UNIVERSITY OF WATERLOO
COMPUTER SCIENCE DEPARTMENT

An Approach to Nonlinear l_∞ Approximation

Andrew R. Conn
Yuying Li

Research Report
CS-89-21

December, 1989

UNIVERSITY OF WATERLOO
COMPUTER SCIENCE DEPARTMENT

AN APPROACH TO NONLINEAR l_∞ APPROXIMATION

ANDREW R. CONN* AND YUYING LI†

Abstract. Recently we have presented a new approach to nonlinear l_∞ approximation that directly exploits generalisations of the characterisation for the classical best linear Chebyshev approximation.

We are able to produce an algorithm that has the ability to recognise the correct active set more rapidly than the more usual nonlinear programming approaches, which are based on equality quadratic programming methods, while avoiding the inefficiencies typically associated with the several inner iterations normally required by an inequality quadratic programming approach.

In addition to summarising the method, we present details of the line search technique, show that certain degenerate problems give rise to a least squares problem with nonnegativity constraints and include certain technical details, required for example, to avoid the Maratos effect. All the proofs of the theorems are omitted to emphasize the main ideas of the algorithm, their proper references are indicated however.

Key Words. nonlinear Chebyshev approximation

AMS(MOS) subject classifications. 41A50, 65D99, 65F20, 65K05

1. Introduction. The underlying problem we wish to consider is to minimize over $\mathbf{R}^n$ the non-smooth function $\psi(x)$ given by the maximum over a finite set, $M = \{1, 2, \cdots, m\}$, of functions $f_i(x) : \mathbf{R}^n \to \mathbf{R}$.

$$\min_{x \in \mathbf{R}^n} \max_{i \in M} f_i(x). \tag{1}$$

In this paper, we concentrate on a special case which is the discrete Chebyshev problem, where $\psi(x)$ is given by

$$\psi(x) = \max_{i \in M} |f_i(x)|.$$

Such problems may have arisen from a discrete approximation to the continuous problem

$$\psi(x) = \max_{t \in T} |f(x, t)|,$$

where T is a compact set.

In any case, the discrete Chebyshev problem,

$$\min_{x \in \mathbf{R}^n} \max_{i \in M} |f_i(x)|, \tag{2}$$

* Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. The research of this author was supported in part by NSERC grant A8639.

† Computer Science Department, Cornell University, Upson Hall, Ithaca, NY, 14853. The research of this author was partially supported by the U.S. Army Research Office through the Mathematical Science Institute, Cornell University.

is the problem of interest in the present article. We are content with finding a *local* minimum of $\psi(x)$ and we assume that the $f_i(x), i \in M$, are twice continuously differentiable.

Most current approaches are based upon the fact that (2) can be transformed into a nonlinear programming problem by adding a single new variable viz.

$$\begin{array}{lll} \min_{(x,z)\in\mathbf{R}^{n+1}} & z & \\ \text{subject to} & z - f_i(x) & \geq 0 \\ & z + f_i(x) & \geq 0 \\ & z & \geq 0. \end{array}$$

Although the structure of this formulation can be exploited to some extent, we are more interested in exploiting directly the structure of an optimal solution to the discrete Chebyshev problem. We are motivated to pursue this latter approach by virtue of the fact that classical Chebyshev theory is able to **characterise** such solutions in the case of continuous linear problems, under certain regularity conditions, and we are able to exploit such a characterisation very successfully in practise (see for example, Barrodale and Phillips [1] and Bartels, Conn and Li [2]).

The basic difficulty is that, in the linear case, there exists a *global* characterisation which is easy for computational exploitation, whereas in the nonlinear case this is not, in general, possible.

In effect, we shall base our algorithm upon *local* attempts at characterisation, which, in the limit, will give the correct characterisation at the solution.

If we consider the one dimensional continuous linear Chebyshev problem to approximate $y(t)$ on the interval $[\alpha, \beta]$ given by

$$\min_{x\in D} \max_{t\in[\alpha,\beta]} |\sum_{i=1}^{n} x_i\phi_i(t) - y(t)|,$$

where $D \subset \mathbf{R}^n$ is a compact set and the ϕ_i's are the 'basis functions' for our approximating set, then we determine an approximation to this continuous problem in t by discretising the interval $[\alpha, \beta]$ into m points, say

$$\alpha = t_1 < t_2 < t_3 \ldots < t_m = \beta.$$

The classical theory gives us the following explicit characterisation (see for example, [17], page 77).

THEOREM 1.1 (CHARACTERISATION THEOREM).

Let $\mathcal{L}$ be an n dimensional linear function subspace of $C[\alpha,\beta]$ that satisfies the Haar condition and let $y(t)$ be a continuous function on $[\alpha,\beta]$. Then $\phi^*(t) \stackrel{\text{def}}{=} \phi(x^*,t)$ is the best minimax approximation from $\mathcal{L}$ to $y(t)$ if and only if there exist $n+1$ points $\{t_i\}_{i=0}^n$ such that the conditions:

$$\alpha \leq t_0 < t_1 < \cdots < t_n \leq \beta$$

and

$$|y(t_i) - \phi^*(t_i)| = \|y(t) - \phi^*(t)\|_\infty, \tag{3}$$

and

$$y(t_{i+1}) - \phi^*(t_{i+1}) = -(y(t_i) - \phi^*(t_i)), \tag{4}$$

are satisfied. Such a set of points $\{t_i\}_{i=0}^n$ is often called an **alternant** *of $\phi^*(t)$.*

There is also an equivalent algebraic characterisation. The following theorem can be found in [17], page 98.

THEOREM 1.2. *Let $\mathcal{L}$ be an n dimensional linear function subspace of $C[\alpha,\beta]$ that satisfies the Haar condition. Furthermore, let $\{t_i\}_{i=0}^n$ be a set of reference points from $[\alpha,\beta]$ that are in ascending order:*

$$\alpha \leq t_0 < t_1 < \cdots < t_n \leq \beta$$

and let $\{\lambda_i\}_{i=0}^n$ be a set of real multipliers that are not all zeroes, and that satisfy

$$\sum_{i=0}^{n} \lambda_i \phi(x,t_i) = 0 \tag{5}$$

for all functions $\phi(x,t) = \sum_{j=1}^n x_j \phi_j(t)$ (i.e. $\phi \in \mathcal{L}$ with basis functions $\phi_j(t)$). Then every multiplier is nonzero, and their signs alternate.

Equation (5) is called the **characteristic equation.**

Given the characteristic equation (5), with the associated multipliers $\{\lambda_i\}_{i=0}^n$, suppose we are approximating the continuous function $y(t)$ by $\phi(x,t)$, with associated error $f(t) = y(t) - \phi(x,t)$, then we have the following definitions.

DEFINITION 1. *The function $\phi(x,t)$ is called a* **reference function** *with respect to the reference $\{t_i\}_{i=0}^n$ and the function $y(t)$ if and only if:*

$$sgn(f_i) = sgn(\lambda_i) \quad \text{for all } i,$$

or

$$sgn(f_i) = -sgn(\lambda_i) \quad \text{for all } i,$$

where $\{\lambda_i\}_{i=0}^n$ *is given by the characteristic equation (5).*

If in addition all the f_i*'s have the same magnitude, called the* **reference deviation**, $\phi(x,t)$ *is a* **levelled reference function**.

Thus, in the linear case we have the following equivalent characterisation.

THEOREM 1.3. *Under the same assumptions of Theorem 1.2, the function of best approximation is the levelled reference function with the maximal reference deviation.*

If we return to algorithms for the linear problem there are two main approaches — dual algorithms, for example Barrodale and Phillips [1] and primal methods, for example Bartels, Conn and Li [2].

The former chooses $n+1$ points, $\{t_i\}_{i=0}^n$ (a reference) and an x^c such that (4) is satisfied. If (3) also holds we are optimal. Otherwise, it is possible to choose a t_j such that the error (value of $|f_j(x^c)|$) is greater than the errors on the reference. We then replace a t_i of the reference by the t_j and iterate.

In contrast, a primal algorithm chooses $n+1$ points and an x^c such that one has $n+1$ activities. If alternation is satisfied, one is optimal. Otherwise, it is possible to find a new x^+ such that n of the residuals that determine the $n+1$ activities at x^c remain active but the $n+1^{st}$ residual is less than the maximum residual $\|y(t)-\phi(x^+,t)\|_\infty$. One can then proceed in a direction that maintains the n activities until a new t_j is determined such that $\|y(t)-\phi(x,t)\|_\infty = |y(t_j)-\phi(x,t_j)|$, thus once again satisfying (3), but with a lower maximum absolute residual.

Thus one might remark that dual methods emphasize the alternating sign property whereas primal method emphasize $n+1$ maximal residuals. When both hold optimality is reached.

Some attempts to generalise the concept of alternant to the nonlinear case have been made (see for example Motzkin [15], Rice [20] and Tornheim [21]), but the results are rather restrictive and difficult to exploit computationally since they depend upon properties that are either not possible to predict a priori or, if they do hold globally, are too strong and are rarely satisfied except for very special cases (for example, linear problems under the Haar condition is one useful instance).

This explains why most techniques for nonlinear discrete Chebshev approximation are based upon the nonlinear programming formulation. We wish to do otherwise.

Now, the characteristic equation (5) can be rewritten as

$$\sum_{j=0}^{n} \lambda_i a_i = 0, \tag{6}$$

where $a_i = [\phi_1(t_i), \cdots, \phi_n(t_i)]^T$. This in turn can be rewritten as

$$\sum_{j=0}^{n} \lambda_j \nabla\phi(x, t_j) = 0, \tag{7}$$

which is independent of x in the linear case.

The point is that this form, although then dependent upon x, can be generalised to the nonlinear case.

2. General Theory. First we require some additional definitions.

DEFINITION 2 ([14]). *At any point x_0, the* **linear gradient space** *$J(t)$ of $\phi(x,t)$ refers to*

$$J(t) = span\{\frac{\partial\phi(x_0,t)}{\partial x_1}, \cdots, \frac{\partial\phi(x_0,t)}{\partial x_n}\}$$

The dimension of this linear function space defined on $t \in [\alpha, \beta]$ is denoted by $d(x_0)$.

For a linear space $\mathcal{L}$ with the classical Haar condition, $d(x) = n$, for all $x \in \mathbf{R}^n$.

In the discrete case, instead of considering the whole gradient function space $J(t)$, we consider the set of vectors corresponding to the columns of the Jacobian matrix

$$J(t_1, t_2, \cdots, t_m) = [\nabla f_1 \cdots, \nabla f_m].$$

Firstly, we remark that $\nabla f_i(x)$ is equivalent to $\nabla\phi(x, t_i)$ and note that our linear characteristic equation (5) can be written as

$$\sum_{j=0}^{n} \lambda_j \nabla f_j(x) = 0. \tag{8}$$

Note: the Haar condition corresponds to any $n \times n$ submatrix of J being non-singular. Thus we are led to consider 'minimal' such sets, via the following important concept.

DEFINITION 3. *The vector set $C = \{\nabla f_{i_0}, \cdots, \nabla f_{i_l}\}$, where the gradients are evaluated at a given fixed point x, is called a* **cadre** *if and only if:*

1. *rank$([\nabla f_{i_0}, \cdots, \nabla f_{i_l}]) = l$;*
2. *for any $\{\nabla f_{j_1}, \cdots, \nabla f_{j_l}\} \subset C$, rank$([\nabla f_{j_1}, \cdots, \nabla f_{j_l}]) = l$.*

Note: the definition is local in that it depends upon x.

LEMMA 2.1 ([8], LEMMA 20). *A vector set $C = \{\nabla f_{i_0}, \cdots, \nabla f_{i_l}\}$ is a cadre if and only if rank(C) $= l$ and there exists $\{\lambda_j \neq 0\}_{j=0}^{l}$ such that*

$$\sum_{j=0}^{l} \lambda_j \nabla f_{i_j} = 0.$$

DEFINITION 4. *If we take for our a_{i_j}, ∇f_{i_j} and normalise the multipliers $\{\lambda_j\}_{j=0}^{l}$ as follows*

$$\begin{array}{lll} \sum_{j=0}^{l} \lambda_j = 1, & \textit{if the sum is nonzero,} & \textit{(cadres of } \textbf{type 1}\textit{)} \\ \lambda_0 = 0, & \textit{otherwise,} & \textit{(cadres of } \textbf{type 2}\textit{)} \end{array} \tag{9}$$

then such a normalised set is unique and we term $\{\lambda_j\}_{j=0}^{l}$, the **cadre multipliers** *associated with the cadre, C.*

These multipliers, although asymptotically related to the Lagrange multipliers associated with the underlying minmax problem, are essentially different from Lagrange multipliers since they are defined for any cadre and are not necessarily based upon maximal functions.

We also now generalise the idea of a reference.

DEFINITION 5. *For continuous Chebyshev problems*

$$\min_{x \in \mathbf{R}^n} \max_{t \in [\alpha,\beta]} |\phi(x,t) - y(t)|$$

the set of points $\{t_{i_j}\}_0^l$ is called a **point cadre**, *at x_0, if and only if $\{\nabla f_{i_j}(x_0)\}_0^l$ is a cadre, where $f_{i_j}(x) = \phi(x, t_{i_j}) - y(t_{i_j})$.*

In the special case of the continuous **linear** Chebyshev problem, Descloux [10] called the above point cadre a cadre. In the linear case ∇f_{i_j} is independent of x, but since in the nonlinear case this is not so we use the term point cadre to emphasize the local structure of this generalisation.

It is clear that we can write (2) as the following minimax problem

$$\min_{x \in \mathbf{R}^n} \max_{1 \le i \le 2m} f_i(x)$$

where $f_{i+m}(x) = -f_i(x)$. For ease of extension to the general minimax problem, we use the above formulation in this paper.

We are now able to extend the notion of a reference function.

DEFINITION 6. *The set of functions $\{f_{i_j}\}_{j=0}^{l}$ are said to locally form a* reference set *of a minmax problem (1) if $C = \{\nabla f_{i_j}\}_{j=0}^{l}$ is a cadre such that*

1. *the cadre multipliers $\{\lambda_j\}_0^l$ satisfy $\lambda_j > 0$, $j = 0, \cdots, l$;*
2. *$\psi(x) f_{i_j}(x) > 0, j = 0, \cdots, l$, where $\psi(x) = \max_{0 \le j \le l} f_{i_j}(x)$.*

The reference set is further called a levelled reference set *if the value of each function is the same, viz., $f_{i_j}(x) = f_{i_k}(x)$, for any $i_j, i_k \in C$.*

Note that, in general, the cadre multipliers may not alternate.

The following is well-known.

THEOREM 2.2 (FIRST ORDER NECESSARY CONDITIONS). *If x^* is a local minimizer of (1), then there exist multipliers $\{\lambda_i\}$ such that*

$$\sum_{i \in \mathcal{A}(x^*,0)} \lambda_i \nabla f_i(x^*) = 0, \tag{10}$$

$$\sum_{i \in \mathcal{A}(x^*,0)} \lambda_i = 1, \tag{11}$$

$$\lambda_i \geq 0, \qquad i \in \mathcal{A}(x^*,0), \tag{12}$$

where $\mathcal{A}(x^*,\epsilon) = \{i \mid \psi(x) - f_i(x) \leq \epsilon, i \in M\}$.

In terms of the reference set these first-order optimality conditions can be restated as follows.

THEOREM 2.3. *There exists a set of $l+1$ functions $\{f_{i_j}(x)\}_{j=0}^{l}$ which is a levelled reference set at x^* on the cadre $C = \{\nabla f_{i_j}(x^*)\}_{j=0}^{l}$ with the maximum deviation.*

We point out that the cadre corresponding to a reference set is of type 1.

3. Main Ideas of the Computational Procedure. From the previous section, finding a local minimum of the Chebyshev problem is equivalent to locating a levelled reference set including all the active functions.

Thus a natural approach to solving the Chebyshev problem is to

1. find a cadre;
2. construct a reference set based on the cadre;
3. level the reference set.

The algorithm we have developed is a descent algorithm with a line search. In addition to maintaining descent, the algorithm proceeds by recognising the structure of cadres. If a cadre is found, descent directions are defined to construct reference sets which are then levelled.

The following two lemmas help us identifying cadres of type 1 and type 2. Their proofs can be found in [7] (Lemma 3.2 and Lemma 3.3).

LEMMA 3.1 (NECESSARY AND SUFFICIENT CONDITIONS FOR LOCATING A CADRE OF TYPE 2). *Suppose $A = [\nabla f_{i_0} - \nabla f_{i_1}, \cdots, \nabla f_{i_0} - \nabla f_{i_{l-1}}]$ is of full rank and that $Z^T \nabla f_{i_0} \neq 0$, where the columns of Z form a basis for the null space of A^T. Then, there exists a cadre $C \subseteq \{\nabla f_{i_0}\}_{j=0}^{l}$ with cadre multipliers summing to zero if and only if $[\nabla f_{i_0} - \nabla f_{i_1}, \cdots, \nabla f_{i_0} - \nabla f_{i_l}]$ is rank deficient.*

LEMMA 3.2 (NECESSARY AND SUFFICIENT CONDITIONS FOR LOCATING A CADRE OF TYPE 1). *Suppose $[\nabla f_{i_0} - \nabla f_{i_1}, \cdots, \nabla f_{i_0} - \nabla f_{i_l}]$ are linearly independent. Then,*

there exists a cadre $C \subseteq \{\nabla f_{i_0}\}_{j=0}^{l}$ *with cadre multipliers summing to one if and only if* $Z^T \nabla f_{i_0} = 0$, *where* $A = [\nabla f_{i_0} - \nabla f_{i_1}, \cdots, \nabla f_{i_0} - \nabla f_{i_l}]$ *and* $Z^T A = 0$.

The cadre structure is monitored through the concept of a **working set**, a collection of indices which are candidates for forming a cadre.

A working set $\mathcal{W} = \{i_0, \cdots, i_l\}$ at a given point, x^c, includes preferentially all the ϵ-active functions (i.e. those functions within ϵ of the maximal functions) but is usually a larger set such that

$$A = [\nabla f_{i_0} - \nabla f_{i_1}, \cdots, \nabla f_{i_0} - \nabla f_{i_l}] \tag{13}$$

is full-rank.

There are different ways of forming working sets. In our algorithm, a working set $\mathcal{W}$ is made up of the indices of the current ϵ-activities plus a subset (possibly not proper) of the working set from the previous iteration.

The motivation for defining the working set in this manner is that it is from the set of the maximum functions that we expect to determine a levelled reference set.

LEMMA 3.3 ([8], LEMMA 38). *Suppose a cadre of type 1 has been located at the point* x *within the working set* $\mathcal{W} = \{\mu, i_0, \cdots, i_l\}$. *Suppose further, we define* v *as the unique least squares solution to*

$$\begin{aligned} &\hat{A}v = -\hat{\Phi}, \quad \textit{where} \\ &\hat{A} = [\nabla f_{i_0} - \sigma_0\sigma_1 \nabla f_{i_1}, \cdots, \nabla f_{i_0} - \sigma_0\sigma_l \nabla f_{i_l}], \\ &\hat{\Phi} = [f_{i_0} - \sigma_0\sigma_1 f_{i_1}, \cdots, f_{i_0} - \sigma_0\sigma_l f_{i_l}], \end{aligned} \tag{14}$$

f_{i_0} *achieves the current maximum deviation, and* $\sigma_j = sgn(f_{i_j})$. *Then* v *is a descent direction for* $\psi(x)$.

If (2) is a linear problem, it can be proven that, at $x + v$, $\{f_{k_0}, \cdots, f_{k_l}\}$ form a levelled reference set where

$$k_j = \begin{cases} i_j & \text{if } \lambda_j > 0 \quad , \\ i_{j+m} & \text{if } \lambda_j < 0 \quad \text{and } i_j \leq m, \\ i_{j-m} & \text{if } \lambda_j < 0 \quad \text{and } i_j > m, \end{cases} \tag{15}$$

and the $\{\lambda_i\}_{i=0}^{l}$ are the cadre multipliers. Thus, v is a desirable direction and the working set is modified to give $\mathcal{W} = \{k_0, \cdots, k_l\}$.

Also, note that the concept of a cadre and cadre multipliers have enabled us, first to consider determining a reference set from an enlarged set and second, to perform multiple dropping — both these concepts are in turn motivated by the generalisation of the characterisation of a solution in the linear case.

We also note that if we have a cadre of type 1 the indices in the working set that correspond to negative cadre multipliers are dropped from the working set. More particularly, if the working set consists uniquely of active functions, indices corresponding to negative cadre multipliers are **automatically** dropped (see Lemma 5.2 of [7]).

Furthermore, if we have a cadre of type 2, we are not in the asymptotic region, as follows directly from the first order conditions for optimality, Theorem 2.2.

On the other hand, if we do not locate a cadre in the working set, we are able to decrease all the active functions and (provided v is not an ascent direction — the usual case) level all the working functions by taking the direction

$$d = \begin{cases} h + v, & \text{if } v \text{ is a descent direction,} \\ h, & \text{otherwise.} \end{cases} \tag{16}$$

Here

$$\begin{aligned} h &= -ZB^{-1}Z^T\nabla f_\mu(x), \\ v &= -A(A^TA)^{-1}\Phi(x), \\ A &= [\nabla f_{i_0} - \nabla f_{i_1}, \cdots, \nabla f_{i_0} - \nabla f_{i_l}], \\ \Phi(x) &= [f_{i_0} - f_{i_1}, \cdots, f_{i_0} - f_{i_l}]^T, \\ Z^TZ &= I_{n-1},\ A^TZ = 0, \end{aligned} \tag{17}$$

i_0 is an index for one of the activities, B is positive definite and the working set, $\mathcal{W}$, is given by $\mathcal{W} = \{i_0, \cdots, i_l\}$.

At a first glance, it seems that the directions h and v depend on the index i_0 and thus are not uniquely determined by the current 'structure', i.e., working set. The following theorem shows that this is not the case.

THEOREM 3.4 (SEE [13], PAGE 61). *Suppose the working set $\mathcal{W}$ is fixed and B is positive definite. Then, the h and v given above by (16) are independent of the choice of i_0.*

If the working set consists only of active functions and we have located a cadre of type 2, all activities change equally with v (up to first order). For a proof, the reader is referred again to Lemma 5.2 of [7]. In addition, for cadres of type 2, all entries in the working set

(up to first order) change equally with h (Lemma 5.1 of [7]). Since little is to be gained by levelling (we do not have the correct type of cadre) we discard v and just take h for our search direction.

Before being able to state the algorithm in some detail two major issues remain to be discussed, namely the line search and the definition of and manner in which we handle degeneracy.

Let us begin with the line search.

4. The Line Search. We use a safeguarded line search that is similar to that of [16]. Thus suppose we have a descent direction d and we wish to find an approximation to

$$\min_{\alpha>0} \psi(x^k + \alpha d), \tag{18}$$

where $\psi(x) = \max_{i\in M} f_i(x)$.

Define

$$\nabla\psi_\epsilon^-(x,d) = \max_{i\in\mathcal{A}(x,\epsilon)\cap\nabla f_i{}^T d<0} \nabla f_i{}^T d. \tag{19}$$

The **acceptance criteria** used in our algorithm is the following:

Given any constants $0 < \delta < \beta < 1$ and $0 < \gamma < 1$, we demand that $x^{k+1} = x^k + \alpha^k d^k$ satisfies:

δ Condition: $\psi(x^{k+1}) \le \psi(x^k) + \delta\alpha^k\nabla\psi_\epsilon^-(x^k,d^k)$
and at least one of the following two:

β Condition: there exists $i \in \mathcal{A}(x^k,\epsilon)$, $\nabla f_i(x^k)^T d^k < 0$ such that

$$\nabla f_i(x^{k+1})^T d^k \ge \beta\nabla f_i(x^k)^T d^k;$$

γ Condition: there exists $i \in \mathcal{A}(x^k,\epsilon)$, $\nabla f_i(x^k)^T d^k \ge 0$ or $i \notin \mathcal{A}(x^k,\epsilon)$ such that

$$f_\mu(x^{k+1}) - f_i(x^{k+1}) \le \gamma[f_\mu(x^k) - f_i(x^k)], \quad \mu \in \mathcal{A}(x^k,0).$$

We require the following additional assumption:

ASSUMPTION 4.1. Each gradient $\nabla f_i(x)$ satisfies the Lipschitz conditions,

$$|\nabla f_i(z) - \nabla f_i(x)| \le L\|z - x\|_2 \quad \text{for all } i \in M.$$

The δ condition ensures that the reduction along each descent direction, d^k, has to be at least $\delta\alpha^k\nabla\psi_\epsilon^-(x^k,d^k)$.

The β and γ conditions essentially enforce that the steplength cannot be smaller than the minimum of $-\zeta\nabla\psi_\epsilon^-(x^k, d^k)$ and η, where ζ and η are positive constants that depend, in general, on the functions being minimized.

The above conditions are generalisation of the stepsize acceptance criteria for smooth minimisation. Similarly we are able to prove that there always exists an interval of the steplengths satisfying the acceptance criteria.

LEMMA 4.1 (SEE [13], PAGE 162). *Assume that d^k is any descent direction for the maximum function $\psi(x)$ at x^k. Then there exists an interval $[\alpha_l, \alpha_r]$ where $\alpha_l < \alpha_r$ such that for all $\alpha \in [\alpha_l, \alpha_r]$, the* δ **Condition** *and either the* β **Condition** *or the* γ **Condition** *is satisfied.*

Note that Lemma 4.1 is independent of the definition of the descent direction.

The next lemma shows that if the acceptance criteria are satisfied, the stepsize cannot be too small.

LEMMA 4.2 (SEE [13], PAGE 160). *Assume either the β Condition or the γ Condition is satisfied with $0 < \beta < 1$ or $0 < \gamma < 1$. Furthermore, assume that the set of descent direction $\{d^k\}$ is bounded and*

$$\nabla f_\mu{}^T d^k - \nabla f_i{}^T d^k = -(f_\mu - f_i) \quad \text{for any } i \in \mathcal{A}(x^k, \epsilon) \text{ and } \nabla f_i{}^T d^k \geq 0. \tag{20}$$

Then there exist positive constants ζ and η such that

$$\alpha^k \geq \min\{-\zeta\psi_\epsilon^-(x^k, d^k), \eta\}$$

is satisfied.

Since the exact minimum of $\psi(x)$ along d^k could occur only at either an intersection of two or more functions or at a minimum of one of the functions, the following result can easily be established.

LEMMA 4.3 (SEE [13], PAGE 167). *Assume that d^k is any descent direction for the maximum function $\psi(x)$ at x^k. Suppose x^{k+1} is the first minimum along the direction d^k. Then, at x^{k+1}, the* δ **Condition** *and either the* β **Condition** *or the* γ **Condition** *are satisfied.*

We are now able to describe the line search procedure.

Line Search Procedure

Step 1 [Initialization] If newflg = true, $\alpha \leftarrow 1$, $j_0 \leftarrow 0$, Go to Step 2.
Compute the leftmost break point if one exists:

$$-\frac{f_\mu - f_{j_0}}{(\nabla f_\mu - \nabla f_{j_0})^T d} = \min\left\{-\frac{f_\mu - f_j}{(\nabla f_\mu - \nabla f_j)^T d} | j \notin \mathcal{W}, (\nabla f_\mu - \nabla f_j)^T d < 0\right\}$$

$$\alpha \leftarrow -\frac{f_\mu - f_{j_0}}{(\nabla f_\mu - \nabla f_{j_0})^T d}.$$

Otherwise, set the initial steplength to one.

$$\alpha \leftarrow 1, \qquad j_0 \leftarrow 0.$$

Step 2 [Evaluation]
Compute the function values and the gradients at $x^{k+1} = x^k + \alpha d^k$. If the acceptance criteria are satisfied at x^{k+1}, stop.

Step 3 [Interpolation]

(i) Do a cubic interpolation for the function $f_{\mu(x^{k+1})}(x^k + \alpha d^k)$, using both the function values and gradients at the points x^k and x^{k+1}. Find its minimum $\alpha_{\mu^{k+1}}$.

(ii) Do a cubic interpolation for $f_{\mu(x^k)}(x)$, using both the function values and gradients at the points x^k and x^{k+1}. Find its minimum α_{μ^k};

(iii) Do two quadratic interpolations for $f_{\mu(x^{k+1})}(x)$ and $f_\mu(x)$, using the function values at the two points x^k and x^{k+1} and the gradients at x^k. Find the intersection α_b.
If $\alpha_{\mu^k} < \alpha_b$, $\alpha^{k+1} \leftarrow \alpha_{\mu^k}$;
If $\alpha_b < \alpha_{\mu^{k+1}}$, $\alpha^{k+1} \leftarrow \alpha_b$;
Otherwise $\alpha_{k+1} \leftarrow \alpha_{\mu^{k+1}}$.

If α_{k+1} is not in the interval $(0, \alpha^k)$, one step of the bisection method is performed. Otherwise $\alpha \leftarrow \alpha_{k+1}$, go to Step 2.

5. Degeneracy. For the discrete Chebyshev problem, degeneracy handling is an important component of any robust algorithm.

DEFINITION 7. *For a general minimax problem (1), the current point x^c is degenerate if and only if there is a cadre $C = \{\nabla f_\mu, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}$ such that $\{\mu, i_1, \cdots, i_l\} \subset \mathcal{A}(x^c, 0)$.*

From the construction of the working set, if the problem is degenerate, we have

$$\mathcal{W}^k \subset \mathcal{A}(x^k, 0).$$

Denote

$$\mathcal{W}^k = \{\mu, i_1, \cdots, i_l\},$$

We have already seen that h given by (17) is a descent direction unless the working set includes a cadre of type 1 (Lemma 3.2). Moreover, even in the degenerate case, if $Z^T \nabla f_\mu \neq 0$, there is no difficulty determining descent.

However, if $Z^T \nabla f_\mu = 0$ and there is more than one cadre $\mathcal{C} = \{\nabla f_\mu, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}$ satisfying $\{\mu, i_1, \cdots, i_l\} \subset \mathcal{A}(x^k, 0)$ it may not be possible to define a search direction such that it decreases the functions in all the cadres, although we know how to define a descending direction on one.

If we consider the cadres which correspond to subsets of active functions, then there can be three types of degenerate points:

Type i) there only exist cadres with cadre multipliers summing to zero;
Type ii) there exists a unique cadre and its cadre multipliers sum to one;
Type iii) there exists more than one cadre and at least one with cadre multipliers summing to one.

For the degenerate points of Type i), there cannot be any reference set consisting of only the active functions. This is because, for any reference set, each of the corresponding cadre multiplier is positive and the sum of them is one. Thus, the current point cannot be optimal. From Lemma 3.3 the vertical direction defined by (14) attempts to construct a levelled reference set from a cadre. Following [8], Lemma 38, for a cadre with cadre multipliers summing to zero, the vertical direction decreases all the functions in the cadre by the same amount. Hence one possible way of constructing a levelled reference set, for degenerate points of Type i), is to decrease all the active functions by the same amount. Note that in this case, the functions in the cadres are all active.

For the degenerate points of Type ii), it is possible that a reference set exists within the active set. If there is such a reference set, then the current point is already a stationary point. Otherwise, a vertical direction can be defined to try to construct a levelled reference set (See [8], Lemma 37 and equation (14), above). In fact, at a degenerate point of Type ii), we can still define a descent direction which attempts to construct such a levelled reference set.

For the degenerate points of Type iii), we do not know a direct way of defining a descent direction and we obtain a descent direction by solving a constrained least squares problem.

The following two lemmas are useful in identifying the type of degeneracy.

LEMMA 5.1 (SEE [13], PAGE 110). *Suppose* $\{\nabla f_\mu - \nabla f_{i_1}, \cdots, \nabla f_\mu - \nabla f_{i_{l-1}}\}$ *are linearly independent and* $\mathcal{W} = \{\mu, i_1, \cdots, i_l\}$. *Assume* $\tilde{Z}^T \nabla f_\mu = 0$ *where the columns of* $\tilde{Z}$ *form a basis for the null space of* $\{\nabla f_\mu - \nabla f_{i_1}, \cdots, \nabla f_\mu - \nabla f_{i_{l-1}}\}$. *Assume further that*

$$(\nabla f_\mu - \nabla f_{i_l}) = \sum_{j=1}^{l-1} \hat{\lambda}_j (\nabla f_\mu - \nabla f_{i_j}).$$

Then, there exists a cadre $\mathcal{C} \subseteq \{\nabla f_\mu, \nabla f_{i_1}, \cdots, \nabla f_{i_{l-1}}\}$ *with cadre multipliers summing to one and there exists at least another cadre including* ∇f_{i_l}.

LEMMA 5.2 (SEE [13], PAGE 112). *Suppose* $\mathcal{W} = \{\mu, i_1, \cdots, i_l\}$ *is a set of indices and* $\{\nabla f_\mu - \nabla f_{i_1}, \cdots, \nabla f_\mu - \nabla f_{i_l}\}$ *are linearly independent. Then, there can exist at most one cadre amongst* $\{\nabla f_\mu, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}$.

Consider the current ϵ-active set $\mathcal{A}(x^k, \epsilon)$. For simplicity of discussion, we assume ϵ is sufficiently small such that $\mathcal{A}(x^k, \epsilon) = \mathcal{A}(x^k, 0)$. This is no loss of generality since we only have a finite number of functions $f_i(x)$. Denote

$$h_R^k = -Z^k Z^{kT} \nabla f_\mu,$$

with

$$A^{kT} Z^k = 0, \quad Z^{kT} Z^k = I$$

and $\mathcal{W}^k = \mathcal{A}(x^k, \epsilon)$.

It is clear that there exists a cadre $\mathcal{C} = \{\nabla f_\mu, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}$ such that $\{\mu, i_1, \cdots, i_l\} \subset \mathcal{A}(x^k, 0)$ if and only if some gradients of active functions are linearly dependent.

From the construction of the working set $\mathcal{W}^k$ (details of which are given in [7], page 8), $\mathcal{W}^k = \{\mu, i_1, \cdots, i_l\}$ is chosen such that the corresponding Jacobian matrix A^k is of full rank and the ϵ-active functions are given priority when forming a working set.

Assume $\mathcal{W}^k \subset \mathcal{A}(x^k, \epsilon)$. In this case, there exists $i_{l+1} \in \mathcal{A}(x^k, \epsilon)$ such that

$$(\nabla f_\mu - \nabla f_{i_{l+1}}) = \sum_{j=1}^{l} \hat{\lambda}_j (\nabla f_\mu - \nabla f_{i_j}). \tag{21}$$

Type i) Degenerate Points: Since there is no cadre with the cadre multipliers summing to one, from Lemma 3.2, we have $h_R^k \neq 0$. Furthermore, by definition of a

degenerate point, there exists a cadre $\mathcal{C}^k$ embedded in the active set $\mathcal{A}(x^k, \epsilon)$ with cadre multipliers summing to zero.

Type i) degeneracy is identified when $\mathcal{W}^k \subset \mathcal{A}(x^k, \epsilon)$ and $\|h_R^k\| > 0$, from (21) and by Lemma 3.1, there exists at least one cadre $\mathcal{C} \subseteq \{\nabla f_\mu, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}$ with cadre multipliers summing to zero. Following Lemma 3.2, there is no cadre with cadre multipliers summing to one. Hence, if $\{\mu, i_1, \cdots, i_l\} \subset \mathcal{A}(x^k, \epsilon)$, x^k is a degenerate point of Type i).

Thus, moving along the direction which decreases all the functions in the cadre $\mathcal{C}^k$ by the same amount (whose existence is assured by [7], Lemma 5.2) is a constructive way of building up a reference set.

Since $h_R^k \neq 0$, $h^k \neq 0$, assuming B^k is positive definite. The horizontal direction is in the null space of $\{\nabla f_\mu - \nabla f_{i_1}, \cdots, \nabla f_\mu - \nabla f_{i_l}\}$. Furthermore, for any other active function f_{i_j} not in the working set $\mathcal{W}^k$,

$$(\nabla f_\mu - \nabla f_{i_j})^T h^k = 0, \quad i_j \notin \mathcal{W}^k.$$

This comes from the fact that

$$(\nabla f_\mu - \nabla f_{i_j}) = \sum_{\nu \in \mathcal{W}^k} \theta_\nu (\nabla f_\mu - \nabla f_\nu), \quad \text{for any } i_j \notin \mathcal{W}^k.$$

Thus h^k actually decreases all the active functions equally (up to the first order) and attempts to build a reference set from $\mathcal{C}^k$.

Hence, in this situation, we just take the horizontal direction as the search direction, i.e., $d^k = h^k$. It is important to realise that if a sequence $\{x^k\}$ converges to a stationary point, then there can only be a finite number of points which are degenerate points of Type i), since, at any stationary point, there exists a cadre with cadre multipliers summing to one.

Type ii) Degenerate Points: If x^k is a degenerate point of Type ii), there exists a unique cadre, based on the current ϵ-active set, with cadre multipliers summing to one. Moreover, $h_R^k = 0$. Since there exists no other cadre, $\mathcal{A}(x^k, \epsilon) \subseteq \mathcal{W}^k$. (Otherwise, using Lemma 5.1, there exists more than one cadre).

Assume zero multipliers are detected when the projected gradient $h_R^k = 0$. From Lemma 3.2, there exists a cadre with cadre multipliers summing to one. If $\mathcal{W}^k = \mathcal{A}(x^k, \epsilon)$, following Lemma 5.2, there does not exist any other cadre, based on the current ϵ-active set. Hence, x^k can only be a Type ii) degenerate point.

In this case, all the ϵ-active functions are in the working set $\mathcal{W}^k$. From the proof of Lemma 3.2, one cadre is given by $\mathcal{C}^k = \{\ \nabla f_{i_j} \mid \lambda_j^k \neq 0, i_j \in \mathcal{W}^k\ \}$, where λ^k is the least

squares solution to

$$\lambda_0 \nabla f_\mu + \sum_{j=1}^{l} \lambda_j \nabla f_{i_j} = 0, \qquad \sum_{j=0}^{l} \lambda_j = 1. \tag{22}$$

If this cadre corresponds to a levelled reference set, then we have found a solution. Otherwise, following a proof similar to Lemma 3.3, it can be shown that the vertical direction, defined by (17), decreases the maximum function $\psi(x)$. Furthermore, this vertical direction attempts to construct a levelled reference set from the cadre C^k.

The multipliers which are the least squares solution to (22) are uniquely defined for this type of degenerate points since A^k has full rank. However, zero multipliers may occur. A zero multiplier in this case indicates that the function does not belong to the cadre which includes the representative function. From (17), the vertical direction will bring the functions with zero multiplers down together if a descending vertical direction is found. However, the functions with zero multiplers are not significant in the definition of the search direction in the sense that whether a descending vertical direction exists or not does not depend on the values of the functions with zero multipliers, since from equation (7.12) of [8],

$$\nabla f_\mu^T v = \sum_{j=0}^{l} \theta_j (f_\mu - \sigma_0 \sigma_j f_{i_j}).$$

Hence, if a zero multiplier occurs, this implies there exist more functions than necessary to form a cadre in the current working set. If $\mathcal{W}^k = \mathcal{A}(x^k, \epsilon)$, then the current point is degenerate. Otherwise, there exists at least one non-ϵ-active function. In this case, it is reasonable to remove a non-ϵ-active function, which is the furtherest away from being active, from the working set, i.e., $\mathcal{W}^k \leftarrow \mathcal{W}^k - I^+$, where

$$I^+ = \begin{cases} \{j_0 \mid f_\mu - f_{j_0} = \max_{j \in \mathcal{W}^k}(f_\mu - f_j)\} & \text{if } \mathcal{A}(x^k, \epsilon) \subset \mathcal{W}^k \\ \emptyset & \text{otherwise.} \end{cases}$$

It is interesting to note that for a degenerate point of Type ii), the definition of the search direction is the same as for a nondegenerate point.

Type iii) Degenerate Points: At a degenerate point of Type iii), amongst the gradients of the ϵ-active functions, there exists more than one cadre and at least one with cadre multipliers summing to one.

Type iii) degeneracy is recognised when $\mathcal{W}^k \subset \mathcal{A}(x^k, \epsilon)$ and $\|h_R^k\| = 0$. By Lemma 3.2, there exists a cadre $\mathcal{C} \subseteq \{\nabla f_\mu, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}$ with cadre multipliers summing to one. Moreover, from (21) and following Lemma 3.1, there exists at least another cadre including

$\nabla f_{i_{l+1}}$. Hence, x^k is a degenerate point of Type iii). There is no obvious way of constructing reference sets for this type of degenerate point.

Assume $\mathcal{A}(x^k, \epsilon) = \{i_0, i_1, \cdots, i_l\}$, and $\mu = i_0$. Following a similar approach to [3], we solve the least squares problem given by

$$\min_{\theta \in \mathbb{R}^{l+1}} \| \sum_{j=0}^{l} \theta_j \nabla f_{i_j} \|_2$$

(23) subject to

$$\sum_{j=0}^{l} \theta_j = 1$$
$$\theta_j \geq 0, \quad j = 0, \cdots, l.$$

Suppose θ^* is a solution to (23). Denote

$$d^k = -\sum_{j=0}^{l} \theta_j^* \nabla f_{i_j}. \tag{24}$$

If the optimum value $\|d^k\| = 0$, the current point x^k is a solution. Otherwise, d^k is the steepest descent direction at the current point in the sense of [9] (page 64), i.e.,

$$\nabla \psi(x^k, d^k) = \min_{\|d\|_2 = 1} \nabla \psi(x^k, d).$$

We modify the working set $\mathcal{W}^k$ as follows:

$$\mathcal{W}^k \leftarrow \{ i_j \mid \theta_j^* > 0 \}.$$

It is clear that $\mathcal{W}^k \subseteq \mathcal{A}(x^k, \epsilon)$.

The least squares problem with linear constraints (23) can be solved by methods described in [12] (page 158). However, we shall exploit its special structure.

The problem (23) is a least squares problem with both equality and nonnegativity constraints.

We are able to show that we can solve (23) via a nonnegativity constrained least squares problem that handles the single equality implicitly.

Denote

$$e_{n+1}^T = [\underbrace{0, \cdots, 0}_{n}, 1], \quad e^T = [\underbrace{1, \cdots, 1}_{l+1}].$$

$$A = [\nabla f_{i_0}, \cdots, \nabla f_{i_l}], \quad \bar{A}^T = [A^T, e].$$

LEMMA 5.3 ([13], PAGE 120). *Suppose λ^* is a solution to the following NNLS (25).*

$$\min_{\lambda \in \mathbb{R}^{l+1}} \|\bar{A}\lambda - e_{n+1}\|_2$$

(25) *subject to*

$$\lambda_i \geq 0, \qquad i = 0, \cdots, l.$$

Then

$$\theta^* = \frac{1}{e^T \lambda^*} \lambda^*$$

is a solution to (23).

In the implementation of the algorithm, we directly use the NNLS algorithm from [12].

6. Maratos Effect. For nondifferentiable optimization problems, difficulties arise when the iterates have to follow a steep sided groove which is a nonlinear curve across which the function has discontinuous first derivatives (change of sign). If we use a linearization of the discontinuity only, limited progress can be made along this linearization if the merit function is to be reduced. Associated with this difficulty, the Maratos effect that some unit steps fail to reduce the merit function may occur even when the iterates $\{x^k\}$ are arbitrarily close to the solution x^*. As a result, it is no longer possible to guarantee superlinear convergence.

In fact, there are examples that indicate that the Maratos effect could occur for the predescribed algorithm where the maximum function $\psi(x)$ has been chosen as the merit function [However, we have not yet seen the Maratos effect *numerically*].

Since in the final iterations of algorithms for nondifferentiable minimization such as a minimax problem, an equivalent nonlinear programming problem is often solved, the Maratos effect is also inevitable unless special strategies are used.

Current available approaches to the Maratos effect include [6], [11] and [4]. The first two are correction methods while [4] use a relaxation technique to allow possible increase of the merit function.

We use the former approach. Thus when close to a stationary point, i.e., a reference set has been found, the horizontal direction is performed first and a vertical direction is conducted afterwards to force the functions to have the same value. This simply amounts to computing the vertical direction by

$$v = -A(A^T A)^{-1} \Phi(x + h) \tag{26}$$

where $\Phi(x)$ is defined as in (17).

We present the entire algorithm now. A user can request to invoke the process designed to avoid the Maratos effect by setting a flag Mflag = 1.

ALGORITHM

Initialization: Suppose an initial point x^0 is given. Set $k \leftarrow 1$, $\mathcal{W}^0 \leftarrow \emptyset$.

Step 1 [QR Decomposition]

Construct the working set (from $\mathcal{A}(x^k, \epsilon) \cup \hat{\mathcal{W}}^{k-1}$), Jacobian A^k and its QR decomposition. Assume the columns of Z^k form a basis for the null space of A^{k^T}.

If $\mathcal{A}(x^k, \epsilon) \subseteq \mathcal{W}^k$ and $\|Z^{k^T} \nabla f_\mu\| \leq \tau_c^k$, go to Step 2;

If $\mathcal{A}(x^k, \epsilon) \subseteq \mathcal{W}^k$ and $\|Z^{k^T} \nabla f_\mu\| > \tau_c^k$, go to Step 3;

If $\mathcal{A}(x^k, \epsilon) \not\subseteq \mathcal{W}^k$ and $\|Z^{k^T} \nabla f_\mu\| > \tau_c^k$, go to Step 4;

If $\mathcal{A}(x^k, \epsilon) \not\subseteq \mathcal{W}^k$ and $\|Z^{k^T} \nabla f_\mu\| \leq \tau_c^k$, go to Step 5;

Step 2 [Cadre "Found" with $\sum_{i \in \mathcal{C}} \lambda_i = 1$] If $\mathcal{W}^k$ is a reference set, obtain $B^k = Z^{k^T} G^k Z^k$, where G^k is a positive definite approximation to the Hessian of $\sum_{i \in \mathcal{C}} \lambda_i f_i(x)$ at x^k. Compute the horizontal direction h^k from (17); If Mflag = 0, compute the vertical direction v^k from (17); Otherwise, compute the vertical direction from (26). Set the search direction $d^k = h^k + v^k$.

If $\mathcal{W}^k$ is not a reference set, compute the vertical direction according to (14). Set $d^k = v^k$. Go to Step 6.

Step 3 [Cadre not Found]

Obtain $B^k = Z^{k^T} G^k Z^k$, where G^k is a positive definite approximation to the Hessian of $\sum_{i \in \mathcal{C}} \lambda_i f_i(x)$ at x^k. Compute the horizontal direction h^k and the vertical direction v^k from (17). If $\nabla f_\mu^T v < 0$, $d^k = h^k + v^k$. Otherwise $d^k = h^k$. Modify $\mathcal{W}^k$ if necessary. Go to Step 6.

Step 4 [Cadre "Found" with $\sum_{i \in \mathcal{C}} \lambda_i = 0$]

Compute $d^k = -Z^k Z^{k^T} \nabla f_\mu^k$. Go to Step 6.

Step 5 [More than One Cadre and at Least One with $\sum_{i \in \mathcal{C}} \lambda_i = 1$]

Compute the search direction d^k using (24). Set $\tau_c^{k+1} \leftarrow \frac{\tau_c^k}{2}$.

Step 6 [Line Search]

Perform a safeguarded line search. Set $k \leftarrow k + 1$. If $\|d^k\|_2 < \tau_s$ and $\mathcal{W}^k$ includes a levelled reference set, stop. Otherwise, go to Step 1. □

7. Conclusion. It is well known that a best linear Chebyshev approximation corresponds to a characteristic structure. It is not so well recognised that a solution of a nonlinear Chebyshev problem also possesses a rich structure and characterisation.

Under the classical Haar condition, the best linear Chebyshev approximation, on the real line, is a levelled reference function with the maximum deviation. There exist exactly $n+1$ distinct and ordered points which achieve the maximum deviation and the signs of the residuals on these points alternate.

Our experience with linear Chebyshev problems indicates that it is important for an efficient algorithm to make use of these special properties of a solution. The famous Remez algorithms [19] & [18] and the descent algorithm given in [2] are examples of such algorithms.

For many Chebyshev problems, such as multidimensional problems, nonlinear problems and discrete problems, however, the classical Haar condition does not hold. Thus, whether there exists some significant properties that can be computationally exploited is of some interest.

Nonlinear Chebyshev approximation theory indicates that, theoretically, for certain classes of nonlinear problems at least, useful characterisations still exist. Nonetheless, these theoretical characterisations are not easily computationally constructable or even recognizable.

The first author's experience with the descent algorithm given in [5], suggested that when there exist intermediate points where two or more activities belong to the same peak of the error curve, the efficiency of the method is impeded. We realise, however, that for usual Chebyshev *solutions*, this cannot happen. This suggests that, if we impose the structure of a solution, the algorithm may be improved.

The idea of a cadre has been introduced in [10] to describe a *linear* Chebyshev solution when the classical Haar condition is absent. Starting from this concept, we have been able generalise it to nonlinear Chebyshev problems. Based on the cadre , for nonlinear Chebyshev problems, we have generalised the reference set and levelled reference set concepts which characterise the property of alternating signs for a linear Chebyshev problem with the classical Haar condition.

We have then proceeded to exploit these results computationally.

The global convergence properties of the algorithm have been analysed through establishing the line search acceptance criteria. We point out that, under certain conditions, the algorithm is globally convergent with a two steps superlinear convergence (see [13], page 191).

The characterisation established for a local minimum of the discrete nonlinear Chebyshev problems is a generalisation of the specific properties of the best approximations for continuous linear Chebyshev problems.

Our algorithm has thus been developed to locate a local minimum of a discrete nonlinear Chebyshev problem by attempting to establish its structure and satisfy the characterisation of the local minimum. The algorithm is a method of successive descent on the maximum function with a line search on this function.

The algorithm builds up the structure through construction of the working set which attempts to approximate a reference set. The concept of the working set plays an important role in the algorithm. The functions in the working set are in general not ϵ-active functions, except near a solution. Along with the search directions, the working set attempts to exploit the geometry of the error curve of the solution. The important levelling process is embodied in the vertical directions.

The algorithm has been implemented in a numerically stable way. Initial numerical testing has indicated the efficacy of the method. Details are given in [7].

With suitable modifications, our approach can be applied to general minimax problems. Additional constraints can also be handled using a rather straightforward extension.

REFERENCES

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

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

[3] S. Busovača, *Handling degeneracy in a nonlinear l_1 algorithm*, Tech. Rep. Tech. Rept. CS-85-34, Univ. of Waterloo, Dept. of Computer Science, Univ. of Waterloo, Waterloo, Ontario N2L 3G1, 1985.

[4] R. M. Chamberlain, C. Lemarechal, H. C. Pedersen, and M. J. D. Powell, *The watchdog technique for forcing convergence in algorithms for constrained optimization*, Math. Prog. Study, 16 (1982), pp. 1–17.

[5] C. Charalambous and A. R. Conn, *An efficient method to solve the minimax problem directly*, SIAM J. Numer. Anal., 15 (1978), pp. 162–187.

[6] T. F. Coleman and A. R. Conn, *Nonlinear programming via an exact penalty function: Asymptotic analysis*, Math. Prog., 24 (1982), pp. 123–136.

[7] A. R. Conn and Y. Li, *An efficient algorithm for nonlinear minimax problems*, Tech. Rep. CS-88-41, Department of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada, 1988.

[8] A. R. Conn and Y. Li, *Structure and characterization of discrete chebyshev problems*, Tech. Rep. CS-88-39, Department of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada, 1988.

[9] Dem'yanov and Malozemov, *Introduction to Minimax*, Keter Publishing House, Jerusalem, 1974.

[10] J. Descloux, *Dégénéresence dans les approximations de Tschebysheff linéaries et discrètes*, Numerische Mathematik, 3 (1961), pp. 180–187.

[11] R. Fletcher, *Second order corrections for non-differentiable optimization*, in Lecture Notes in Mathematics 912, G. Watson, ed., Springer Verlag, 1981, pp. 85–114.

[12] C. L. Lawson and R. J. Hanson, *Solving Least Square Problems*, Prentice-Hall, 1974.

[13] Y. Li, *An Efficient Algorithm for Nonlinear Minimax Problems*, PhD thesis, University of Waterloo, 1988.

[14] G. Meinardus, *Approximation of Functions: Theory and Numerical Methods*, Springer Verlag, 1967. translated by Larry, L. Schumaker.

[15] T. S. Motzkin, *Approximation by curves of a unisolvent family*, Bull. Amer. Math. Soc., 55 (1949), pp. 789–793.

[16] W. Murray and M. L. Overton, *Steplength algorithms for minimizing a class of nondifferentiable functions*, Computing, 23 (1979), pp. 309–331.

[17] M. J. D. Powell, *Approximation Theory and Methods*, Cambridge University Press, 1981.

[18] Remez, *Sur le calcul effectif des polynomes d'approximation de tchebichef*, Competes Rendues, 199 (1934), pp. 337–340.

[19] ——, *Sur un procédé convergent d'approximations successives pour déterminer les polynômes d'approximation*, Competes Rendues, 198 (1934), pp. 2063–2065.

[20] J. R. Rice, *On the existence and characterisation of best nonlinear Tchebyshev approximation*, Tran. Am. Math. Soc., 110 (1964), pp. 88–97.

[21] L. TORNHEIM, *On n-parameter families of functions and associated convex functions*, Trans. Amer. Math. Soc., 69 (1950), pp. 457–467.