods for linear programming raises the question: can such methods be tailored to the linear l_1 problem? Indeed, an interior point strategy applied to the dual of (3) has been suggested (see [12] and [8]) Such methods can be viewed as generating an *infinite* sequence of approximate solutions $\{x_k\}$. [1] However, despite numerous connections to Newton-like processes (e.g., [11]), global affine scaling methods, such as Meketon's algorithm, do not possess second-order convergence, even in the limit.

In this paper we propose a global affine scaling algorithm, for the l_1 problem, which is ultimately quadratically convergent.

1.1. Notations and Definitions. We use the sign function sgn(v), where v is a vector, in the following sense: if $w = sgn(v)$,

$$w_i = \begin{cases} 1 & \text{if } v_i \geq 0, \\ -1 & \text{otherwise.} \end{cases} \tag{4}$$

We use e_i to denote the ith elementary vector. We let $\mathcal{A}(x)$ denote the indices of zero residuals at any point x, i.e.,

$$\mathcal{A}(x) = \{\ i \mid a_i^T x - b_i = 0\} \tag{5}$$

[1] One can also take a finite view of such algorithms - assuming integer data and exact arithmetic. This view leads to a complexity analysis; e.g., is the number of steps bounded by a polynomial in the size of the problem ? This is not our concern in this paper.

and $\mathcal{A}^c(x)$ denotes the complementary set. (We will suppress the argument when it is clear from context.) We use the subscripts 1 and 0 to denote the compressed vector whose components correspond to the nonactivities and activities respectively, e.g., $g_1 = g_{\mathcal{A}^c}, g_0 = g_{\mathcal{A}}$. In some cases this denotation will refer to the current point; otherwise, it will refer to the limit point. The usage will be clear from context.

The dual of (3) is

$$\begin{aligned} & \max_{\lambda} b^T\lambda \\ \text{subject to} \quad & A\lambda = 0 \\ & -1 \le \lambda_i \le 1,\ i = 1,\cdots,m. \end{aligned} \tag{6}$$

DEFINITION 1. *We say an l_1 problem is* **primal nondegenerate** *if and only if, at any point x the vectors a_i, $i \in \mathcal{A}(x)$, are linearly independent.*

DEFINITION 2. *We call λ a* **dual basic point** *if and only if $A\lambda = 0$ and $|\{\lambda_i : |\lambda_i| = 1\}| \ge m - n$. We say an l_1 problem is* **dual nondegenerate** *if and only if, at any dual basic point λ, $|\{\lambda_i : |\lambda_i| = 1\}| = m - n$.*

The optimal solution can be characterized in various ways. For example, x is optimal if and only if there exists $\lambda \in R^{|\mathcal{A}|}$ such that

$$\sum_{i\in\mathcal{A}^c} \operatorname{sgn}(a_i^T x - b_i)a_i = \sum_{i\in\mathcal{A}} \lambda_i a_i$$

where

$$-1 \le \lambda_i \le 1, \quad \forall i \in \mathcal{A}.$$

It is easy to verify that this formulation is equivalent to the following: x is optimal if and only if there exists $\lambda \in R^m$ such that

$$(a_i^T x - b_i) * (sgn(a_i^T x - b_i) - \lambda_i) = 0, \quad i = 1:m \tag{7}$$

$$A\lambda = 0 \tag{8}$$

$$-1 \le \lambda_i \le 1, \quad \forall\ i \in \mathcal{A}. \tag{9}$$

This characterization is especially interesting because the nonlinear equation (7) suggests the possibility of a Newton method. We explore this possibility in Section 3.

Remark: Primal and dual nondegeneracy imply at solution (x, λ),

$$|\mathcal{A}(x)| = \operatorname{rank}(A_0(x)) = n, \quad \text{and}$$

$$|\mathcal{A}^c(x)| = |\{\lambda_i : |\lambda_i| = 1\}| = m - n.$$

Moreover, the linear l_1 problem has a unique minimizer, (x^*, λ^*), under these nondegeneracy assumptions, e.g., [10].

1.2. An Equivalent Formulation. Following Seneta and Steiger [13], we consider an alternative (but equivalent) formulation of the l_1 problem. Let Z denote matrix whose rows form a basis for the null space of A. Hence Z has dimensions $(m - rank(A)) \times m$, $rank(Z) = m - rank(A)$, and

$$AZ^T = 0.$$

Then, defining $r = b - A^T x$, the linear l_1 problem is equivalent to the following constrained l_1 problem with m variables r:

$$\begin{aligned} &\min_{r \in \Re^m} \|r\|_1 \\ \text{subject to} \quad &Zr = Zb. \end{aligned} \tag{10}$$

This formulation of the l_1 problem is useful for the derivation of our algorithms; however, implementation of our methods, as discussed in Section 4, does not necessarily involve the computation of matrix Z.

Once again optimality conditions can take different forms. For example, r is a solution if and only if there exists $\mu \in R^{|\mathcal{A}|}$ and $w \in R^{m-|\mathcal{A}|}$ such that

$$\sum_{\mathcal{A}^c} \text{sgn}(r_i) e_i = \sum_{\mathcal{A}} e_i \mu_i + Z^T w \tag{11}$$

$$Z(r - b) = 0$$

$$-1 \leq \mu_i \leq 1, \quad \forall i \in \mathcal{A}.$$

Clearly this implies that $|(Z^T w)_i| \leq 1, \quad \forall i \in \mathcal{A}$. Therefore, if we define $\lambda = Z^T w$ an alternative expression of optimality conditions is

$$r_i(sgn(r_i) - \lambda_i) = 0, \quad i = 1 : m \tag{12}$$

$$Z(r - b) = 0 \tag{13}$$

$$-1 \leq \lambda_i \leq 1, \quad i \in \mathcal{A}. \tag{14}$$

This formulation is important to our design of a quadratic algorithm: equations (12) and (13) lead to a Newton process, which we introduce in Section 3.

2. An Affine Scaling Method. In this section we present a new affine scaling method for the linear l_1 problem. Our approach maintains feasibility with respect to the equality constraints in (10); the scaling involved in the computation of a descent direction is determined by the distance the current point is from the lines of nondifferentiability (i.e., $r_i = 0$). It is this notion that replaces the more standard

" distance from infeasibility " definition used in most interior methods. However, since the differentiable region in (10) is not connected, and since we do not know how to immediately identify a connected differentiable region adjacent to the optimal point, our algorithm must allow for the ability to cross lines of nondifferentiability.

Let D be a positive diagonal matrix: e.g., $D = \text{diag}\{|r_i|^{\frac{1}{2}}\}$.

Assuming we are in the differentiable region (i.e., $r_i \neq 0, \quad \forall i$), we can define a descent direction by the following "trust region problem":

$$\begin{aligned} & \min_{d \in R^m} g^T d \\ \text{subject to} \quad & Zd = 0 \\ & \|D^{-1} d\|_2 \leq \delta \end{aligned} \tag{15}$$

where δ is a positive number reflecting the trust region size; $g = g(r)$ is the gradient of $\sum |r_i|$; the matrix D defines the shape of the ellipsoid. Therefore with our suggested choices of D, the ellipsoid is short in directions corresponding to components of r_i close to zero, and long in directions corresponding to relatively large $|r_i|$. The solution to (15) is of the form $d_* = \alpha d$ where

$$d = -D^2(g - Z^T(ZD^2Z^T)^{-1}ZD^2g), \tag{16}$$

and $\alpha = \alpha(\delta)$, for small $\delta > 0$.

Rather than choose δ a priori, we prefer to compute d as suggested by (16) and then define α using a piecewise linear minimization technique along the ray d (allowing for the ability to cross lines of nondifferentiability). Hence we must determine all nonnegative breakpoints,

$$J = \{\alpha_i : \ \alpha_i = -\frac{r_i}{d_i}, \ r_i d_i < 0\},$$

and using this information determine the minimizer in direction d:

$$\alpha_* = \min_{\alpha > 0} \psi(r + \alpha d) = \min_{\alpha \in J} \psi(r + \alpha d).$$

This is done by considering each breakpoint in J in turn, adjusting the gradient to reflect a step just beyond the breakpoint and then determining if d continues to be a descent direction. For example, if α_j is the smallest positive breakpoint, then a step just beyond this point yields the following gradient:

$$g^+ = g - 2g_j e_j.$$

If $(g^+)^T d < 0$ the next breakpoint is considered, etc. Of course we cannot step all the way to the minimizer (since points of nondifferentiability must be avoided) and therefore we compute $\tilde{\alpha} = \max_{J \cup \{0\}}\{\alpha_i : \ 0 \leq \alpha_i < \alpha_*\}$ and take as our steplength

$$\alpha = \tilde{\alpha} + \tau(\alpha_* - \tilde{\alpha}) \tag{17}$$

where $0 < \tau < 1$.

The linear algorithm follows. Let r^0 be an initial differentiable point satisfying $Zr^0 = Zb$; $k \leftarrow 0$.

Algorithm 1
Step 1 Define $D^k = \text{diag}\{|r_i^k|^{\frac{1}{2}}\}$ and $g^k = \text{sgn}(r^k)$.
Step 2 Compute

$$d^k = -(D^k)^2\{g^k - (Z^k)^T[Z^k(D^k)^2(Z^k)^T]^{-1}Z^k(D^k)^2g^k\}.$$

Step 3 Do a line search on the piecewise linear function $\psi(\alpha)$, as described above, to determine α^k. Then

$$r^{k+1} \leftarrow r^k + \alpha^k d^k, \quad k \leftarrow k+1.$$

Remarks:
i) Note that the linesearch maintains $D^k > 0$; therefore, d^k is well-defined and the linesearch guarantees $\psi(r_{k+1}) < \psi(r_k)$.
ii) We have chosen not to use the more standard $D^k = \text{diag}\{|r_i^k|\}$ as our scaling matrix: there are two reasons for this. First, with our choice we are able to define a hybrid method, described in later sections, that connects this linear algorithm with a Newton process; we have been unable to produce such a smooth transition using the more standard scaling. Second, it is not clear that the standard choice leads to a globally convergent process - certainly our convergence proof does not apply in this case. The difficulty appears to be that the standard choice, $D^k = \text{diag}\{|r_i^k|\}$, produces directions too nearly tangential to the near activities thus inhibiting the algorithm's ability to cross lines of nondifferentiability. This is not an issue in a standard interior point method since, in that setting, constraints are never crossed due to the maintenance of feasibility.

3. A Local Newton Process. A Newton process for the l_1 problem can be defined by considering conditions (12) and (13). Most importantly, consider (12) which we rewrite as

$$\text{diag}(r)(\text{sgn}(r) - Z^T w) = 0. \tag{18}$$

In general this system is not differentiable due to the discontinuities caused when $r_i = 0$, for any i. However, system (18) can certainly be differentiated at any point with no zero residual. Moreover, in a neighbourhood of a solution r^*, $\text{sgn}(r_i^*)$ will remain constant for any $i \in \mathcal{A}^c(r^*)$. Therefore, in a neighbourhood of r^*, discontinuities will be due only to the active equations, $i \in \mathcal{A}(r^*)$. As we formally establish in Section 6, these discontinuities do not impede the local quadratic convergence behaviour of a Newton process. Therefore, define $g = \text{sgn}(r)$, $D_r = diag(r)$, and $D_\lambda = diag(g - Z^T w)$; differentiate (18) and (13) to define a Newton correction,

$$\begin{bmatrix} D_\lambda & -D_r Z^T \\ Z & 0 \end{bmatrix} \begin{bmatrix} \triangle r \\ \triangle w \end{bmatrix} = \begin{bmatrix} -D_r(g - Z^T w) \\ 0 \end{bmatrix}. \tag{19}$$

Define $d_N = \triangle r$; it is easy to prove that the Newton step with respect to r is equal to

$$d_N = -A^T(AD_r^{-1}D_\lambda A^T)^{-1}Ag. \tag{20}$$

Note that the linear step (16) is equivalent to

$$d = -A^T(AD^{-2}A^T)^{-1}Ag. \tag{21}$$

It is this similarity in form that yields a smooth transition from the linear algorithm to the Newton method. Before presenting this procedure, it is worth noting that the matrix $AD_r^{-1}D_\lambda A^T$ is positive definite when (r, w) is close enough to the solution (r^*, w^*) and $\prod r_i \neq 0$.

LEMMA 1. *Assume (r^*, w^*) is a solution and the l_1 problem is both primal and dual nondegenerate. Then there exists a neighbourhood of (r^*, w^*) such that when (r, w) is within this neighbourhood, and no component of r is equal to zero, the matrix $AD_r^{-1}D_\lambda A^T$ is positive definite.*

Proof. First we note that provided the diagonal of D_r is zero–free,

$$AD_r^{-1}D_\lambda A^T = A_0{D_r}_0^{-1}D_{\lambda 0}A_0^T + A_1{D_r}_1^{-1}D_{\lambda 1}A_1^T$$

where the zero subscript refers to the active set at r^*, i.e., $r_i^* = 0$, and the unity subscript refers to the inactive set at r^*. Hence, by definition, D_{r1}^* is zero-free and since $D_{\lambda 1}^* = 0$, by complementary slackness (18), it follows that

$$A_1(D_{r1}^*)^{-1}D_{\lambda 1}^*A_1^T = (0).$$

Primal and dual nondegeneracy yield $|A_0| = n$ and therefore, using complementary slackness (18),

$$|\lambda_i^*| < 1, \quad \forall i \in \mathcal{A}(r^*).$$

Therefore, for $i \in \mathcal{A}(r^*)$, there exists a neighbourhood around (r^*, w^*) such that $(\frac{\mathrm{sgn}(r_i)-\lambda_i}{r_i})$ is positive and strictly bounded away from zero for all $i \in \mathcal{A}^*$ and $r_i \neq 0$. Since A_0 is full rank (by primal nondegeneracy) the result follows.

■

Lemma 1 implies that Newton directions are descent directions in a neighbourhood of the solution. It is also the case that as the iterates converge, for each $i \in \mathcal{A}(r^*)$,

$$(D_\lambda D_r^{-1})_i \to +\infty.$$

This unbounded behaviour may appear to be cause for concern; however, in Sections 5 and 6 we prove that theoretically this is not a problem. Moreover, our numerical experience suggests that reliable numerical performance can be attained provided sufficient care is taken with respect to implementation details.

4. A Hybrid Method. The Newton and linear directions, introduced in the previous sections, have similar forms but differ in their definitions of diagonal matrix. Consider now a third choice:

$$D^2 = |D_r D_\theta^{-1}| \tag{22}$$

where $D_r = \text{diag}(r)$, $D_\theta = \theta \cdot \text{diag}(g) + (1-\theta)D_\lambda$, and $0 \le \theta \le 1$. Clearly if $\theta = 1$ then $D^2 = |D_r|$ and d is the linear step proposed in Section 2, i.e., Algorithm 1. On the other hand, as $\theta \to 0$,

$$A^T D^{-2} A^T \to A D_r^{-1} D_\lambda A^T,$$

(which, by Lemma 1, is positive definite in a neighbourhood of the solution) and so the Newton direction is approached asymptotically (i.e., as $\theta \to 0$).

Our remaining task is to define $\theta \in [0,1]$ so that $\theta \to 0$ if and only if $(r,\lambda) \to (r^*,\lambda^*)$. Our idea is to let θ encapsulate the optimality conditions. Two possible choices are:

$$\theta = \frac{\frac{\sqrt{\sum_{i=1}^m (r_i(g_i-\lambda_i))^2}}{\psi(r^0)} + e^T \max\{|\lambda| - e, 0\}}{\gamma + \frac{\sqrt{\sum_{i=1}^m (r_i(g_i-\lambda_i))^2}}{\psi(r^0)} + e^T \max\{|\lambda| - e, 0\}}, \tag{23}$$

$$\theta = \frac{\max\{\max(\frac{|r_i(g_i-\lambda_i)|}{\psi(r^0)}), \max\{\max\{|\lambda| - e, 0\}\}}{\gamma + \max\{\max(\frac{|r_i(g_i-\lambda_i|}{\psi(r^0)})), \max\{\max\{|\lambda| - e, 0\}\}}, \tag{24}$$

where $0 < \gamma < 1$. Clearly θ is bounded above by 1; assuming $Z(r-b) = 0$ then $\theta = 0$ if and only if $(r,\lambda) = (r^*,\lambda^*)$. The following result indicates that, unless optimal, D_θ has a zero-free diagonal. Consequently, when $\theta \neq 0$ and no component of r is zero, D^2 is a *positive* diagonal matrix and so d is a descent direction.

LEMMA 2. *Suppose* $0 < \gamma < 1$. *Assume* θ *is defined by either (23) or (24). Then* D_θ *satisfies*

$$|D_\theta| \ge \min\{\theta, (1-\gamma)\theta\} I.$$

Proof. By definition of the matrix D_θ, we can write

$$D_\theta = \text{diag}(g) - (1-\theta)\text{diag}(Z^T w).$$

Consider the ith diagonal element of $D_\theta = \theta_i$. If $|\lambda_i = Z_i^T w| \le 1$, it is clear that

$$|g_i - (1-\theta)\lambda_i| \ge \theta.$$

If $|\lambda_i| > 1$ and θ is defined by (23), we have

$$\frac{scs + e^T \max(|Z^T w| - e, 0)}{scs + e^T \max(|Z^T w| - e, 0) + \gamma} = \theta,$$

where *scs*, *s*caled *c*omplementary *s*lackness, is given by

$$scs = \frac{\sqrt{\sum_{i=1}^{m}(r_i(g_i - \lambda_i))^2}}{\psi(r^0)}.$$

Hence

$$(1-\theta)e^T \max(|Z^T w| - e, 0) = (\theta - 1)scs + \gamma\theta.$$

Therefore, we have

$$\begin{aligned}(1-\theta)(|\lambda_i| - 1) &\leq (\theta-1)scs + \gamma\theta \\ &\leq \gamma\theta.\end{aligned}$$

(It is straightforward to prove that the above is true when θ is defined by (24) and $scs = max\frac{|r_i(g_i-\lambda_i)|}{\psi(r_0)}$.) Thus

$$\begin{aligned}|\lambda_i| &\leq 1 + \frac{\gamma\theta}{1-\theta} \\ &= \frac{1-(1-\gamma)\theta}{1-\theta}\end{aligned}$$

which yields

$$(1-\theta)|\lambda_i| \leq (1-(1-\gamma)\theta).$$

Therefore

$$|g_i - (1-\theta)\lambda_i| \geq (1-\gamma)\theta,$$

and the result is established. ∎

Therefore, D_θ is nonsingular when $\theta \neq 0$; consequently, either definition of θ leads to a descent direction. Moreover, θ induces a smooth transition from the linear step to the Newton step.

4.1. Computation of the Hybrid Step. We consider three different methods for computing the hybrid step.

4.1.1. Full Space Implementation. The hybrid step can be viewed as an approximate Newton step. Assuming (r, w) is our current guess, where $Zr = Zb$, the approximate Newton step, in analogy with (19), can be expressed as

$$\begin{bmatrix} \mathrm{diag}(g)|D_\theta| & -D_r Z^T \\ Z & 0 \end{bmatrix} \begin{bmatrix} d \\ d_w \end{bmatrix} = \begin{bmatrix} -D_r(g - Z^T w) \\ 0 \end{bmatrix}.$$

Note that d is the hybrid direction defined above. The dimension of this system is somewhat daunting, $(2m-n)$-by-$(2m-n)$, and can be reduced by implicitly satisfying $Zd = 0$:

$$Solve \quad [\mathrm{diag}(g)|D_\theta|A^T, -D_r Z^T] \begin{bmatrix} d_x \\ d_w \end{bmatrix} = -D_r(g - Z^T w) \tag{25}$$

and then

$$d \leftarrow A^T d_x. \tag{26}$$

After solving this full-space system and obtaining (d_x, d_w), the dual variables can be updated $w^+ \leftarrow w + D_r Z^T d_w$; the primal variables are updated using a linesearch along direction d, $r^+ \leftarrow r + \alpha d$.

Note that as $\theta \rightarrow 0$, i.e., as we converge to the solution, the matrix in (25) approaches $(D_{\lambda^*} A^T, -D_{r^*} Z^T)$ which, under nondegeneracy assumptions, is full rank (and bounded). This is the advantage of using the full-space implementation: the limit matrix is well-behaved. The major disadvantage is that the system is still large, m-by-m, and requires the computation of Z.

It is possible to reduce the dimension of the system to be solved (and still compute the same correction) by realizing that what is needed is either d_x and $Z^T d_w$ or w and $A^T d_x$ but not both d_x and d_w. More generally, consider the following simple observations.

LEMMA 3. *Let $M = [M_1, M_2]$ be full rank, $U = null(M_2^T)$, $V = null(M_1^T)$. Then the matrix $[U, V]$ has full rank.*

THEOREM 4. *Assume $M = [M_1, M_2]$ has full rank, $U = null(M_2^T)$, $V = null(M_1^T)$. Then the system*

$$Mv = b$$

is equivalent to

$$\begin{bmatrix} U^T M_1 & 0 \\ 0 & V^T M_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} U^T b \\ V^T b \end{bmatrix}.$$

Proof. From Lemma 3, $[U, V]$ has full rank. Hence $Mv = b$ is equivalent to

$$[U, V]^T M v = [U, V]^T b.$$

■

COROLLARY 5. *To obtain $[v_1, M_2 v_2]$ such that*

$$[M_1, M_2] \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = b$$

the following algorithm suffices:

1. *Determine $U = null(M_2^T)$;*
2. *Solve $U^T M_1 v_1 = U^T b$;*
3. *$M_2 v_2 \leftarrow b - M_1 v_1$.*

4.1.2. Range Space Implementation. If we define $M_1 = \text{diag}(g)|D_\theta|A^T$ and $M_2 = -D_r Z^T$ then we can solve (25) using Corollary 5:

$$Solve \quad D^{-1}A^T d_x \overset{\text{l.s.}}{=} Dg \tag{27}$$

and

$$d \leftarrow -A^T d_x, \quad \lambda^+ \leftarrow g + D^{-2}d. \tag{28}$$

The main computational work is the solution of the linear least squares problem of order m-by-n. Note that Z is not required; this approach is particularly attractive when n is small or A is sparse. The primal variables are updated via a linesearch: $r^+ \leftarrow r + \alpha d$.

A null space implementation is obtained from Corollary 5 by letting $M_1 = -D_r Z^T$ and $M_2 = \text{diag}(g)|D_\theta|A^T$.

4.1.3. Null Space Implementation. This method first requires the computation of Z and then

$$Solve \quad DZ^T w^+ \overset{\text{l.s.}}{=} Dg \tag{29}$$

and

$$A^T d_x \overset{def}{=} d \leftarrow -D^2(g - Z^T w^+). \tag{30}$$

In this case an m-by-$(m-n)$ least squares system is solved at each iteration; hence, this approach is attractive when n is large. Note that Z is computed only once; moreover, it is often possible to compute a sparse Z given that A is sparse (e.g., [5,6]).

We conclude this section with a presentation of the simple hybrid algorithm. Bear in mind that there are a number of numerical concerns, such as unequal row scaling and stopping criteria, that must be taken care of before this algorithm can be reliably used. These issues are briefly discussed in Section 7.

The linesearch algorithm differs slightly from the procedure presented in Section 2: update (17) is replaced with

$$\alpha = \tilde{\alpha} + \max\{\tau, 1 - \theta^k\}(\alpha_* - \tilde{\alpha}). \tag{31}$$

Let r^0 be an initial differentiable point satisfying $Z(r^0 - b) = 0$; $k \leftarrow 0$; Compute an initial point w^0.

Algorithm 2

Step 1 Compute θ^k from (23) or (24). Define D^k from (22) and $g^k \leftarrow \text{sgn}(r^k)$.

Step 2 Compute d^k and w^{k+1} using one of the three methods above .

Step 3 Do a line search on the piecewise linear function $\psi(\alpha)$ (as described in Section 2 and using (31)) to determine α^k,

$$r^{k+1} \leftarrow r^k + \alpha^k d^k, \quad k \leftarrow k + 1.$$

5. Global Convergence. In this section, we establish global convergence of the linear and hybrid methods, Algorithms 1 and 2. We make the following global assumption:

The n-by-m matrix A has full row rank n.

Let P^k be the orthogonal projector onto $null(ZD^k)$, i.e.,

$$P^k = I - D^k Z^T (Z(D^k)^2 Z^T)^{-1} Z D^k.$$

Both algorithms use a diagonal matrix $D^k \stackrel{def}{=} diag(s_i^k)$ to yield the search direction d^k:

$$\begin{aligned} d^k &= -D^k P^k D^k g^k \\ &= -(D^k)^2 (g^k - Z^T w^{k+1}), \end{aligned} \tag{32}$$

where w^{k+1} is the least squares solution to

$$D^k Z^T w^{k+1} \stackrel{\text{l.s.}}{=} D^k g^k,$$

and

$$s_i^k = \begin{cases} |r_i^k|^{\frac{1}{2}} & \text{for Algorithm 1} \\ |r_i^k|^{\frac{1}{2}} |\theta^k g_i^k + (1-\theta^k)(g_i^k - Z_i^T w^k)|^{-\frac{1}{2}} & \text{for Algorithm 2.} \end{cases}$$

The first major step in the convergence proof is to show that $\|P^k D^k g^k\| \to 0$. This is established in Lemma 10 after several preliminary results.

Definition: The positive scalar α corresponds to a *breakpoint* in the direction d if, for some i, $r_i \neq 0$ and $r_i + \alpha d_i = 0$.

LEMMA 6. *Assume $\{d^k\}$ is defined by Algorithm 1 or Algorithm 2. Then*

$$\lim_{k\to\infty} \bar{\alpha}^k \|P^k D^k g^k\|_2^2 = 0 \tag{33}$$

where $\bar{\alpha}^k$ corresponds to the first breakpoint in direction d^k.

Proof. Since $\|r^k\|_1$ is monotonically decreasing and bounded below, $\|r^k\|_1$ converges; therefore,

$$\lim_{k\to\infty} (\|r^k\|_1 - \|r^{k+1}\|_1) = 0. \tag{34}$$

From (32),

$$g^{k^T} d^k = -\|P^k D^k g^k\|_2^2;$$

therefore, since $r^{k+1} = r^k + \alpha^k d^k$,

$$\begin{aligned} & \|r^k\|_1 - \|r^{k+1}\|_1 \\ = \; & g^{k^T} r^k - g^{k+1^T} r^k - \alpha^k g^{k+1^T} d^k \\ = \; & (g^k - g^{k+1})^T r^k + \alpha^k (g^k - g^{k+1})^T d^k + \alpha^k \|P^k D^k g^k\|_2^2. \end{aligned}$$

However, if $g_i^k \neq g_i^{k+1}$ then $g_i^k = -g_i^{k+1}$; hence,

$$\begin{aligned} & \|r^k\|_1 - \|r^{k+1}\|_1 \\ = \; & \sum_{g_i^k \neq g_i^{k+1}} 2(g_i^k r_i^k + \alpha^k g_i^k d_i^k) - (\alpha^k - \bar{\alpha}^k) g^{k^T} d^k + \bar{\alpha}^k \|P^k D^k g^k\|_2^2. \end{aligned} \tag{35}$$

For $g_i^k \neq g_i^{k+1}$, define α_i^k to be the breakpoint for equation i in direction d_i^k: $\alpha_i^k = \frac{-r_i^k}{d_i^k}$. Therefore, for $g_i^k \neq g_i^{k+1}$,

$$r_i^k + \alpha^k d_i^k = (\alpha^k - \alpha_i^k) d_i^k. \tag{36}$$

Consequently, substituting (36) into (35),

$$\begin{aligned} & \|r^k\|_1 - \|r^{k+1}\|_1 \\ = \; & \sum_{g_i^k \neq g_i^{k+1}} 2(\alpha^k - \alpha_i^k) g_i^k d_i^k - (\alpha^k - \bar{\alpha}^k) g^{k^T} d^k + \bar{\alpha}^k \|P^k D^k g^k\|_2^2. \\ = \; & -(\alpha^k - \bar{\alpha}^k)(g^{k^T} d^k - 2 \sum_{g_i^k \neq g_i^{k+1}} g_i^k d_i^k) - 2 \sum_{g_i^k \neq g_i^{k+1}} (\alpha_i^k - \bar{\alpha}^k) g_i^k d_i^k + \bar{\alpha}^k \|P^k D^k g^k\|_2^2. \end{aligned}$$

Next we establish that each term in the expression above is positive. The line-search stops before optimality is reached in the direction d^k; hence, $g^{k+1^T} d^k < 0$. But $g^{k+1^T} d^k = g^{k^T} d^k - 2 \sum_{g_i^k \neq g_i^{k+1}} g_i^k d_i^k$ and so it follows that

$$(g^{k^T} d^k - 2 \sum_{g_i^k \neq g_i^{k+1}} g_i^k d_i^k) < 0,$$

and therefore the first term is nonnegative. But, $\bar{\alpha}^k \leq \alpha_i^k$ and so

$$\sum_{g_i^k \neq g_i^{k+1}} (\alpha_i^k - \bar{\alpha}^k) g_i^k d_i^k \leq 0,$$

which yields a nonnegative second term. Thus

$$0 \leq \bar{\alpha}^k \|P^k D^k g^k\|_2^2 \leq \|r^k\|_1 - \|r^{k+1}\|_1,$$

and therefore, applying (34),

$$\lim_{k \to \infty} \bar{\alpha}^k \|P^k D^k g^k\|_2^2 = 0.$$

■

Recall: We use the subscripts 1 and 0 to denote the components corresponding to the nonactivities and activities respectively. So, for example, Z_1 is the submatrix of Z consisting of columns of Z corresponding to the nonzero components of r.

LEMMA 7. *Assume r is any primal nondegenerate point. Then Z_1 is of full row rank. Moreover, if we assume that $(r, \lambda = Z^T w)$ satisfy complementary slackness (11), then λ is a dual basic solution. In addition, if we assume dual nondegeneracy, then $|\mathcal{A}(r)| = n$.*

Proof. Assume that there exist y_1 such that

$$Z_1^T y_1 = 0. \tag{37}$$

From $AZ^T = 0$, we have

$$A_0 Z_0^T + A_1 Z_1^T = 0.$$

Hence

$$A_0 Z_0^T y_1 + A_1 Z_1^T y_1 = 0.$$

From (37), we have $A_0 Z_0^T y_1 = 0$; from the primal nondegeneracy assumption, A_0 is of full column rank. Therefore $Z_0^T y_1 = 0$. Thus

$$Z^T y_1 = 0.$$

But Z^T has full column rank; therefore, $y_1 = 0$ and we have established that Z_1 is of full row rank. Now, let us further assume complementary slackness (11); hence,

$$g_1 = Z_1^T w$$

where $\lambda = Z^T w$. But $AZ^T = 0$ which yields $A\lambda = 0$ and so we obtain dual feasibility (i.e., λ is a dual basic point). Finally, dual nondegeneracy and full row rank of Z_1 result in $|\mathcal{A}| = n$. ■

LEMMA 8. *Under primal and dual nondegeneracy assumptions, J is nonsingular at any point $(r, \lambda = Z^T w)$, where*

$$J = \begin{bmatrix} diag(g)|D_\theta| & -D_r Z^T \\ Z & 0 \end{bmatrix}.$$

Proof. For any $y = [y_0^T, y_1^T]^T$ such that $Jy = 0$, we have

$$\mathrm{diag}(g)|D_\theta| y_0 - D_r Z^T y_1 = 0 \tag{38}$$

$$Z y_0 = 0. \tag{39}$$

Hence $y_0 = A^T z$, for some z. Thus

$$\text{diag}(g)|D_\theta|A^T z - D_r Z^T y_1 = 0. \tag{40}$$

First assume $\theta \neq 0$. Then, by Lemma 2, D_θ is nonsingular. Multiplying the above by the full rank matrix $Z(\text{diag}(g))|D_\theta^{-1}|$ yields

$$Z \text{ diag}(g) \ |D_\theta|^{-1} D_r Z^T y_1 = 0,$$

which implies

$$Z_1 \text{ diag}(g_1) \ |D_{\theta 1}|^{-1} D_{r1} Z_1^T y_1 = 0.$$

From the primal nondegeneracy assumption and Lemma 7, it is clear that $y_1 = 0$. From (40), we have $z = 0$. Thus, $y_0 = 0$. Hence J is nonsingular.

Second, assume $\theta = 0$. Complementary slackness and (40) yield

$$\begin{aligned} &|D_{\lambda 0}|A_0^T z = 0, \quad \text{and} \\ &D_{r1} Z_1^T y_1 = 0. \end{aligned}$$

From primal and dual nondegeneracy and Lemma 7, $|D_{\lambda 0}|A_0^T$ and $D_{r1} Z_1^T$ are nonsingular, hence

$$y_1 = 0, z = 0.$$

Therefore $y_0 = 0$ and so J is nonsingular. ∎

The next result establishes that, in Algorithm 2, the multipliers are bounded in size. The key to the proof is that $\{D_\theta^k\}$ is bounded.

LEMMA 9. *Assume that an l_1 problem is both primal and dual nondegenerate and $\{\lambda^k = Z^T w^k\}$ is obtained by Algorithm 2. Then there exists $M > 0$ such that*

$$\|\lambda^k\| \leq M.$$

Proof. By definition, there exists d_r^k such that (d_x^k, w^{k+1}) is the solution to the following linear system:

$$[\text{diag}(g^k)|D_\theta^k|A^T, -D_r^k Z] \begin{bmatrix} d_x^k \\ w^{k+1} \end{bmatrix} = -D_r^k g^k.$$

Assume that $\{\|\lambda^k\|\}$ is unbounded. Then there exists a subsequence such that $\{\|\lambda^{k_j}\|\} \to +\infty$ and $\{r^{k_j}\} \to \bar{r}$ for some limit point $\bar{r}$. Now consider

$$[\text{diag}(g^k)|D_\theta^{k_j-1}|A^T, -D_r^{k_j-1} Z] \begin{bmatrix} d_x^{k_j-1} \\ w^{k_j} \end{bmatrix} = -D_r^{k_j-1} g^{k_j-1}. \tag{41}$$

First, assume $\{\theta^{k_j-1}\}$ is bounded away from zero. We first prove that $\{D_\theta^{k_j-1}\}$ is bounded above. Consider the diagonal component $\delta_i^{k_j-1}$ of $D_\theta^{k_j-1}$. By definition

$$\delta_i^{k_j-1} = g_i^{k_j-1} - (1-\theta^{k_j-1})\lambda_i^{k_j-1}.$$

If $\{\lambda_i^{k_j-1}\}$ is bounded then since $\{\theta^k\}$ is bounded, $\{\delta_i^{k_k-1}\}$ is also. Suppose that $\{\lambda_i^{k_j-1}\} \to \infty$. But from the definition (23) of θ, we have for k sufficiently large,

$$|(1-\theta^{k_j-1})\lambda_i^{k_j-1}| \leq |\frac{\lambda_i^{k_j-1}}{\gamma + |\lambda_i^{k_j-1}| - 1}|$$

Hence, $\{D_\theta^{k_j-1}\}$ is bounded.

Now consider a convergent subsequence $\{D_\theta^{k_{j_l}-1}\} \to \bar{D}_\theta$. Since, by assumption, $\{\theta^{k_j-1}\}$ is bounded away from zero, $(\bar{D}_\theta)^{-1}$ exists, by Lemma 2; the coefficient matrix in (41) is nonsingular in the limit (this is easily proved along the lines of the first half of the proof of Lemma 8). It follows that $\{\|\lambda^{k_{j_l}}\|\}$ is bounded, a contradiction.

Second, assume there exists a subsequence such that the corresponding sequence $\{\theta^{k_j-1}\}$ converges to zero. Then, by definition of θ, $\{\lambda^{k_j-1}\} \to \lambda^*$ and clearly $\{\|\lambda^{k_{j_l}}\|\}$ is bounded. ∎

We can now state the first major result.

LEMMA 10. *Assume $\{d^k\}$ is defined by Algorithm 1 or 2; assume primal and dual nondegeneracy. Then*

$$\lim_{k\to\infty} \|P^k D^k g^k\|_2 = 0.$$

Proof. Using Lemma 6, we know that

$$\lim_{k\to\infty} \bar{\alpha}^k \|P^k D^k g^k\|_2^2 = 0. \tag{42}$$

(Recall: $\bar{\alpha}^k$ is the step to the first breakpoint from r^k in the direction d^k.) From Lemma 9, there exists $M' > 0$, such that

$$\|(g^k - Z^T w^{k+1})\| \leq M'. \tag{43}$$

Now assume there exists a subsequence, which we denote with the use of an underscore, satisfying,

$$\{-\|\underline{P}^k \underline{D}^k \underline{g}^k\|_2^2\} \to c_1 < 0. \tag{44}$$

We now prove that the corresponding subsequence of first breakpoints, $\{\underline{\bar{\alpha}}^k\}$, satisfies

$$\underline{\bar{\alpha}}^k > c_4, \quad c_4 > 0,$$

and this will lead to the obvious contradiction.

From the definition of θ^k and (44), there exists $c_2 \in (0,1)$ such that $\bar{\theta^k} > c_2$. Recall, from Lemma 2, that for any i,

$$(45) \qquad \begin{aligned} |\underline{\theta}^k \underline{g_i}^k - (1-\underline{\theta}^k)(\underline{g_i}^k - \underline{\lambda_i}^k)| &\geq \min\{\underline{\theta}^k, (1-\gamma)\underline{\theta}^k\} \\ &\geq \min\{c_2, (1-\gamma)c_2\} \\ &\stackrel{def}{=} c_3 > 0. \end{aligned}$$

However, if we let j denote the index of the first breakpoint in direction $\bar{d}^k$,

$$\begin{aligned} \bar{\underline{\alpha}}^k &\geq -\frac{\underline{r}_j^k}{\underline{d}_j^k} \\ &= \frac{|\underline{\theta}^k \underline{g}_j^k + (1-\underline{\theta}^k)(\underline{g}_j^k - \underline{\lambda}_j^k)|}{(|\underline{g}_j^k - Z_j^T \underline{w}^k|)}, \quad \text{using (30)} \\ &\geq \frac{c_3}{M'}, \quad \text{using (43,45)}, \\ &\stackrel{def}{=} c_4 > 0. \end{aligned}$$

Therefore, by Lemma 6, $\|P^k D^k g^k\| \to 0$. ∎

The next result says that if the primal variables converge then so do the duals. The point of convergence, $(\bar{r}, \bar{\lambda} = Z^T \bar{w})$, satisfies all optimality conditions except possibly for $|\bar{\lambda}_0| < e$.

LEMMA 11. *Suppose $\{r^k\}$ and $\{w^k\}$ are obtained by Algorithm 1 or 2 and assume $\{r^k\} \to \bar{r}$, a limit point. Further, assume primal and dual nondegeneracy. Then $\{w^k\}$ converges; i.e., there exists a point $\bar{w}$ such that*

$$(46) \qquad \lim_{k \to \infty} w^k = \bar{w},$$

where w^{k+1} is defined from

$$(47) \qquad D^k Z^T w^{k+1} \stackrel{\text{l.s.}}{=} D^k g^k.$$

Moreover,

$$(48) \qquad |\mathcal{A}(\bar{r})| = n, \quad A\bar{\lambda} = 0, \quad |\bar{\lambda}_1| = e, \quad \text{where} \quad \bar{\lambda} = Z^T \bar{w}.$$

Proof. Using Lemma 10 and (32) we know that

$$D^k(g^k - Z^T w^{k+1}) \to 0.$$

Without loss of generality, assume

$$D_r = \begin{bmatrix} D_{\bar{r}1} & 0 \\ 0 & D_{r0} \end{bmatrix}, Z^T = \begin{bmatrix} Z_1^T \\ Z_0^T \end{bmatrix},$$

where $D_{\bar{r}1}$ is a diagonal matrix with zero–free diagonal, and $D_{\bar{r}0}$ is the zero matrix. But $D^k = |D_r^k[D_\theta^k]^{-1}|$ and so

$$\lim_{k\to\infty}[D_{\theta 1}^k]^{-1}(g_1^k - Z_1^T w^{k+1}) = 0.$$

But, the proof of Lemma 9 established the boundedness of $|D_\theta^k|$; hence, there exists $\bar{M} > 0$ such that $|[D_{\theta 1}^k]^{-1}| > \bar{M}$. Therefore, for any convergent subsequence of w^k, we have

$$\bar{g}_1 = Z_1^T \bar{w}, \tag{49}$$

where $\bar{w}$ is the limit point of the subsequence. (Note: $\bar{g} = \mathrm{sgn}(\bar{r})$.) By primal nondegeneracy and Lemma 7, Z_1^T is of full column rank; therefore, the solution to (49) is unique. Since the limit point of any convergent subsequence of $\{w^k\}$ is a solution of (49), and since by Lemma 9 $\{w^k\}$ is bounded, $\{w^k\}$ converges to $\bar{w}$.

Finally, to show (48) note that

$$[A_0, A_1]\bar{\lambda} = 0, \quad \text{where } \bar{\lambda}^T \stackrel{def}{=} (\bar{w}^T Z_0, \bar{w}^T Z_1)$$

since $AZ^T = 0$. But, by (49), the magnitude of each component of $\bar{\lambda}_1$ is unity; therefore, by Lemma 7, A_0 is order n. ∎

Theorem 14 below is the next major result: it says that if the primal variables converge, they converge to the optimal point. Before proving this theorem two technical results are provided, the second of which we state without proof (the proof is relatively straightforward and unenlightening).

LEMMA 12. *Assume $\{d^k\}$ is obtained by either Algorithm 1 or Algorithm 2. Suppose there exist vectors u and v such that*

$$A_1 u = -A_0 v.$$

Then, if $g_1^k = u$,

$$g_1^{kT} d_1^k = -v^T d_0^k. \tag{50}$$

Proof. From the definition of the direction d^k (26), we know there exists d_x^k such that

$$d^k = A^T d_x^k.$$

Thus

$$\begin{aligned} g_1^{kT} d_1^k &= g_1{}^T A_1^T d_x^k \qquad (\text{since } g_1^k = g_1) \\ &= -v^T A_0^T d_x^k \qquad (\text{from (50)}) \\ &= -v^T d_0^k. \end{aligned}$$

■

LEMMA 13. *Suppose* $\theta \in (0,1)$ *and* $-1 < \beta < 1 < \eta$. *Then,*

$$\frac{1+(1-\theta)\beta}{1+\beta} > \frac{1+(1-\theta)\eta}{1+\eta}.$$

THEOREM 14. *Assume* $\{r^k\}$ *is obtained from Algorithm 1 or Algorithm 2. Assume primal and dual nondegeneracy. If the sequence* $\{r^k\}$ *converges to a point* r^*, *then* r^* *is optimal.*

Proof. From Lemma 11, $|\mathcal{A}(r^*)| = n$, $\{w^k\} \to w^*$ and if λ^* is defined $\lambda^* = Z^T w^*$ then $A\lambda^* = 0$ with $|\lambda_1^*| = e$. Therefore, to establish optimality we need only show that $\|\lambda_0^*\|_\infty \leq 1$.

Assume the contrary, i.e., for some i with $r_i^* = 0$, $\lambda_i^* > 1$. Then, there exists k_1 such that for $k > k_1$,

$$|\lambda_i^k| > 1, \text{ for all } i \text{ such that } |\lambda_i^*| > 1$$

and

$$\lambda_j^k r_j^k > 0, \text{for all } j \text{ such that } |\lambda_j^*| = 1.$$

If there exists $|\lambda_i^*| > 1$ and $\lambda_i^{k+1} r_i^k > 0$ where $k > k_1$, from

$$r_i^{k+1} = r_i^k - \alpha^k s_i^{k^2}(g_i^k - \lambda_i^{k+1})$$

it follows that $|r_i^{k+1}| > |r_i^k| > 0$, for all $k > k_1$. Hence, we see immediately that $\{r_i^k\} \not\to 0$ which contradicts that $\{r_i^k\} \to r_i^* = 0$.

Now assume there does not exist $k_2 > k_1$ such that $r_i^{k_2}\lambda_i^{k_2+1} > 0$, for any $|\lambda_i^*| > 1$. In Algorithm 1, the breakpoint corresponding to equation l is given by

$$\alpha_l^k = \frac{1}{g_l^k(g_l^k - \lambda_l^{k+1})},$$

whereas for Algorithm 2,

$$\alpha_l^k = \frac{r_l^k|g_l^k - (1-\theta^k)\lambda_l^k|}{|r_l^k|(g_l^k - \lambda_l^{k+1})} = \frac{|g_l^k - (1-\theta^k)\lambda_l^k|}{g_l^k(g_l^k - \lambda_l^{k+1})}$$

This latter expression can be simplified if we consider two separate cases. For $|\lambda_j^*| < 1$, we have

$$\begin{aligned} \alpha_j^k &= \frac{g_j^k(g_j^k - (1-\theta^k)\lambda_j^k)}{g_j^k(g_j^k - \lambda_j^{k+1})} \\ &= \frac{1-(1-\theta^k)g_j^k\lambda_j^k}{1-g_j^k\lambda_j^{k+1}}. \end{aligned}$$

Hence, for $l \in \mathcal{A}^c$, $\alpha_l^k \to \infty$. For $|\lambda_i^*| > 1$, since $g_i^k = -\text{sign}(\lambda_i^{k+1}) = -\text{sign}(\lambda_i^*)$, we have

$$\begin{aligned} \alpha_i^k &= \frac{|g_i^k - (1-\theta^k)\lambda_i^k|}{(1 - \lambda_i^{k+1} g_i^k)} \\ &= \frac{1 + (1-\theta^k)|\lambda_i^k|}{1 + |\lambda_i^{k+1}|} \end{aligned}$$

Using Lemma 13, we know that

$$\frac{1 + (1-\theta^*)|\lambda_i^*|}{1 + |\lambda_i^*|} < \frac{1 + (1-\theta^*)\lambda_j^*}{1 + \lambda_j^*}$$

and

$$\frac{1 + (1-\theta^*)|\lambda_i^*|}{1 + |\lambda_i^*|} < \frac{1 - (1-\theta^*)\lambda_j^*}{1 - \lambda_j^*}$$

since $\theta^* \neq 0$.

Hence, when k is sufficiently large, we obtain, for both algorithms,

$$\alpha_i^k < \alpha_j^k, \qquad \text{for any } \; j \neq i, |\lambda_j^*| < 1. \tag{51}$$

Now let g^+ denote the new gradient after passing all the functions with $|\lambda_i^*| > 1$. Then

$$\begin{aligned} g^{+T} d^k &= -\sum_{|\lambda_i^*|>1, i\in\mathcal{A}} g_i^k d_i^k + \sum_{|\lambda_j^*|<1, j\in\mathcal{A}} g_j^{kT} d_j^k + g_1^{kT} d_1^k \\ &= -\sum_{|\lambda_i^*|>1, i\in\mathcal{A}} g_i^k d_i^k + \sum_{|\lambda_j^*|<1, j\in\mathcal{A}} g_j^{kT} d_j^k - \lambda_0^{*T} d_0^k \qquad \text{(using Lemma 12)} \\ &= \sum_{|\lambda_i^*|>1, i\in\mathcal{A}} (-\lambda_i^* - g_i^k) d_i^k + \sum_{|\lambda_j^*|<1, j\in\mathcal{A}} (g_j^k - \lambda_j^*) d_j^k \end{aligned}$$

But

$$-\lambda_i^* d_i^k < g_i^k d_i^k < 0, \qquad \text{for } |\lambda_i^*| > 1, r_i^* = 0.$$

and

$$-|\lambda_j^* d_j^k| > g_j^k d_j^k, \qquad \text{for } |\lambda_j^*| < 1, r_j^* = 0.$$

Hence

$$g^{+T} d^k < 0.$$

Therefore, the linesearch will yield $r_i^{k+1}\lambda_i^{k+2} > 0$, a contradiction. ∎

THEOREM 15. *If an l_1 problem is both primal nondegenerate and dual nondegenerate, then the sequence $\{(r^k, \lambda^k)\}$, generated by either Algorithm 1 or 2, is convergent.*

Proof. It is clear that $\{r^k\}$ is bounded. From Lemma 9, $\{\lambda^k\}$ is bounded. Hence, there exists a limit point $(\bar{r}, \bar{\lambda})$, $\bar{\lambda} = Z^T\bar{w}$. Moreover, from Lemma 10, $\lim_{k\to\infty} D^k(g^k - \lambda^{k+1}) = 0$. Therefore, either $\lim_{k\to\infty} \mathrm{diag}(r^k)(g^k - \lambda^k) = 0$ or $\lim_{k\to\infty} \mathrm{diag}(r^k)(g^k - \lambda^{k+1}) = 0$. Hence

$$\bar{g}_1 - Z_1^T\bar{w} = 0, \quad \text{where } \bar{g}_1 = \mathrm{sign}(\bar{r}_1).$$

For $\delta > 0$, let

$$B_\delta = \{(r, w) \mid |w - \bar{w}| < \delta\}.$$

Let S be the set of points satisfying complementary slackness, i.e., $S = \{(r,w)|\ r_i(g_i - \lambda_i) = 0,\ i = 1:m,\ \lambda = Z^Tw\}$. The nondegeneracy assumptions imply that S is a finite set of vertices. Hence, any limit point of $\{(r^k, Z^Tw^k)\}$ belongs to S. By dual nondegeneracy assumption, there exists $\epsilon > 0$ such that

$$0 < \epsilon < \min\{1 - |\bar{\lambda}_i| \mid \bar{r}_i = 0\}.$$

Since each point in S is isolated, there exists $\delta > 0$ such that

$$B_{2\delta} \cap S = (\bar{r}, \bar{w})$$

and

$$\epsilon < \min\{1 - |\lambda_i| \mid r_i = 0\}, \text{ when } (r,w) \in B_\delta.$$

Let

$$C_\delta = \{(r,w) \mid (r,w) \in B_\delta, \|w^+ - w\| < \delta, w^+ \text{ solves } (52)\}$$

where

$$[D_\theta A^T, -D_r Z]\begin{bmatrix} d_x \\ w^+ \end{bmatrix} = -D_r g. \tag{52}$$

Note that, since $Z_1\bar{w} = \bar{g}_1$, $(d_x = 0, \bar{w})$ solves (52).

Thus, when $(r,w) \in B_\delta$ is sufficiently close to $(\bar{r}, \bar{w})$, $(r^+, w^+) \in C_\delta$. For any $(r,w) \in C_\delta$, from (52),

$$|T_w(r,w) - \bar{w}| = |w - \bar{w} + w^+ - w| < 2\delta$$

Thus $T(C_{\delta,\epsilon}) \subset B_{2\delta}$.

Now, since $(\bar{r}, \bar{w})$ is a limit point, $\{(r^k, w^k)\}$ belongs to C_δ infinitely often. Suppose $\{(r^k, w^k)\}$ has another limit point, e.g. $(\hat{r}, \hat{w})$. From the nondegeneracy assumption, $\bar{w} \neq \hat{w}$. Then the sequence $\{(r^k, w^k)\}$ must leave C_δ infinitely often. Each time it leaves, it generates at least one point in $B_{2\delta} - C_\delta$. Since $\{(r^k, w^k)\}$ is bounded, we see that the subsequence of $\{(r^k, w^k)\}$ belonging to $B_{2\delta} - C_\delta$ must have a limit point. But every limit point has to belong to S which is disjoint from the closure of $B_{2\delta} - C_\delta$. This is a contradiction and consequently there can be only one limit point. ∎

It is now clear that under nondegeneracy assumptions, Algorithms 1 and 2 generate points that converge to the optimum point. This follows from Lemma 11, and Theorems 14,15.

6. Quadratic Convergence. In this section we establish that Algorithm 2 produces a sequence $\{(r^k, w^k)\}$ that converges to (r^*, w^*) at a *quadratic* rate. The primary difficulty in establishing this is the discontinuous nature of the first derivatives of ψ. This problem is circumvented by considering a finite set $\mathcal{F}$ of systems of nonlinear equations, where each system in $\mathcal{F}$ has the following form:

$$\begin{cases} \hat{F}_\nu(y) \stackrel{def}{=} D_r(g(\nu) - Z^T w) = 0, \\ Zr = 0, \end{cases} \tag{53}$$

where

$$y = \begin{bmatrix} r \\ w \end{bmatrix},$$

$$g(\nu)_i = \begin{cases} \text{sign}(r_i^*) & \text{if } r_i^* \neq 0 \\ \nu_i & \text{if } r_i^* = 0. \end{cases}$$

and ν_i can be either 1 or -1. Note that $y^* = (r^*, w^*)$ is a solution to each system; each system is continuously differentiable in a neighbourhood of y^* .

The Newton step at y^k, for each of the above systems F_ν, is defined by

$$J_\nu^k d_N^k = -F_\nu(y^k)$$

where

$$J_\nu^k = \begin{bmatrix} \text{diag}(g(\nu) - \lambda^k) & -D_r^k Z^T \\ Z & 0 \end{bmatrix}, \quad \lambda^k = Z^T w^k,$$

and

$$F_\nu(y^k) = \begin{bmatrix} \hat{F}_\nu(y^k) \\ 0 \end{bmatrix}. \tag{54}$$

Note that the hybrid step d^k satisfies

$$B_{\nu^k} d^k = -F_{\nu^k}(y^k), \tag{55}$$

where

$$B^k = \begin{bmatrix} \text{diag}(g^k)|D_\theta^k| & -D_r^k Z^T \\ Z & 0 \end{bmatrix}, \tag{56}$$

and $\nu^k = g^k$. It is clear that $F_{\nu^k} \in \mathcal{F}$ (i.e., d^k is a Newton–like step for some $F_{\nu^k} \in \mathcal{F}$). Of course $F_{\nu^k} \neq F_{\nu^{k+1}}$, in general, and therefore quadratic convergence is not automatic; however, a slight modification of Theorem 3.4 in [7] yields a viable approach.

THEOREM 16. *Let $\mathcal{F} = \{F_\nu : \Re^m \to \Re^m\}$ be a finite set of functions satisfying the assumptions that each F_ν is continuously differentiable in an open convex set D and there is a y^* in D such that $F_\nu(y^*) = 0$ and each $\nabla F_\nu(y^*)$ is nonsingular; let $\{B^k\}$ in $L(\Re^m)$ be a sequence of nonsingular matrices. Suppose that for some y^0 in D the sequence*

$$y^{k+1} = y^k - (B^k)^{-1} F_{\nu^k}(y^k), \qquad k = 0, 1, \cdots,$$

remains in D, $y^k \neq y^*$ for $k > 0$, and converges to y^*. Moreover, assume

$$\|B^k - \nabla F_{\nu^k}(y^*)\| = O(\|y^k - y^*\|). \tag{57}$$

Then $\{y^k\}$ converges quadratically to y^*.

We now show that Algorithm 2 can be described in a manner consistent with Theorem 16 and therefore quadratic convergence is achieved. Specifically, (57) must be established: the next four results establish several preliminary bounds.

First we show that the change in the dual variables, λ, is bounded be the distance to the solution. Recall the full space implementation (26) of Algorithm 2: the hybrid step d^k solves

$$[\mathrm{diag}(g^k)|D_\theta^k|A^T \ - D_r^k Z^T] \begin{bmatrix} d_x^k \\ d_w^k \end{bmatrix} = -D_r^k (g^k - Z^T w^k), \tag{58}$$

and $d^k = A^T d_x^k$. We denote the $m \times m$ matrix C^k by

$$C^k = [\mathrm{diag}(g^k)|D_\theta^k|A^T \ - D_r^k Z^T].$$

LEMMA 17. *Assume that $\{(r^k, w^k)\}$ is any subsequence, convergent to $y^* = (r^*, w^*)$, obtained by Algorithm 2. Then,*

$$\|\lambda^{k+1} - \lambda^k\| = \|d_w^k\| = O(\|y^k - y^*\|), \tag{59}$$

where $\lambda = Z^T w$.

Proof. Since

$$\begin{bmatrix} d_r^k \\ d_w^k \end{bmatrix} = -C^{k^{-1}} D_r^k (g^k - Z^T w^k),$$

we immediately obtain

$$\|w^{k+1} - w^k\| \ \leq \ L\|C^{k^{-1}}\|\|y^k - y^*\|$$

and since $\{\|(C^k)^{-1}\|\}$ is bounded, the result ensues. ∎

Next we show that θ is bounded by the distance to the solution.

LEMMA 18. *Suppose* $\{\theta^k\}$ *is defined as in (23) or (24). Then,*

$$\theta^k \leq L_1 \|y^k - y^*\|. \tag{60}$$

Proof. By definition of θ^k,

$$\theta^k \leq L_0 \sqrt{\sum_{i=1}^{m} r_i^{k2} (g_i^k - \lambda_i^k)^2}.$$

Thus, from (54) and (53), we have

$$\begin{aligned} \theta^k &\leq L_0 \|F_{\nu^k}(y^k)\|_2 \\ &= L_0 \|F_{\nu^k}(y^k) - F_{\nu^k}(y^*)\|_2 \\ &\leq L_1 \|y^k - y^*\|_2, \end{aligned}$$

for positive constants L_0, L_1. The last inequality comes from continuity and finiteness of $\mathcal{F}$. ∎

LEMMA 19. *Assume that an* l_1 *problem is primal and dual nondegenerate and that the sequence* $\{(r_k, w_k)\}$ *is generated by Algorithm 2. Then, for each* i *such that* $r_i^* = 0$

$$\alpha_i^k - 1 = O(\|y^k - y^*\|). \tag{61}$$

Moreover, the sequence of optimal stepsizes, $\{\alpha_*^k\}$, *converges to unity.*

Proof. First we establish (61). For any i corresponding to an activity at the solution,

$$\begin{aligned} \alpha_i^k - 1 &= \frac{-r_i^k}{d_i^k} - 1 \\ &= \left(\frac{r_i^k (g_i^k - (1-\theta^k)\lambda_i^k)}{r_i^k (g_i^k - \lambda_i^{k+1})} - 1\right), \quad \textit{by} \quad (30), \\ &= \frac{\lambda_i^{k+1} - \lambda_i^k + \theta^k \lambda_i^k}{g_i^k - \lambda_i^{k+1}}. \end{aligned}$$

But since i corresponds to an active constraint, then, by dual nondegeneracy, $g_i^k - \lambda_i^{k+1}$ is bounded from zero. Therefore, since $\theta^k \to 0$ and Lemma 17 holds, we have $\alpha_i^k - 1 = O(\|y^k - y^*\|)$ and $\lim_{k\to\infty} \alpha_i^k = 1$.

Next we establish that the optimal stepsize, α_*^k, converges to unity. First we note that the stepsizes to the breakpoints of the inactive constraints go to infinity. That is, if i corresponds to an inactive constraint at the solution,

$$\lim_{k\to\infty} \frac{-r_i^k}{d_i^k} = \infty. \tag{62}$$

This is true since $d_i^k \to 0$ but r_i^k is bounded from zero when i corresponds to an inactivity at the solution. Next we prove that the stepsizes to the breakpoints of the activities all go to unity. To establish this we first prove that [2]

$$|g_1^{kT} d_1^k| < |g_0^{kT} d_0^k|.$$

To see this recall, from dual feasibility and dual nondegeneracy, that

$$g_1^* A_1 = -A_0 \lambda_0^*, \quad \text{where } \max\{|\lambda_0^*|\} < 1.$$

Moreover, for k sufficiently large, it is clear that $g_1^k = g_1^*$. Hence, from Lemma 12, we have

$$g_1^{kT} d_1^k = -\lambda_0^{*T} d_0^*.$$

Since $|\lambda_0^*| < 1$, we have

$$-g_0^{kT} d_0^k \leq \sum_{i \in \mathcal{A}} |d_i^k|, \quad \text{for sufficiently large } k.$$

Therefore,

$$\frac{|g_1^{kT} d_1^k|}{|g_0{}^T d_0^k|} \leq \max\{|\lambda_0^*|\} < 1.$$

Thus, when k is sufficiently large

$$\frac{|g_1^{kT} d_1^k|}{|g_0{}^T d_0^k|} < 1,$$

which implies

$$|g_1^{kT} d_1^k| \leq |g_0^{kT} d_0^k| = -g_0^{kT} d_0^k, \quad \text{for sufficiently large } k. \tag{63}$$

Now let g^{k+} denote the gradient after crossing all the break points corresponding to the activities. Then, by (63),

$$\begin{aligned} g^{k+T} d^k &= -g_0^{kT} d_0^k + g_1^{kT} d_1^k \\ &> 0, \quad \text{for sufficiently large } k. \end{aligned}$$

This, together with (62), implies that the optimal steplength equals one of the steplengths corresponding to an activity. Thus, the optimal steplength, α_k^*, converges to unity. ■

[2] Notation: In this section subscript 0 refers to components corresponding to activities at the *solution*, i.e., $r_i^* = 0$. Subscript 1 refers to components corresponding to inactivities at the *solution*, i.e., $r_i^* \neq 0$.

LEMMA 20. *Assume* $\{(r^k, w^k)\}$ *is generated by Algorithm 2 where* $\{\alpha^k\}$ *is the stepsize; assume primal and dual nondegeneracy. Then*

$$|\alpha^k - 1| \leq L_2 \|y^k - y^*\|. \tag{64}$$

Proof. From the computation of the steplength (31) and convergence, it is clear that

$$\alpha^k = \alpha_i{}^k - \theta^k \alpha_i{}^k, \quad \text{for } k \text{ sufficiently large,}$$

for some $i \in \mathcal{A}$ (from the proof of Lemma 19).

From (59),

$$|\alpha_i{}^k - 1| = O(\|y^k - y^*\|); \tag{65}$$

therefore,

$$\begin{aligned} |\alpha^k - 1| &= |\alpha_i^k - 1 - \theta^k \alpha_i^k| \\ &\leq |\alpha_i^k - 1| + \theta^k \alpha_i^k \\ &\leq L_2 \|y^k - y^*\| \qquad \text{(using (65) and Lemma 18).} \end{aligned}$$

■

Denote $B^k = B_{\nu^k} S^k$ where

$$S^k = \operatorname{diag}\left(\begin{bmatrix} \frac{1}{\alpha^k} e_m \\ e_{m-n} \end{bmatrix} \right)$$

where B_{ν^k} is defined as in (56). But $y^{k+1} = y^k + S^{k-1} d^k$ where d^k is defined by (55); therefore, from Lemma 20, we have

$$\|S^k - I\| = O(\|y^k - y^*\|).$$

The hybrid step defined by Algorithm 2 satisfies

$$B^k(y^{k+1} - y^k) = -F_{\nu^k}(y^k), \qquad \text{for } k \text{ sufficiently large.}$$

LEMMA 21. *Assume* $\{y^k\}$ *is obtained through Algorithm 2; assume primal and dual nondegeneracy. Then*

$$\|B^k - \nabla F_{\nu^k}(y^*)\| \leq L \|y^k - y^*\|$$

for some $F_{\nu^k} \in \mathcal{F}$.

Proof. From continuity and Lemma 20, it is clear that

$$\begin{aligned} B^k - J^*_{\nu^k} &= (B_{\nu^k} - J^k_{\nu^k}) S^k + (J^k_{\nu^k} - J^*_{\nu^k}) S^k + (S^k - I) J^*_{\nu^k} \\ &= O(\|B^k_{\nu^k} - J^k_{\nu^k}\|) + O(\|y^k - y^*\|) + O(\|y^k - y^*\|) \end{aligned}$$

It is clear that, for k sufficiently large, $g_0^k = \text{sign}(g_0^k - \lambda_0^k)$, $g_1^k = \lambda_1^*$. Hence

$$\begin{aligned} & \|\text{diag}(g^k)|g^k - (1-\theta^k)\lambda^k| - (g^k - \lambda^k)\|_2 \\ = \; & \|\text{diag}(g_0^k)|g_0^k - (1-\theta^k)\lambda_0^k| - (g_0^k - \lambda_0^k)\|_2 + \|\text{diag}(g_1^k)|g_1^k - (1-\theta^k)\lambda_1^k| - (g_1^k - \lambda_1^k)\|_2 \\ \leq \; & \|\theta^k \lambda_0^k\|_2 + \|\lambda_1^* - \lambda_1^k\|_2 + \theta^k \|\lambda_1^k\|_2 + \|\lambda_1^* - \lambda_1^k\|_2 \end{aligned}$$

Thus

$$\|\text{diag}(g^k)|g^k - (1-\theta^k)\lambda^k - (g^k - \lambda^k)\|_2 = O(\|y^k - y^*\|_2).$$

Hence

$$B_{\nu^k} - J_{\nu^k}^k = \begin{bmatrix} \text{diag}(g^k)|D_\theta^k| - D_\lambda^k & 0 \\ 0 & 0 \end{bmatrix}.$$

Hence,

$$\|B_{\nu_k}^k - J_{\nu_k}^k\| = O(\theta^k).$$

But by Lemma 18,

$$\theta^k = O(\|y^k - y^*\|);$$

therefore,

$$\|B^k - J_{\nu_k}^*\| = O(\|y^k - y^*\|).$$

■

The assumptions of Theorem 16 are now established; quadratic convergence of Algorithm 2 follows immediately.

THEOREM 22. *Suppose the l_1 problem is primal and dual nondegenerate. Assume the sequence $\{(r^k, w^k)\}$ is obtained from the Algorithm 2. Then $y^k = \begin{bmatrix} r^k \\ w^k \end{bmatrix}$ converges quadratically to $y^* = \begin{bmatrix} r^* \\ w^* \end{bmatrix}$.*

7. Numerical Testing. In this section we provide numerical results concerning Algorithm 2, the hybrid method (*New*). In particular, we compare our range-space implementation of this algorithm - see Section 4 - with our implementation of the interior point algorithm described in [8]. We denote the latter method by *Dual*. Our comparisons are based on the number of iterations: since the dominant work in both implementations is the solution of a linear least squares problem at each iteration, of size m-by-n,the number of iterations accurately reflects the overall relative computational cost.

The dependent variable θ^k is a measure of the distance from optimality: our stopping criterion for both algorithms is:

$$\theta^k < 10^{-13},$$

where machine precision on our system is approximately 10^{-16}. In our experiments, we compute θ^k as defined by (24). The origin is a natural starting point for algorithm *Dual*; the starting point for *New* is computed as follows:

$$r^0 \leftarrow b - A^T x^0, \text{ where } A^T x^0 \overset{\text{l.s.}}{=} b$$

$$\lambda^0 \leftarrow \frac{\tau}{max\{abs\{r\}\}} * r.$$

The settings of the parameters for Algorithm 2, *New*, are:

$$\tau \leftarrow .975, \quad \gamma \leftarrow .99$$

Algorithm *Dual* also uses $\tau = .975$ in the linesearch algorithm; there are no other parameters.

We have implemented the methods in PRO-MATLAB [9] using SUN 3/50 and 3/160 workstations. The linear least squares subproblems are solved with the orthogonal QR-factorization using row interchanges for greater stability (row interchanges are advisable when a least squares problem involves widely varying row scalings [14]. No account was made of sparsity in our experiments.

Our experiments are not exhaustive; our purpose here is to determine the viability of our approach. Specifically, does the ultimate quadratic convergence appear to yield significantly fewer iterations in practise?

We have generated several kinds of test problems. First, since l_1 minimization is often used in a function approximation context [10], we have tried several such problems. We have also generated several random problems of varying dimensions. Finally, despite lack of theory for degenerate problems, we have experimented with degenerate and near-degenerate problems.

Problem 1. Approximate $f(z)$, evaluated at $z = 0, \frac{1}{m}, ..., 1$, by a polynomial of degree $n-1$:

$$\phi(z) \;=\; \sum_{j=1}^{n} \alpha_j \; z^{j-1} \;,$$

where $x = (\alpha_1, ..., \alpha_n)$.

$n = 5, f_1(z) = \exp(z)$.

Number of Steps		
m	Dual	New
100	18	8
200	19	11
400	21	10
600	21	12

$n = 6, f_2(z) = \sin(z)$

Number of Steps		
m	Dual	New
100	15	9
200	16	8
400	16	9
600	17	9

Problem 2. Approximate $f(z)$, evaluated at $z = 0, \frac{1}{m}, ..., 1$, by a polynomial of degree $n = 10$:

$$\phi(z) = \sum_{j=1}^{n} \alpha_j \, z^{j-1} \ ,$$

where $x = (\alpha_1, \cdots, \alpha_n)$.

$$f_3(z) = \exp(z) + \begin{cases} 1 & \text{if } 0.1 < z \leq 0.2 \\ 0 & \text{otherwise .} \end{cases}$$

Number of Steps		
m	Dual	New
40	30	13
100	27	12
200	29	12
500	33	12

Approximation problems often result in a very large number of rows, m. The number of iterations required by algorithm *New* is essentially insensitive to the growth of m, as demonstrated below.

Number of Steps			
m	n	f(z)	New
1000	5	$f_1(z)$	13
1000	6	$f_2(z)$	10
800	10	$f_3(z)$	12
1000	10	$f_3(z)$	15

Problem 3. Random problems: We generated the elements of matrix A and right-hand-side b in a uniform random manner.

$m = 50$

Number of Steps		
n	Dual	New
10	25	9
20	25	12
30	25	10
40	23	9

$m = 100$

Number of Steps		
n	Dual	New
10	25	10
20	26	13
40	26	13
50	26	11
70	25	11
90	23	10

$m = 150$

Number of Steps		
n	Dual	New
10	26	10
20	27	13
50	26	17
70	27	11
90	27	13
135	24	12
140	23	9

$m = 200$

Number of Steps		
n	Dual	New
10	26	13
20	27	17
50	27	16
70	27	19
100	26	15
140	26	14
160	25	14
190	24	9

We do not have any theoretical results for degenerate problems; however, both algorithms can still be used on such problems - performance appears to be unaffected by degeneracy.

Problem 4. Dual degenerate problems: In the following table results obtained on dual degenerate problems are presented. The parameter *ndeg* is the number of activities with corresponding multipliers of unit size. In each problem the number of rows, m, is equal to 100.

Number of Steps			
n	ndeg	Dual	New
50	5	27	14
50	10	27	13
70	15	25	11
70	20	27	10

Primal degenerate problems: In the following table results obtained on primal degenerate problems are presented. At the solution, there are *ndeg* redundant activities. The number of rows, m, equals 100.

Number of Steps			
n	ndeg	Dual	New
20	5	25	9
40	10	25	12
70	15	24	10
90	9	9	2

It turns out that near-degeneracy is more troublesome than exact degeneracy for algorithm *New*. The reason is probably related to the existence of a small radius of convergence for Newton's method. The following table reflects performance of algorithm *New* on primal near-degenerate problems. The problems are primal near-degenerate in that two of the inactive equations have small residuals at the solution, i.e., residual values are about 10^{-9}.
Problem 5.

$m = 50$

Number of Steps		
n	Dual	New
10	25	14
20	25	15
30	25	16
40	25	18

$m = 100$

Number of Steps		
n	Dual	New
10	25	16
20	26	21
40	26	21

Remarks: In general, the quadratically convergent Algorithm 2 (*New*) provides a clear advantage over the (linear) interior point method (*Dual*) on our test collection. The number of iterations required by *New* is always fewer, usually by a factor between two and three. *New* wins by the smallest margin on near-degenerate problems.

The number of iterations required by *New* is not quite as consistent, over different problems and different dimensions, as the number required by *Dual*. For example, on the random problems, *Dual* varies between 23 and 27 iterations whereas *New* ranges between 9 and 17. This greater variation is probably due to different radii of convergence for Newton's method. Nevertheless, the number of required iterations is almost never more than 15 and is considerably more consistent than number of iterations required by a finite pivoting algorithm.

8. Conclusions. We have presented a new global and quadratic algorithm for the linear l_1 problem. Computationally, the algorithm appears to be consistently superior to the interior point approach [8], typically requiring a factor of two to three fewer iterations. (The iterations are comparable in cost and so the overall running times compare in a similar way.)

The quadratic convergence property stems from the nonlinear expression for complementary slackness and the resulting Newton-like process; the integration of the Newton process with a (linear) global affine scaling method comes in part through the use of a "square-root" scaling matrix to define the linear algorithm (Algorithm 1). The use of this scaling is compatible with the fact that "constraint lines", i.e., $r_i = 0$, must be crossed, in general.

The linear l_1 algorithm proposed here is probably most applicable to large-scale problems. There are several reasons for this. First, Algorithm 2 (*New*) seems fairly insensitive to problem size: it appears that about 15 iterations will be required, regardless of dimension. Therefore, this method will be increasingly attractive, relative to finite pivoting algorithms, as the dimension grows. Second, for sparse problems the technology required to exploit sparsity is only that required for solving a sequence of sparse linear least squares problems: this is a heavily studied problem with techniques available and research ongoing. Finally, the new method is attractive for parallel computation in that the number of outer iterations (sequential steps) is modest (about 15); moreover, parallel techniques for solving linear least-squares problems have been developed (e.g, [4] and [3]) and research is continuing in this area.

9. Acknowledgement. The authors would like to thank their colleague, Mike Todd, for numerous helpful suggestions.

REFERENCES

[1] I. Barrodale and F. Roberts, *An improved algorithm for discrete l_1 linear approximation*, SIAM J. Num. Anal., 10 (1972), pp. 839–848.

[2] R. H. Bartels, A. R. Conn, and J. W. Sinclair, *Minimization techniques for piecewise differentiable functions: the l_1 solution to an overdetermined linear system*, Siam J. Numer. Anal., 15 (1978), pp. 224–240.

[3] C. Bischof, *Qr factorization algorithms for coarse-grained distributed systems*, Tech. Rep. 88-939, Cornell University, 1988.

[4] T. F. Coleman and P. Plassmann, *Solution of nonlinear least-squares problems on a multiprocessor*, Tech. Rep. CS-87-923, Computer Science Department, Cornell University, 8987.

[5] T. F. Coleman and A. Pothen, *The null space problem I: Complexity*, SIAM Journal on Algebraic and Discrete Methods, 7 (1987), pp. 527–537.

[6] ——, *The null space problem II: Algorithms*, SIAM Journal on Algebraic and Discrete Methods, 8 (1987), pp. 544–563.

[7] J. E. Dennis, Jr. and J. J. Moré, *Quasi-newton methods, motivation and theory*, SIAM Review, 19 (1977), pp. 46–89.

[8] M. S. Meketon, *Least absolute value regression*, tech. rep., AT & T Bell Laboratory, 1987.

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

[10] M. R. Osborne, *Finite Algorithms in Optimization and Data Analysis*, John Wiley & Sons, 1985.

[11] J. Renegar, *A ploynomial-time algorithm, based on newton's method, for linear programming*, Mathematical Programming, (1988), pp. 59-93.

[12] S. A. Ruzinsky and E. T. Olsen, *l_1 and l_∞ minimization via a variant of Karmarkar's algorithm*, IEEE Transactions on Ascoustics Speech and Signal Processing, 37 (1989), pp. 245-253.

[13] E. Seneta and W. L. Steiger, *A new lad curve-fitting algorithm: Slightly overdetermined equation systems in l_1*, Discrete Applied Mathematics, 7 (1984), pp. 79-91.

[14] C. Van Loan, *On the method of weighting for equality-constrained least squares problems*, SIAM Journal on Numerical Analysis, 22 (1985), pp. 851-864.