A Global and Quadratically-Convergent Method for Linear L_∞ Problems

Thomas Coleman*
Yuying Li**

TR 90-1121
April 1990

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

*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.

**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.

A GLOBAL AND QUADRATICALLY-CONVERGENT METHOD FOR LINEAR L_∞ PROBLEMS

THOMAS F. COLEMAN* AND YUYING LI†

April 20, 1990

Abstract. We propose a new global and quadratically convergent algorithm for the linear l_∞ problem. This method works on the piecewise linear l_∞ problem directly by generating descent directions - via a sequence of weighted least squares problems - and using piecewise linear linesearches to ensure a decrease in the l_∞ function at every step. We prove that ultimately full Newton-like steps are taken where the Newton step is based on the complementary slackness condition holding at the solution. Numerical results suggest a very promising method; the number of iterations required to achieve high accuracy is relatively insensitive to problem size.

Key Words. Linear l_∞ estimation, linear programming, interior-point algorithm, simplex method, affine scaling method, Karmarkar, discrete Chebyshev problem

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

Abbreviated Title. A quadratic method for linear l_∞ problems

* 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. Given a matrix $A \in \Re^{n \times m}$ and a set of data $b \in \Re^m$, a common problem is to find a vector $x \in \Re^n$ such that the linear model $A^T x$ closely matches the data b. Therefore, the following problem is important:

$$\min_{x \in \Re^n} \|A^T x - b\|$$

where the most often used measures are 2–norm, 1–norm, and ∞–norm. The 2-norm solution, by far the most popular choice, can be obtained in a single step, e.g., using a QR factorization of A^T.

There are situations where it is preferable to use either $\|\cdot\|_1$ or $\|\cdot\|_\infty$; however, the resulting numerical problems are much more difficult. For example, the piecewise linear functions can be minimized by forming an equivalent linear programming problem with special structure. A tailored simplex method can then be used (e.g. [1], [2]). Alternatively, the linear programming formulations can be solved using an interior point method [9].

In both approaches indicated above, the solution techniques are iterative; however, the approaches differ in that in the first case the sequence of points generated is finite, whereas in the second, assuming exact real arithmetic, an infinite sequence is generated (theoretically). The sequence produced by an interior point method converges at best linearly: this is one place where an improvement can be made. Indeed, Coleman and Li [4] have developed a global and quadratically convergent method for the $\|\cdot\|_1$ case. In [5], this algorithm has been subsequently extended to solve linear programming problems with upper and lower bounds on all variables. (A related algorithm for minimization of a convex quadratic function, subject to bounds on the variables, is given by Coleman and Hulbert [3]). This approach bears some resemblance to the interior point methods, a sequence of weighted least-squares problems is solved, but it also has some distinct differences. For example, the iterates are not feasible, the l_1-function is decreased at each iteration, piecewise linear minimization is performed, and the ultimate convergence rate is quadratic.

The purpose of this paper is to propose an algorithm for l_∞-minimization that is similar in spirit to the l_1 algorithm [4].

The l_1 algorithm proposed in [4] is a descent direction algorithm. Defining a good descent direction is nontrivial due to the hyperplanes of nondifferentiability, $a_i^T x = b_i$. The manner in which the Coleman-Li l_1 algorithm deals with nondifferentiability can be summarized as follows:

1. The l_1 algorithm generates differentiable points. Therefore, the gradient direction is defined at each iteration.
2. When far from the solution the negative gradient direction is scaled by the squareroot of the distances to the "constraint boundaries". This is done by first globally transforming the variables so that the new variables correspond to the distances to the constraints. This scaling helps avoid small steps.
3. Given a descent direction, lines of nondifferentiability are crossed provided the l_1 function continues to decrease.
4. Asymptotically, unit Newton steps are taken (Newton steps are defined with respect to the complementary slackness condition) thus ensuring quadratic convergence.

The beauty of this approach is that the first-order direction and the Newton step can be combined in a smooth and automatic way [4].

Here we develop a similar approach for the l_∞ problem; however, this is not a trivial extension because there is no global transformation similar to that referred to above (in which the new variables themselves correspond to the distances to the planes of nondifferentiability). This is because the index to the maximum residual, j say, can change from iteration to iteration and the planes of nondifferentiability are defined, locally, by

$$|a_i^T x - b_i| = |a_j^T x - b_j|, \quad i \neq j$$

(unlike the l_1 case where $a_i^T x - b_i = 0$ indicates nondifferentiability). Without a global transformation the adaptation of the l_1 approach is not obvious.

In a nutshell, this problem can be overcome by using *local* transformations: at step k we define a matrix T^k that transforms the current "residuals" to variables. In this paper we show that this can be done efficiently; moreover, the resulting method is global and ultimately quadratically convergent.

The linear l_∞ problem is:

$$\min_{x \in \Re^n} \max_{1 \leq i \leq m} |a_i^T x - b_i|.$$

We use the sign function $\mathrm{sgn}(v)$, where v is a vector, in the following sense: if $w = sgn(v)$,

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

Denote the residual vector $A^Tx - b$ by r and $\sigma = sgn(r)$. At any point r, let $\mathcal{A}(r)$ denote the indices of residuals with the maximum magnitude, i.e.,

$$\mathcal{A}(r) = \{\ l \mid |r_l| = \max_{1 \le i \le m} |r_i| \}.$$

DEFINITION 1. *We say an l_∞ problem is* **primal nondegenerate** *if and only if, at any point r, $\{\sigma_j a_j - \sigma_i a_i \mid i \in \mathcal{A} - \{j\}\}$ is of full rank for some $j \in \mathcal{A}(r)$.*

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

Note: If the matrix A satisfies the Haar condition, the problem is both primal and dual nondegenerate.

2. An Affine Scaling Algorithm. In this section we define a new method for the l_∞-problem. This method does not possess second-order convergence; however, it is new, it is globally convergent, and it can be combined with the Newton ideas, discussed in the next section, to ultimately yield a global and second-order method (Section 3).

The objective function $\psi(r)$, defined in (2.1) below, is not differentiable at the points where more than one residual has maximum magnitude, i.e., $|r_l| = |r_j| = \max_{1 \le i \le m} |r_i|$, $l \ne j$. However, if there is only a single maximum function then the negative gradient of ψ is well-defined and corresponds to a descent direction for ψ. Unfortunately, it may be a poor descent direction since it can lead to a very small step if a line of nondifferentiability is immediately encountered. This can happen when there are near-activities, i.e., several residuals are near maximum in magnitude. Therefore, it is preferable to introduce a scaling to (partially) avoid near-active functions. To do this we define a local transformation.

Let Z denote an $(m - n) \times m$ matrix, with rank $m - n$, such that

$$AZ^T = 0.$$

Then the l_∞ problem is equivalent to the following constrained l_∞ problem with m variables r:

$$\min_{r \in \Re^m} \psi(r) \stackrel{def}{=} \|r\|_\infty \qquad \text{subject to} \quad Zr = Zb. \tag{2.1}$$

At our current point r, assume $|r_j| = \max_{1\leq i\leq m} |r_i|$, and there is no other maximum valued residual. Therefore, $\|r\|_\infty$ is differentiable in a neighbourhood of the current point r, and the nearby nondifferentiable region is (locally) defined by the hyperplanes $|r_j| - |r_i| = 0$. If we define vector s as

$$s_i = \sigma_j r_j - \sigma_i r_i, \quad i \neq j \tag{2.2}$$

$$s_j = \sigma_j r_j, \tag{2.3}$$

then component i of s represents the distance to hyperplane $|r_i| - |r_j| = 0$, $i \neq j$. Alternatively, s can be written $s = T^{-1}r$ where T is a simple elementary matrix:

$$T = [-\sigma_1 e_1, \cdots, -\sigma_{j-1}e_{j-1}, \sigma, -\sigma_{j+1}e_{j+1}, \cdots, -\sigma_m e_m]. \tag{2.4}$$

Note that

$$T^{-1} = [-\sigma_1 e_1, \cdots, -\sigma_{j-1}e_{j-1}, \sigma_j e, -\sigma_{j+1}e_{j+1}, \cdots, -\sigma_m e_m].$$

Now problem (2.1) becomes

$$\begin{array}{rl} & \min_{s\in\Re^m} \|Ts\|_\infty \\ \text{subject to} & ZTs = Zb. \end{array} \tag{2.5}$$

Locally, the nondifferentiable points are simply $s_i = 0$ for some i, $i \neq j$.

2.1. The Search Direction. Assume r is a differentiable point and let $g = \nabla\psi = \sigma_j e_j$, $D = \mathrm{diag}\{s_i^{\frac{1}{2}}\}$, and $T = T(r)$ is defined by (2.4). We solve the following subproblem to determine a descent direction:

$$\begin{array}{rl} & \min_{\hat{d}_s\in\Re^m} g^T T\hat{d}_s \\ \text{subject to} & ZT\hat{d}_s = 0 \\ & \|D^{-1}\hat{d}_s\|_2 \leq \delta. \end{array} \tag{2.6}$$

Then $\hat{d}_s = \alpha d_s$ where

$$\begin{aligned} d_s &= -T^{-1}A^T(AT^{-T}D^{-2}T^{-1}A^T)^{-1}Ag \\ &= -D^2T^T(g - Z^Tw^+) \end{aligned} \tag{2.7}$$

or

$$\begin{aligned} d_r &= -A^T(AT^{-T}D^{-2}T^{-1}A^T)^{-1}Ag \\ &= -TD^2T^T(g - Z^Tw^+) \end{aligned} \tag{2.8}$$

where w^+ is defined by

$$DT^TZ^Tw^+ \overset{\text{l.s.}}{=} DT^Tg.$$

So, for example, we can compute the search direction d_r by

$$\text{(2.9)} \qquad \begin{cases} D^{-1}T^{-1}A^Td_x \overset{\text{l.s.}}{=} DT^Tg \\ d_r = -A^Td_x \\ \lambda^+ = g - T^{-T}(D)^{-2}T^{-1}d_r \end{cases}$$

where $\lambda^+(= Z^Tw^+)$ denotes the dual variables.

2.2. The Algorithm. We compute $d = d_r$ as suggested by (2.9) and then define α using a piecewise linear minimization technique along the ray d (allowing for the ability to cross lines of nondifferentiability). This is done by considering each breakpoint (intersection of a residual $|r_i + \alpha d|$ with the maximum residual $|r_j + \alpha d|$) 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_l is the smallest positive breakpoint, then a step just beyond this point yields the following gradient: $g^+ = \sigma_l^+ e_l$. If $(g^+)^T d = \sigma_l^+ d_l < 0$, the intersections with $|r_l|$ are considered, etc.

The breakpoints corresponding to intersections with r_j can be computed as ,

$$\text{(2.10)} \qquad J = \{\alpha_i : \ \alpha_i = -\frac{|r_j| - |r_i|}{\sigma_j d_j - \sigma_i d_i}, \text{ or } \alpha_i = -\frac{|r_j| + |r_i|}{\sigma_j d_j + \sigma_i d_i}\}.$$

Asymptotically, the breakpoint will be computed by the first formula because the second one corresponds to the case where r_i changes sign and then intersects with $|r_j|$. From (2.7), it is clear that $\alpha_i = (\sigma_i \lambda_i^+)^{-1}$. Hence the stepsize is actually determined by the dual multipliers.

The convergence of the algorithm requires the ability to cross at least one intersecting hyperplane if the current maximum residual r_j is not active at the solution. This means that the descent direction has to remain a descent direction after crossing at least one of the intersecting hyperplanes. This is easier to manage if we can ensure that the first intersecting hyperplane is distinctly closer than the rest. However, if all multipliers have the same value at some nonoptimal vertex, then the stepsizes to the breakpoints, $(\alpha_i \lambda_i^+)^{-1}$ may become indistinguishable. Therefore, to ensure separation of a first breakpoint from the rest, we introduce another diagonal scaling matrix, D_I. Usually $D_I = I$; however, when it appears that the first two breakpoints are converging to the same point (i.e., if equations i and j define breakpoints very close to each other along our search direction, then $|\frac{\sigma_i\lambda_i^+}{\sigma_j\lambda_j^+} - 1|$ will be small) one of the entries of D_I is set to $\frac{1}{2}$. The effect of this perturbation (*Step* 3) is to generate a direction, in the next iteration, in which the first breakpoint is clearly separated from the rest.

Finally, we restrict our step to be just shy of the true minimizing point along our search direction in order to avoid a point of nondifferentiability. The parameter τ is used for this purpose; τ is typically a number less than but very close to unity, e.g., $\tau = .975$.

Line Search Procedure 1: given τ, $\alpha_\ell \leftarrow 0$;

Step 1 Compute the set of breakpoints J with the current maximum residual r_j by (2.10);

Step 2 Determine the smallest breakpoint $\alpha_l = \min\{\alpha_i : \alpha_i > \alpha_\ell\}$. If $(g^+)^T d = \sigma_l^+ d_l < 0$, $\alpha_\ell \leftarrow \alpha_l$, $j \leftarrow l$, go to Step 1. Otherwise, continue;

Step 3 Let $\alpha_\iota = \min\{\alpha_i : \alpha_i > \alpha_l\}$. If $\alpha_\ell = 0$ and

$$|\frac{\sigma_l\lambda_l^+}{\sigma_\iota\lambda_\iota^+} - 1| < 1 - \tau,$$

we modify D_I:

$$(D_I)_{ll} \leftarrow \frac{1}{2} \tag{2.11}$$

Step 4 Compute the stepsize:

$$\alpha = \alpha_\ell + \tau(\alpha_l - \alpha_\ell) \tag{2.12}$$

return;

The (infinite) algorithm follows. In practice the loop is terminated when we are deemed close enough to the solution: see Section 7 for more details.

Algorithm 1: Let r^0 be an initial differentiable point satisfying $Zr^0 = Zb$; $k \leftarrow 0$; $D_I^0 \leftarrow I$.

Step 1 Define $T^k = T(r^k)$, $s^k \leftarrow (T^k)^{-1}r^k$, $D^k = \text{diag}\{(s_i^k)^{\frac{1}{2}}\} \cdot D_I^k$, and $g^k = \sigma_j^k e_j$;

Step 2 Compute d^k and λ^{k+1} by (2.9);

Step 3 Set $D_I^{k+1} = I$. Do a line search on the piecewise linear function $\psi(r^k + \alpha d^k)$, as described above, to determine α^k. Then

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

3. A Local Newton Process. Close inspection of the optimality conditions for the linear l_∞ problem can yield a local, quadratically convergent, Newton process.

Vector x is optimal if and only if there exists a vector μ such that

$$\sum_{i\in\mathcal{A}(x)} a_i\sigma_i\mu_i = 0 \tag{3.1}$$

$$\sum_{i\in\mathcal{A}(x)} \mu_i = 1, \quad \mu_i \geq 0. \tag{3.2}$$

Note that the constraint $\sum_{i\in\mathcal{A}}\mu_i = 1$ is artificially imposed to obtain a unique definition for the multipliers μ_j: this is standard.

Let $j \in \mathcal{A}$ such that $\mu_j > 0$. Then, $\mu_j = 1 - \sum_{i\in\mathcal{A}-\{j\}}\mu_i > 0$ and optimality conditions can be (equivalently) stated:

$$\sigma_j a_j = \sum_{i\in\mathcal{A}-\{j\}} \mu_i(\sigma_j a_j - \sigma_i a_j) \tag{3.3}$$

$$\mu_i \geq 0, \; 1 - \sum_{i\in\mathcal{A}-\{j\}} \mu_i > 0. \tag{3.4}$$

Note that, if T is defined by (2.4), (3.3) can be expressed

$$\sigma_j a_j = \sum_{i\in\mathcal{A}-\{j\}} \mu_i(AT^{-T})e_i. \tag{3.5}$$

Alternatively, consistent with the formulation in (2.1), r is optimal if and only if there exist vectors μ, w such that

$$T^T g = \sum_{i\in\mathcal{A}-\{j\}} \mu_i e_i + T^T Z^T w \tag{3.6}$$

$$Zr = Zb \tag{3.7}$$

$$\mu_i \geq 0, \quad 1 - \sum_{i\in\mathcal{A}-\{j\}} \mu_i > 0. \tag{3.8}$$

This equivalence can be seen by comparing (3.5) with (3.6): e.g., multiply both sides of (3.6) by AT^{-T} and recall that $g = \sigma_j e_j$. Note also from (3.6), assuming $\lambda = Z^T w$: if $i \in \mathcal{A}$ then $\lambda_i = \sigma_i\mu_i$; otherwise, $\lambda_i = 0$.

Now here is the key: system (3.6) can be viewed as a system of nonlinear equations. In particular, (3.6) is equivalent to:

$$D_s T^T(g - Z^T w) = 0 \tag{3.9}$$

where $D_s = \mathrm{diag}(s)$. Equivalently, in terms of r,

$$D_r T^T(g - Z^T w) = 0 \tag{3.10}$$

where $D_r \stackrel{def}{=} \text{diag}(|r_j|e - |r| + \xi e_j)$ where $\xi = \|r\|_\infty$.

A (local) Newton iteration can be based on the nonlinear system (3.10) and the linear constraints (3.7). Nevertheless, note that T is dependent on the choice of active function j used to define it; this choice need not stabilize. However, in a neighborhood of the optimal solution (r^*, w^*), there are only finite number of possiblities for j (there are $n+1$ activities when the problem is nondegenerate) and each nonlinear equation, corresponding to fixed index j, has the common root (r^*, w^*). As we formally establish in Section 6, the local quadratic convergence behavour of a Newton process is maintained under these circumstances.

Define $g = \nabla\psi$, $D_s = \text{diag}(s)$, and $D_\lambda = \text{diag}(T^T(g - Z^T w))$; differentiate (3.9) to yield a Newton correction,

$$\begin{bmatrix} D_\lambda & -D_s T^T Z^T \\ Z & 0 \end{bmatrix} \begin{bmatrix} \Delta s \\ \Delta w \end{bmatrix} = \begin{bmatrix} -D_s T^T (g - Z^T w) \\ 0 \end{bmatrix}, \tag{3.11}$$

or

$$\begin{bmatrix} D_\lambda T^{-1} & -D_r T^T Z^T \\ Z T^{-1} & 0 \end{bmatrix} \begin{bmatrix} \Delta r \\ \Delta w \end{bmatrix} = \begin{bmatrix} -D_r T^T (g - Z^T w) \\ 0 \end{bmatrix}.$$

Define $d_N = \Delta r$; it is easy to prove that the Newton step d_N can be expressed

$$d_N = -A^T (A T^{-T} D_r^{-1} D_\lambda T^{-1} A^T)^{-1} A g \tag{3.12}$$

LEMMA 1. *Assume (r^*, w^*) is a solution and the l_∞ 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 $|r_i| < |r_j| = \|r\|_\infty$, for all $i \neq j$, the matrix $A T^{-T} D_r^{-1} D_\lambda T^{-1} A^T$ is positive definite.*

Proof. The proof is similar to Lemma 1 in [4]. ∎

Therefore, the Newton direction becomes a descent direction in the neighborhood of the solution.

4. Globalization. The similarity in form between the linear step, Section 2.1, and the Newton step, Section 3, suggests a possible hybrid method. The linear step can be expressed

$$d_r = -A^T (A T^{-T} \underbrace{D_r^{-1}} T^{-1} A^T)^{-1} A g$$

whereas, from (3.12), the Newton step equals

$$d_N = -A^T(AT^{-T}\underbrace{D_r^{-1}D_\lambda}T^{-1}A^T)^{-1}Ag.$$

Our idea is to define a matrix D_θ such that the matrix $AT^{-T}D_r^{-1}D_\theta T^{-1}A^T$ goes smoothly from $AT^{-T}D_r^{-1}T^{-1}A^T$ to $AT^{-T}D_r^{-1}D_\lambda T^{-1}A^T$, as we converge.

Define

$$D_\theta = \text{diag}((1-\theta)|T^T(g-\lambda)| + \theta I) \tag{4.1}$$

where $\lambda = Z^T w$. Let

$$d = -A^T(AT^{-T}\underbrace{D_r^{-1}D_\theta}T^{-1}A^T)^{-1}Ag. \tag{4.2}$$

It is clear that as $\theta \to 0$, vector d converges to a Newton step; when $\theta = 1$, $d = d_r$.

Our remaining task is to define $\theta \in [0,1]$ so that $\theta \to 0$ if and only if $(r,\lambda) \to (r^*,\lambda^*)$. We let θ encapsulate the optimality conditions. One possible choice is: define vector v: $v_i = max\{-\sigma_i\lambda_i, 0\}$;

$$\theta = \frac{\frac{\|D_rT^T(g-\lambda)\|_2}{\psi(r^0)} + \|v\|_\infty}{\gamma + \frac{\|D_rT^T(g-\lambda)\|_2}{\psi(r^0)} + \|v\|_\infty} \tag{4.3}$$

where $0 < \gamma < 1$. Clearly θ is bounded above by 1; assuming $Z(r-b) = 0$, then $\theta = 0$ if and only if $(r,\lambda) = (r^*,\lambda^*)$.

In order to obtain final quadratic convergence, we must have $\alpha \to 1$ sufficiently fast. Fortunately, if r^k converges to r^*, then the breakpoints of the activities with r_k^k converge to unity at a fast enough rate (see Section 6). Thus, when θ^k is small, the stepsize is essentially determined by the breakpoints along d^k.

The stepsize procedure follows. The parameter τ plays the same role as in Procedure 1: τ is initially less than but close to unity, e.g., $\tau = .975$. Logical variable *mod* indicates whether or not a perturbation to the diagonal of D_θ should be made (to help bypass nonoptimal vertices - see Section 2 for more discussion). In particular, if $\{r^k\}$ converges and $\bar{\theta}$ is the limit point of $\{\theta^k\}$ then for each function active at the limit point:

$$\alpha_i \to \frac{\bar{\theta} + (1-\bar{\theta})\bar{\sigma}_i\bar{\lambda}_i}{\bar{\sigma}_i\bar{\lambda}_i}, \qquad i \neq j$$

where α_i corresponds to the breakpoint for function i. Therefore, the breakpoints may not be well-separated in the neighbourhood of an nonoptimal point when all multipliers, corresponding to active functions, are equal. We introduce a slight perturbation, identified when $mod = true$, to avoid this (rare) case.

Line Search Procedure 2 : Given τ, set $\alpha_\ell \leftarrow 0$, mod= false

Step 1 Compute the set of break points J with the current maximum residual r_j by (2.10);

Step 2 Determine the smallest break point α_l, $\alpha_l = \min\{\alpha_i : \alpha_i > \alpha_\ell\}$.
If we continue to descend (i.e. ${g^+}^T d < 0$) and we are not close to a solution (e.g., $\theta > 0.01$), $\alpha_\ell \leftarrow \alpha_l$, $j \leftarrow l$, go to Step 1;
If we continue to descend but we are close to a solution (e.g., $\theta \leq 0.01$), then $\alpha_\ell \leftarrow \alpha_l$, $\alpha_l = \min\{\alpha_i : \alpha_i > \alpha_\ell\}$, go to Step 4;

Step 3 Let $\alpha_\iota = \min\{\alpha_i : \alpha_i > \alpha_\ell\}$. If $\alpha_\ell = 0$ and

$$\left|\frac{\sigma_l \lambda_l^+}{\sigma_\iota \lambda_\iota^+} - 1\right| < 1 - \max(\tau, 1-\theta),$$

Set $mod = true$, $index = l$;

Step 4 Compute the stepsize:

$$\alpha = \alpha_\ell + \max(\tau, 1-\theta)(\alpha_l - \alpha_\ell). \tag{4.4}$$

return;

The (infinite) algorithm follows; assume $mod = true$ initially. In Section 7 we give our stopping criteria.

Algorithm 2 : Let r^0 be an initial differentiable point satisfying $Zr^0 = Zb$; $k \leftarrow 0$;

Step 1 Let j be the index to the maximum residual; Compute θ^k from (4.3);
Set D_θ^k by (4.1);

Step 2 If $mod = true$ then $l \leftarrow index$, $(D_\theta^k)_{ll} \leftarrow (D_\theta^k)_{ll} - \frac{\theta^k}{2}$;

Step 3 $D^k = (\text{diag}(s^k)(D_\theta^k)^{-1})^{\frac{1}{2}}$; $g^k \leftarrow \sigma_j^k e_j$;

Step 4 Compute d^k and λ^{k+1}:

$$\begin{cases} (D^k)^{-1}(T^k)^{-1}A^T d_x^k \stackrel{\text{l.s.}}{=} D^k(T^k)^T g^k; \\ d^k \leftarrow -A^T d_x^k; \end{cases} \tag{4.5}$$

$$\lambda^{k+1} = g^k - {T^k}^{-T}(D^k)^{-2}(T^k)^{-1}d^k;$$

Step 5 Do a line search on the piecewise linear function $\psi(r^k+\alpha d^k)$ (as described above) to determine α^k:

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

Note: There are alternative ways to compute the search direction. For example, the following extended system can be used, provided Z is available:

$$\begin{bmatrix} D_\theta^k(T^k)^{-1} & -D_r^k(T^k)^T Z^T \\ Z & 0 \end{bmatrix} \begin{bmatrix} d^k \\ \Delta w^k \end{bmatrix} = \begin{bmatrix} -D_r^k(T^k)^T(g^k - Z^T w^k) \\ 0 \end{bmatrix}. \tag{4.6}$$

and $w^{k+1} \leftarrow w^k + \Delta w^k$, $\lambda^{k+1} \leftarrow Z^T w^{k+1}$.

5. Global Convergence. In this section, we establish global convergence of the linear and hybrid methods, Algorithms 1 and 2. All the proofs are similar in spirit to those for l_1 [4]. The main difference comes from the line search technique which is reflected in Lemma 2, Lemma 7, and Theorem 8.

We make the following global assumption:

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

Let $S^k = T^k D^k$ and P^k be the orthogonal projector onto null(ZS^k), i.e.,

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

The diagonal matrix $D^k = D_r^k (D_\theta^k)^{-1}$ where D_θ^k is defined by (4.1) and (4.2) for Algorithm 2, (let $D_\theta^k = D_I^k$ for Algorithm 1).

Both algorithms generate the search direction d^k:

$$\begin{aligned} d^k &= -S^k P^k S^{k^T} g^k \\ &= -(S^k S^{k^T})(g^k - Z^T w^{k+1}) \\ &= -(T D^{k^2} T^T)(g^k - Z^T w^{k+1}) \end{aligned} \tag{5.1}$$

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

$$S^{k^T} Z^T w^{k+1} \overset{\text{l.s.}}{=} S^{k^T} g^k.$$

Alternatively, we can compute the step d by $d^k = A^T d_x^k$ where

$$[D_\theta^k T^{-1} A^T, \ -D_r^k T^T Z^T] \begin{bmatrix} d_x^k \\ w^{k+1} \end{bmatrix} = -D_r^k T^T g^k. \tag{5.2}$$

The first major step in the convergence proof is to show that $\|P^kS^kg^k\| \to 0$. This is established in Lemma 5 after several preliminary results.

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

$$\lim_{k\to\infty} \bar{\alpha}^k \|P^kS^{k^T}g^k\|_2^2 = 0 \tag{5.3}$$

where $\bar{\alpha}^k$ corresponds to the first breakpoint in direction d^k (starting from point r^k).

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

$$\lim_{k\to\infty} (\|r^k\|_\infty - \|r^{k+1}\|_\infty) = 0. \tag{5.4}$$

Assume l corresponds to the index of the maximum residual at $r^k + \alpha^k d^k$. Then, it is clear that $\sigma_l^{k+1}d_l^k < 0$. Thus

$$\begin{aligned}
&\|r^k\|_\infty - \|r^{k+1}\|_\infty \\
= \ & g^{k^T}r^k - g^{k+1^T}r^k - \alpha^k g^{k+1^T}d^k && (\text{since } r^{k+1} = r^k + \alpha^k d^k), \\
= \ & (g^k - g^{k+1})^T r^k + \alpha^k (g^k - g^{k+1})^T d^k + \alpha^k \|P^kS^{k^T}g^k\|_2^2 && (g^{k^T}d^k = -\|P^kS^{k^T}g^k\|_2^2); \\
= \ & \sigma_j^k r_j^k - \sigma_l^{k+1}r_l^k + \alpha^k(\sigma_j^k d_j^k - \sigma_l^{k+1}d_l^k) + \alpha^k\|P^kS^{k^T}g^k\|_2^2 && (\text{ note} : g^k = \sigma_j^k e_j\ ) \\
= \ & \sigma_j^k r_j^k - \sigma_l^{k+1}r_l^k + \bar{\alpha}^k(\sigma_j^k d_j^k - \sigma_l^{k+1}d_l^k) - (\alpha^k - \bar{\alpha}^k)\sigma_l^{k+1^T}d_l^k + \bar{\alpha}^k\|P^kS^{k^T}g^k\|_2.
\end{aligned}$$

From

$$\sigma_j^k r_j^k - \sigma_l^{k+1}r_l^k + \bar{\alpha}^k(\sigma_j^k d_j^k - \sigma_l^{k+1}d_l^k) \geq 0$$

we claim that

$$0 \leq (1-\tau)\bar{\alpha}^k\|P^kS^kg^k\|_2^2 \leq \|r^k\|_\infty - \|r^{k+1}\|_\infty. \tag{5.5}$$

To establish (5.5) note that $\sigma_l^{k+1}d_l^k < 0$. If $\alpha^k < \bar{\alpha}^k$, then $l = j$, $\alpha^k \geq (1-\tau)\bar{\alpha}^k$ and (5.5) holds; on the other hand, if $\alpha^k > \bar{\alpha}^k$, it is clear that again (5.5) follows.

From (5.5) and (5.4),

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

The proof is completed. ∎

The proofs of Lemmas 3 and 4 are essentially identical to those of Lemmas 8 and 9 in [4]. So they are omitted here.

LEMMA 3. *Assume D_θ (D_I) is defined by Algorithm 2 (Algorithm 1). Under primal and dual nondegeneracy assumptions, J is nonsingular at any point $(r, \lambda = Z^T w)$, where*

$$J = \begin{bmatrix} D_\theta T^{-1} & -D_r T^T Z^T \\ Z & 0 \end{bmatrix}.$$

Moreover, $C = [D_\theta T^{-1} A^T, \ -D_r T^T Z^T]$ is also nonsingular everywhere.

Using the fact that $\{D_\theta\}$ is bounded above for any λ, $0 \le \theta \le 1$, and C is nonsingular everywhere, it is easy to prove the following result (see Lemma 9 in [4] for details).

LEMMA 4. *Assume that an l_∞ problem is both primal and dual nondegenerate and $\{\lambda^k = Z^T w^k\}$ is obtained by Algorithm 1 or 2. Then there exists $M > 0$ such that $\|\lambda^k\| \le M$.*

We can now state the first major result.

LEMMA 5. *Assume $\{d^k\}$ is defined by Algorithm 1 or 2; assume primal and dual nondegeneracy. Then*

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

Proof. Using Lemma 2, we know that

$$\lim_{k \to \infty} \bar{\alpha}^k \|P^k S^k g^k\|_2^2 = 0. \tag{5.6}$$

(Recall: $\bar{\alpha}^k$ is the step to the first breakpoint from r^k in the direction d^k.) From Lemma 4, there exists $M_1 > 0$, such that

$$\|(g^k - Z^T w^{k+1})\| \le M_1. \tag{5.7}$$

Now assume there exists a subsequence, which we still denote with k for simplicity, satisfying,

$$\{-\|P^k S^k g^k\|_2^2\} \to c_1 < 0. \tag{5.8}$$

From $|d_j^k| = \|P^k S^{k^T} g^k\|_2^2 = \|S^k(g^k - Z^T w^k)\|_2^2$, we have $\|d^k\|_2 \not\to 0$. Hence from (4.6) and Lemma 3, we know that

$$D_r^k T^T (g^k - \lambda^k) \not\to 0.$$

Thus there exists $c_2 > 0$ such that $\theta^k > c_2$.

We now prove that the sequence of first breakpoints, $\{\bar{\alpha}^k\}$ is bounded away from zero, satisfying $\bar{\alpha}^k > c_4 > 0$, and this will lead to the obvious contradiction.

However, if we let l denote the index of the first breakpoint in direction d^k, then $\bar{\alpha}^k = -\frac{|r_j^k|-|r_l{}^k|}{\sigma_j^k d_j^k - \sigma_l^k d_l^k}$,

$$
\begin{aligned}
\bar{\alpha}^k &\geq \frac{\frac{\theta^k}{2} + (1-\theta^k)|\lambda_l^k|}{(|Z_l^T w^k|)}, \\
&\geq \frac{c_2}{M_1}, \\
&\stackrel{def}{=} c_4 > 0.
\end{aligned}
$$

Using the facts that $|r_j^k|$ and D_θ^k are bounded away from zero and $|\lambda^k|$ is bounded above, we can similarly prove that when $\bar{\alpha}^k = -\frac{|r_j^k|+|r_l{}^k|}{\sigma_j^k d_j^k + \sigma_l^k d_l^k}$, $\bar{\alpha}^k$ is also bounded away from zero which is a contradiction.

Therefore, by (5.6), $\|P^k S^k g^k\| \to 0$. ■

From Lemma 5, it follows that the point of convergence, $(\bar{r}, \bar{\lambda} = Z^T\bar{w})$, satisfies (3.6); this along with the nondegeneracy assumption, implies that if the primal variables converge then so do the duals. We state this formally in Lemma 6 (the proof is essentially identical to that of Lemma 11 in [4] and therefore we omit it here).

LEMMA 6. *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 $\{w^k\} \to \bar{w}$. Moreover,*

$$
(5.9) \quad |\mathcal{A}(\bar{r})| = n+1, \quad A\bar{\lambda} = 0, \quad \sum_{i\in\mathcal{A}} \bar{\sigma}_i\bar{\lambda}_i = 1, \quad \bar{\lambda}_i = 0 \ \forall\, i \in \bar{\mathcal{A}}^c, \quad \bar{\lambda}_i \neq 0, \ \forall\, i \in \bar{\mathcal{A}}
$$

where $\bar{\lambda} = Z^T\bar{w}$.

The next result is crucial: it is established that it is not possible to have convergence to a nonoptimal point satisfying $\bar{\sigma}_j\bar{\lambda}_j < 0$, if the maximum residual, j, has stabilized. From this it is easy to establish (Theorem 8) that a point of convergence must be optimal. Lemma 7 applies to both Algorithm 1 and 2; however, for simplicity we give the proof only for Algorithm 2. The proof is trivially adapted to Algorithm 1: replace D_θ with D_I; define $diag(\delta_\theta^k) = D_I$.

LEMMA 7. *Assume the conditions of Lemma 6 are satisfied. Furthermore, assume that for all k sufficiently large the maximum residual is fixed, say function j, and $\bar{\sigma}_i\bar{\lambda}_i > 0$, $i \in \bar{\mathcal{A}} - \{j\}$. Then $\bar{\sigma}_j\bar{\lambda}_j > 0$.*

Proof. Lemma 6 immediately implies that (5.9) holds with $\bar{\lambda}_i \neq 0, \forall i \in \bar{\mathcal{A}}$. Assume then that (5.9) holds, $\bar{\sigma}_i\bar{\lambda}_i > 0$, $i \in \bar{\mathcal{A}} - \{j\}$, but $\bar{\sigma}_j\bar{\lambda}_j < 0$. Hence $\bar{\theta} > 0$. In the remainder of this proof we establish a contradiction; the proof breaks into three parts.

Part 1. We show that, for any k suficiently large, there exists some $i \in \bar{\mathcal{A}} - \{j\}$, $\sigma_i^k d_i^k < 0$.

From (5.9) and $d = A^T d_x$, we have

$$(5.10) \qquad (\bar{\sigma}_j\bar{\lambda}_j)(\bar{\sigma}_j d_j^k) = -\sum_{i\in\bar{\mathcal{A}}-\{j\}} \bar{\sigma}_i\bar{\lambda}_i(\bar{\sigma}_i d_i^k),\ \bar{\sigma}_j\bar{\lambda}_j = 1 - \sum_{i\in\bar{\mathcal{A}}-\{j\}} \bar{\sigma}_i\bar{\lambda}_i < 0.$$

If $\bar{\sigma}_i d_i^k > 0$, for all $i \in \bar{\mathcal{A}} - \{j\}$, since $\bar{\sigma}_i\bar{\lambda}_i > 0$ for any $i \in \mathcal{A} - \{j\}$, we have $(\bar{\sigma}_j\bar{\lambda}_j)(\bar{\sigma}_j d_j^k) < 0$. But $\bar{\sigma}_j\bar{\lambda}_j < 0$, hence $\bar{\sigma}_j d_j^k > 0$ which contradicts the fact that d is a descent direction. Thus, there exists some $i \in \bar{\mathcal{A}} - \{j\}$ such that $\bar{\sigma}_i d_i^k < 0$. Hence, for k sufficiently large, $\sigma_i^k d_i^k < 0$.

Part 2. We show that the first breakpoint in the search direction d^k is distinct from the other breakpoints for large enough k.

By definition, if stepsize α_i^k corresponds to an intersection with the maximum function j, then

$$\alpha_i^k = \frac{\delta_\theta^k}{|\lambda_i^{k+1}|}, \quad i \in \bar{\mathcal{A}},\ i \neq j,$$

where $diag(\delta_\theta^k) \stackrel{def}{=} D_\theta^k$; So

$$\delta_{\theta_i}^k = \begin{cases} \theta^k + (1-\theta^k)|\lambda_i^k| & \text{if } mod = false \\ \frac{\theta^k}{2} + (1-\theta^k)|\lambda_i^k| & \text{if } mod = true\ . \end{cases}$$

Hence it is clear that $\bar{\alpha}_i = \infty, i \in \bar{\mathcal{A}}^c$.

Assume that ℓ (a function of $k, \ell \in \bar{\mathcal{A}}$) denotes the index of the first break point. Next we establish that the first breakpoint, α_ℓ, will be distinct from the rest.

First, assume for all k sufficiently large, D_θ^k is not modified, i.e., $mod = false$. Then, either constraint ℓ is crossed or constraint ℓ is not crossed and $|\frac{|\lambda_\ell^k|}{|\lambda_\imath^k|} - 1| > 1 - \max(\tau, 1 - \theta^k) > 0$ where $\imath \in \bar{\mathcal{A}}$ denotes the second breakpoint. But in the first case the maximum function must change which contradicts our assumption. Assume then the second case.

But this means that ℓ and some $\imath \in \bar{\mathcal{A}}$ satisfy $\bar{\alpha}_\ell < \bar{\alpha}_\imath$. Therefore the first breakpoint is distinct from the rest.

Second, assume δ_θ^k is modified an infinite number of times. Again we claim that $\bar{\alpha}_\ell < \bar{\alpha}_\imath$. Suppose l is such that

$$|\bar{\lambda}_l| = \max_{i \in \bar{\mathcal{A}} - \{j\}} (|\bar{\lambda}_i|).$$

Let $\bar{\tau}$, $0 < \bar{\tau} < 1$, denote the limit point of $\max\{\tau, 1 - \theta^k\}$, and define

$$\mathcal{E} = \{i \ : \ |\frac{|\bar{\lambda}_l|}{|\bar{\lambda}_i|} - 1| \leq 1 - \bar{\tau}\}.$$

Since δ_θ^k is modified infinite number of times, the first break point corresponds to an index $l \in \mathcal{E}$ infinitely often. From

$$\alpha_\ell^k = \frac{\frac{\theta^k}{2} + (1 - \theta^k)|\lambda_\ell^k|}{|\lambda_\ell^{k+1}|} \rightarrow \frac{\bar{\theta}}{|2\bar{\lambda}_\ell|} + 1 - \bar{\theta} < \frac{\bar{\theta}}{|\bar{\lambda}_i|} + 1 - \bar{\theta} = \bar{\alpha}_i, \quad \forall i \neq l,$$

it is clear that, for all k sufficiently large, the first breakpoint in the direction d^k, ℓ, remains fixed and $\ell \in \mathcal{E}$. Therefore, $\bar{\alpha}_\ell < \bar{\alpha}_\imath$ and the first breakpoint separates from the rest.

Part 3. We now establish the required result: $\sigma_\ell^k d_\ell^k < 0$ and therefore the maximum residual is not fixed, a contradiction.

From (2.7), it is easy to see that $d_{s_i} = -\frac{s_i}{\alpha_i}$. Hence

$$s_\ell^{k+1} = s_\ell^k - \tau \alpha_l^k \frac{s_\ell^k}{\alpha_\ell^k} = s_\ell^k(1 - \tau h_\ell^k).$$

Similarly

$$s_\imath^{k+1} = s_\imath^k(1 - \tau h_\imath^k).$$

Define

$$h_i^k = \frac{\alpha_\ell^k}{\alpha_i^k}, \qquad \forall i \in \bar{\mathcal{A}}.$$

Thus, for $\bar{\alpha}_\ell < \bar{\alpha}_\imath$ and large enough k, we have

$$\frac{1 - \tau h_\ell^k}{1 - \tau h_\imath^k} < 1 - \rho, \quad 0 < \rho < 1.$$

Then

$$\frac{s_\ell^k}{s_i^k} = \frac{s_\ell^{k-1}}{s_i^{k-1}}\frac{1-\tau h_\ell^{k-1}}{1-\tau h_i^{k-1}} < \frac{s_\ell^{k-1}}{s_i^{k-1}}(1-\rho) < \cdots < \frac{s_\ell^{k_2}}{s_i^{k_2}}(1-\rho)^{k-k_2}.$$

Since $(1-\rho)^{k-k_2} \to 0$, we have

$$\lim_{k\to\infty} \frac{s_\ell^k}{s_i^k} = 0.$$

Thus, from $d_{s_i} = -\frac{s_i}{\alpha_i}$, we have

$$\lim_{k\to\infty} \frac{d_{s_\ell}^k}{d_{s_i}^k} = 0.$$

From

$$\bar{\sigma}_j d_j^k = \sum_{i\in\bar{\mathcal{A}}-\{j\}} \bar{\sigma}_i \bar{\lambda}_i d_{s_i}^k,$$

we have

$$\bar{\sigma}_\ell d_\ell^k = \sum_{i\in\bar{\mathcal{A}}-\{j\}} \bar{\sigma}_i \bar{\lambda}_i d_{s_i}^k - d_{s_\ell}^k < 0, \qquad \text{for sufficiently large } k.$$

For k sufficiently large, $d_{s_i} < 0, \forall i \in \bar{\mathcal{A}}$. Hence, $\bar{\sigma}_\ell d_\ell^k < 0$ which implies, by convergence, that $\sigma_\ell^k d_\ell^k < 0$. Therefore, function j does not remain the maximum function, a contradiction. ∎

Theorem 8 below is the next major result: it says that if the primal variables converge, they converge to the optimal point.

THEOREM 8. *Assume $\{r^k\}$ is obtained from Algorithm 1 or 2. Assume primal and dual nondegeneracy. If the sequence $\{r^k\}$ converges to a point r^*, then r^* is optimal.*

Proof. Under the nondegeneracy assumption, we know that $r_i^* \neq 0, i \in \mathcal{A} = \mathcal{A}(r^*)$. From Lemma 6, $|\mathcal{A}(r^*)| = n+1$, $\{w^k\} \to w^*$ and, if λ^* is defined $\lambda^* = Z^T w^*$, then $A\lambda^* = 0$ with $\lambda_i^* = 0$, $i \in \mathcal{A}^c$. Therefore, to establish optimality we need only show that $\sigma_i^* \lambda_i^* > 0$, $\forall i \in \mathcal{A}$.

Assume the contrary, i.e., for some $i \in \mathcal{A}$, $\sigma_i^* \lambda_i^* < 0$. Then $\theta^* > 0$ and there exists k_1 such that for $k > k_1$, $r_i^k \lambda_i^k < 0$, for all i such that $r_i^* \lambda_i^* < 0$, and $r_i^k r_i^* > 0$, for all $i \in \mathcal{A}$. Thus, for $k > k_1$, $\sigma_i^{k+1} = \sigma_i^k, i \in \mathcal{A}$.

Therefore, from $d_s = -D^2 T^T (g - Z^T w^+)$, we have

$$|r_j^{k+1}| - |r_i^{k+1}| = |r_j^k| - |r_i^k| - \alpha^k \delta_i^{k^2} (\sigma_i^k \lambda_i^{k+1}), \quad \text{for } i \in \mathcal{A},$$

where $D^k = \mathrm{diag}(\delta^k)$.

First, if there exists $k_2 > k_1$ such that $i \neq j$, i.e. $|r_i^{k_2}| < \|r^{k_2}\|_\infty$; since $\sigma_i^k \lambda_i^{k+1} < 0$ for $k > k_1$, it follows that $\|r^{k+1}\|_\infty - |r_i^{k+1}| > \|r^k\|_\infty - |r_i^k| > 0$, for all $k > k_2$. Hence, we see immediately that $\{|r_i^k|\} \not\to \|r^*\|_\infty$ which contradicts that $i \in \mathcal{A}$. Therefore, we know that the only case $\sigma_i^* \lambda_i^* < 0$ is possible is when $i = j$ is the index of the maximum residual for all $k > k_1$. But by Lemma 7 this is impossible. ■

Proof that our sequence converges, as stated in the following theorem, is identical to the proof of Theorem 15 in [4].

THEOREM 9. *If an l_∞ problem is both primal nondegenerate and dual nondegenerate, then the sequence $\{(r^k, \lambda^k)\}$, generated by either Algorithm 1 or 2, is convergent.*

It is now clear that under nondegeneracy assumptions, Algorithms 1 and 2 generate points that converge to the optimum point. This follows from Lemma 6, Theorem 8 and 9.

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 main difficulty is that our Newton steps come from different nonlinear systems depending on the index of the maximum residual. Similar to our approach in [4], the 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:

$$\left\{ \begin{array}{l} \hat{F}_j(y) \stackrel{def}{=} D_r T^T(g(j) - Z^T w) = 0, \\ Zr = 0, \end{array} \right. \tag{6.1}$$

where $y = [r^T, w^T]^T$, $j \in \mathcal{A}^*$ and $g(j) = \sigma_j^* e_j$. Note that T corresponds the transformation defined by j, the maximum residual: i.e., T depends on j as well. It is trivial to see that (r^*, w^*) is a solution to each system; each system is continuously differentiable in a neighbourhood of $y^* = (r^*, w^*)$.

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

$$J_{j^k}^k d_N^k = -F_{j^k}(y^k)$$

where

$$J_{j^k}^k = \begin{bmatrix} D_\lambda^k T^{-1} & -D_r^k T^T Z^T \\ Z & 0 \end{bmatrix}, \quad \lambda^k = Z^T w^k,$$

and

$$(6.2) \qquad F_{j^k}(y^k) = \begin{bmatrix} \hat{F}_j(y^k) \\ 0 \end{bmatrix}.$$

Note that the hybrid step d^k satisfies

$$(6.3) \qquad B_{j^k}^k d^k = -F_{j^k}(y^k),$$

where

$$(6.4) \qquad B_{j^k}^k = \begin{bmatrix} D_\theta^k T^{-1} & -D_r^k T^T Z^T \\ Z & 0 \end{bmatrix},$$

and j^k corresponds to the index of the maximum residual at the iteration k . It is clear that $F_{j^k} \in \mathcal{F}$ (i.e., d^k is a Newton–like step for some $F_{j^k} \in \mathcal{F}$). Of course $F_{j^k} \neq F_{j^{k+1}}$, in general, and therefore quadratic convergence is not automatic; however, a slight modification of Theorem 3.4 in [6] yields a viable approach.

THEOREM 10. *Let $\mathcal{F} = \{F_j : \Re^m \rightarrow \Re^m\}$ be a finite set of functions satisfying the assumptions that each F_j is continuously differentiable in an open convex set D and there is a y^* in D such that $F_j(y^*) = 0$ and each $\nabla F_j(y^*)$ 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*

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

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

$$(6.5) \qquad \|B^k - \nabla F_{j^k}(y^*)\| = O(\|y^k - y^*\|).$$

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

We now show that Algorithm 2 can be described in a manner consistent with Theorem 10 and therefore quadratic convergence is achieved. Specifically, (6.5) must be established: the next four results establish several preliminary bounds.

The next two lemma establishes that the dual multipliers and θ^k are bounded by $\|y^k - y^*\|$. The proofs are omitted because they are copies of the proofs for Lemmas 17 and 18 in [4].

LEMMA 11. *Assume that $\{(r^k, w^k)\}$ is any subsequence, convergent to (r^*, w^*), obtained by Algorithm 2. Then,*

$$(6.6) \qquad \|w^{k+1} - w^k\| = \|d_w^k\| = O(\|y^k - y^*\|);$$

Consequently, $\|\lambda^{k+1} - \lambda^k\| = O(\|y^k - y^*\|)$.

LEMMA 12. *Suppose* $\{\theta^k\}$ *is defined as in (4.3). Then,*

$$\theta^k \le L_1\|y^k - y^*\|. \tag{6.7}$$

LEMMA 13. *Assume that an* l_∞ *problem is primal and dual nondegenerate and that the sequence* $\{(r_k, w_k)\}$ *is generated by Algorithm 2. Then,*

$$\alpha^k - 1 = O(\|y^k - y^*\|). \tag{6.8}$$

Proof. Since $\theta^k \to 0$, from Line Search Procedure 2, we know the stepsize $\alpha^k = (1-\theta^k)\alpha_l^k$ or $\alpha^k = \alpha_i^k + (1-\theta^k)(\alpha_l^k - \alpha_i^k), \alpha_l^k > \alpha_i^k$, where $l, i \in \mathcal{A}^*$. From Lemma 11 and $\lambda_i^* \ne 0$, it is clear that

$$\alpha_i^k - 1 = \frac{\theta^k + (1-\theta^k)|\lambda_i^k|}{|\lambda_i^{k+1}|} - 1 = O(\|y^k - y^*\|); \tag{6.9}$$

similarily,

$$\alpha_l^k - 1 = O(\|y^k - y^*\|). \tag{6.10}$$

From Lemma 12

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

therefore, using (6.9, 6.10),

$$\alpha^k - 1 = O(\|y^k - y^*\|).$$

The proof is completed. ■

Denote $B^k = B_{j^k}^k \Omega^k$ where

$$\Omega^k = \mathrm{diag}(\begin{bmatrix} \frac{1}{\alpha^k} e_m \\ e_{m-n} \end{bmatrix})$$

and $e_m(e_{m-n})$ denotes a m-vector ($(m-n)$-vector) with each entry equal to unity, B_{j^k} is defined as in (6.4). But $y^{k+1} = y^k + {\Omega^k}^{-1} d^k$ where d^k is defined by (6.3); therefore, from Lemma 13, we have

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

The hybrid step defined by Algorithm 2 satisfies

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

LEMMA 14. *Assume $\{y^k\}$ is obtained through Algorithm 2; assume primal and dual nondegeneracy. Then*

$$\|B^k - \nabla F_{j^k}(y^*)\|_2 \leq L\|y^k - y^*\|_2$$

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

Proof. From continuity and Lemma 13, it is clear that

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

From

$$B^k_{j^k} - J^k_{j^k} = \begin{bmatrix} (D^k_\theta - D^k_\lambda)T^{-1} & 0 \\ 0 & 0 \end{bmatrix},$$

we have

$$\begin{aligned} \|B^k_{j^k} - J^k_{j^k}\| &= O(\|D^k_\lambda T^{-1} - D^k_\theta T^{-1}\|) \\ &= O(\|((1-\theta^k)|D^k_\lambda T^{-1}| + \theta^k I - D^k_\lambda T^{-1}\|) \\ &= O(\||D^k_\lambda T^{-1}| - D^k_\lambda T^{-1}\|) + O(\theta^k) \\ &= O(\|y^k - y^*\|_2) + O(\theta^k) \\ &= O(\|y^k - y^*\|) \qquad \text{(from Lemma 12).} \end{aligned}$$

Hence,

$$\|B^k_{j_k} - J^k_{j_k}\|_2 = O(\|y^k - y^*\|_2),$$

and therefore,

$$\|B^k - J^*_{j_k}\|_2 = O(\|y^k - y^*\|_2).$$

The proof is completed. ∎

The assumptions of Theorem 10 are now established; quadratic convergence of Algorithm 2 follows immediately.

THEOREM 15. *Suppose the l_∞ problem is primal and dual nondegenerate. Assume the sequence $\{(r^k, w^k)\}$ is obtained from the Algorithm 2. Then $\{(r^k, w^k)\}$ converges quadratically to (r^*, w^*).*

7. Numerical Testing. In this section we provide preliminary numerical results concerning Algorithm 2, the hybrid method (*New*). Our experiments are not exhaustive; our purpose here is to determine the viability of our approach. Is quadratic convergence observed? Is high accuracy achieved? Does the method hold promise for problems of increasing dimensions?

We have implemented the method in PRO-MATLAB [7] using SUN 3/50 and 3/160 workstations. The linear least squares subproblems are solved using orthogonal QR-factorizations with row interchanges for greater stability [10]. No account was made of sparsity in our experiments. starting point for *New* is computed as follows:

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

$$\lambda^0 \leftarrow \frac{\tau}{\|r^0\|_\infty} * r.$$

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

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

We have generated two classes of test problems. First we generated several random problems of varying dimensions. Second, since l_∞ minimization is often used in a function approximation context [8], we have tried several such problems.

On each test problem we compare the number of iterations to that achieved by the popular Barrodale-Phillips algorithm [1]. We do this to get a general feeling for the relative standing of the two algorithms and to examine the relative sensitivities of the two algorithms to problem size and problem class in terms of number of iterations required. We do not compare running times: we do not yet have a sparse (or parallel) implementation of our method (this is the setting in which we expect our method will prosper).

The entries in the tables below represent the total number of required major iterations.

[1] For simplicity we choose λ^0 without requiring $\lambda^0 = Z^T w^0$ for some w^0. However, the computation of λ^k for $k = 1, 2, \ldots$ ensures $\lambda^k = Z^T w^k$ for $k = 1, 2, \ldots$.

Problem 1. Random l_∞ problems: We generated the elements of matrix A and right-hand-side β in a uniform random manner.

$m = 50$

Number of Steps		
n	New	BP
10	10	24
20	9	40
30	10	47
40	9	64

$m = 100$

Number of Steps		
n	New	BP
10	8	25
20	11	69
30	12	68
40	13	116
50	10	113
70	12	153
90	12	124

$m = 200$

Number of Steps		
n	New	BP
10	7	25
20	12	68
30	12	113
50	13	193
70	13	249
100	13	340
140	22	382
160	15	388
190	14	302

New exhibits little variation with m and n: for fixed m there is a mild increase in number of required iterations as n increases (until n gets close to m). On the other hand BP requires significantly more iterations as problem size increases.

Problem 2. Determine $x = (\alpha_1, ..., \alpha_n)$ so that

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

is a best l_∞ fit to $f(z)$ on the points $z = 0, \frac{1}{m}, ..., 1$.

$n = 5, f1(z) = \exp(z)$.

Number of Steps		
m	BP	New
100	10	7
200	11	8
400	12	8
600	12	8
800	12	8
1000	12	8
1200	12	9
1500	12	10
1800	12	9
2000	13	9

$n = 8, f(z) = \exp(z)$.

Number of Steps		
m	BP	New
100	15	7
200	18	9
400	18	10
600	21	10
800	19	9
1000	18	10
1200	22	10
1500	21	10
1800	21	11
2000	20	10

The relative performance of BP is much improved for approximation problems compared to random problems. As observed in [2], this is largely due to the clever starting point procedure available to the BP algorithm for approximation problems.

8. Conclusions. In this paper, we have presented a new iterative method for solving l_∞ problems. The algorithm is appealing because, similar to affine scaling approach for linear programming, the number of iterations required to solve a problem is relatively insensitive to the problem size. Moreover, since the algorithm is quadratically convergent, a solution can be obtained to high accuracy quickly (thus comparable to a solution obtained by a simplex type algorithm).

The algorithm is easy to implement: At each iteration, the major computation is a weighted least squares solve. Finally, we remark that any technique available to speed up least squares solving - e.g., exploitation of structure, sparsity, parallelism - will benefit this l_∞ algorithm directly.

REFERENCES

[1] I. Barrodale and C. Phillips, *An improved algorithm for discrete Chebychev linear approximation*, in Proc. 4th Manitoba Conf. on Numer. Math., U. of Manitoba, Winnipeg, Canada, 1974, pp. 177–190.

[2] R. H. Bartels, A. R. Conn, and Y. Li, *Primal methods are better than dual methods for solving overdetermined linear systems in the l_∞ sense?*, SIAM J. Numer. Anal., 26 (1989), pp. 693–726.

[3] T. F. Coleman and L. Hulbert, *A superlinear algorithm to solve the simply bound quadratic programming problem*, Tech. Rep. In preparation, Computer Science Department, Cornell University, 1990.

[4] T. F. Coleman and Y. Li, *A global and quadratic affine scaling method for linear l_1 problems*, Tech. Rep. 89–1026, Computer Science Department, Cornell University, 1989.

[5] ——, *A quadratic algorithm for the linear programming problem with lower and upper bounds*, in Proceedings on Large-scale Optimization Workshop, SIAM, 1989. Mathematical Sciences Institute, Cornell University.

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

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

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

[9] 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.

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