ALGORITHMS FOR SOLVING OVERDETERMINED SYSTEMS OF LINEAR EQUATIONS IN THE ℓ_p SENSE

Thesis for the Degree of Ph. D.
MICHIGAN STATE UNIVERSITY
ROBERT WILLIAM OWENS
1975

ABSTRACT

ALGORITHMS FOR SOLVING OVERDETERMINED SYSTEMS OF LINEAR EQUATIONS IN THE ℓ_p SENSE

By

Robert William Owens

In this thesis, we investigate methods for approximating a solution of an overdetermined system of linear equations. A best approximate solution of the linear system $Ax = b$ is taken to be a vector x which minimizes the length of the error vector $\eta(x) = b - Ax$. We consider the two classes of approximation problems obtained when we determine the length of $\eta(x)$ first by a smooth strictly convex norm, and second by an ℓ_p metric for $0 < p < 1$.

For each of the two approximation problems, we study a dual problem whose solution leads directly to a solution of the original problem. Algorithms for solving the dual problems are presented, and numerical results from several ℓ_p approximation problems, $0 < p < \infty$, are discussed.

ALGORITHMS FOR SOLVING OVERDETERMINED SYSTEMS OF LINEAR EQUATIONS IN THE ℓ_p SENSE

By

Robert William Owens

A THESIS

Submitted to

Michigan State University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

Department of Mathematics

1975

ACKNOWLEDGMENTS

I would like to express my sincere thanks to Professor V.P. Sreedharan, without whose direction and encouragement I could not have written this thesis. I also thank the other members of my committee, Professors Harvey S. Davis, Norman L. Hills, Edward C. Ingraham, and Mary J. Winter, whose helpful comments and suggestions aided in the final preparation of the thesis. Finally, special thanks are reserved for Jill Shoemaker for typing the final manuscript.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The problem that we shall study can be stated somewhat generally as follows:

> Given a linear subspace K of $\mathbb{R}^m$ and a point b not in K, find $x \in K$ which is closest to b.

For any metric d on $\mathbb{R}^m$, the above approximation problem makes sense mathematically, while in practical problems only a few of these are of interest. Moreover, even in the presence of certain well behaved naturally occurring norms on $\mathbb{R}^m$ as the metric, there are still many theoretical and computational difficulties.

The ℓ_p norms, $1 \leq p \leq \infty$, given by

$$\|x\|_p = \left(\sum_{j=1}^{m} |x_j|^p\right)^{1/p} \qquad 1 \leq p < \infty \quad \text{and} \tag{1.1}$$

$$\|x\|_\infty = \max_{1 \leq j \leq m} |x_j| \tag{1.2}$$

have received the most consideration with the ℓ_1, ℓ_2, and ℓ_∞ cases often being singled out for special attention because of their many applications.

Since this thesis is concerned with ℓ_p approximation problems in the form just mentioned along with extensions to

the case where $0 < p < 1$, we briefly review what is known about the subject.

First, however, we rephrase the original problem. Let A be an $m \times n$ real matrix with $m > n$ and $b \in \mathbb{R}^m \setminus \text{Image}(A)$. We are interested in finding $x \in \mathbb{R}^n$ minimizing the number $\|b - Ax\|_p$. We shall refer to this as the ℓ_p problem. In the case of $p = \infty$, we often refer to it as the Chebyshev problem.

The case where $p = 2$ is the easiest to solve since this is just the problem of finding the usual Euclidean distance from a point to a subspace. If the matrix A has rank n, then $x = (A^TA)^{-1}A^Tb$ is the solution of the ℓ_2 problem, and the only difficulties that arise are computational ones stemming from the fact that A^TA may be ill conditioned.

A number of essentially different methods have been developed to handle the Chebyshev problem and to a lesser degree the ℓ_1 problem. Since each of these norms is neither strictly convex nor smooth, i.e. both have "flat spots" and "sharp corners", most of the powerful theorems from approximation theory are not applicable and special techniques must be devised.

Since the unit ℓ_1 and ℓ_∞ balls are convex polyhedra, the methods of linear programming are applicable for solving the ℓ_1 and Chebyshev problems. Stiefel [22] presented such an algorithm to solve the Chebyshev problem, and more recently Barrodale and Young [4], Barrodale and

Roberts [5], and Abdelmalek [1] have given improved algorithms for handling both problems by linear programming.

It has been conjectured that the solution of the ℓ_∞ problem is the limit of the solutions of the ℓ_p problem as $p \to \infty$, and similarly for the ℓ_1 problem letting $p \to 1^+$. It is well known that if these limits exist, then they are exactly the solution that one would expect, but the question of convergence is the crucial point. Descloux [8] established the convergence of the solutions of the ℓ_p problem as $p \to \infty$ providing a justification for this method of solving the Chebyshev problem. The question concerning the convergence of the solutions of the ℓ_p problems as $p \to 1^+$, however, is still unanswered. Abdelmalek [1] incorrectly applied a theorem of Hoel [13] to conclude that the solutions of the ℓ_p problems converge as $p \to 1^+$. Moreover, even if such a theorem were established, there would remain the problem of computing the solutions of the ℓ_p problems when p is near 1. This, as shall be mentioned, is another significant difficulty. Methods of solving the Chebyshev problem by trying to find the limit of the solutions of the ℓ_p problems as $p \to \infty$ have been presented by Fletcher, Grant, and Hebden [12]. Attempts to solve the ℓ_1 problem by locating a limit of the solutions of the ℓ_p problems as $p \to 1^+$ have been investigated by Abdelmalek [1].

Another method used to solve the Chebyshev problem was developed by Lawson [14]. Using a result of Motzkin and Walsh [16], Lawson was able to compute the solution of the

Chebyshev problem as the limit of the solutions of a sequence of weighted ℓ_p problems where p may be held fixed and the weights change in each iteration. Since weighted ℓ_2 problems are almost as easy to solve as ℓ_2 problems, Lawson chose $p = 2$ throughout his algorithm. An analysis of the rate of convergence of the Lawson algorithm was made by Cline [7].

Another theorem of Motzkin and Walsh [16] similar to the one guaranteeing that Lawson's algorithm converges to the solution of the Chebyshev problem also says that the solution of the ℓ_1 problem is identically the same as the solution of a weighted ℓ_p problem for some choice of weight functions depending upon p. So far, however, no one has discovered an algorithm for computing those weights either directly or iteratively even in the case when $p = 2$.

The best known method for solving the Chebyshev problem is the exchange algorithm as can be found in Cheney [6]. It exploits the fact that the $m \times n$ Chebyshev problem given by the overdetermined linear system $Ax = b$ has exactly the same solution as some particular $(n+1) \times n$ Chebyshev problem determined by $\bar{A}x = \bar{b}$, where $\bar{A}$ is composed of $n + 1$ rows of A and $\bar{b}$ is the $n + 1$ vector of the corresponding entries from b. In fact, of all such $(n+1) \times n$ problems, the one yielding the solution of the original problem is that one for which the error $\min_{x \in \mathbb{R}^n} \|\bar{A}x - b\|$ is the largest.

Although most of their work deals with the ℓ_p problem for $1 < p < \infty$, Duris and Sreedharan [10] have also presented algorithms for solving the Chebyshev and ℓ_1 problems that use only the solutions of least squares problems.

Much more can usually be said about the solutions of the ℓ_p problem when $1 < p < \infty$ since in that case the norm is both strictly convex and smooth, i.e. the boundary of the unit ball contains no line segment and at each point on the boundary of the unit ball there is a unique hyperplane supporting the unit ball. Also, the duality of the spaces ℓ_p and ℓ_q where, $\frac{1}{p} + \frac{1}{q} = 1$, can be exploited.

Sreedharan [19,20,21], Duris [9], and Anton and Duris [3] have developed several algorithms for solving the approximation problem for an arbitrary smooth strictly convex norm and in particular for the ℓ_p norm when $1 < p < \infty$. In each case, the solution of a dual problem is obtained iteratively by carrying out an orthogonal projection and solving a single nonlinear equation in one real variable at each iteration. Numerical results indicate that the algorithms converge slowly when p is very near 1 or very large, but work satisfactorily in all other cases.

In Chapter 2 of this thesis, another algorithm for solving the given approximation problem with a smooth strictly convex norm is presented. The algorithm uses ideas developed by Sreedharan in [20,21]. Numerical results on ℓ_p problems are presented in Chapter 4.

Rice and Usow [18] have extended the previously mentioned Lawson algorithm to include ℓ_p problems for $2 \leq p < \infty$ and they have developed a method for accelerating the convergence of the algorithms. However, they observed that for p large the convergence is still quite slow. Moreover, the algorithm is not applicable at all for $1 < p < 2$.

Another algorithm applicable for $2 \leq p < \infty$ but not for $1 < p < 2$ has been presented by Fletcher, Grant, and Hebden [11]. Although primarily designed for L_p approximation problems, i.e. continuous rather than discrete approximation, the modification is immediate. For $p \geq 3$, second order convergence is proven, but numerical results are few and inconclusive regarding this algorithm.

A final method of solving the ℓ_p problem is by minimizing the differentiable function of n real variables

$$f(x) = \|b - Ax\|_p \tag{1.3}$$

It is sufficient to find a zero of the gradient ∇f. Newton's method and the method of steepest descent are applicable to speed convergence to a root of ∇f if p is large enough to guarantee enough derivatives of f. Unfortunately, $p \geq 3$ is required so again the case $1 < p < 2$ is left untreated.

In summary, the situation is roughly as follows:

$p = 1$: Only a very few algorithms are available.

$1 < p < 1 + \epsilon$, ϵ small: Almost no effective method is known for treating these ℓ_p problems.

p near 2, i.e. not too close to 1 nor very large: There are many algorithms available - few difficulties.

p very large: There are a couple of good algorithms available to choose from, but not very many.

$p = \infty$ (Chebyshev problem): There are several efficient algorithms to solve this problem.

With a slight modification in the definition of $\|\cdot\|_p$, the ℓ_p problem for $0 < p < 1$ also makes mathematical sense and we are consequently led to investigate solutions of that problem also. With the exception of Hoel [13] and Motzkin and Walsh [15,16,17], no one has considered this question.

When we choose $0 < p < 1$, $\|\cdot\|_p$ suitably defined turns out to be a p-homogeneous metric but not a norm, and the unit ball is not convex. In Chapter 3 we consider the ℓ_p problem for $0 < p < 1$ establishing theoretical and computational methods for finding a solution. In Chapter 4 some numerical results are presented.

CHAPTER II

APPROXIMATION WITH A SMOOTH STRICTLY CONVEX NORM

In this chapter, we shall study the system of linear equations

$$Ax = b$$

where A is an $m \times n$ matrix, $m > n$, x is a n-vector, and b is an m-vector. We assume all numbers are real. Let $\|\cdot\|$ be a smooth strictly convex norm on $\mathbb{R}^m$. The problem that we shall be concerned with, referred to as problem (P), is

(P): Find $x \in \mathbb{R}^n$ minimizing $\{\|b - Ax\| \mid x \in \mathbb{R}^n\}$.

In [21], a dual problem (P*) was considered, the correspondence between problems (P) and (P*) established, and two algorithms for constructing the solution of problem (P*) were presented.

Here we review problem (P*), the previous results concerning that problem and (P), and then present a new algorithm for solving problem (P).

Before actually beginning this development, we set down some assumptions, definitions, and notation that will remain standard throughout this chapter.

$(\cdot|\cdot)$ denotes the usual inner product on $\mathbb{R}^m$, i.e.

$$(x|y) = \sum_{j=1}^{m} x_j y_j$$

We denote the transpose of a matrix A by A^T. (v) means the linear span of the vector v.

$$K = \text{Image } (A) = \{Ax | x \in \mathbb{R}^n\}$$

$$K^\perp = \text{Ker } A^T = \{x \in \mathbb{R}^m | (k|x) = 0 \ \forall k \in K\}$$

$E: \mathbb{R}^m \to K^\perp$ is the orthogonal projection of $\mathbb{R}^m$ onto $K^\perp$, where orthogonal means with respect to the inner product $(\cdot|\cdot)$ given above.

$$s = Eb$$

$$\rho = \inf\{\|b - k\| \mid k \in K\}$$

We assume that $\rho > 0$ since the problem is trivial otherwise.

Recall that a norm $\|\cdot\|$ is strictly convex if and only if

$$\|x\| = \|y\| = \tfrac{1}{2}\|x + y\| \text{ implies that } x = y,$$

and $\|\cdot\|$ is smooth if and only if through each point of unit norm there passes a unique hyperplane supporting the closed unit ball $B = \{x \in \mathbb{R}^m \mid \|x\| \leq 1\}$.

Given $v \neq 0$, we define the vector $v^* \in \mathbb{R}^m$ by

$$(v^*|v) = \|v\| \text{ and } \max\{(u|v^*) \mid \|u\| \leq 1\} = 1.$$

By Lemma 2.1 of [21], the correspondence $v \mapsto V^*$ is a continuous function from $\mathbb{R}^m \setminus \{0\}$ into itself. Moreover, in the special case when $\|\cdot\|$ is the ℓ_p norm with $1 < p < \infty$, v^* takes on the particularly simple form

$$v_j^* = \frac{|v_j|^{p-1}}{\|v\|_p^{p-1}} \operatorname{sgn} v_j .$$

In [21], Sreedharan introduced the following dual problem (P*)

(P*): Find $z \in K\oplus(s)$, $\|z\| = 1$ maximizing $(s|w)$ over all $w \in K\oplus(s)$, $\|w\| = 1$.

2.1 Lemma. (i) There exists a unique $k \in K$ minimizing

$$\{\|b - y\| \mid y \in K\} . \tag{2.1.1}$$

(ii) Problem (P*) has a unique solution.

Proof: (i) Let $x \in K$ and set $\delta = \|b - x\|$. $S = \{y \in K \mid \|b - y\| \leq \delta\}$ is a compact set, and $f:K \to \mathbb{R}$ by $f(z) = \|b - z\|$ is continuous on S, so f must achieve its minimum value on S. Suppose $x,y \in S$ such that

$$f(x) = \|b - x\| = \min\{\|b - z\| \mid z \in S\} = \|b - y\| = f(y). \tag{2.1.2}$$

$$\begin{aligned} \tfrac{1}{2}\|2b - (x + y)\| &= \tfrac{1}{2}\|(b - x) + (b - y)\| \\ &\leq \tfrac{1}{2}(\|b - x\| + \|b - y\|) \\ &= \|b - x\| = \|b - y\| \end{aligned}$$

By the strict convexity of $\|\cdot\|$, $b - x = b - y$, i.e. $x = y$. Thus there is a unique $k \in K$ minimizing (2.1.1).

(ii) It is clear that problem (P*) has a solution. Suppose y,z both solve problem (P*) and $y \neq z$. Since $\|\cdot\|$ is strictly convex, $\|y\| = \|z\| = 1$, and $y \neq z$,

$$\|y + z\| < 2. \tag{2.1.3}$$

Since $(s|y) = (s|z) > 0$, $y + z \neq 0$. Let $w = \frac{y+z}{\|y+z\|}$. Then $w \in K \oplus (s)$, $\|w\| = 1$, and

$$(s|w) = \frac{(s|y) + (s|z)}{\|y+z\|} = \frac{2(s|y)}{\|y+z\|} > (s|y) \quad \text{by (2.1.3).}$$

But this contradicts the assumption that y solves problem (P*). Thus the solution of problem (P*) must be unique.

The key results of [21] used to solve problems (P) and (P*) are

2.2 Theorem. If $z \in K \oplus (s)$ with $\|z\| = 1$ and $z^* \in K^\perp$, then

$$b - (b|z^*)z \in K \quad \text{and} \quad \rho = (b|z^*).$$

2.3 Corollary. Let z be as above. Then $(b|z^*) = \frac{(s|s)}{(s|z)}$.

2.4 Theorem. Let z solve problem (P*). Then

$$b - \frac{(s|s)}{(s|z)} z \in K \quad \text{and} \quad \rho = \frac{(s|s)}{(s|z)}.$$

2.7 Lemma. Let $z,w \in \mathbb{R}^m$ be linearly independent. Then $\alpha \in \mathbb{R}$ satisfies

$$\|z - \alpha w\| \leq \|z - \lambda w\| \quad \forall \lambda \in \mathbb{R}$$

if and only if

$$((z - \alpha w)^* | w) = 0.$$

Next we give a new algorithm for solving problem (P).

2.2 Algorithm.

Step 1. Set $k = 0$ and $y_k = \frac{s}{\|s\|}$

Step 2. Compute y_k^* and $v_k = AA^T y_k^*$.

Step 3. If $v_k = 0$, go to step 6; otherwise continue.

Step 4. Choose $\alpha_k > 0$ such that $((y_k - \alpha_k v_k)^* | v_k) = 0$.

Step 5. Set $y_{k+1} = \frac{z}{\|z\|}$ where $z = y_k - \alpha_k v_k$.
Increment k by 1 and return to step 2.

Step 6. Using y_k, the solution of problem (P*), solve problem (P).

2.3 Theorem. Either algorithm 2.2 solves problem (P) in a finite number of steps or the y_k converge to the solution of problem (P*).

Before proving theorem 2.3, we make a couple of observations about algorithm 2.2 and establish a few facts to be used in the proof. In step 1, we could have selected y_0 to be any element of $K \oplus (s)$ of norm one with $(y_0 | s) > 0$. Our choice is just a very convenient one. In step 3, one might choose to stop the computations if $\|v_k\|$ or $\|\alpha_k v_k\|$ is

small. The reasons for these stop rules will be apparent in the course of the proof of the convergence of the algorithm. With respect to step 6, Theorem 2.4 of [21] says that $Ax = b - \frac{(s|s)}{(s|y_k)} y_k = b - \rho y_k$ has a solution, say $\bar{x}$. Since $\|b - A\bar{x}\| = \rho$, $\bar{x}$ solves problem (P).

We next prove a couple of propositions essentially saying that algorithm 2.2 makes sense. In particular, $v_k = 0$ implies that y_k solves problem (P*), and the choice of α_k specified in step 4 is always possible.

2.4 Lemma. Let $v_k \neq 0$.

(i) There exists $\alpha > 0$ such that $\|y_k - \alpha v_k\| < 1$.

(ii) If y_k, v_k are linearly independent, then for every $\alpha > 0$ such that $\|y_k - \alpha v_k\| < 1$, we have $0 < \|y_k - \alpha v_k\|$.

(iii) If y_k, v_k are linearly dependent, then there is a unique $\alpha > 0$ such that $\|y_k - \alpha v_k\| = 0$.

(iv) There is a unique $\beta \in \mathbb{R}$ such that $\forall \lambda \in \mathbb{R}$ $\|y_k - \beta b_k\| \leq \|y_k - \lambda v_k\|$, and $\beta > 0$.

(v) $((y_k - \beta v_k)^* | v_k) = 0$.

Proof: By the definition of y_k^*,

$$\max\{(u|y_k^*) \mid \|u\| \leq 1\} = 1.$$

$$\|y_k - \alpha v_k\| = \|y_k - \alpha v_k\| \max\{(u|y_k^*) \mid \|u\| \leq 1\}$$

$$\geq \|y_k - \alpha v_k\| \left(\frac{y_k - \alpha v_k}{\|y_k - \alpha v_k\|} \mid y_k^*\right)$$

$$= (y_k - \alpha v_k | y_k^*)$$

$$= (y_k | y_k^*) - \alpha (y_k^* | v_k)$$

$$= \|y_k\| - \alpha (y_k^* | v_k)$$

$$= 1 - \alpha (y_k^* | v_k). \qquad (2.4.1)$$

Observe that

$$(y_k^* | v_k) = (y_k^* | AA^T y_k^*) = (A^T y_k^* | A^T y_k^*) = \|A^T y_k^*\|_2^2 . \qquad (2.4.2)$$

Together with the definition $v_k = AA^T y_k^*$, we have that

$$v_k = 0 \text{ if and only if } A^T y_k^* = 0. \qquad (2.4.3)$$

By assumption, $v_k \neq 0$ so from (2.4.1)

$$\|y_k - \alpha v_k\| > 1 \text{ for all } \alpha < 0. \qquad (2.4.4)$$

Suppose that y_k, v_k are linearly dependent. Then there is a $\beta \in \mathbb{R}$ such that $\|y_k - \beta v_k\| = 0$. By (2.4.4), $\beta > 0$. More specifically, $\beta = \frac{1}{\|v_k\|}$ proving (iii). Suppose y_k, v_k are linearly independent. If there were no $\alpha > 0$ such that $\|y_k - \alpha v_k\| < 1$, then

$$\|y_k - \alpha v_k\| \geq 1 = \|y_k\| \text{ for all } \alpha \in \mathbb{R}. \qquad (2.4.5)$$

By Lemma 2.7 of [21], (2.4.5) implies that $(y_k^* | v_k) = 0$, which by (2.4.2) and (2.4.3) means that $v_k = 0$, contradicting the hypothesis that $v_k \neq 0$. Consequently, there must be an $\alpha > 0$

such that $\|y_k - \alpha v_k\| < 1$ establishing (i) and (ii). Consider the strictly convex function

$$f:\mathbb{R} \to \mathbb{R} \quad \text{by} \quad f(\lambda) = \|y_k - \lambda v_k\| .$$

$f(0) = 1$, $f(\lambda) \to \infty$ as $\lambda \to \infty$, and by (i) there exists an $\alpha > 0$ such that $f(\alpha) < 1$. By the continuity and strict convexity of f, there is a unique $\bar{\lambda} > 0$ such that $f(\bar{\lambda}) = 1$, and by the strict convexity of f, there is a unique β, $0 < \beta < \bar{\lambda}$, minimizing f. (iv) is proven. Finally, from (iv) and lemma 2.7 of [21], it follows that

$$((y_k - \beta v_k)^* | v_k) = 0$$

for the β found in (iv).

This completes the proof of Lemma 2.4.

We summarize the important points of Lemma 2.4 in

2.5 Proposition. Assume that y_k and v_k are linearly independent. Then there exist a unique $\alpha > 0$ such that

$$((y_k - \alpha v_k)^* | v_k) = 0. \tag{2.5.1}$$

Moreover, with this choice of α,

$$0 < \|y_k - \alpha v_k\| < 1. \tag{2.5.2}$$

As an immediate consequence of proposition 2.5, step 4 of algorithm 2.2 is guaranteed to make sense.

2.6 Proposition. For any $k \geq 0$, $y_k \in K \oplus (s)$, $\|y_k\| = 1$, and $(y_k | s) > 0$.

Proof: The assertion is true for $k = 0$ by construction. Assume that the proposition holds for $k = 0,1,\ldots,n$. We shall now verify its validity for $k = n+1$.

If $v_n = 0$, then $y_{n+1} = \frac{y_n}{\|y_n\|} = y_n$ so the assertion is true for $k = n+1$. Assume that $v_n \neq 0$. Since $v_k = AA^T y_k^*$, $v_k \in K \ \forall k$. $s \in K^\perp$, $y_n \in K \oplus (s)$ with $(y_n|s) > 0$, and $v_n \in K$ implies that y_n and v_n are linearly independent. So by proposition 2.5,

$$0 < \|y_n - \alpha_n v_n\| . \tag{2.6.1}$$

Since $y_n \in K \oplus (s)$ and $v_n \in K$, $y_{n+1} = \frac{y_n - \alpha_n v_n}{\|y_n - \alpha_n v_n\|} \in K \oplus (s)$, and $\|y_{n+1}\| = 1$. Finally,

$$\begin{aligned} (y_{n+1}|s) &= \left(\frac{y_n - \alpha_n v_n}{\|y_n - \alpha_n v_n\|} \,|\, s\right) \\ &= \frac{(y_n|s) - \alpha_n (v_n|s)}{\|y_n - \alpha_n v_n\|} \\ &= \frac{(y_n|s)}{\|y_n - \alpha_n v_n\|} \qquad (2.6.2) \\ &> 0 \quad \text{by the induction hypothesis.} \end{aligned}$$

This completes the proof of proposition 2.6.

We next show that steps 3 and 6 of the algorithm do not lead us astray by proving

2.7 Proposition. If $v_k = 0$, then y_k solves problem (P*).

Proof: By (2.4.3), $v_k = 0$ implies that $y_k^* \in \text{Ker } A^T = K^\perp$. Also, by Lemma 2.6, $y_k \in K \oplus (s)$ and $\|y_k\| = 1$. Applying theorem 2.2 of [21], we have

$$b - (b|y_k^*)y_k \in K \quad \text{and} \quad \rho = (b|y_k^*). \tag{2.7.1}$$

(2.7.1) together with corollary 2.3 of [21] and lemma 2.1(i) yields that

$$b - \frac{(s|s)}{(s|y_k)}\, y_k \tag{2.7.2}$$

is the unique point in K closest to b.

Let z solve problem (P*). Then by Theorem 2.4 of [21],

$$b - \frac{(s|s)}{(s|z)}\, z \tag{2.7.3}$$

is also the unique point in K closest to b.

Equating (2.7.2) and (2.7.3), we have

$$\frac{z}{(s|z)} = \frac{y_k}{(s|y_k)} \,. \tag{2.7.4}$$

Taking the norm of both sides of (2.7.4) and using the facts that $\|z\| = \|y_k\| = 1$ and $(y_k|s)\,,\ (s|z) > 0$, we have

$$(s|z) = (s|y_k)\,. \tag{2.7.5}$$

Lemma 2.1(ii) now says that $y_k = z$, i.e. y_k solves problem (P*).

We now prove Theorem 2.3, i.e. that algorithm 2.2 solves problem (P). Since the proof is somewhat lengthly, we first give an outline of the steps to be taken. We show that

(i) $\forall k \geq 0 \quad 0 < \rho_k < \rho_{k+1}$, where $\rho_k = (y_k|s)$,

(ii) $\lim_{k\to\infty} \|y_k - \alpha_k v_k\| = 1$,

(iii) $\lim_{k\to\infty} v_k = 0$,

(iv) any limit point of $\{y_k \mid k \geq 0\}$ solves problem (P*), and

(v) $\lim_{k\to\infty} y_k = y$ which solves problem (P*).

Because of Proposition 2.7, we can assume that $v_k \neq 0 \ \forall k \geq 0$.

2.8 Proof of Theorem 2.3. Let

$$\rho_k = (y_k \mid s). \tag{2.8.1}$$

(i) Claim: $0 < \rho_k < \rho_{k+1} \quad \forall k \geq 0$.

$$\rho_0 = \left(\frac{s}{\|s\|} \mid s\right) > 0$$

$$\rho_{k+1} = (y_{k+1} \mid s)$$

$$= \frac{(y_k \mid s)}{\|y_k - \alpha_k v_k\|} \qquad \text{by (2.6.2)},$$

$$> (y_k \mid s) \qquad \text{by (2.6.1)}.$$

Thus $0 < \rho_k < \rho_{k+1} \quad \forall k \geq 0$.

(ii) Claim: $\lim_{k\to\infty} \|y_k - \alpha_k v_k\| = 1$.

From (i),

$$0 < \rho_k < \rho_{k+1} \leq \max\{(w \mid s) \mid w \in K \oplus (s), \|w\| = 1\}$$

and from (2.6.2),

$$\frac{\rho_k}{\rho_{k+1}} = \|y_k - \alpha_k v_k\| .$$

$$\text{Thus} \quad \lim_{k\to\infty} \|y_k - \alpha_k v_k\| = \lim_{k\to\infty} \frac{\rho_k}{\rho_{k+1}} = 1.$$

(iii) Claim: $\lim_{k\to\infty} v_k = 0$.

Suppose the claim were false. Then there exists $\delta > 0$ such that, by taking a subsequence if necessary,

$$\|v_k\| \geq \delta \quad \text{for all} \quad k \geq 0 . \tag{2.8.2}$$

$$1 > \|y_k - \alpha_k v_k\| \geq \|\alpha_k v_k\| - \|y_k\|$$

$$= \alpha_k \|v_k\| - 1$$

$$\geq \alpha_k \delta - 1$$

from which one obtains

$$0 < \alpha_k < \frac{2}{\delta} . \tag{2.8.3}$$

Again by taking subsequences if necessary, we can assume that

$$\lim_{k\to\infty} \alpha_k = \alpha \quad \text{and} \quad \lim_{k\to\infty} y_k = y . \tag{2.8.4}$$

By Lemma 2.1 of [21], the mapping $x \mapsto x^*$ is continuous on $\mathbb{R}^m \setminus \{0\}$, so

$$\lim_{k\to\infty} v_k = \lim_{k\to\infty} AA^T y_k^* = AA^T y^* \equiv v .$$

$$\|v\| \geq \delta \qquad \text{by (2.8.2)} .$$

Also, by the continuity of $x \to x^*$ and $\|\cdot\|$, and since

$$((y_k - \alpha_k v_k)^* | v_k) = 0 \qquad \forall k \geq 0,$$

we have that

$$((y - \alpha v)^* | v) = 0, \tag{2.8.5}$$

$$\|y - \alpha v\| = 1 = \|y\|. \tag{2.8.6}$$

Either $\alpha = 0$ or $\alpha > 0$. We show that each of these possibilities leads to a contradiction forcing us to conclude that $\lim_{k \to \infty} v_k = 0$.

If $\alpha = 0$, then (2.8.5) becomes $(y^* | v) = 0$. Together with (2.4.2) and (2.4.3), this means that $v = 0$, contradicting our assumption that $v \neq 0$. Suppose $\alpha > 0$. First note that y, v are linearly independent since $(v|s) = 0$ and $(y|s) = \lim_{k \to \infty} (y_k|s) = \lim_{k \to \infty} \rho_k > 0$. By Lemma 2.7 of [21], (2.8.5) implies that

$$\|y - \alpha v\| \leq \|y - \lambda v\| \quad \text{for all} \quad \lambda \in \mathbb{R}. \tag{2.8.7}$$

But (2.8.6) and the strict convexity of $\|\cdot\|$, forces

$$\|y - \frac{\alpha}{2} v\| < 1$$

which contradicts (2.8.7). Hence we must conclude that $\lim_{k \to \infty} v_k = 0$.

(iv) Claim: any limit point of $\{y_k | k \geq 0\}$ solves problem (P*).

By taking a subsequence and reindexing if necessary, we can assume that $\lim_{k\to\infty} y_k = y$. From (iii) and the continuity of the mapping $x \mapsto x^*$, we have that

$$v = \lim_{k\to\infty} v_k = 0, \tag{2.8.8}$$

which by (2.4.3) implies that $y^* \in \text{Ker } A^T = K^\perp$. Also, by proposition 2.6,

$$y \in K \oplus (s) \quad \text{and} \quad \|y\| = 1. \tag{2.8.9}$$

Applying Theorem 2.2 and Corollary 2.3 of [21] and Lemma 2.1(i),

$$b - \frac{(s|s)}{(s|y)}\, y$$

is the unique point in K closest to b.

Suppose z solves problem (P*). Then by theorem 2.4 of [21],

$$b - \frac{(s|s)}{(s|z)}\, z = b - \frac{(s|s)}{(s|y)}\, y \ , \quad \text{i.e.}$$

$$\frac{z}{(s|z)} = \frac{y}{(s|y)} \ . \tag{2.8.10}$$

Taking the norm of both sides of (2.8.10) and using $\|z\| = \|y\| = 1$, $(s|z) > 0$, $(s|y) = \lim_{k\to\infty} (s|y_k) = \lim_{k\to\infty} \rho_k > 0$, it follows that $(s|z) = (s|y)$. Finally, Lemma 2.1(ii) says that $z = y$, i.e. y solves problem (P*).

(v) Claim: $\lim_{k\to\infty} y_k = y$, the unique solution of problem (P*).

By the uniqueness of the solution of problem (P*) any convergent subsequence of $\{y_k \mid k \geq 0\}$ converges to the unique solution of problem (P*). Due to the compactness of $B = \{x \in K \oplus (s) \mid \|x\| = 1\}$, $\lim_{k \to +\infty} y_k = y$.

This completes the proof of Theorem 2.3 establishing that algorithm 2.2 is guaranteed to find the solution of problem (P).

CHAPTER III

THE ℓ_p PROBLEM, $0 < p < 1$

We shall be considering the system of linear equations

$$Ax = b$$

where A is an $m \times n$ matrix, $m > n$, b is an m-vector, and x is an n-vector. All numbers are assumed to be real. For $y \in \mathbb{R}^m$ and $0 < p < 1$, let

$$\|y\|_p = \sum_{j=1}^{m} |y_j|^p.$$

Given A,b, and p, the problem that we shall be concerned with, referred to as problem (P), is

$$(P): \text{ Find } x \in \mathbb{R}^n \text{ minimizing } \{\|b - Ax\|_p | x \in \mathbb{R}^n\}.$$

In order to solve the given problem, we will introduce a dual problem, to be denoted (P*). A relation between problems (P) and (P*) is proven, and a characterization of the solutions of problem (P*) established. Although in general problem (P*) can not be solved in a computationally feasible manner, it can always be solved in a finite number of steps and efficient algorithms for solving particular cases are given. Exchange type algorithms for finding local solutions

of problem (P*) are given and some of the difficulties encountered in searching for a global solution of the general problem (P*), reviewed. Problem (P*), when $\mathbb{R}^m$ is equipped with a norm, was considered in [21].

It should be noted that $\|\cdot\|_p$, $0 < p < 1$, is not a norm on $\mathbb{R}^m$. We shall, however, refer to $\|\cdot\|_p$ as the ℓ_p-norm and problem (P) as the ℓ_p problem. Although not a norm, $\|\cdot\|_p$ does satisfy the following properties:

(i) $\|x\|_p \geq 0 \quad \forall x \in \mathbb{R}^m$ with equality if and only if $x = 0$.

(ii) $\|x + y\|_p \leq \|x\|_p + \|y\|_p \quad \forall x, y \in \mathbb{R}^m$.

(iii) $\|\alpha x\|_p = |\alpha|^p \|x\|_p \quad \forall x \in \mathbb{R}^m, \ \forall \alpha \in \mathbb{R}$.

The usual ℓ_p norm when $p \geq 1$, $\|x\|_p = \left(\sum_{j=1}^{m} |x_j|^p\right)^{1/p}$, fails to satisfy the triangle inequality if one chooses $0 < p < 1$ and consequently fails to be even a metric. We regain the triangle inequality at the expense of substituting p-homogeneity, see (iii), for 1-homogeneity by defining $\|\cdot\|_p$ as above for $0 < p < 1$. More generally, a function $\|\cdot\|$ satisfying (i), (ii), (iii) above has been called a p-homogeneous norm.

Let $(\cdot|\cdot)$ denote the usual inner product on $\mathbb{R}^m$, i.e.

$$(x|y) = \sum_{j=1}^{m} x_j y_j .$$

Set

$$K = \text{Image } (A) = \{Ax \mid x \in \mathbb{R}^n\}, \text{ and}$$

$$K^{\perp} = \text{Ker } A^T = \{x \in \mathbb{R}^m \mid (x|k) = 0 \quad \forall k \in K\}.$$

Let $E:\mathbb{R}^m \to K^\perp$ be the orthogonal projection of $\mathbb{R}^m$ onto $K^\perp$, where orthogonal means with respect to the inner product given above, and set $s = Eb$. Let

$$\rho = d(b,K) = \inf\{\|b-k\|_p \,|\, k \in K\}.$$

We assume that $b \notin K$, or equivalently $\rho > 0$, since otherwise the problem is trivial. Finally, let (v) denote the linear span of the vector v.

Observe that $s - b \in K$ since $0 = s - Eb = Es - Eb = E(s-b)$. As a result, $d(s,K) = \inf\{\|s-k\|_p \,|\, k \in K\} = \inf\{\|b-k+(s-b)\|_p \,|\, k \in K\} = \inf\{\|b-k\|_p \,|\, k \in K\} = d(b,K)$.

The existence of a solution of problem (P) follows immediately from the continuity of the ℓ_p norm and the finite dimensionality of the subspace K.

Given problem (P), we associate a dual problem

(P*): Find $z \in K \oplus (s)$, $\|z\|_p = 1$ maximizing $(s|w)$ over all $w \in K \oplus (s)$, $\|w\|_p = 1$.

The relation between problem (P) and problem (P*) is given in the following theorem which extends Theorem 2.4 of [21].

3.1 Theorem. Let z solve problem (P*). Then

$$\text{(i)} \quad \rho^{1/p}(s|z) = (s|s), \tag{3.1.1}$$

and

$$\text{(ii)} \quad b - \rho^{1/p} z \in K. \tag{3.1.2}$$

Proof: (i) $(s|z) = \max\{(s|w) \,|\, w \in K \oplus (s),\ \|w\|_p = 1\}$

$= \max\{(s|k+\beta s) \,|\, k \in K,\ \beta \in \mathbb{R},\ \|k+\beta s\|_p = 1\}$

$$= (s|s)\max\{\beta \in \mathbb{R} \mid \|k + \beta s\|_p = 1,\ k \in K\}$$

$$= (s|s)\max\{\beta \in \mathbb{R} \setminus \{0\} \mid \|k + s\|_p = \frac{1}{\beta^p},\ k \in K\}$$

$$= (s|s)\max\{\frac{1}{\|k + s\|_p^{1/p}} \mid k \in K\}$$

$$= (s|s)\frac{1}{\min\{\|k + s\|_p^{1/p} \mid k \in K\}}$$

$$= (s|s)\frac{1}{\rho^{1/p}} .$$

Thus $\rho^{1/p}(s|z) = (s|s)$ which is (3.1.1).

(ii) Let $t \in K^{\perp}$. We shall show that $(t|b - \rho^{1/p}z) = 0$. Suppose $(t|s) = 0$. Since $z \in K \oplus (s)$ and $t \in K^{\perp} \cap (s)^{\perp}$, $(t|z) = 0$. Also $(t|b) = (Et|b) = (t|Eb) = (t|s) = 0$. Thus

$$(t|b - \rho^{1/p}z) = 0 \quad \text{if} \quad (t|s) = 0. \tag{3.1.3}$$

Next suppose that $t = s$. Since E is the orthogonal projection of $\mathbb{R}^m$ onto $K^{\perp}$ and $s = Eb$, we have that

$$(s|b) = (s|s). \tag{3.1.4}$$

$$(t|b - \rho^{1/p}z) = (s|b - \rho^{1/p}z),$$

$$= (s|b) - \rho^{1/p}(s|z),$$

$$= (s|s) - (s|s), \text{ by (3.1.4) and (3.1.1)},$$

so

$$(t|b - \rho^{1/p}z) = 0 \quad \text{if} \quad t \in (s). \tag{3.1.5}$$

Let F be the orthogonal projection of $\mathbb{R}^m$ onto (s), and let $t \in K^\perp$. Write $t = t_1 + t_2$, where $t_1 = t - Ft$ and $t_2 = Ft$. Clearly, $(t_1|s) = 0$ and $t_2 \in (s)$, so by (3.1.3) and (3.1.5), $(t|b - \rho^{1/p}z) = 0$. Since $t \in K^\perp$ was arbitrary, $b - \rho^{1/p}z \in K$. This completes the proof of the theorem 3.1.

Observe that $\|b - (b - \rho^{1/p}z)\|_p = \|\rho^{1/p}z\| = \rho\|z\|_p = \rho$, so $b - \rho^{1/p}z$ is a point of K of minimal distance from b.

In the statement and proof of Theorem 3.1, we used only the p-homogeneity of the ℓ_p norm on $\mathbb{R}^m$.

Figure 1 on the following page gives a geometric interpretation of theorem 3.1. Given K and b as in the theorem, we are to find $b' \in K$ which minimizes $\{\|b - k\|_p | k \in K\}$. Since the ℓ_p-norm is translation invariant, this is equivalent to finding $s' \in K$ minimizing $\{\|s - k\|_p | k \in K\}$. Let $\rho = d(b,K)$ and let z solve (P*) as indicated in the figure. By expanding the unit ball by a factor of ρ, which by the p-homogeneity of $\|\cdot\|_p$ means multiplying the unit ball by $\rho^{1/p}$, multiplying by -1, and translating it by s, one can easily see that z is taken by the above operations to the point s', i.e. $s' = s - \rho^{1/p}z$. Thus $b' = b - \rho^{1/p}z$.

Figure 1 also indicates in which direction one should look for a characterization of the solutions of problem (P*). By the symmetry and concavity of $\|\cdot\|_p$, a solution of problem (P*) should lie at a "corner" or on an "edge" of the unit ball, and these correspond to points at which a certain number of coordinates of the given point are zero. Before making

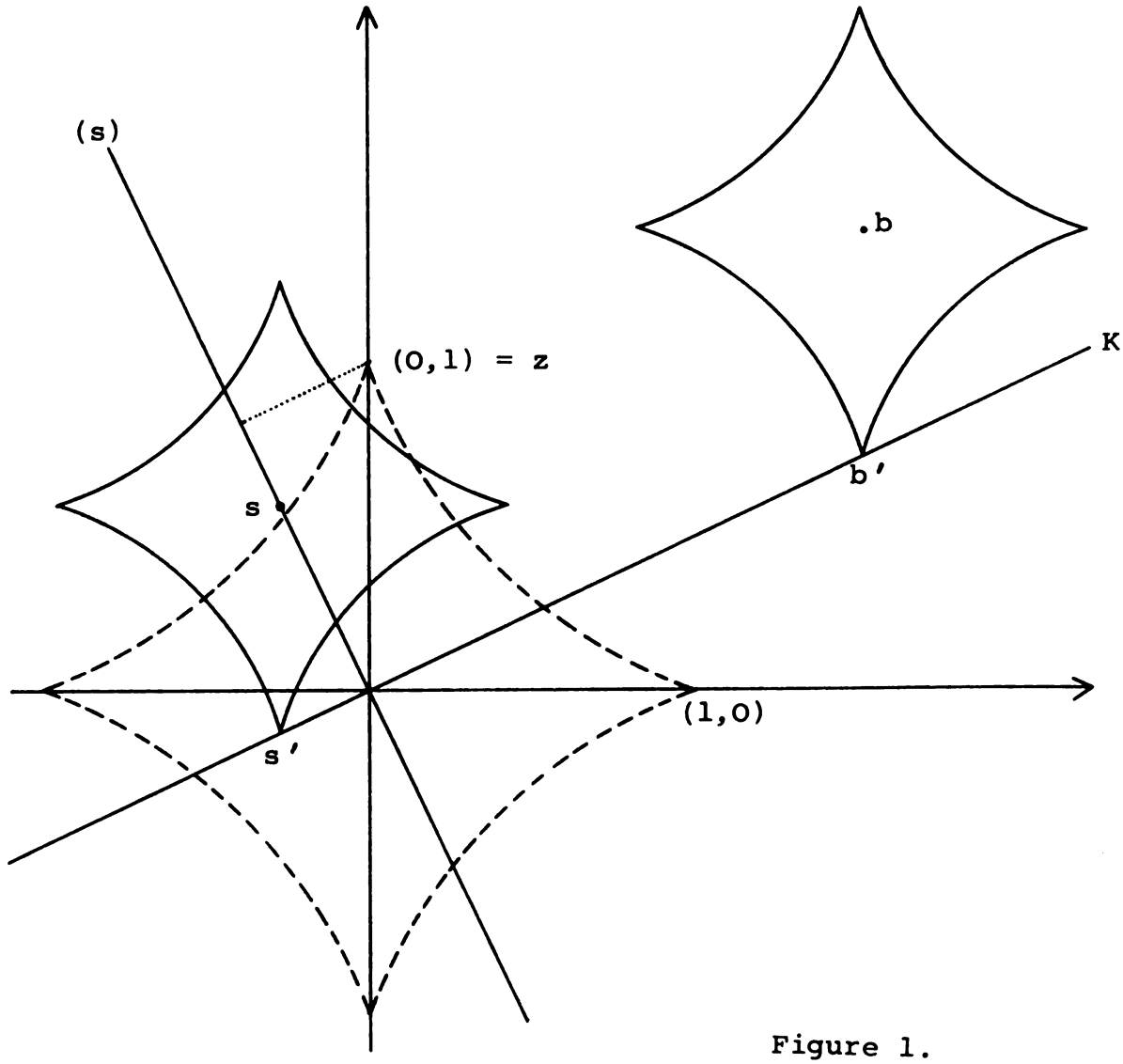

Figure 1.

Geometric interpretation of theorem 3.1.

these intuitive ideas precise, we introduce some terminology and establish a couple of lemmas.

Let X be a real linear metric space, i.e. a real vector space on which a translation invariant metric is defined so that the metric space structure is compatible with the linear space structure. Denote by X^* the algebraic dual of X, i.e. the space of all linear functions on X.

3.2 Definition. Let X be a real linear metric space, $A \subset X$, $a \in A$, and H a non-trivial homogeneous hyperplane in X, i.e. $H = \{x \in X | f(x) = 0\}$, where $f \in X^* \setminus \{0\}$. We say that $H + a = \{h + a | h \in H\}$ supports A at a if either

$$f(x) \geq f(a) \quad \forall x \in A \quad \text{or} \quad f(x) \leq f(a) \quad \forall x \in A. \tag{3.2.1}$$

3.3. Lemma. Let X be a real linear metric space, $f \in X^* \setminus \{0\}$, $A = f^{-1}(0)$, $z \in X \setminus A$, Z a subspace of X with $z \in Z$ and $\dim Z > 0$, and $A_1 = A \cap Z$. Then A_1 is a hyperplane in Z.

Proof: $A_1 = A \cap Z = \{x \in Z | f(x) = 0\}$ being the kernel of a linear functional can fail to be a hyperplane in Z only if either $\dim Z = 0$, which is not so by hypothesis, or if $f \equiv 0$ on Z. The latter is also impossible since $z \in Z \setminus A$ by hypothesis so $f(z) \neq 0$. Thus A_1 is a hyperplane in Z.

3.4 Lemma. Let K be a subspace of $\mathbb{R}^m$, f a linear functional on K with $f \not\equiv 0$ on K, $H = f^{-1}(0) = \{x \in K | f(x) = 0\}$, and $B = \{x \in K \mid \|x\|_p \leq 1\}$. Let $z \in K$ satisfy (i) $\|z\|_p = 1$, (ii) $z_i \neq 0$, $i = 1,2,\ldots,m$, (iii) $H + z$ supports B at z. Then $\dim K = 1$.

Proof: $\dim K \geq 1$ since $z \in K \setminus \{0\}$. Suppose $\dim K > 1$. First we show that $z \notin H$. If $z \in H$, then $f(z) = 0$. Since $\dim H = \dim K - 1 > 0$, there exists $x \in K \setminus H$ with $\|x\|_p \leq 1$, i.e. $x \in B$ and $f(x) > 0$. Hence $f(x) > f(z) > f(-x)$. Since $-x \in B$ this contradicts the hypothesis that $H + z$ supports B at z. Thus $z \notin H$. Choose $x \in H \setminus \{0\}$ and define

$$\delta_i = \begin{cases} 1 & \text{if } x_i = 0 \\ \left|\frac{z_i}{x_i}\right| & \text{if } x_i \neq 0 \end{cases}, \quad i = 1,\ldots,m \tag{3.4.1}$$

$$\delta = \tfrac{1}{2} \min\{\delta_i \mid i = 1,\ldots,m\}. \tag{3.4.2}$$

Then $|z_i + \varepsilon x_i| > 0$ for $i = 1,\ldots,m$ and $\forall \varepsilon \in (-\delta, \delta)$. Let

$$g(\varepsilon) = \|z + \varepsilon x\|_p, \quad -\delta < \varepsilon < \delta . \tag{3.4.3}$$

$$\frac{d^2 g}{d\varepsilon^2} = p(p-1) \sum_{j=1}^{m} x_j^2 |z_j + \varepsilon x_j|^{p-2} < 0 \tag{3.4.4}$$

since not all of the $x_j = 0$ and $0 < p < 1$. This shows that g cannot have a local minimum for $\varepsilon = 0$, i.e.

$$\|z + \tilde{\delta} x\|_p < 1 \quad \text{for} \quad 0 < |\tilde{\delta}| < \delta. \tag{3.4.5}$$

Choose any such $\tilde{\delta}$. Then there must exist $\varepsilon > 0$ such that

$$\|\alpha z + \tilde{\delta} x\|_p \leq 1 \quad \forall\, \alpha \in (1 - \varepsilon, 1 + \varepsilon). \tag{3.4.6}$$

Let $u = (1 - \frac{\varepsilon}{2})z + \tilde{\delta} x$ and $v = (1 + \frac{\varepsilon}{2})z = \tilde{\delta} x$. By construction, $u, v \in B$ and by the linearity of f

$$f(u) = (1 - \tfrac{\varepsilon}{2}) f(z) \quad \text{and} \quad f(v) = (1 + \tfrac{\varepsilon}{2}) f(z). \tag{3.4.7}$$

Since we have already verified that $f(z) \neq 0$, either

$$f(u) < f(z) < f(v) \quad \text{or} \quad f(u) > f(z) > f(v).$$

In either case, we have a contradiction of our hypothesis that $H + z$ supports B at z. Thus the assumption that $\dim K > 1$ must be wrong. In other words, $\dim K = 1$ which is what we wanted to prove.

We are now ready to establish a theorem which guarantees that we need only check a finite number of points in order to find a solution of problem (P*).

3.5 Theorem. Let K be an n dimensional subspace of $\mathbb{R}^m$, $n < m$, with basis $b_1, \ldots, b_n$. Let D be the matrix $(b_1, \ldots, b_n) = (r_1, \ldots, r_m)^T$, where r_i is the $i^{\underline{th}}$ row of the $m \times n$ matrix D. Let $B = \{x \in K \mid \|x\|_p \leq 1\}$. Suppose that f is a non-zero linear functional on K such that $H + z$ supports B at z, where $\|z\|_p = 1$ and $H = \{x \in K \mid f(x) = 0\}$. Denote by I the index set of i's such that $z_i \neq 0$, and $J = \{1, 2, \ldots, m\} \setminus I$. Let A be the matrix with rows r_j, $j \in J$, and set $N = \{x \in \mathbb{R}^n \mid Ax = 0\}$. Then $\dim N = 1$.

Proof: First of all, $\dim N \neq 0$. To see this note that there exists $\beta \in \mathbb{R}^n$ such that $D\beta = z$. By the definition of the set J, it follows that $A\beta = 0$, i.e. $\beta \in N$. Now $\beta \neq 0$ since $z \neq 0$. Hence $\dim N = q \geq 1$.

Without loss of generality, let $I = \{1, \ldots, \mu\}$ and $J = \{\mu + 1, \ldots, m\}$, and put $K^* = \{Dx \mid x \in N\}$. Denote by f^* the restriction of f to K^*, and let $H^* = H \cap K^*$ and $B^* = B \cap K^*$. Clearly, $z \in B^*$.

We claim that $\dim K^* = q$. To see this, observe that the rank of the matrix D is n, and hence by the rank-nullity

theorem of linear algebra, $Dx = 0$ implies that $x = 0$. Hence $\dim K^* = \dim N = q$.

By Lemma 3.3, H^* is a hyperplane in K^*. Finally, $H^* + z$ supports B^* at z since $f^* = f|K^*$. At this point we pause to observe that each $k \in K^*$ satisfies $k_{\mu+1} = \cdots = k_m = 0$. To see this, recall that $k \in K^*$ if and only if there exists an $x \in N$ such that $k = Dx$. But $x \in N$ implies that $Ax = 0$, i.e. $(Dx)_j = 0$ for $j \in J = \{\mu+1,\ldots,m\}$.

We now in essence drop the $m - \mu$ trailing zero coordinates and consider our problem in $\mathbb{R}^\mu$. To make this precise, let

$$\tilde{K} = \{x \in \mathbb{R}^\mu \mid (x_1,\ldots,x_\mu,0,\ldots,0) \in K^*\},$$

$$\tilde{f}:\tilde{K} \to \mathbb{R} \quad \text{by} \quad \tilde{f}(x) = f^*(x_1,\ldots,x_\mu,0,\ldots,0),$$

$$\tilde{H} = \{x \in \tilde{K} \mid \tilde{f}(x) = 0\},$$

$$\tilde{B} = \{x \in \tilde{K} \mid \|x\|_p \leq 1\}, \text{ where now } \|x\|_p = \sum_{j=1}^{\mu} |x_j|^p,$$

$$\tilde{z} = (z_1,\ldots,z_\mu).$$

Notice that $\tilde{z}$ has no coordinates equal to zero, and $\sum_{j=1}^{\mu} |\tilde{z}_j|^p = 1$. Also, $\tilde{H}$ is a hyperplane in $\tilde{K}$ and $\tilde{H} + \tilde{z}$ supports $\tilde{B}$ at $\tilde{z}$. Finally, $\dim \tilde{K} = \dim K^* = q$. But by lemma 3.4, $q = 1$, which is precisely what we wanted to prove, namely $\dim N = 1$.

3.6 Definition. With an eye toward the future and a certain dislike of repetition, we define the phrase "the usual n dimensional situation" to be the following:

K is an n dimensional subspace of $\mathbb{R}^m$ with basis $b_1,\ldots,b_n$; $n < m$; $s \in K^{\perp} \setminus \{0\}$ where $K^{\perp} = \{x \in \mathbb{R}^m \mid (x|k) = 0 \ \forall k \in K\}$; $D = (b_1,\ldots,b_n,s) = (r_1,\ldots,r_m)^T$ where r_i is the $i^{\underline{th}}$ row of the $m \times (n+1)$ matrix D; $B = \{x \in K \oplus (s) \mid \|x\|_p \leq 1\}$; $0 < p < 1$.

3.7 Definition. Given "the usual n dimensional situation", a point $z \in B$ is called a corner point or simply a corner if $\dim\{x \in \mathbb{R}^{n+1} \mid Ax = 0\} = 1$ where $A = (r_{i_1},\ldots,r_{i_k})^T$, and $\{i_1,\ldots,i_k\} = \{i \mid z_i = 0\}$.

Observe that if we are working in "the usual n dimensional situation" and if z is a corner point, then at least n coordinates of z must be zero. Using the same notation as earlier, since $\dim\{x \in \mathbb{R}^{n+1} \mid Ax = 0\} = 1$, A must have at least n rows, and hence z must have at least n coordinates equal to zero.

We next show that the solution of problem (P*) must be a corner point, and that there are only a finite number of such points. We will have then reduced our problem to that of finding and checking a finite number of points.

Recall that problem (P*) requires us to find $z \in K \oplus (s)$, $\|z\|_p = 1$, such that $(z|s) = \max\{(w|s) \mid w \in K \oplus (s), \|w\|_p = 1\}$, i.e. find $z \in B$ such that $K + z$ supports B at z. Using this formulation of problem (P*) and assuming that z is a solution, we see that all of the hypotheses of

Theorem 3.5 are satisfied, so z must be a corner point. Thus we have proven

3.8 Theorem. Given "the usual n dimensional situation" and that z solves problem (P*), then z is a corner point.

3.9 Corollary. Under the hypotheses of Theorem 3.8, z has at least n coordinates equal to zero.

Proof: This is just the observation made earlier.

3.10 Corollary. If, in addition to the hypotheses of Theorem 3.8, D satisfies the Haar Condition, i.e. each $n+1$ rows of D are linearly independent, then a solution z of problem (P*) has exactly n coordinates equal to zero.

Proof: We again use the notation of Theorem 3.5. Since $\dim\{x \in \mathbb{R}^{n+1} \mid Ax = 0\} = 1$, the Haar Condition forces A to be an $n \times (n+1)$ matrix. Hence z has exactly n coordinates equal to zero.

3.11 Corollary. Assume that "the usual n dimensional situation" holds, that $n = m-1$, and that $|s_j| = \max\{|s_i| \mid i = 1,\ldots,m\}$. Then a solution of problem (P*) is $z = e_j \operatorname{sgn} s_j$, where e_j is the usual unit basis vector in $\mathbb{R}^m$.

Proof: By Corollary 3.9, z must have $n = m-1$ coordinates equal to zero, and since $\|z\|_p = 1$, z must be one of the vectors $\pm e_i$, $1 \leq i \leq m$. $(\pm e_i \mid s) = \pm s_i$ is clearly maximized

by $e_j \operatorname{sgn} s_j$ where $|s_j| = \max\{|_i| \mid i = 1,\ldots,m\}$. Hence $e_j \operatorname{sgn} s_j$ solves problem (P*).

This case in which $K \oplus (s) = \mathbb{R}^m$ and the trivial case where $K = \{0\}$ for which $z = \frac{s}{\|s\|_p^{1/p}}$ solves problem (P*) are rare in that they can be completely solved directly with little effort. Most of the others, as we shall see, require considerably more work.

3.12 Lemma. Suppose that we have "the usual n dimensional situation", and that x,y are corner points with $x = D\beta$, $y = D\gamma$, $I = \{i \mid x_i = 0\}$, $J = \{j \mid y_j = 0\}$, and $J \subset I$. Then $x = \pm y$.

Proof: Let $M = \{u \in \mathbb{R}^{n+1} \mid (u|r_i) = 0, i \in I\}$ and $N = \{v \in \mathbb{R}^{n+1} \mid (v|r_j) = 0, j \in J\}$. Since $J \subset I$, $M \subset N$. And since x and y are both corner points, $\dim M = \dim N = 1$. Thus $M = N = (\beta)$. Finally, $\|x\|_p = \|y\|_p = 1$ implies that $x = \pm y$.

3.13 Theorem. In "the usual n dimensional situation", there are at most $\binom{m}{n}$ different corner points. Moreover, if D satisfies the Haar Condition, then there are exactly $\binom{m}{n}$ different corner points. (By different corner points we mean that if z is a corner point, then -z will not be considered as a corner point also.)

Proof: Recall that z is a corner point if and only if $\dim\{x \in \mathbb{R}^{n+1} \mid (x|r_i) = 0 \ \forall i$ such that $z_i = 0\} = 1$, where

r_i is the i^{th} row if the matrix D given in Definition 3.6. Since there are $\sum_{j=n}^{m-1}\binom{m}{j}$ ways of choosing at least n but no more than $m-1$ coordinates of a point $z \in \mathbb{R}^m$ to be zero, there are at most $\sum_{j=n}^{m-1}\binom{m}{j}$ corner points, i.e. the number of corner points is finite.

Let the set of $\binom{m}{n}$ distinct n element subsets of $\{1,\ldots,m\}$ be denoted by E_1. Let $c(1),\ldots,c(q)$ be a complete enumeration of all the different corner points. We shall show that to each corner point $c(i)$ we can assign a distinct $I \in E_1$ establishing that the number of different corner points is at most $\binom{m}{n}$.

For $i = 1,\ldots,q$, define $G_i = \{j | c_j(i) = 0\}$, where $c_j(i)$ is the j^{th} coordinate of $c(i)$. We now define F_i and E_i, $i = 1,\ldots,q$, iteratively by setting $F_i = \{ I \in E_i | I \subset G_i\}$ and $E_{i+1} = E_i \setminus F_i$. We claim that $F_i \neq \emptyset$, $i = 1,\ldots,q$. For suppose the claim were not true. Then for some k, $1 \leq k \leq q$, $F_k = \emptyset$. Since $c(K)$ is a corner point, $\dim\{x \in \mathbb{R}^{n+1} | (x|r_j) = 0, j \in G_k\} = 1$. Thus there must be n linearly independent rows of D, say r_i, $i \in I = \{i_1,\ldots, i_n\} \subset G_k$, such that $\dim\{x \in \mathbb{R}^{n+1} | (x|r_i) = 0, i \in I\} = 1$. Note that $c_j(k) = 0$ for all $j \in I$ since $I \subset G_k$. By assumption,

$$\emptyset = F_k = E_1 \setminus (F_1 \cup \ldots \cup F_{k-1}).$$

So $I \in F_\ell$ for some ℓ, $1 \leq \ell \leq k-1$, i.e. $I \subset G_\ell$ and hence $c_i(\ell) = 0$, $i \in I$. Since $\dim\{x \in \mathbb{R}^{n+1} | (x|r_i) = 0,$

$i \in I\} = 1 = \dim\{x \in \mathbb{R}^{n+1} \mid (x|r_j) = 0, j \in G_\ell\}$ and $I \subset G_\ell$, we conclude that

$$\{x \in \mathbb{R}^{n+1} \mid (x|r_i) = 0, i \in I\} = \{x \in \mathbb{R}^{n+1} \mid (x|r_j) = 0, j \in G_\ell\}.$$

Moreover, since $\|c(k)\|_p = \|c(\ell)\|_p$, $c(k) = \pm c(\ell)$ contradicting the assumption that $c(k)$ and $c(\ell)$ are different corner points. Hence $F_i \neq \emptyset$, $i = 1, \ldots, q$.

By construction, the F_i are mutually exclusive, so to each corner point $c(i)$ we can assign a distinct $I \in F_i$. Thus there are at most $\binom{m}{n}$ different corner points.

If D satisfies the Haar Condition, then by Corollary 3.10 each corner point has exactly n coordinates equal to zero. Each of the $\binom{m}{n}$ choices of n coordinates from the m yields an $n \times (n+1)$ matrix A for which $\dim\{x \in \mathbb{R}^{n+1} \mid Ax = 0\} = 1$, and hence each of the $\binom{m}{n}$ possible choices produce a different corner point. This completes the proof of Theorem 3.13.

Corollary 3.9 was first proven by Motzkin and Walsh [15,Theorem 6] in the following form:

> "Let E consist of the real points $x_1, \ldots, x_m$ $(m \geq n+1)$, let $F(x)$ be defined on E, let p $(0 < p < 1)$ be given, and let the functions $\psi_1(x), \ldots, \psi_{n+1}(x)$ satisfy Condition A. Then every function $P(x) \equiv \sum_{j=1}^{m+1} \alpha_\nu \psi_\nu(x)$ of best approximation measured by the deviation

$$\sum_{k=1}^{m} \mu_k |F(x_k) - P(x_k)|^p \qquad (\mu_k > 0)$$

coincides with $F(x)$ in at least $n + 1$ points of E."

Condition A says that the $m \times (n+1)$ matrix $(\Psi_j(x_i))$ has rank $n+1$.

In the same paper, Motzkin and Walsh observe that "Theorem 6 implies that every extremal polynomial $P(x)$ is found by interpolation to $F(x)$ in $n+1$ points of E; there exist but a finite number of polynomials interpolating to $F(x)$ in $n+1$ points of E, so every extremal polynomial can be found merely by comparing their measures of approximation."

Without making a further assumption about the matrix $\Psi = (\Psi_j(x_i))$, namely that it satisfies the Haar Condition, there need not be only a finite number of polynomials interpolating $F(x)$ at $n+1$ points of E. The suggested procedure for finding a polynomial of best approximation, consequently, need not be finite. Applying Corollary 3.10, we can prove that the observation of Motzkin and Walsh is correct if Ψ satisfies the Haar Condition. When the Haar Condition is violated, however, we can easily construct, even in $\mathbb{R}^3$, counterexamples to the assertion of Motzkin and Walsh. Suppose $m = 3$, $n = 1$, $F(x_1) = F(x_2) = 0$, $F(x_3) = \Psi_1(x_1) = \Psi_1(x_2) = \Psi_2(x_1) = \Psi_2(x_2) = \Psi_2(x_3) = 1$, and $\Psi_1(x_3) = -2$. The matrix $\Psi = (\Psi_j(x_i)) = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ -2 & 1 \end{pmatrix}$ clearly has rank $n+1 = 2$,

so Condition A is satisfied. For any $\alpha \in \mathbb{R}$,

$$P(x) = -\alpha\psi_1(x) + \alpha\psi_2(x)$$

satisfies $P(x_1) = P(x_2) = 0$, so that $P(x)$ interpolates $F(x)$ in $n+1 = 2$ points of E. Clearly, there are an infinite number of these interpolating polynomials showing that the observation of Motzkin and Walsh is incorrect.

The solution of the given approximation problem is indeed included among all those functions which interpolate $F(x)$ in at least $n+1$ points, but this collection of interpolating functions need not be finite as claimed by Motzkin and Walsh. Of course, this counterexample is possible only because ψ does not satisfy the Haar Condition. By Theorems 3.8 and 3.13, we are guaranteed that the given approximation problem can always be solved in a finite number of steps.

Before proceeding any further toward a solution of problem (P*), we pause to consider some of the difficulties that lie ahead. Essentially we want to find a point on an $n+1$ dimensional cross section of the ℓ_p unit ball in $\mathbb{R}^m$ at which a translation of a fixed subspace is tangent to the ball. So far we have reduced the problem to one of considering only a finite number of points. As one can see in the following figures, these points correspond closely to corners of a polyhedron, hence the name corner points. In the figures, the corner points have been connected by straight lines rather than by the curved arcs that one would obtain when the ℓ_p unit ball is intersected with the specified plane K.

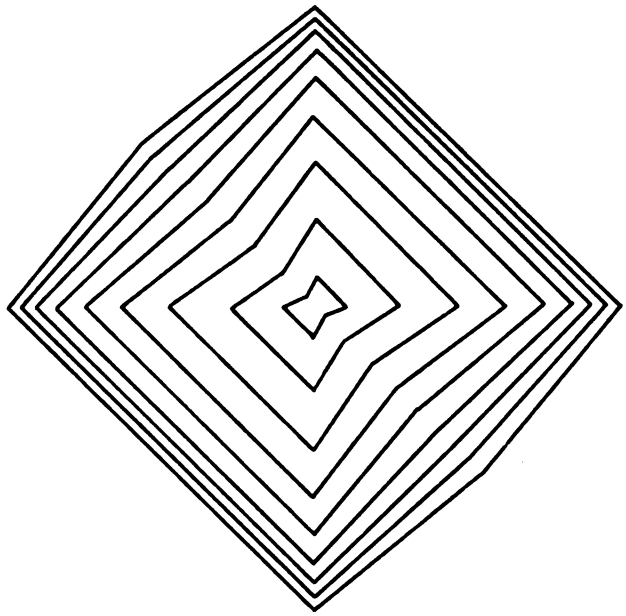

Figure 2(a). m = 3

$p = 1.0, .9, .8, \ldots, .2$

$$K = \text{span}\left\{\begin{bmatrix} 0 \\ -1 \\ 10 \end{bmatrix}, \begin{bmatrix} 10 \\ 1 \\ 0 \end{bmatrix}\right\}$$

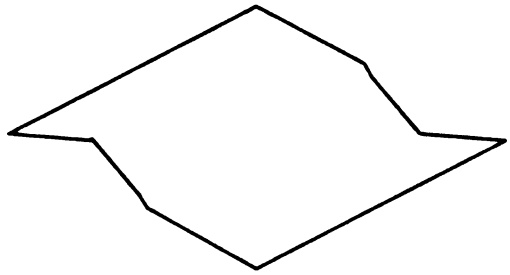

Figure 2(b). m = 5, p = .25

$$K = \text{span}\left\{\begin{bmatrix} -1 \\ 0 \\ 1 \\ 2 \\ 30 \end{bmatrix}, \begin{bmatrix} 30 \\ 1 \\ 2 \\ 3 \\ 0 \end{bmatrix}\right\}$$

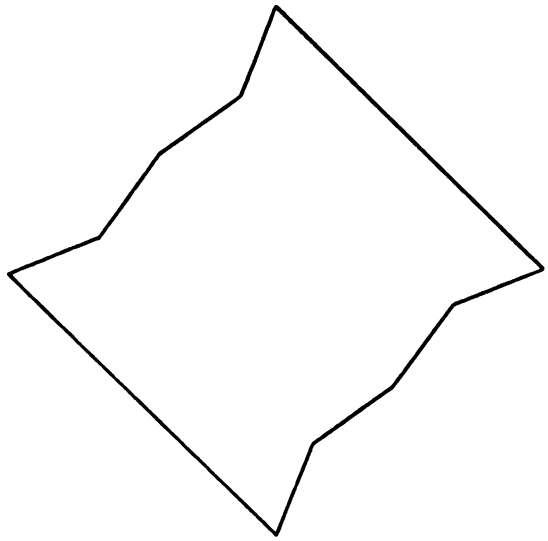

Figure 2(c). m = 5, p = .2

$$K = \text{span}\left\{\begin{bmatrix} 50 \\ 1 \\ 5 \\ 5 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -5 \\ -1 \\ -1 \\ 50 \end{bmatrix}\right\}$$

Our approach to the problem up to now has been similar to the idea behind linear programming. In linear programming problems, a linear functional defined on a finite dimensional space is maximized subject to certain linear constraints on the same space. In this case, one also proceeds by reducing the number of points at which the solution may occur from an infinite number to a finite one. In linear programming, however, the points one is left with actually are corners of a convex polyhedron. To solve the linear programming problem, one moves from one corner to an adjacent one always increasing (at least not decreasing) the value of the objective function, until a corner point is reached where all adjacent corners give no higher values for the objective function. By the convexity of the polyhedron, one then concludes that such a point is in fact a solution of the linear programming problem. The advantage of such an algorithm is that one may not have to check all of the corner points all of the time. In practice, the number of points actually computed and checked is usually far less than the total number of corners, though there are examples where the usual simplex algorithm would end up checking all of the corner points.

If possible, we would like to develop a similar exchange type algorithm to obtain a solution of problem (P*). There is no difficulty in moving from one corner to another continually increasing the value of the objective function until a point is reached all of whose neighbors yield no higher value of the objective function. The crucial step at which

we encounter trouble is trying to conclude that such a point solves problem (P*). Since the ℓ_p unit ball is not convex, some of the corner points often lie inside the convex hull of all of the corner points. In figure 2(a), as p decreases the cross section of the unit ball changes from convex to non convex. Figures 2(b) and 2(c) are further examples of non convex cross sections of various ℓ_p unit balls. Moreover, it is quite possible that a local maximum of the objective function may not be a global maximum. In other words, we may not have solved problem (P*). Figure 2(c) shows a situation in which this may occur

Keeping this obstacle of non convexity clearly in view, we begin to consider the practical task of actually solving problem (P*). We now turn to the construction of algorithms for solving problem (P*).

3.14 Remarks. Given any point $z \in K \oplus (s)$, there exists a $\beta \in \mathbb{R}^{n+1}$ such that $z = D\beta$, where the definition of D appears in 3.6. Also, since $s \in K^{\perp}$, $(z|s) = (D\beta|s) = \beta_{n+1}(s|s)$. Problem (P*) requires us to

maximize $(w|s)$ over all $w \in K \oplus (s)$, $\|w\|_p = 1$, i.e.
maximize $\beta_{n+1}(s|s)$ over all $\beta \in \mathbb{R}^{n+1}$, $\|D\beta\|_p = 1$, i.e.
maximize β_{n+1} over all $\beta \in \mathbb{R}^{n+1}$, $\|D\beta\|_p = 1$.

Using this last formulation of problem (P*) and Theorem 3.8, we see that the solution of problem (P*) entails at most the following:

(1) find all those $\beta \in \mathbb{R}^{n+1}$ for which $D\beta$ is a corner point;

(2) select from this collection the one with the largest $n+1^{st}$ coordinate, say β. Then

(3) $z = D\beta$ solves problem (P*).

If D satisfies the Haar Condition, then it is quite easy to find all $\beta \in \mathbb{R}^{n+1}$ such that $D\beta$ is a corner point. From Theorem 3.13 and Corollary 3.10, we know that the solution of problem (P*) must be among the $\binom{m}{n}$ distinct corner points each of which have n coordinates equal to zero and $m-n$ coordinates non-zero. Suppose we wish to find the corner point z for which $\{i|z_i = 0\} = \{i_1,\ldots,i_n\}$. Select any $i_{n+1} \in \{1,\ldots,m\} \setminus \{i_1,\ldots,i_n\}$, and with $A = (r_{i_1},\ldots,r_{i_{n+1}})^T$, solve $Ax = e_{n+1}$ where $e_{n+1} = (0,\ldots,0,1)^T \in \mathbb{R}^{n+1}$. The Haar Condition insures that x exists and is unique. Set $\beta = \frac{x}{\|Dx\|_p^{1/p}}$. Then $z = D\beta$ has exactly n coordinates equal to zero, viz. $z_{i_1} = \cdots = z_{i_n} = 0$, and $\|z\|_p = 1$. In this manner, all of the corner points can be found.

The above algorithm lacks three important features: (1) a method for determining if the Haar Condition is satisfied; (2) an orderly way of selecting the $\binom{m}{n}$ points to be computed; and (3) the ability, at least theoretically, to ignore some of the corner points some of the time.

The first of these is lacking for the very good reason that checking for the Haar Condition involves more work than

solving the original problem itself. The second shortcoming will be corrected shortly, and the third problem will be considered afterwards.

Since the set of points in $K \oplus (s)$ of unit norm and having exactly n coordinates equal to zero is identically the same as the set of all corner points when the Haar Condition holds, having that condition satisfied seems like an ideal setting in which to work. The tremendous problem of verifying that the Haar Condition holds diminishes that ideal somewhat. An even more damaging blow is leveled by Lemma 3.12 which says that the greatest number of possible corner points occurs only when the Haar Condition holds. In other words, that situation requires the most work to find all of the corner points since there are more corners to find. Our algorithms for finding all of the corner points will work whether the Haar Condition holds or not. In fact, it is actually faster without that condition present.

3.15 Definition. Let $U = \{u \in \mathbb{N}^n \mid 1 \leq u_1 < u_2 < \cdots < u_n \leq m\}$ and $T = \{1, \ldots, \binom{m}{n}\}$. Define $\Psi: T \to U$ by the following rules. Given $t \in T$,

(1) Set $t_0 = t$, $u_0 = 0$, and $i = 1$.

(2) Find $u_i \in \mathbb{N}$ such that $u_{i-1} < u_i \leq m - n + i$ and $1 \leq t_{i-1} - \sum_{j=1+u_{i-1}}^{u_i - 1} \binom{m-j}{n-i} \leq \binom{m-u_i}{n-i}$.

(3) Set $t_i = t_{i-1} - \sum_{j=1+u_{i-1}}^{u_i - 1} \binom{m-j}{n-i}$ and increment i by 1.

(4) Repeat steps 2 and 3 until u_n has been found.

Then $\Psi(t) = u \in U$ where the components $u_1,\ldots,u_n$ of u were found above. We adopt the convention that $\sum_{j=\mu}^{\nu}(\cdot) = 0$ if $\mu > \nu$.

We must verify that Definition 3.15 makes sense, i.e. that u can be found as claimed. We begin this task by proving the following lemma.

3.16 Lemma. Let $m,n,i,k \in \mathbb{N}$, $i \leq n < m$, $k \leq m+i-n-1$.

Then $\sum_{j=1+k}^{m-n+i} \binom{m-j}{n-i} = \binom{m-k}{n-i+1}$

Proof: Since $\binom{p}{q} = \binom{p-1}{q-1} + \binom{p-1}{q}$ for all $p,q \in \mathbb{N}$, $p > q$,

$$\begin{aligned}
\binom{m-k}{n-i+1} &= \binom{m-k-1}{n-i} + \binom{m-k-1}{n-i+1}, \\
&= \binom{m-k-1}{n-i} + \left[\binom{m-k-2}{n-i} + \binom{m-k-2}{n-i+1}\right], \\
&\cdots \\
&= \binom{m-k-1}{n-i} + \binom{m-k-2}{n-i} + \cdots + \binom{n-i+1}{n-i} + \binom{n-i+1}{n-i+1}, \\
&= \binom{m-k-1}{n-i} + \binom{m-k-2}{n-i} + \cdots + \binom{n-i+1}{n-i} + \binom{n-i}{n-i}, \\
&= \sum_{j=1+k}^{m-n+i} \binom{m-j}{n-i}.
\end{aligned}$$

3.17 Proposition. Ψ is well defined.

Proof: First observe that for any i, $1 \leq i \leq n$, if u_i exists then it is unique. For suppose that both μ,ν satisfy

the conditions for u_i specified in (2) of definition 3.15, and suppose $\mu < \nu$. Then

$$1 \leq t_{i-1} - \sum_{j=1+u_{i-1}}^{\nu-1} \binom{m-j}{n-i},$$

$$= t_{i-1} - \sum_{j=1+u_{i-1}}^{\mu-1} \binom{m-j}{n-i} - \binom{m-\mu}{n-i} - \sum_{j=\mu+1}^{\nu-1} \binom{m-j}{n-i},$$

$$\leq t_{i-1} - \sum_{j=1+u_{i-1}}^{\mu-1} \binom{m-j}{n-i} - \binom{m-\mu}{n-i} \leq 0,$$

since $1 \leq t_{i-1} - \sum_{j=1+u_{i-1}}^{u_i-1} \binom{m-j}{n-i} \leq \binom{m-\mu}{n-i}$ by hypothesis. This contradiction leads us to conclude that if u_i exists then it is unique.

We show inductively that the u_i exists. Since $u_0 = 0$, we first construct u_1. For this we must show that for some integer k, $0 < k \leq m - n + 1$,

$$1 \leq t - \sum_{j=1}^{k-1} \binom{m-j}{n-1} \leq \binom{m-k}{n-1}.$$

By hypothesis, $t \leq \binom{m}{n}$, and by Lemma 3.16, $\binom{m}{n} = \sum_{j=1}^{m-n+1} \binom{m-j}{n-1}$. Thus there is a smallest integer k, $0 < k \leq m - n + 1$, such that

$$\sum_{j=1}^{k-1} \binom{m-j}{n-1} < t \leq \sum_{j=1}^{k} \binom{m-j}{n-1}. \qquad (3.17.1)$$

If $k > 1$, the left inequality of (3.17.1) follows from the definition of k. If $k = 1$, then $\sum_{j=1}^{k-1} \binom{m-j}{n-1} = 0$, and by

hypothesis, $t > 0$. Thus we conclude from (3.17.1) that

$$1 \leq t - \sum_{j=1}^{k-1} \binom{m-j}{n-1} \leq \binom{m-k}{n-1},$$

which is what we were to show. We let $u_1 = k$ proving the existence of u_1. We have already established the uniqueness of u_1.

Suppose that $u_1,\ldots,u_{i-1}$ have been found satisfying the conditions specified in Definition 3.15. We next show that there exists a unique u_i, $u_{i-1} < u_i \leq m - n + i$, such that

$$1 \leq t_{i-1} - \sum_{j=1+u_{i-1}}^{u_i-1} \binom{m-j}{n-i} \leq \binom{m-u_i}{n-i}.$$

By the induction hypothesis on u_{i-1}, i.e. that u_{i-1} satisfies the conditions of Definition 3.15,

$$t_{i-1} = t_{i-2} - \sum_{j=1+u_{i-2}}^{u_{i-1}-1} \binom{m-j}{n-i+1} \leq \binom{m-u_{i-1}}{n-i+1}.$$

And by Lemma 3.16, $\binom{m-u_{i-1}}{n-i+1} = \sum_{j=1+u_{i-1}}^{m-n+i} \binom{m-j}{n-i}$. Thus

$$t_{i-1} \leq \sum_{j=1+u_{i-1}}^{m-n+1} \binom{m-j}{n-i}.$$

So there is a smallest integer k, $u_{i-1} < k \leq m - n + i$, such that

$$\sum_{j=1+u_{i-1}}^{k-1} \binom{m-j}{n-i} < t_{i-1} \leq \sum_{j=1+u_{i-1}}^{k} \binom{m-j}{n-i}. \qquad (3.17.2)$$

If $k > u_{i-1} + 1$, the left inequality of (3.17.2) follows from the definition of k. If $k = u_{i-1} + 1$, then $\sum_{j=1+u_{i-1}}^{k-1} \binom{m-j}{n-i} = 0$, and by the induction hypothesis on u_{i-1}, $t_{i-1} \geq 1$. Hence from (3.17.2) we have

$$1 \leq t_{i-1} - \sum_{j=1+u_{i-1}}^{k-1} \binom{m-j}{n-i} \leq \binom{m-k}{n-i} .$$

By the uniqueness of u_2 already proven, we conclude that there exists a unique u_i satisfying (2) of Definition 3.15. This concludes the proof of Proposition 3.17.

3.18 Proposition. Ψ is a one-to-one function.

Proof: Let $1 \leq t, t' \leq \binom{m}{n}$ such that $\Psi(t) = \Psi(t') = u$. Using the notation of Definition 3.15, observe that

$$1 \leq t_n \leq \binom{m-u_n}{n-n} = \binom{m-u_n}{0} = 1$$

implying that $t_n = t'_n = 1$. Suppose that $t_{n-k} = t'_{n-k}$ for $k = 0,1,\ldots,i$. We shall show that $t_{n-(i+1)} = t'_{n-(i+1)}$. By (3) of definition 3.15,

$$t_{n-(i+1)} = t_{n-i} + \sum_{j=1+u_{n-i-1}}^{u_{n-i}-1} \binom{m-j}{i} ,$$

$$= t'_{n-i} + \sum_{j=1+u_{n-i-1}}^{u_{n-i}-1} \binom{m-j}{i} ,$$

$$= t'_{n-(i+1)} .$$

Hence $t_k = t'_k$, $k = n, n-1, \ldots, 0$. But $t = t_0$ and $t'_0 = t'$ so that $t = t'$. Thus Ψ is a one-to-one function.

Both T and U have $\binom{m}{n}$ elements by construction, and since $\Psi: T \to U$ is one-to-one, it must also be onto. Hence $\Psi^{-1}: U \to T$ exists. Moreover, by rearranging (3) of definition 3.15 and iterating, we have

$$\Psi^{-1}(u) = 1 + \sum_{i=1}^{n} \left[\sum_{j=1+u_{i-1}}^{u_i - 1} \binom{m-j}{n-i} \right],$$

where we let $u_0 = 0$ and $\sum_{j=\mu}^{\nu} (\cdot) = 0$ if $\mu > \nu$.

3.19 Definition. Let T and U be as in definition 3.15. Define $\Phi: U \to T$ by $\Phi = \Psi^{-1}$.

An element $u \in U$ corresponds to the possible corner point $z \in \mathbb{R}^m$ for which $z_i = 0$, $i = u_1, \ldots, u_n$. The correspondence $U \leftrightarrow T$ is not as arbitrary as it may appear at first. An example with $m = 6$, $n = 3$ appears in Table 3.1. Notice that the first zero coordinates appear as high as possible for a point in $\mathbb{R}^m$ on the left and proceed downward to the right.

We are now prepared to present an algorithm for solving problem (P*). Assume that we have "the usual n dimensional situation".

3.20 Algorithm. Step 1 Set $q = 1$, $p_i = 1$, $\beta_i = 0$ for $i = 1, \ldots, \binom{m}{n}$.

Step 2 Compute $\Psi(q) = (k_1, \ldots, k_n)^T$.

$\Phi(u) = t =$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	4
$u = \Psi(t) =$	2	2	2	2	3	3	3	4	4	5	3	3	3	4	4	5	4	4	5	5
	3	4	5	6	4	5	6	5	6	6	4	5	6	5	6	6	5	6	6	6
Corresponding point in $\mathbb{R}^m$	o	o	o	o	o	o	o	o	o	o										
	o	o	o	o							o	o	o	o	o	o				
	o				o	o	o				o	o	o				o	o	o	
		o			o			o	o		o			o	o		o	o		o
			o			o		o		o		o		o		o	o		o	o
				o			o		o	o			o		o	o		o	o	o
					$\binom{5}{2}$								$\binom{4}{2}$				$\binom{3}{2}$			$\binom{2}{2}$

Table 3.1

<u>Example 1</u>: Compute $\Psi(15)$. Given: $u_0 = 0$ $t_0 = 15$

Compute: $u_1 = 2$ $t_1 = 5$

$u_2 = 4$ $t_2 = 2$

$u_3 = 6$ $t_3 = 1$

$\Psi(15) = \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix}$, which agrees with the table above.

<u>Example 2</u>: Compute $\Phi(u)$ for $u = (1,4,5)^T$

$$\Phi(u) = 1 + \sum_{i=1}^{n} \left[\sum_{j=l_{i-1}+1}^{u_i - 1} \binom{m-j}{n-i} \right] = 1 + \sum_{i=1}^{3} \left[\sum_{j=1+u_{i-1}}^{u_i - 1} \binom{6-j}{3-i} \right]$$

$$= 1 + \sum_{j=1}^{0} \binom{6-j}{2} + \sum_{j=2}^{3} \binom{6-j}{1} + \sum_{j=5}^{4} \binom{6-j}{0} = 1 + 0 + \binom{4}{1} + \binom{3}{1} + 0$$

$= 1 + 4 + 3 = 8$, which again agrees with the table above.

Step 3 Construct $A = \begin{bmatrix} r_{k_1} \\ \vdots \\ r_{k_n} \end{bmatrix}$ and set $I = \{k_1, \ldots k_n\}$.

Step 4 Select $i \in \{1, \ldots, m\} \setminus I$ and form $C = \begin{pmatrix} A \\ r_i \end{pmatrix}$.

Step 5 Does C contain $n+1$ linearly independent rows? If yes, then go to Step 7. If no, then continue.

Step 6 Set $A = C$, $I = I \cup \{i\}$ and return to Step 4.

Step 7 Solve $Ax = 0$, $(r_i | x) = 1$.

Step 8 Compute Dx and $\|Dx\|_p^{1/p} = \gamma$, where D is given in definition 3.6.

Step 9 Set $\beta_q = \dfrac{|x_{n+1}|}{\gamma}$ and $z(q) = \dfrac{\operatorname{sgn} x_{n+1}}{\gamma} Dx$. Let $J = \{j | (Dx)_j = 0\}$.

Step 10 Form all possible sets containing exactly n different elements of J.

Step 11 For each set $\{j_1, \ldots, j_n\}$ found in Step 10, with $1 \le j_1 < j_2 < \ldots < j_n \le m$ and $j = (j_i)_{i=1}^n$, compute $\phi(j) = t$ and set $p_t = 0$.

Step 12 If $p_i = 0$, $i = 1, \ldots, \binom{m}{n}$, then go to Step 13. Otherwise, let q be the smallest integer k, $1 \le k \le \binom{m}{n}$ such that $p_k = 1$. Return to Step 2.

Step 13 Select k, $1 \leq k \leq \binom{m}{n}$, such that $\beta_k = \max\{\beta_i \mid 1 \leq i \leq \binom{m}{n}\}$. Then $z(k)$ solves problem (P*) and $\max\{(s|w) \mid w \in K \oplus (s), \|w\|_p = 1\} = \beta_k(s|s)$.

In Step 5, one must eventually answer the question in the affirmative since the row rank of D equals the column rank of D which is $n+1$ by hypothesis. The question itself can be answered in a number of ways. For example, one might orthogonalize the rows of C and check whether any zero rows occur. This method will also help when step 7 is reached since one then knows which rows of C yield a nonsingular matrix G with which to solve $Gx = e_n$. Steps 10 and 11 are present to exploit Lemma 3.12 which says that some of the original $\binom{m}{n}$ possible corner points may in fact be redundant. In Step 9, one need not save all of the β_i and $z(i)$ but only the current largest β_i and the corresponding $z(i)$ which would make Step 13 unnecessary.

Algorithm 3.20 solves problem (P*) for any choice of positive integers m,n with $m > n$. The price paid for this flexibility is a rather complex algorithm involving a considerable amount of index manipulation. In the case where n is very small or nearly equal to m, one can avoid much of this work by developing special algorithms designed to solve only problems with a particular fixed choice of dim K. As with algorithm 3.20, Corollary 3.9 provides the basis for each algorithm.

Recall that if $\dim K = n$, then z, the solution of problem (P*), has at most $m-n$ nonzero coordinates, and $z \in K \oplus (s)$ implies that z must also satisfy $m-(n+1)$ orthogonality constraints. For $n = m-1, m-2, m-3$, these facts lead to simpler algorithms for solving problem (P*).

3.21 Algorithm. Corollary 3.8 has already given the solution z when $n = m-1$. In that case, $z = e_j \operatorname{sgn} s_j$, where e_j is the usual unit basis vector in $\mathbb{R}^m$ and j is determined by $|s_j| = \max\{|s_i| \mid 1 \le i \le m\}$.

3.22 Algorithm. Suppose that $n = m-2$ and that $a \in (K \oplus (s))^{\perp} \setminus \{0\}$. With $s = (s_1, \ldots, s_m)^T$, $a = (a_1, \ldots, a_m)^T$, and $z = (z_1, \ldots, z_m)^T$, problem (P*) reduces to

$$\text{Maximize } f_{ij},\ 1 \le i, j \le m,\ i \ne j,$$

$$f_{ij} = s_i z_i + s_j z_j, \tag{3.22.1}$$

$$\text{subject to } 0 = a_i z_i + a_j z_j, \tag{3.22.2}$$

$$1 = |z_i|^p + |z_j|^p. \tag{3.22.3}$$

Let i and j be fixed. Then

$$|a_i z_i| = |a_j z_j|, \text{ by (3.22.2)},$$

$$|a_i|^p |z_i|^p = |a_j|^p |z_j|^p,$$

$$= |a_j|^p (1 - |z_i|^p), \text{ by (3.22.3)}.$$

Solving for $|z_i|$ we find that

$$|z_i| = \frac{|a_j|}{(|a_i|^p + |a_j|^p)^{1/p}} . \qquad (3.22.4)$$

Similarly,

$$|z_j| = \frac{|a_i|}{(|a_i|^p + |a_j|^p)^{1/p}} . \qquad (3.22.5)$$

Noticing in (3.22.1) that f_{ij} is maximized by taking $\operatorname{sgn} z_i = \operatorname{sgn} s_i$ and $\operatorname{sgn} z_j = \operatorname{sgn} s_j$, we conclude that for fixed i,j the maximum value of f_{ij} is

$$f_{ij} = \frac{|a_j s_i| + |a_i s_j|}{(|a_i|^p + |a_j|^p)^{1/p}} . \qquad (3.22.6)$$

And the corner point z for which $(z|s) = f_{ij}$ has coordinates

$$z_i = \frac{|a_j| \operatorname{sgn} s_i}{(|a_i|^p + |a_j|^p)^{1/p}} ,$$

$$z_j = \frac{|a_i| \operatorname{sgn} s_j}{(|a_i|^p + |a_j|^p)^{1/p}} ,$$

$$z_k = 0 , \quad k \neq i,j, \quad 1 \leq k \leq m.$$

We now compute the $\binom{m}{2}$ values f_{ij}. If $f_{\mu\nu}$ is the largest of the f_{ij}, then the solution of problem (P*) is

$$z = z_\mu e_\mu + z_\nu e_\nu .$$

3.23 Algorithm. In a similar manner, the solution of problem (P*) can be found directly when $n = m - 3$. Let $(a_1,\ldots,a_m)^T$, $(b_1,\ldots,b_m)^T \in (K \oplus (s))^{\perp}$ be linearly independent. Problem (P*) can be stated as

$$\text{Maximize } g_{ijk},\ 1 \le i,j,k \le m, i \ne j \ne k \ne i,$$

$$g_{ijk} = s_i z_i + s_j z_j + s_k z_k \tag{3.23.1}$$

$$\text{subject to } 0 = a_i z_i + a_j z_j + a_k z_k, \tag{3.23.2}$$

$$0 = b_i z_i + b_j z_j + b_k z_k, \tag{3.23.3}$$

$$1 = |z_i|^p + |z_j|^p + |z_k|^p. \tag{3.23.4}$$

Let i,j,k be fixed. Then (3.23.2) and (3.23.3) can be rewritten, after eliminating the z_k from (3.23.2) and the z_j term from (3.23.3), as

$$A_i z_i + A_j z_j = 0, \tag{3.23.5}$$

$$B_i z_i + B_k z_k = 0, \tag{3.23.6}$$

where $A_i = \begin{vmatrix} a_i & a_k \\ b_i & b_k \end{vmatrix}$, $A_j = \begin{vmatrix} a_j & a_k \\ b_j & b_k \end{vmatrix}$, $B_i = \begin{vmatrix} A_j & A_i \\ b_j & b_i \end{vmatrix}$,

and $B_k = b_k A_j$. We have

$$\begin{aligned} |A_i|^p |z_i|^p &= |A_j|^p |z_j|^p \text{ , by (3.23.5),} \\ |B_i|^p |z_i|^p &= |B_k|^p |z_k|^p \text{ , by (3.23.6).} \end{aligned} \tag{3.23.7}$$

Also

$$|A_jB_k|^p = |A_jB_k|^p|z_i|^p + |A_jB_k|^p|z_j|^p + |A_jB_k|^p|z_k|^p, \text{ by (3.23.4)},$$

$$= |A_jB_k|^p|z_i|^p + |A_iB_k|^p|z_i|^p + |A_jB_i|^p|z_i|^p, \text{ by (3.23.7)},$$

$$= (|A_jB_k|^p + |A_iB_k|^p + |A_jB_i|^p)|z_i|^p. \qquad (3.23.8)$$

Let $D = (|A_jB_k|^p + |A_iB_k|^p + |A_jB_i|^p)^{1/p}$, so that

$$|z_i| = \frac{|A_jB_k|}{D} \quad , \text{ by (3.23.8)}.$$

Similarly, one can show that $|z_j| = \frac{|A_iB_k|}{D}$ and $|z_k| = \frac{|A_jB_i|}{D}$.

From (3.23.1), it is clear that for i,j,k fixed, g_{ijk} is maximized by taking $\operatorname{sgn} z_\mu = \operatorname{sgn} s_\mu$, $\mu = i,j,k$, with the maximum value being

$$g_{ijk} = \frac{|A_jB_ks_i| + |A_iB_ks_j| + |A_jB_is_k|}{D} \quad .$$

The corner point z for which $(z|s) = g_{ijk}$ has coordinates

$$z_i = |A_jB_k| \operatorname{sgn} s_i/D \ ,$$

$$z_j = |A_iB_k| \operatorname{sgn} s_j/D \ ,$$

$$z_k = |A_jB_i| \operatorname{sgn} s_k/D \ ,$$

$$z_\ell = 0, \quad \ell \in \{1,\ldots,m\} \setminus \{i,j,k\}.$$

We now compute the $\binom{m}{3}$ values g_{ijk} corresponding to the values of the objective function at all of the corner points. If $g_{\lambda\mu\nu}$ is the largest of the g_{ijk}, then the solution of problem (P*) is

$$z = z_\lambda e_\lambda + z_\mu e_\mu + z_\nu e_\nu .$$

As these special algorithms indicate, the amount of work required to find the solution of problem (P*) by this technique increases rapidly as dim $K^\perp$ increases. One could, however, try the same general approach for small values of $n = \dim K$.

3.24 Algorithm. If K is one dimensional, say $K = (a)$ where $a = (a_1,\ldots,a_m)^T$, then z, the solution of problem (P*), by Corollary 3.9, has at least one coordinate equal to zero. We can also assume that $|a_i| + |s_i| \neq 0$ for $1 \leq i \leq m$ since otherwise the entire problem takes place in $\mathbb{R}^{m-k}$, $k > 0$. Problem (P*) can be stated as

$$\text{Maximize } f_i,\ 1 \leq i \leq m,$$

$$f_i = (\alpha_k a + \beta_i s | s), \tag{3.24.1}$$

$$\text{subject to } 0 = \alpha_i a_i + \beta_i s_i, \tag{3.24.2}$$

$$1 = \|\alpha_i a + \beta_i s\|_p . \tag{3.23.3}$$

If $s_i = 0$, then $\alpha_i = 0$ by (3.24.2), and hence $|\beta_j| = \|s\|_p^{-1/p}$ by (3.24.3). If $a_i = 0$, then $\beta_i = 0$ by (3.24.2), and hence $|\alpha_i| = \|a\|_p^{-1/p}$. If $a_i s_i \neq 0$, then $\beta_i = -\frac{a_i \alpha_i}{s_i}$ by (3.24.2), so that $1 = \|\alpha_i a - \frac{a_i \alpha_i}{s_i} s\|_p = |\alpha_i|^p \|a - \frac{a_i}{s_i} s\|_p$ by (3.24.3). Thus $|\alpha_i| = 1/\|a - \frac{a_i}{s_i} s\|_p^{1/p} = \frac{|s_i|}{\|as_i - a_i s\|_p^{1/p}}$. Similarly, if $a_i s_i \neq 0$, $|\beta_i| = \frac{|a_i|}{\|as_i - a_i s\|_p^{1/p}}$. In all three cases,

$$|\alpha_i| = \frac{|s_i|}{\|as_i - a_i s\|_p^{1/p}}, \text{ and } |\beta_i| = \frac{|a_i|}{\|as_i - a_i s\|_p^{1/p}} \qquad (3.24.4)$$

Since $(a|s) = 0$, $f_i = \beta_i(s|s)$. To maximize f_i, we will choose $\operatorname{sgn} \beta_i = 1$ whenever $\beta_i \neq 0$. Thus by (3.24.2) we conclude that $\operatorname{sgn} \alpha_i = -\operatorname{sgn}(s_i a_i)$ if $a_i \neq 0$, and of no interest if $a_i = 0$ since $\beta_i = 0$ then. Thus by (3.24.4)

$$\alpha_i = \frac{-s_i \operatorname{sgn} a_i}{\|as_i - a_i s\|_p^{1/p}}, \text{ and } \beta_i = \frac{|a_i|}{\|as_i - a_i s\|_p^{1/p}}.$$

By evaluating the m β_i's, we have essentially - up to a factor of $(s|s)$ - evaluated the objective function at all of the corner points. Consequently, if β_μ is the largest of the β_i, then the solution of problem (P*) is

$$z = \alpha_\mu a + \beta_\mu s.$$

Algorithm 3.20 and, to a lesser degree, the four special algorithms just described have two distinguishing features - one good and the other bad. On the one hand, they always work, i.e. they give the correct solution of problem (P*). On the other hand, algorithm 3.20 in particular can involve a tremendous amount of work since every corner point must be computed. Consequently, unless m and n are fairly small numbers or the Haar Condition is so flagrantly violated that the actual number of corner points is reasonably small, algorithm 3.20 does not represent a computationally feasible method for finding the solution of problem (P*).

3.25 Definition. Let x,y be corner points with $x_i = 0$, $i \in I$, $y_i = 0$, $i \in I'$, $x_j \neq 0$ $j \in J$, $y_j \neq 0$ $j \in J'$, and $I \cup J = I' \cup J' = \{1,\ldots,m\}$. We say that x and y are adjacent if $\{r_i | i \in I \cap I'\}$ contains $n-1$ linearly independent vectors.

The idea behind this definition is most readily seen if we assume that the Haar Condition holds since otherwise the idea can be easily lost among the subscripts. In this situation, x,y being adjacent (corner) points implies that I and I' both have exactly n elements and $\{r_i | i \in I \cap I'\}$ contains $n-1$ linearly independent row vectors, i.e. $I \cap I'$ contains exactly $n-1$ elements. Thus there is an $i \in I$ and $j \in I'$ such that $I = J \cup \{i\} \setminus \{j\}$ and $I' = I \cup \{j\} \setminus \{i\}$. In terms of coordinates, all but one of the zero coordinates of either x or y is also a zero coordinate of the other.

It should be noted that adjacent points can be much farther apart than one might expect a term like adjacent to allow. For example, if K is a one dimensional subspace, then each two corner points are adjacent. This follows immediately since corner points need only have one coordinate equal to zero.

3.26 Definition. A corner point z is a local solution of problem (P*) if $(z|s) \geq (x|s)$ for all x adjacent to z.

It follows from the definition of adjacent corner points that there can exist corner points which are not adjacent. Consequently, a local solution of problem (P*) need not be a solution of problem (P*).

We can now remedy the shortcoming of excessive computations found in algorithm 3.20 by presenting an exchange algorithm similar to that used in linear programming. The solution found in this manner may, however, only be a local solution of problem (P*). As always, we assume "the usual n dimensional situation".

3.27 Algorithm. Step 1. Find a corner point $z = D\beta$, and set $I = \{i \mid z_i = 0\}$.

Step 2. Select n linearly independent rows $r_{i_1}, \ldots, r_{i_n}$ of D with $i_1, \ldots, i_n \in I$.

Step 3. Pick $\mu \in \{i_1, \ldots, i_n\}$.

Step 4. Relabel the $r_{i_1}, \ldots, r_{i_n}$ as $\bar{\rho}_1, \ldots, \bar{\rho}_n$ with $\bar{\rho}_n = r_\mu$.

Step 5. Orthogonalize the $\bar{\rho}_j$ by (i) $\rho_1 = \bar{\rho}_1$, and (ii) for $i = 2, \ldots, n$ $\rho_i = \bar{\rho}_i - \sum_{j=1}^{i-1} \frac{(\bar{\rho}_j \mid \rho_i)}{(\rho_j \mid \rho_j)} \rho_j$.

Step 6. Pick $k \in \{1, \ldots, m\} \setminus I$.

Step 7. Set $\tilde{\beta} = \gamma_k \|D\gamma_k\|^{-1/p} \operatorname{sgn}(\gamma_k)_{n+1}$, where $\gamma_k = \rho_n - \frac{(\rho_n \mid r_k)}{(\beta \mid r_k)} \beta$.

We claim that $\tilde{z} = D\tilde{\beta}$ is a corner point. To verify this, we show that $\tilde{z}$ satisfies the definition of a corner point.

By construction, $\|\tilde{z}\|_p = 1$. Let $J = \{j \mid \tilde{z}_j = 0\}$ and $N = \{x \in \mathbb{R}^{n+1} \mid (x|r_j) = 0, j \in J\}$. We need to show that $\dim N = 1$. By construction, $\{r_i, \ldots, r_{i_m}, r_k\}$ are $n+1$ linearly independent vectors. $\{i_1, \ldots, i_n, k\} \setminus \{\mu\} \subset J$, so $\dim N \leq 1$. But $\dim N \geq 1$ since $\tilde{\beta} \in N \setminus \{0\}$. Thus $\tilde{z} = D\tilde{\beta}$ is a corner point.

Step 8. (i) If $\tilde{\beta}_{n+1} > \beta_{n+1}$, replace β by $\tilde{\beta}$ and z by $\tilde{z}$, and then return to step 2. Notice that we have n linearly independent rows on hand from above.

(ii) If $\tilde{\beta}_{n+1} \leq \beta_{n+1}$, return to step 6 and try another $k \in \{1, \ldots, m\} \setminus I$ until all of these have been tried.

(iii) When all of the $k \in \{1, \ldots, m\} \setminus I$ have been tried in (ii), return to step 3 to choose another $\mu \in \{i, \ldots, i_n\}$.

(iv) When all of the $\mu \in \{i, \ldots, i_n\}$ have been tried in (iii), z is a local solution of problem (P*) with value β_{n+1}.

Step 1 can be accomplished in the same manner that corner points were found in algorithm 3.20. Experience with a few examples seems to indicate that a good starting corner point to find in step 1 is that z which has zero coordinates where the coordinates of the vector s are the smallest in absolute value. In many cases, this corner point actually solves problem (P*).

Step 8(i) insures that algorithm 3.27 eventually terminates since the value of the objective function $\beta_{n+1} = (s|z)$ is non decreasing and there are only a finite number of corner points. Lemma 3.12 guarantees that in step 8(ii) we need only check the n coordinates listed rather than all of the zero coordinates. The assertion in step 8(iv) that the point z found by algorithm 3.27 is a local solution of problem (P*) follows directly from the definition of a local solution.

If $\dim (K \oplus (s))$ is close to $\dim \mathbb{R}^m$, then the computations involved in choosing $n+1$ linearly independent row vectors (step 2) and orthogonalizing them (step 5) can become tedious. In this case, it may be advantagous to exploit the fact that when n is nearly equal to m, then $\dim\{(K \oplus (s))^{\perp}\} = m-n-1$ is quite small. Before presenting such an algorithm with all of its details, a brief example clarifying the key ideas involved is sketched.

Assume "the usual n dimensional situation". First find $m-n-1$ linearly independent vectors $b_1,\ldots,b_{m-n-1} \in \mathbb{R}^m$ which span $(K \oplus (s))^{\perp}$, i.e. $D^T b_j = 0$ $j = 1,\ldots,m-n-1$. Letting $x_0 = (x|s)$, we can reformulate problem (P*) as:

maximize x_0 subject to the constraints

$$-x_0 + x_1 s_1 + x_2 s_2 + \cdots + x_m s_m = 0,$$

$$x_1 b_{1_j} + x_2 b_{2_j} + \cdots + x_m b_{m_j} = 0, \qquad j = 1,\ldots,m-n-1,$$

$$\|x\|_p = 1.$$

By relabelling indices if necessary, we can assume that the linear constraints can be put in the following simpler form:

$$
\begin{aligned}
-x_0 \qquad\qquad & + c_{m-n} x_{m-n} + \cdots + c_m x_m = 0, \\
x_1 \qquad\qquad & + a_{m-n,1} x_{m-n} + \cdots + a_{m,1} x_m = 0, \\
x_2 \qquad\qquad & + a_{m-n,2} x_{m-n} + \cdots + a_{m,2} x_m = 0, \\
\cdots \quad x_{m-n-1} & + a_{m-n,m-n-1} x_{m-n} + \cdots + a_{m,m-n-1} x_m = 0.
\end{aligned}
$$

If we set the n variables $x_{m-n+1} = \cdots = x_m = 0$, we would be left with

$$-x_0 + c_{m-n} x_{m-n} = 0,$$

$$x_j + a_{m-n,j} x_{m-n} = 0, \qquad j = 1, \ldots, m-n-1.$$

Substituting these restrictions on x_{m-n} into the equation $\|x\|_p = 1$, we find that $|x_{m-n}| = (1 + \sum_{j=1}^{m-n-1} |a_{m-n,j}|^p)^{-1/p}$, and thus we conclude that $|x_0| = |c_{m-n}|(1 + \sum_{j=1}^{m-n-1} |a_{m-n,j}|^p)^{-1/p}$. Similarly we could have set all of x_{m-n} through x_m equal to zero except x_{m-n+k}, $0 \le k \le n$, and found that $|x_0| = |c_{m-n+k}|(1 + \sum_{j=1}^{m-n-1} |a_{m-n+k,j}|^p)^{-1/p}$ in that case. Since we are interested in maximizing x_0, choose q to be that subscript which maximizes this last expression, and call the maximum value x_0^*. If we now interchange the roles of x_q and some x_i, $1 \le i \le m-n-1$, we can repeat the above steps to obtain another $\bar{x}_0^*$. If $\bar{x}_0^* \le x_0^*$, then the objective function has not been increased, so we try another $1 \le i \le m-n-1$ until all have been considered. When that occurs,

x is a local solution of problem (P*) with $(x|s) = x_0^*$. If $\bar{x}_0^* > x_0^*$, then set $x_0^* = \bar{x}_0^*$ and repeat the computations.

With this general scheme in mind, we present

3.28 Algorithm.

Step 1. Find $m-n-1$ linearly independent vectors $b_1,\ldots,b_{m-n-1} \in \mathbb{R}^m$ such that $D^T b_j = 0$, $j = 1,\ldots,m-n-1$.

Step 2. Select $I \subset \{1,\ldots,m\}$ containing $m-n-1$ indices $i_1,\ldots,i_{m-n-1}$, so that x_{i_j} can be eliminated from $-x_0 + (s|x) = 0$ and from all but the j^{th} equation of the system $(b_k|x) = 0$, $k = 1,\ldots,m-n-1$, in which x_{i_j} has coefficient 1. Call the resulting system $-x_0 + (c|x) = 0$,

$$(a_j|x) = 0, \quad j = 1,\ldots,m-m-1.$$

Step 3. Find $k \in \{1,\ldots,m\} \setminus I$ which maximizes

$$x_k = \left(1 + \sum_{j=1}^{m-n-1} |a_{k,j}|^p\right)^{-1/p}.$$

Step 4. Set $x^* = c_k x_k$.

Step 5. For each $i_j \in I$, form $I_j = I \cup \{k\} \setminus \{i_j\}$. Interchange the roles of x_{i_j} and x_k, i.e. eliminate x_k from $-x_0 + (c|x) = 0$ and from all of $(a_i|x) = 0$, $i = 1,\ldots,m-n-1$, except where $i = j$ in which x_k has

coefficient 1. Call the resulting system

$$-x_0 + (c(j)|x) = 0,$$

$$(a_i(j)|x) = 0, \quad i = 1,\ldots,m-n-1.$$

Step 6. Find $\mu \in \{1,\ldots,m\} \setminus I_j$ which minimizes

$$x_\mu(j) = (1 + \sum_{i=1}^{m-n-1} |a_{\mu,i}(j)|^p)^{-1/p},$$

and set $x_j^* = x_\mu(j)c_\mu(j)$.

Step 7. (i) If $|x_j^*| \leq |x^*|$ for all $i_j \in I$, then the vector x found in steps 2 and 3 is a local solution of problem (P*).

(ii) If any $|x_j^*| > |x^*|$, let x_q^* be the largest one in absolute value. Set $I = I_q$, $a_j = a_j(q)$, $j = 1,\ldots,m-n-1$, $c = c(q)$, $k = q$, and $x^* = x_q^*$. Return to step 5.

To see that algorithms 3.27 and 3.28 compute the same local solution of problem (P*), observe the following eqivalence of steps in the two algorithms:

Algorithm 3.28	Algorithm 3.27
Step 2	Steps 2,5
Step 3	Steps 6,7,8(ii)
Step 5	Step 8(iii)
Step 7(i)	Step 8(iv)
Step 7(ii)	Step 8(i)

Before leaving the subject of algorithms for solving the ℓ_p problem, $0 < p < 1$, one unsuccessful attempt to solve problem (P*) may be of interest. Rather than moving from vertex to vertex in an exchange algorithm, one might try to solve a series of problems in which the dimension of the subspace K changes by one at each step.

Since one can find the global solution if either $K \oplus (s) = \mathbb{R}^m$ or $\dim(K \oplus (s)) = 2$, hopefully one could start with the solution to one of those two problems, and by successively decreasing or increasing the dimension of K by 1, obtain the solution of the given n dimensional problem. Any such algorithm would depend upon getting some information about the solution of an $n \pm 1$ dimensional problem from a known solution of a related n dimensional problem. In some sense, the solutions should not be very far apart.

Assume K, $\bar{K}$ satisfy "the usual $n, n+1$ dimensional situations" respectively with K a subspace of $\bar{K}$, that $z, \bar{z}$ solve problem (P*) for $K, \bar{K}$, that $z_i = 0$, $i \in I$, $\bar{z}_i = 0$, $i \in \bar{I}$, $z_j \neq 0$, $j \in J$, $z_j \neq 0$, $j \in \bar{J}$, and $I \cup J = \bar{I} \cup \bar{J} = \{1, \ldots, m\}$. We can also assume that the Haar Condition holds in both cases so that $|I| = n$, $|\bar{I}| = n+1$, $|J| = m-n$, $|\bar{J}| = m-n-1$. Under these conditions, it was conjectured that $I \subset \bar{I}$, or equivalently, $\bar{J} \subset J$. In $\mathbb{R}^3$ this means that the solution of the two dimensional ℓ_p problem must lie on one of the four edges of the unit ball passing through the vertex that was the solution of the three dimensional problem. Alternatively, the solution of the three dimensional problem is one

of the two end points of the edge on which the solution of the two dimensional ℓ_p problem lies.

Unfortunately, this conjecture is not true even in $\mathbb{R}^3$ for either $0 < p < 1$ or $p = 1$.

3.29 Counterexample. Let $p = 1/2$, $s = \begin{bmatrix} 1 \\ 1.01 \\ 1 \end{bmatrix}$, and $K = (s)^{\perp}$. By corollary 3.11, e_2 solves problem (P*). But if K is taken to be (v), where $v = \begin{bmatrix} .97 \\ 1.01 \\ .02 \end{bmatrix}$, the solution of problem (P*) is approximately $\begin{bmatrix} .009 \\ 0 \\ .9 \end{bmatrix}$, which is on an edge that does not have e_2 as either of its end points.

3.30 Counterexample. Let $p = 1$, $s = \frac{1}{25}\begin{bmatrix} 11 \\ 10 \\ 4 \end{bmatrix}$, and $K = (s)^{\perp}$. Corollary 3.11 is true if $p = 1$ also, so e_1 solves this three dimensional problem. But if $K = (v)$, where $v = \begin{bmatrix} -110 + 66\epsilon \\ 137 - 177\epsilon \\ -40 + 261\epsilon \end{bmatrix}$, with $0 < \epsilon < \frac{3}{150}$, then the solution of the two dimensional problem is $\begin{bmatrix} 0 \\ 1 - \epsilon \\ \epsilon \end{bmatrix}$, which is on an edge that does not have e_1 as either of its end points.

We conclude this section by pointing out that many of the results obtained for $0 < p < 1$ also hold for the case where $p = 1$, thus providing a method of solving the ℓ_1 problem. With $p = 1$, Theorem 3.1 is a special case of Theorem 2.4 in [21]. If we alter Lemma 3.4 so that (iii) reads $H + z$ properly supports B at z, i.e. either $f(x) < f(z)$ for all $x \in B \setminus \{z\}$ or $f(x) > f(z)$ for all $x \in B \setminus \{z\}$, then Lemma 3.4 is true for $p = 1$. In the proof,

choose x, δ, $\bar{\delta}$, and ϵ as before, but obtain a contradiction either to the hypothesis that $H + z$ supports B at z or to the strictness of the support. Theorem 3.5 then follows immediately for $p = 1$ if we again assume that $H + z$ properly supports B at z. Theorem 3.5 and all three of its corollaries, Lemma 3.12 and Theorem 3.13 all hold as previously stated for $p = 1$.

Algorithms 3.27 and 3.28 both solve problem (P*) when $p = 1$ since the ℓ_1 unit ball being convex eliminates the possibility of finding a local solution that is not a global solution of problem (P*). Consequently, in the ℓ_1 case, algorithm 3.20 is unnecessary since all three algorithms find the same solution while algorithm 3.20 requires much more computational effort to do it.

CHAPTER IV

NUMERICAL RESULTS

In this chapter we consider some of the computational limitations placed upon the algorithms presented in chapters II and III. Since the main application of algorithm 2.2 is to the ℓ_p spaces, $1 < p < \infty$, we discuss the computational difficulties encountered in that context, and then present numerical results from two examples. Examples of ℓ_p approximation, $0 < p \leq 1$, conclude the chapter.

As we noted in chapter II, in the case of the ℓ_p spaces, $1 < p < \infty$, if $y \neq 0$, then

$$y_j^* = \frac{|y_j|^{p-1}}{\|y\|_p^{p-1}} \operatorname{sgn} y_j .$$

This particularly simple form for y^* makes the evaluation of y_k^* in step 2 and $(y_k - \alpha_k v_k)^*$ in step 4 of algorithm 2.2 immediate.

The two main computational difficulties occurring in algorithm 2.2 are finding α_k such that $((y_k - \alpha_k v_k)^* | v_k) = 0$ and solving the original problem (P) once the solution of problem (P*) has been found. Since values obtained on a computer are seldom exactly correct, each of these difficulties

brings up a related one also. How close to α_k is close enough, and how close to $v_k = 0$ is close enough to call y_k the solution of problem (P*)? Since the questions related to knowing when problem (P*) has been solved are the easier, we dispose of those first.

By Theorem 2.4 of [21], if y solves problem (P*), then

$$z = b - \frac{(s|s)}{(s|y)} y$$

is in the image of A, and $x \in \mathbb{R}^n$ such that $Ax = z$ solves problem (P).

Let the rank of A be k. Form the $m \times k$ matrix M composed of the k linearly independent columns $a_{i_1},\ldots,a_{i_k}$ of A. Then

$$w = (M^T M)^{-1} M^T z$$

satisfies

$$Mw = z.$$

Let $x \in \mathbb{R}^m$ be given by

$$x_i = \begin{cases} w_i & \text{if} \quad i \in \{i_1,\ldots,i_k\} \\ 0 & \text{if} \quad i \notin \{i_1,\ldots,i_k\} \end{cases},$$

and let $y \in \mathbb{R}^m$ be any solution of

$$Ay = 0.$$

Then $u = x + y$ satisfies

$$Au = z,$$

and hence u solves problem (P). If the rank of A is n, then the solution of problem (P) is unique, and in general, the solutions of problem (P) are the translates of an n-rank (A) dimensional subspace of $\mathbb{R}^m$.

The answer to the question of how small v_k must be to accept y_k as the solution of problem (P*) is somewhat dependent upon the matrix A. Since

$$y_{k+1} = \frac{y_k - \alpha_k v_k}{\|y_k - \alpha_k v_k\|},$$

$$\lim_{k \to \infty} \|y_k - \alpha_k v_k\| = 1, \text{ and}$$

$$\|y_k\| = 1 \quad \text{for all} \quad k \geq 0,$$

$|(\alpha_k v_k)_j|$ is almost exactly how much the j^{th} coordinate of y_k is changing at each iteration. Consequently, in our examples the algorithm terminated if

$$\max\{|(\alpha_k v_k)_j| \mid 1 \leq j \leq m\} < 10^{-7},$$

and our computed solutions agreed with published solutions for the same problems.

The related problems of how to compute and how accurately to compute the α_k in step 4 pose the greatest problem in algorithm 2.2. Where no confusion arises, we drop the subscript k since the iteration being considered is usually

irrelevant. We assume throughout the discussion that y,v are linearly independent since the problem is trivial otherwise.

Let $f(\alpha) = ((y - \alpha v)^* | v)$

$$= \frac{1}{\|y - \alpha v\|_p^{p-1}} \sum_{j=1}^{m} v_j |y - \alpha v_j|^{p-1} \operatorname{sgn}(y_j - \alpha v_j),$$

and define

$$g(\alpha) = \sum_{j=1}^{m} v_j |y_j - \alpha v_j|^{p-1} \operatorname{sgn}(y_j - \alpha v_j).$$

Clearly, f and g have the same roots so we consider only the simpler function g. By Proposition 2.5, g has a unique root α and $\alpha > 0$. Moreover, $g(0) > 0$ since

$$f(0) = (y^* | v) > 0$$

by (2.4.2) and (2.4.3). g is clearly continuous since $1 < p < \infty$.

With this information, a number of techniques for finding the root α of g are available. The following four methods were used.

1. <u>The method of bisection</u>.

Choose $\alpha_1 > 0$ arbitrarily and compute $g(\alpha_1)$. Since $g(0) > 0$ and the root α is unique,

$$\text{if } g(\alpha_1) > 0, \quad \text{set } \alpha_2 = 2\alpha_1 ,$$

$$\text{if } g(\alpha_1) < 0, \quad \text{set } \alpha_2 = \tfrac{1}{2}\alpha_1 .$$

Of course, if $g(\alpha_1) = 0$, then we are finished. Repeat this procedure until two numbers, say α^+ and α^-, have been found such that

$$g(\alpha^-) < 0 < g(\alpha^+).$$

Then let

$$\alpha = \frac{\alpha^+ + \alpha^-}{2} .$$

If $g(\alpha)$ is positive (negative), replace α^+ (α^-) by α, and repeat this bisection of the interval $[\alpha^+, \alpha^-]$ until the root of g is obtained.

2. The secant method.

Begin as in the bisection method by finding α^+, α^- such that

$$g(\alpha^-) < 0 < g(\alpha^+).$$

Instead of taking the midpoint of the interval $[\alpha^+, \alpha^-]$, find the point at which the line through the points $(\alpha^+, g(\alpha^+))$ and $(\alpha^-, g(\alpha^-))$ crosses the x-axis, i.e.

$$\alpha = \frac{\alpha^- g(\alpha^+) - \alpha^+ g(\alpha^-)}{\alpha^+ - \alpha^-} .$$

As above, if $g(\alpha)$ is positive (negative), replace α^+ (α^-) by α and repeat until the root of g is found.

3. The secant-bisection method.

After finding α^+, α^- as described above, alternate one iteration of the bisection method and one step of the secant method. This is similar to Dekker's algorithm.

4. Newton's method.

Observe that g is a differentiable function of α except at those α_j such that

$$y_i = \alpha_j v_j , \quad j = 1,\ldots,m.$$

After checking if any of these α_j is the desired root, one can apply Newton's method to find α such that $g(\alpha) = 0$. Choose $\alpha_1 > 0$ arbitrarily, and compute

$$\alpha_{n+1} = \alpha_n - \frac{g(\alpha_n)}{g'(\alpha_n)} , \quad n = 1,2,\ldots .$$

It should be noted that some care must be taken with the use of Newton's method. In our examples, it often failed to locate α for want of a sufficiently good initial guess. For p small, both the bisection method and the secant method had trouble converging to the root in some instances. It occasionally happened that one but not the other had difficulties handling a specific situation. The mixed secant-bisection method worked quite successfully in these cases enjoying the benefits of each while avoiding many of their shortcomings.

For values of p roughly between 1.25 and 200, few difficulties arise with any of these four methods for locating α. For small or very large values of p, however, g can become somewhat unruly.

While in theory algorithm 2.2 always solves problem (P), one should expect a little less in practice. Recall that

$$g(\alpha) = \sum_{j=1}^{m} v_j |y_j - \alpha v_j|^{p-1} \operatorname{sgn}(y_j - \alpha v_j).$$

The cause of our difficulties is the exponent $p-1$. If p is near 1, then

$$|y_j - \alpha v_j|^{p-1} \operatorname{sgn}(y_j - \alpha v_j) \simeq \operatorname{sgn}(y_j - \alpha v_j)$$

independent of the magnitude of $|y_j - \alpha v_j|$. A number that is supposed to be zero, but because of roundoff errors is actually 10^{-15} on the computer, will be greater than .7 after being raised to the $p-1$ power when $p = 1.01$, and approximately .966 when $p = 1.001$. Similarly, when p is very large,

$$|y_j - \alpha v_j|^{p-1} \simeq \begin{cases} 0 & \text{if } |y_j - \alpha v_j| < 1 \\ 1 & \text{if } |y_j - \alpha v_j| = 1 \\ \infty & \text{if } |y_j - \alpha v_j| > 1 . \end{cases}$$

The point of these comments is that one should not expect algorithm 2.2 to be computationally feasible throughout $1 < p < \infty$.

With p near 1, the thirty-two figures of double precision machine accuracy was sometimes not sufficient to determine α such that $|g(\alpha)| < 10^{-5}$. The question of how accurately one must know α becomes of interest at this

point. The question will be taken up when the actual examples are discussed.

A second unpleasant feature of the function g must also be considered when p is near 1. We know that y_k converges to some y and that v_k converges to 0, but the behavior of α_k is not known from the theory. Numerical results indicate that the sequence $\{\alpha_k | k \geq 0\}$ has two limit points with $\{\alpha_{2n}\}$ and $\{\alpha_{2n+1}\}$ approaching two numbers that sometimes differ considerably. For p near 1, one of the two limits appears to be 0 making those $\alpha_k v_k$ go to 0 rapidly while the other $\alpha_k v_k$ approach 0 more slowly.

At the other end of the range $1 < p < \infty$, our program came to a halt not because the algorithm was sensitive to large values of p, which it is, but because the computer can not store exponents that are too large. More to the point is that we could not compute the ℓ_p norm when p grew too large.

Two examples were programmed in FORTRAN IV for a CDC 6500. The first is taken from Barrodale and Young [4]. The linear system to be solved is

$$
\begin{aligned}
x &= 1.52 \\
x + y &= 1.025 \\
x + 2y &= 0.475 \\
x + 3y &= 0.01 \\
x + 4y &= -.475 \\
x + 5y &= -1.005
\end{aligned}
$$

P	x	y	ρ	Iterations
∞	1.5	-.5	.025	
128	1.499972	-.499955	.0251982	26
64	1.499944	-.499909	.0253977	21
32	1.499890	-.499817	.025801	21
16	1.499894	-.499671	.0266422	21
8	1.500651	-.499637	.0285313	23
4	1.503757	-.500057	.032874	53
2	1.514762	-.502571	.043216	1
1.5	1.520005	-.503800	.050790	30
1.4	1.520126	-.5037866	.0532758	45
1.3	1.520215	-.503744	.056436	47
1.2	1.520187	-.5036206	.060543	37
1.1	1.520037	-.5033896	.0659792	27
1.04	1.5200001	-.503333	.0701003	364
1	1.52	-.503	.073	

Table 4.1

The best ℓ_p approximate solution was computed for

$$p = 1.04, 1.06, \ldots, 1.5, 2, 4, 8, \ldots, 128.$$

The search for α_k was terminated when $|g(\alpha)| < 10^{-6}$ or after 32 iterations of the mixed bisection-secant method whichever occurred first. All of the solutions together were computed in less than forty seconds. Some of the solutions are listed in Table 4.1.

The second example appears in Cheney [6, p.44]. The overdetermined system of linear equations is

$$\begin{aligned} x + y &= 3 \\ x - y &= 1 \\ x + 2y &= 7 \\ 2y + 4y &= 11.1 \\ 2x + y &= 6.9 \\ 3x + y &= 7.2 \end{aligned}$$

This system poses special difficulties because the solution of the ℓ_1 problem is not unique. All points on the segment joining

$$P_1 = (1.77, 1.89) \quad \text{and} \quad P_2 = (2.516667, 1.516667)$$

solve the ℓ_1 problem with a minimal ℓ_1 error vector of length 4.7.

The ℓ_p problem was solved for

$$p = 1.06, 1.08, \ldots, 1.5, 2, 4, 6, 20, 40, 100, 400.$$

P	x	y	Iterations
1.5	2.0883483	1.7400827	15
1.46	2.0889511	1.7365094	16
1.42	2.0893571	1.7337896	16
1.38	2.0895666	1.7319498	14
1.34	2.0896032	1.7309038	10
1.3	2.0895121	1.7304580	8
1.26	2.0893464	1.7303686	34
1.22	2.0891500	1.7304290	22
1.18	2.088926	1.7305371	11
1.14	2.0884417	1.7307791	13
1.10	2.0880841	1.7309614	13
1.06	2.0872254	1.7313904	23

Table 4.2

The computed results for $p \geq 1.5$ agree with those published by Cheney [6] and Duris [9]. For $1.06 \leq p < 1.5$, no published figures are available, but as p decreases (see Table 4.2), the solution of the ℓ_p problems approximately equals a solution of the ℓ_1 problem.

For $1 < p < 1.06$, algorithm 2.2 consistently obtained solutions about

$$Q = (2.04615, 1.75193),$$

which is quite far from the computed limit [1] of the solutions of the ℓ_p problems as $p \to 1$, (2.0883,1.7309). It should be noted, however, that Q is an ℓ_1 solution of the problem since

$$Q = \lambda P_1 + (1 - \lambda) P_2, \text{ where } \lambda = .369839.$$

The problem of determining how accurately one must know α in step 2 was particularly troublesome with this example. The smaller $|g(\alpha)|$ is forced, the longer the algorithm takes and the greater is the possibility that the computer is not capable of locating the desired α. To solve the $\ell_{1.06}$ problem, it was necessary to know α such that $|g(\alpha)| < 10^{-24}$ in each iteration, and the computations took 28.5 seconds. The inability of the computer to store numbers exactly and the presence of many solutions of the ℓ_1 problem near the unique solution of the ℓ_p problem apparently teamed up to render algorithm 2.2 ineffective for values of p less than about 1.06.

On the topic of ℓ_p approximation when $0 < p \leq 1$, algorithm 3.20 was programmed in FORTRAN IV for a CDC 6500, and two examples were studied.

The example taken from Cheney [6, p.44] that we discussed earlier was tested with

$$p = \frac{n}{10}, \quad n = 1,2,\ldots,10.$$

For $p = 1$, the algorithm found both corner point solutions mentioned previously. All points on the line segment joining those two points are also solutions of the ℓ_1 problem. For $0 < p < 1$, however, the original problem has the unique solution $(2.516667, 1.516667)$. For each case, the algorithm took less than a second to compute the solutions.

For a second example, we chose K to be the subspace spanned by the single vector $\begin{bmatrix} 5 \\ 1 \\ -5 \end{bmatrix}$, and took b to be $\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$. The intersection of $K \oplus (s)$, where $s = b$ since $b \in K^{\perp}$, and the three dimensional ℓ_p unit ball for $p = \frac{2}{10}, \frac{3}{10}, \ldots, \frac{10}{10}$ are shown in Figure 2(a) of Chapter III. s runs diagonally from the upper left to the lower right passing through the corner points shown. As Table 4.3 indicates, the solution of problem (P*) "jumps" from the corner along s to the ones on either side for $p \simeq .777$ as the picture indicates.

p	β			y_1	y_2	y_3	ρ
1	.4545	.5000	.4545	.5000	.0000	.5000	2.000
.9	.4383	.4629	.4383	.4629	0	.4629	2.000
.8	.4160	.4204	.4160	.4204	0	.4204	2.000
.78	.4107	.4112	.4107	.4112	0	.4112	2.000
.775	.4093	.4089	.4093	0	-.08185	.8185	1.998
				.8185	-.08185	0	1.998
.7	.3856	.3715	.3856	0	-.07711	.7711	1.949
				.7711	-.07711	0	1.949
.6	.3442	.3150	.3442	0	-.06883	.6883	1.896
				.6883	-.06883	0	1.896
.5	.2886	.2500	.2886	0	-.05772	.5772	1.861
				.5772	-.05772	0	1.861
.4	.2163	.1768	.2163	0	-.04327	.4327	1.845
				.4327	-.04327	0	1.845
.3	.1291	.0992	.1291	0	-.02582	.2582	1.848
				.2582	-.02582	0	1.848
.2	.0433	.0313	.0433	0	-.008665	.08665	1.873
				.08665	-.008665	0	1.873
.1	.0014	.0010	.0014	0	-.0002891	.002891	1.923
				.002891	-.0002891	0	1.923

Table 4.3

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Abdelmalek, N.H., Linear L_1 approximation for a discrete point set and L_1 solutions for overdetermined linear equations, J. ACM 18 (1971), 41-47.

[2] Abdelmalek, N.H., On the discrete linear L_1 approximation and L_1 solutions of overdetermined linear equations, J. Approximation Theory 11 (1974), 38-53.

[3] Anton, H. and Duris, C.S., An algorithm for best approximate solutions to $Av = b$ in normed linear spaces, J. Approximation Theory 8 (1973), 133-141.

[4] Barrodale, I. and Young, A., Algorithms for best L_1 and L_∞ linear approximations on a discrete set, Numer. Math. 8 (1966), 295-306.

[5] Barrodale, I. and Roberts, F.D.K., An improved algorithm for discrete ℓ_1 linear approximation, SIAM J. Numer. Anal. 10 (1973), 839-848.

[6] Cheney, E.W., Introduction to Approximation Theory, McGraw Hill, New York, N.Y. 1966.

[7] Cline, A.K., Rate of convergence of Lawson's algorithm, Math. Comp. 26 (1972), 167-176.

[8] Descloux, J., Approximations in L^p and Chebyshev approximations, J. Soc. Indust. Appl. Math. 11 (1963), 1017-1026.

[9] Duris, C.S., An algorithm for solving overdetermined linear equations in the ℓ^p-sense, Mathematical Report 70-10, Drexel University, 1970.

[10] Duris, C.S. and Sreedharan, V.P., Chebyshev and ℓ^1-solutions of linear equations using least squares solutions, SIAM J. Numer. Anal. 5 (1968), 491-505.

[11] Fletcher, R., Grant, J.A., and Hebden, M.D., The calculation of linear L_p approximations, Comput. J. 14 (1971), 276-279.

[12] Fletcher, R., Grant, J.A., and Hebden, M.D., Linear minimax approximation as the limit of best L_p-approximation, SIAM J. Numer. Anal. 11 (1974), 123-136.

[13] Hoel, P.G., Certain problems in the theory of closest approximation, Amer. J. Math. 57 (1935), 891-901.

[14] Lawson, C.L., Contributions to the Theory of Linear Least Maximum Approximation, Ph.D. Thesis, University of California, Los Angeles, 1961.

[15] Motzkin, T.S. and Walsh, J.L., Least p^{th} power polynomials on a real finite point set, Trans. AMS 78 (1955), 67-81.

[16] Motzkin, T.S. and Walsh, J.L., Polynomials of best approximations on a real finite point set. I, Trans. AMS 91 (1959), 231-245.

[17] Motzkin, T.S. and Walsh, J.L., Polynomials of best approximation on an interval, Proc. N.A.S. 45 (1959), 1523-1528.

[18] Rice, J.R. and Usow, K.H., The Lawson algorithm and extensions, Math. Comp. 22 (1968), 118-127.

[19] Sreedharan, V.P., Solutions of overdetermined linear equations which minimize error in an abstract norm, Numer. Math. 13 (1969), 146-151.

[20] Sreedharan, V.P., Least squares algorithms for finding solutions of overdetermined linear equations which minimize error in an abstract norm, Numer, Math. 17 (1971), 387-401.

[21] Sreedharan, V.P., Least squares algorithms for finding solutions of overdetermined systems of linear equations which minimize error in a smooth strictly convex norm, J. Approximation Theory 8 (1973), 46-61.

[22] Stiefel, E., Note on Jordan elimination, linear programming, and Tchebycheff approximation, Numer. Math. 2 (1960), 1-17.