The Computational Structure and Characterization of Nonlinear Discrete Chebyshev Problems

Andrew R. Conn*
Yuying Li**

88-956
December 1988

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

*The research of this author was supported in part by NSERC grant A8639.

**The research of this author was partially supported by the U. S. Army Research Office through the Mathematical Science Institute, Cornell University.

THE COMPUTATIONAL STRUCTURE AND CHARACTERIZATION OF NONLINEAR DISCRETE CHEBYSHEV PROBLEMS

ANDREW R. CONN * AND YUYING LI †

Abstract. We present the generalisation of some concepts in linear Chebyshev theory to the nonlinear case. We feel these generalisations capture the inherent structure and characteristics of the best Chebyshev approximation and that they can be usefully exploited in the computation of a solution to the discrete Chebyshev problem.

Key Words. nonlinear Chebyshev approximation

Subject Classification. 41A50, 65D99, 65K05, 65K10

1. Introduction.

A problem of the form:

$$\min_{x\in\Re^n}\max_{i\in M}|f_i(x)| \tag{1.1}$$

is called a *discrete minimax problem*, where M is a finite index set. We seek to find the minimum value for the maximum absolute function $\max_{i\in M}|f_i(x)|$.

In this paper, we are content to find a local minimum of (1.1) and we assume that each $f_i(x)$ is continuously differentiable.

It is clear that a discrete Chebyshev problem (1.1) could be regarded as a special case of a general minimax problem with:

$$\min_{x\in\Re^n}\max_{i\in M} f_i(x) \tag{1.2}$$

where

$$M = \{1,2,\cdots,m,m+1,\cdots,2m\},$$

$$f_{i+m}(x) = -f_i(x), \qquad i=1,\cdots,m.$$

It is well known that problem (1.1) is equivalent to:

$$\begin{aligned} &\min_{(z,x)} z \\ (1.3)\qquad \text{subject to}\quad & \\ & z - f_i(x) \ge 0, \qquad i\in M, \end{aligned}$$

where z is an additional variable. The dimension of the variable space now becomes $n+1$.

---

* 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, Ithaca, New York, 14853. The research of this author was partially supported by the U.S. Army Research Office through the Mathematical Science Institute, Cornell University.

Numerical methods for the discrete nonlinear Chebyshev/minimax problem are less prolific than for the linear problem. Moreover, the maximum function $\psi(x) = \max_{i\in M} f_i(x)$ is not differentiable at kinks that arise whenever $f_i(x) = f_j(x),\ i,j \in M,\ i \neq j$. Therefore, traditional gradient type methods cannot be applied directly.

Existing numerical methods for discrete nonlinear Chebyshev problems could be classified as *first order methods*, for which only the first derivatives are used, and *second order methods* for which the Hessians of the functions whose value is equal to the current maximum function are included. As we might expect, methods that use approximations to these Hessians are also available. Most of these algorithms have relied upon some general linear/nonlinear programming technique to solve a linear/quadratic programming subproblem at each iteration.

A major class of first order methods establishes a linear subproblem at each iteration and an algorithm for the linear Chebyshev problem is then invoked. The methods of [1], [16] and [21] belong to this class.

The method of [7] is a first order method which uses an active set strategy. A different approach was used in [2], where each subproblem is a differentiable least-pth approximation problem.

A first order method is generally not sufficiently sophisticated for nonlinear problems because of its slow convergence rate. Second order methods have been considered by various authors. Examples include [9], [12], [13], [14], [20] and [32].

If we omit the details of the different techniques for the line search and the second order information approximation, the aforementioned second order methods can be loosely classified into two classes.

The first class applies the sequential equality constrained quadratic programming approach to the equivalent nonlinear programming problem (1.3). Examples are [9], [20] and [32]. In [9], the subproblem is formulated in a more direct way in which the objective function is taken from one of the maximum functions and the constraints essentially make all the ϵ-active functions (i.e. within ϵ of being active) equivalent. Each time the subproblem is solved as an equality constrained quadratic programming problem, with an additional Newton step that attempts to make ϵ-active functions exactly active when close to a stationary point.

The methods of [13] and [30] are both based on the corresponding first order method to solve the linear subproblems. The second order information is used when one is *near* a stationary point and some equations established from the first order optimality conditions are solved by Newton's method. These methods require a switch back and forth between solving the linear subproblem and the system of nonlinear equations and they suffer the defect of being nondescent algorithms. Furthermore, computationally, these methods are expensive in the sense that the calculation of the search direction at each iteration requires the *solution* of a constrained linear minimax problem. This normally requires an extensive number of inner iterations.

In summary, all the methods mentioned so far emphasize the role played by the underlying mathematical programming problem.

However, the discrete Chebyshev problem has a rich structure. In [9] and [20], the special structure of the equivalent nonlinear programming problem (1.3) has been emphasized in the sense that the original problem has a natural merit function, the maximum function $\psi(x)$.

The importance of the original structure of the Chebyshev problem in the design of algorithms for linear Chebyshev problems has been recognised (e.g. [5] and [27]). We would like to similarly exploit the structure of nonlinear Chebyshev problems. Hence, one of the main objectives of this paper has been to recognise the characterisation of the local solution from the computational point of view. The characterisation obtained has a strong connection with Chebyshev theory and is equivalent to the Chebyshev theory in the linear case.

2. Basic Notation. We introduce some notation.

We always use x and d to denote column vectors in $\Re^n$ while we use t to denote scalars in $\Re$.

We also use ∇ to denote the gradient operator with respect to x.

For a minimax problem (1.2), we define the maximum function by

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

At any point x, we also refer to $\psi(x)$ as the current *maximum deviation.*

We use $x \stackrel{\text{def}}{=} y$ to mean x is defined by y.

$C[\alpha, \beta]$ denotes the set of continuous functions defined on $[\alpha, \beta]$.

DEFINITION 1. *At any point x, the functions that achieve the maximum deviation are called the* **active functions**. *The* **active set** *is indexed by*

$$\mathcal{A}(x, 0) = \{\ l \mid \psi(x) - f_l(x) = 0,\ \ l \in M\}. \tag{2.1}$$

The **ϵ-active** *functions are the functions that achieve the maximum deviation within a tolerance of ϵ, a small positive constant. The set of ϵ-active functions are*

$$\mathcal{A}(x, \epsilon) = \{\ l \mid \psi(x) - f_l(x) \le \epsilon,\ \ l \in M\}. \tag{2.2}$$

Given a vector set $\{v_i \mid v_i \in \Re^n, i = 1, \cdots, m\}$, the maximum number of vectors within the vector set which are linearly independent is called the *rank* of the vector set.

The sign function $\text{sgn}(x)$ is defined as follows

$$\text{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}.$$

We use $\mathcal{A} \subseteq \mathcal{B}$ to denote $\mathcal{A}$ is a subset of $\mathcal{B}$. Moreover, if $\mathcal{A} \subseteq \mathcal{B}$ and $\mathcal{A} \neq \mathcal{B}$, we use the notation $\mathcal{A} \subset \mathcal{B}$.

Subsequent notation will be introduced when it is needed.

3. Chebyshev Approximation Theory.

One common source of discrete Chebyshev problems (1.1) comes from function approximation.

A general one dimensional *Chebyshev approximation* problem can be written as

$$\min_{x \in \Re^n} \max_{t \in T} |\phi(x,t) - y(t)|, \tag{3.1}$$

where T is a closed interval.

Discretising the problem on a finite set $\{t_1, \cdots, t_m\}$, we have:

$$\min_{x \in \Re^n} \max_{i \in M} |f_i(x)|. \tag{3.2}$$

Here, $M = [1, 2, \cdots, m]$. Each $f_i(x)$ could be interpreted as $\phi(x, t_i) - y(t_i)$. Hence, the discretised Chebyshev problem bears a close relation to the continuous approximation.

3.1. Linear Chebyshev Approximation Theory.

When $\phi(x,t)$ is in a finite dimensional linear function subspace, i.e., $\phi(x,t) = \sum_{i=1}^{n} x_i \phi_i(t)$ where $t \in [\alpha, \beta]$, the problem (3.1) is termed a *linear Chebyshev problem.* In this case, there is an elegant theory for existence, uniqueness and characterisation of the solution. Furthermore, there exist efficient algorithms to solve linear Chebyshev problems (e.g., [3] and [4]). In this section, the results which are relevant to our discussion are summarised.

DEFINITION 2. *An n dimensional linear function subspace $\mathcal{L} = span\{\phi_1(t), \cdots, \phi_n(t)\}$ of $C[\alpha, \beta]$ is said to satisfy the* **Haar condition** *if and only if, for every nonzero $\phi(x,t) \in \mathcal{L}$, the number of roots of the equation $\phi(x,t) = 0$, as a function of t, within $[\alpha, \beta]$ is less than the dimension of $\mathcal{L}$.*

There is a well-known theorem that characterises the best linear Chebyshev approximation explicitly (see for example, [23], page 77).

THEOREM 3. *[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:

$$\alpha \le t_0 < t_1 < \cdots < t_n \le \beta \tag{3.3}$$

and

$$|y(t_i) - \phi^*(t_i)| = \|y(t) - \phi^*(t)\|_{\infty}, \; i = 0, 1, \cdots, n, \tag{3.4}$$

and

$$y(t_{i+1}) - \phi^*(t_{i+1}) = -(y(t_i) - \phi^*(t_i)), \; i = 0, 1, \cdots, n-1, \tag{3.5}$$

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

DEFINITION 4. *A* **reference** *is by definition a set* $\{t_i\}_{i=0}^{n}$ *of* $n+1$ *distinct ordered points*

$$\alpha \le t_0 < t_1 < \cdots < t_n \le \beta \tag{3.6}$$

Under the Haar condition, it can be proved that the corresponding values $\{\phi(x,t_i)\}$ of any function $\phi(x,t)$ in $\mathcal{L}$ are related by a linear relation, called the *characteristic relation:*

$$\lambda_0\phi(x,t_0) + \lambda_1\phi(x,t_1) + \cdots + \lambda_n\phi(x,t_n) = 0 \tag{3.7}$$

with multipliers $\lambda_j \neq 0$ for $j = 0,1,\cdots,n$, *where the* λ_j's *are independent of* x .

DEFINITION 5. *Let* $\phi(x,t)$ *be any function in* $\mathcal{L}$ *and* $f_i = y(t_i) - \phi(x,t_i)$ *be the errors of the approximation at the points* t_i *of the reference* $\{t_i\}$. *The function* $\phi(x,t)$ *is called a* **reference function** *with respect to the reference* $\{t_i\}$ *and the function* $y(t)$ *if and only if:*

$$\begin{aligned} sgn(f_i) &= sgn(\lambda_i) && \text{for all } i, \text{ or} \\ sgn(f_i) &= -sgn(\lambda_i) && \text{for all } i. \end{aligned}$$

The **levelled reference** *function with respect to a given reference* $\{t_i\}$ *is characterised by the property that the errors* $\{f_i\}$ *have the same absolute value. The common absolute value* $|f_i|$ *of the approximation errors is called the* **reference deviation** .

The reference deviation is always less than or equal to the maximum error [27].
The following theorem has been extracted from [23](page 98) since it is useful for later discussions.

THEOREM 6. *Let* $\mathcal{L}$ *be an* n *dimensional linear function subspace of* $C[\alpha,\beta]$ *that satisfies the Haar condition, let* $\{t_i\}_{i=0}^{n}$ *be a set of reference points from* $[\alpha,\beta]$ *that are in ascending order:*

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

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

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

for all functions $\phi(x,t)$ *in* $\mathcal{L}$. *Then every multiplier is nonzero, and their signs alternate.*

The following lemma illustrates that the reference function embodies the notion of alternating signs.

LEMMA 7. *Suppose $\phi(x,t) = \sum_{i=1}^{n} x_i\phi_i(t)$ and the Haar condition is satisfied. Suppose further $\phi(\bar{x},t)$ is a reference function with respect to a reference $\mathcal{C}$ and a function $y(t)$. Then, $\phi(\bar{x},t) - y(t)$ alternates in sign on the reference.*

Proof. Follows directly from Theorem 6 and Definition 5. ■

In terms of the reference, Theorem 3 can be stated as follows:

THEOREM 8. *The function of the best approximation is the levelled reference function with the maximal reference deviation.*

Computational methods for linear discrete Chebyshev problems have complied with the linear Chebyshev theory in the sense that the numerical methods have deliberately or naturally searched for an approximation with the characteristics of the best approximation.

3.2. One Dimensional Nonlinear Chebyshev Theory. The one dimensional continuous nonlinear Chebyshev problem is described by

$$\min_{x\in\mathcal{D}} \max_{t\in[\alpha,\beta]} |\phi(x,t) - y(t)|, \tag{3.8}$$

where $\phi(x,t)$ is nonlinear in x and $\mathcal{D} \subset \Re^n$ is a parameter space. We further assume that $\phi(x,t)$ is continuously differentiable with respect to x.

Nonlinear Chebyshev problems have been much more intractable. The existence of a best approximation is not guaranteed because, even when $\mathcal{D} \subseteq \Re^n$ is compact, the set $\{\phi(x,t) \mid x \in \mathcal{D}\}$ is often not compact under pointwise convergence (see, for example [25]). The existence of a solution is not easily determined. An extended linear Chebyshev theory holds only for a special class of nonlinear problems for which the decidability of a problem belonging to the class is also difficult.

A first step in extending the linear Chebyshev theory was made by [19] and [28] and further developed by [25]. The main results are described in this section, following the terminology of [24] and [26].

As has been discussed in §3.1, any best linear Chebyshev approximation has an alternant of degree $n+1$ under the classical Haar condition. This important characterisation has been generalised by [24] and [26] via the following five definitions.

DEFINITION 9. *$\phi(x,t)$ has* **Property Z of degree m** *at x_0 if for every $x \neq x_0$, $\phi(x_0,t) - \phi(x,t)$ has at most $m-1$ zeros for $t \in [\alpha,\beta]$.*

A function $\phi(x,t)$ with Property Z of degree m at x_0 is a function that has a unique interpolant over m points for the curve $\phi(x_0,t)$, if one exists. For linear Chebyshev problems, the degree of Property Z is always n, under the classical Haar condition, where $\mathcal{D} = \Re^n$ and, of course, the interpolant always exists.

DEFINITION 10. *A function $\phi(x,t)$ is said to be* **locally solvent of degree m** *at $x_0 \in \mathcal{D}$ on $[\alpha,\beta]$ if given a set $\alpha \leq t_1 < t_2 < \cdots < t_m \leq \beta$ and $\epsilon > 0$, there is a*

$\delta = \delta(x_0, \epsilon, t_1, t_2, \cdots, t_m) > 0$ *such that, for any set of m arbitrary numbers* $\{y_j\}_1^m$,

$$|\phi(x_0, t_j) - y_j| < \delta, j = 1, \cdots, m,$$

implies that there exists a solution $x \in \mathcal{D}$ *to the system* $\phi(x, t_j) = y_j, \quad j = 1, \cdots, m,$ *with*

$$\max_{t \in [\alpha, \beta]} |\phi(x, t) - \phi(x_0, t)| < \epsilon.$$

The degree of local solvence describes a property of a function $\phi(x, t)$, namely that for any m points $\{y_i\}_1^m$, close to $\{\phi(x_0, t_i)\}_1^m$, there always exists a $\phi(x, t)$ in the neighbourhood of the curve $\phi(x_0, t)$ that interpolates these m points. For the linear Chebyshev family under the classical Haar condition, the degree of solvency is always n, for $\mathcal{D} = \Re^n$.

DEFINITION 11. *A* **varisolvent function** *is a function* $\phi(x, t)$ *which possesses both of the properties in the Definition 9 and 10 with the same degree for each* $x \in \mathcal{D}$. *Thus* $\phi(x, t)$ *has Property Z of degree* $m(x)$ *at* x, *and* $\phi(x, t)$ *is solvent of degree* $m(x)$ *at* x. *The* **degree** *of* $\phi(x, t)$ *at* x *is the common degree of Property Z and local solvence.*

It is clear that a linear Chebyshev family satisfying the Haar condition always has local solvence of degree n. As a natural extension of the linear Chebyshev theory, the following class of functions has been of interest.

DEFINITION 12. *An approximating function* $\phi(x, t)$, *where* $\mathcal{D} \subset \Re^n$ *and* $t \in [\alpha, \beta]$, *is said to be* **solvent of degree m** *if, given a set of distinct points* $\{t_i\}_1^m \subset [\alpha, \beta]$ *and a set of m arbitrary numbers* $\{y_i\}_1^m$, *there exists* $x \in \mathcal{D}$ *such that*

$$\phi(x, t_i) = y_i, \qquad i = 1, \cdots, m.$$

Definition 12 differs from Definition 10 in that the solvability is global because the set of numbers $\{y_i\}_1^m$ is not necessarily close to the curve $\phi(x_0, t)$.

DEFINITION 13. *The approximating function* $\phi(x, t)$, *where* $\mathcal{D} \subset \Re^n$ *and* $t \in [\alpha, \beta]$, *is said to be a* **unisolvent function** *if (1) it is solvent of degree n, and (2) it has property Z of degree n.*

The linear Chebyshev theory has been extended to the family of unisolvent approximating functions as follows.

THEOREM 14. *[24] (page 73)*
Let $\phi(x, t)$ *be a unisolvent function on* $[0, 1]$ *with n parameters, and let* $y(t)$ *be continuous on* $[0, 1]$. *Then*

(i) $y(t)$ *possesses a best Chebyshev approximation* $\phi(x^*,t)$*;*

(ii) $\phi(x^*,t)$ *is a best approximation to* $y(t)$ *if and only if there are* $n+1$ *points upon which the magnitudes of the errors achieve the maximum and the signs of these errors alternate;*

(iii) the best approximation is unique.

For the varisolvent approximating function, the following characterisation for a best Chebyshev approximation exists.

THEOREM 15. *[26] (page 10)*
Let $\phi(x,t)$ *be a varisolvent function with degree* $m(x)$ *at* $x \in \mathcal{D}$*. Assume* T *is any compact subset of* $[0,1]$*. Then the function* $\phi(x,t)$ *is a best approximation of* $y(t)$ *on* T *if and only if there are* $m(x)+1$ *points upon which the magnitudes of the errors achieve the maximum and the signs of these errors alternate.*

Note that using the interval $[0,1]$ in Theorem 15 implies no loss of generality.

Under certain conditions, the varisolvent approximating function is the only type of function that has a degree which gives a characteristic number of alternations of best approximation. Furthermore, it has been shown that unisolvent functions are the only approximating functions satisfying a certain closure hypothesis to which the elegant linear theory may be extended. In this case, an approximation having an alternant of $n+1$ points is necessary and sufficient for a best approximation (see, for example, [25]).

Rational functions and exponential functions are varisolvent functions. However, in general, determining whether a function belongs to the class of varisolvent functions is not trivial.

Thus, under certain conditions, the best nonlinear approximation also possesses special properties which are similar to those of best linear approximation. In such cases, theorems 14 and 15 give explicit characterisation of the best nonlinear Chebyshev approximation.

4. Structure of the Solution to Minimax Problems. For certain class of nonlinear Chebyshev approximation problems, the existing nonlinear Chebyshev theory gives explicit descriptions of the properties of the solution. However, since the characterisation depends on the degree of solvency, which is unknown and varies with the parameter x, it is difficult to exploit these properties computationally. Furthermore, the nonlinear theory is restrictive in the sense that many approximating functions are not varisolvent and moreover, the theory cannot be applied to discrete Chebyshev problems.

We have stated that, under the classical Haar condition, the concepts of reference and reference function together with extreme points uniquely describe a best linear Chebyshev approximation.

The concept of reference is essential in describing the structure because it provides us with a template upon which the characteristic equation is satisfied. Further requiring an approximating function with the signs of the errors the same as those of the corresponding multipliers completes the characterisation.

We feel that the reference and reference function characterise the best linear Chebyshev approximation more fundamentally than the explicit alternating sign property. As long as the approximating function satisfies certain conditions, the alternating sign property becomes a natural consequence of the fact that the signs of the multipliers in the characteristic equation alternate.

In this section, we extend the concepts of linear Chebyshev approximation to a general nonlinear Chebyshe problem (discrete or continuous). The extension also applies to minimax problems.

Let us again consider an n dimensional linear function subspace

$$\mathcal{L} = \text{span}\{\phi_1(t), \cdots, \phi_n(t)\}.$$

For any function $\phi(x,t) \in \mathcal{L}$,

$$\phi(x,t) = \sum_{i=1}^{n} x_i \phi_i(t).$$

Given a reference $t_0 < t_1 < \cdots < t_n$, the characteristic equation (3.7) can be equivalently written as

$$\lambda_0 a_0 + \lambda_1 a_1 + \cdots + \lambda_n a_n = 0,$$

where

$$a_j = [\phi_1(t_j), \cdots, \phi_n(t_j)]^T.$$

Since

$$\nabla\phi(x,t_j) = a_j,$$

the characteristic equation can also be written as

$$\sum_{j=0}^{n} \lambda_j \nabla\phi(x,t_j) = 0. \tag{4.1}$$

The gradient space with respect to the parameter $x \in \Re^n$

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

is equivalent to

$$J = \text{span}\{\phi_1(t), \cdots, \phi_n(t)\}.$$

Therefore, the gradient space uniquely determines the family of approximating functions for linear Chebyshev problems.

We now consider a continuous one dimensional nonlinear Chebyshev problem written as in (3.8).

DEFINITION 16. *[18] At any point x_0, the* **linear gradient space** $J(t)$ *of* $\phi(x,t)$ *refers to*

$$\begin{aligned} J(t) &= \{ \sum_{i=1}^{n} \lambda_i \frac{\partial \phi(x_0,t)}{\partial x_i}, \quad \lambda_i \in \Re \} \\ &= span\{ \frac{\partial \phi(x_0,t)}{\partial x_1}, \cdots, \frac{\partial \phi(x_0,t)}{\partial x_n}.\} \end{aligned}$$

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 $x \in \Re^n$.

In contrast to a linear Chebyshev problem, the gradient space of a nonlinear Chebyshev problem does not uniquely describe the family of approximating functions.

In the discrete case, where $f_i(x) = y(t_i) - \phi(x, t_i)$, instead of considering the whole gradient function space $J(t)$, we consider the following set of vectors denoted by the rows of the Jacobian matrix:

$$\begin{aligned} J(t_1, t_2, \cdots, t_m)^T &= \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \cdot & \cdots & \cdot \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \\ &= \begin{bmatrix} \nabla f_1^T \\ \vdots \\ \nabla f_m^T \end{bmatrix}. \end{aligned}$$

Here we consider the gradients of the error functions $f_i(x)$ instead of those of the approximating functions $\phi(x, t_i)$ because, for a discrete problem, we often do not have the corresponding $\phi(x, t_i)$. Since the gradients are with respect to the parameter x, it is easy to see their equivalence. For linear Chebyshev approximation, if the classical Haar condition is satisfied, then any $n \times n$ submatrix of J is nonsingular.

An important concept is now introduced.

DEFINITION 17. *The vector set $\mathcal{C} = \{\nabla f_{i_0}, \cdots, \nabla f_{i_l}\}$, 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 $\{\nabla f_{j_1}, \cdots, \nabla f_{j_l}\} \subset \mathcal{C}$, $rank([\nabla f_{j_1}, \cdots, \nabla f_{j_l}]) = l$.*

Thus the concept of a cadre is a local one although frequently the argument x will be omitted, its value being understood from the context.

The vector set $\mathcal{C}$ which is a cadre according to Definition 17 is sometimes referred to as the minimum H-set [6].

DEFINITION 18. *For continuous Chebyshev problems (3.8), the set of points* $\{t_{i_j}\}_0^l$ *is called a* **point cadre**, *at* x, *if and only if* $\{\nabla f_{i_0}, \cdots, \nabla f_{i_l}\}$ *is a cadre, where* $f_{i_j}(x) = y(t_{i_j}) - \phi(x, t_{i_j})$, $j = 0, \cdots, l$.

REMARK 19. *Consider the linear Chebyshev approximation where* $\mathcal{L} = span\{\phi_1(t), \cdots, \phi_n(t)\}$. *If* $\phi(x,t) = \sum_{i=1}^n x_i\phi_i(t) \in \mathcal{L}$, *then*

$$J(t_1, \cdots, t_m)^T = \begin{bmatrix} \phi_1(t_1), & \cdots, & \phi_n(t_1) \\ \vdots & \vdots & \vdots \\ \phi_1(t_m), & \cdots, & \phi_n(t_m) \end{bmatrix}.$$

Thus, a point set $\{t_{i_0}, \cdots, t_{i_l}\}$ *is called a point cadre if and only if the gradient vectors* $\{v_0, \cdots, v_l\}$ *form a cadre where*

$$v_j = \begin{bmatrix} \phi_1(t_{i_j}) \\ \vdots \\ \phi_n(t_{i_j}) \end{bmatrix}.$$

Of course, here the cadre is independent of x. *In [11], the point set* $\{t_{i_0}, \cdots, t_{i_l}\}$ *has been called a cadre for a continuous linear Chebyshev problem. Descloux [11] introduced the concept of the cadre in the linear Chebyshev problem to describe the optimum when* $\{\phi_1(t), \cdots, \phi_n(t)\}$ *is not a Chebyshev set (i.e. the Haar condition is not satisfied). Since we want to generalise Definition 17 to nonlinear discrete Chebyshev problems in* $\Re^n$, *we prefer to refer to a cadre as a vector set and use the point cadre, in the continuous case, to signify the set of points whose corresponding gradients form a cadre.*

For one dimensional linear Chebyshev problems, a *reference* is defined to be any $n+1$ distinct points on the real line. Suppose that the classical Haar condition is satisfied, then any reference is a point cadre by Definition 18. Therefore the concept of point cadre could be considered to be a generalisation of the reference.

The rank of the cadre has essentially the same role as the largest degree of local solvency for continuous Chebyshev problems. Assuming the approximating function $\phi(x,t)$ is differentiable, if $\phi(x,t)$ is a unisolvent function of degree n, then, at any point x, the gradients at any $n+1$ points $\{\nabla\phi(x, t_{i_0}), \cdots, \nabla\phi(x, t_{i_n})\}$ is always a cadre. If $\phi(x,t)$ is a varisolvent function, at any point x, the gradients at any l points, where l is the degree of solvency at x, is a cadre of rank $l-1$.

The following lemma provides a way of identifying a cadre.

LEMMA 20. *A vector set* $C = \{a_{i_0}, \cdots, a_{i_l}\}$ *is a cadre if and only if rank(C)* $= l$ *and there exists* $\{\lambda_j \neq 0, j = 0, 1, \cdots, l\}$ *such that*

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

Proof. Necessity follows directly from the definition of the cadre.

Without loss of generality, suppose $\{a_{i_1}, \cdots, a_{i_l}\}$ is linearly independent. Assume $\mathcal{C}$ is not a cadre. Then a subset of cardinality l is rank deficient. Assume that $\{a_{i_0}, \cdots, a_{i_{l-1}}\}$ are linearly dependent. Since $\{a_{i_1}, \cdots, a_{i_{l-1}}\}$ are linearly independent, we have

$$a_{i_0} = \sum_{j=1}^{l-1} \theta_j a_{i_j}.$$

By assumption, however,

$$\sum_{j=0}^{l} \lambda_j a_{i_j} = 0, \qquad \lambda_j \neq 0.$$

Thus

$$a_{i_0} = \sum_{j=1}^{l} \hat{\lambda}_j a_{i_j}, \quad \text{with} \quad \hat{\lambda}_j \neq 0, \quad j = 1, \cdots, l.$$

Hence, we have

$$\sum_{j=1}^{l-1} (\theta_j - \hat{\lambda}_j) a_{i_j} - \hat{\lambda}_l a_{i_l} = 0.$$

This is a contradiction since $\hat{\lambda}_l \neq 0$.
The lemma has been proved. ∎

DEFINITION 21. *Assume $\mathcal{C} = \{\nabla f_{i_0}, \cdots, \nabla f_{i_l}\}$ is a cadre. The multipliers $\{\lambda_j\}$ satisfying*

$$\sum_{j=0}^{l} \lambda_j \nabla f_{i_j} = 0, \quad \lambda_j \neq 0, \quad j = 0, 1, \cdots, l, \tag{4.2}$$

are unique up to scaling. We normalise the multipliers as follows

$$\begin{cases} \sum_{j=0}^{l} \lambda_j = 1 & \textit{if } \sum_{j=0}^{l} \lambda_j \neq 0; \\ \lambda_0 = 1 & \textit{if } \sum_{j=0}^{l} \lambda_j = 0; \end{cases} . \tag{4.3}$$

We call such a normalised set $\{\lambda_j\}_0^l$, which is unique, the **cadre multipliers** *associated with $\mathcal{C}$.*

The relation (4.2) is also called the **characteristic relation** (cf. (3.7)).

Cadre multipliers are essentially different from the Lagrangian multipliers which are used in optimization. The Lagrangian multipliers are only defined at a stationary point. The cadre multipliers are associated with any cadre, and are not even necessarily based upon the active functions. Hence, we deliberately use the term cadre multipliers instead of just multipliers in order to differentiate them from the Lagrangian multipliers.

In the linear Chebyshev theory on $\Re^n$, it is clear that, in addition to the concept of degree of solvency which is always n, the alternating sign property is also important. This is captured by the concept of the reference function in the linear Chebyshev case. Now, we generalise this concept as follows.

DEFINITION 22. *A function* $\phi(t) \stackrel{\text{def}}{=} \phi(x_0, t)$ *is called a* **reference function** *at the point* x_0, *for a one dimensional dimensional continuous Chebyshev problem (3.8), on the point cadre* $C = \{t_{i_0}, \cdots, t_{i_l}\}$, *if and only if*

$$\begin{cases} sgn(f_{i_j}(x_0)) = sgn(\lambda_j) & \text{for } j = 0, \cdots, l, \qquad \text{or} \\ sgn(f_{i_j}(x_0)) = -sgn(\lambda_j) & \text{for } j = 0, \cdots, l, \end{cases}$$

where $\{\lambda_j\}_0^l$ *is the set of the cadre multipliers and* $f_{i_j}(x_0)$ *is the residual* $y(t_{i_j}) - \phi(x_0, t_{i_j})$. *A reference function is called a* **levelled reference function** *if the absolute residuals at each point in the cadre are the same.*

In general, it is not known whether the alternating sign property holds for a reference function on the cadre on which it is defined. Intuitively, it can be considered that we wish to know the number of "peaks" that characterise the optimal solution and moreover, from the geometry, it is natural that these peaks alternate in sign because in some sense one needs to average the maximum peaks.

In the discrete case, a similar concept is introduced.

DEFINITION 23. *The set of functions* $\{f_{i_j}(x),\ j = 0, 1, \cdots, l\}$ *is called a* **reference set** *of the discrete Chebyshe problem at* y *if and only if* $C = \{\nabla f_{i_0}(y), \cdots, \nabla f_{i_l}(y)\}$ *is a cadre and*

$$\begin{cases} sgn(f_{i_j}(y)) = sgn(\lambda_j) & \text{for } j = 0, \cdots, l, \qquad \text{or} \\ sgn(f_{i_j}(y)) = -sgn(\lambda_j) & \text{for } j = 0, \cdots, l, \end{cases}$$

where $\{\lambda_j\}_0^l$ *is the set of the cadre multipliers. Furthermore, the reference set is called a* **levelled reference set** *if the magnitude of each function is the same, viz.,*

$$|f_{i_j}(y)| = |f_{i_k}(y)|, \qquad \text{for any } i_j, i_k \in C.$$

REMARK 24. *The concept of reference set is totally different from the concept of reference. A reference set is a set of functions of the parameter* x. *A reference however, refers to a set of values of the variable* t. *A reference set corresponds to a special cadre on which each residual has the same sign as the corresponding cadre multiplier. A cadre however, is an extension of the reference concept to nonlinear problems.*

REMARK 25. *Suppose we have a discretised problem. A reference set* $\{f_{i_j}(x)\}$ *then corresponds to the functions* $\{\phi(x, t_{i_j})\}$ *obtained from a reference function* $\{\phi(x, t)\}$ *on the point cadre* $\{t_{i_j}\}$. *A reference set is therefore a discretisation of a reference function on the cadre.*

The definition of a reference set can also be extended to the general minimax problem.

DEFINITION 26. *The functions* $\{f_{i_j}(x),\ j = 0, \cdots, l\}$ *are said to be locally forming a* **reference set** *of a minimax problem (1.2) if* $\mathcal{C} = \{\nabla f_{i_0}, \cdots, \nabla f_{i_l}\}$ *is a cadre such that*

1. *The cadre multipliers* $\{\lambda_j\}$ *satisfy* $\lambda_j > 0,\ j = 0, \cdots, l$*;*
2. *The functions* $\{f_{i_j}(x)\}$ *all have the same sign.*

The reference set is further called a **levelled reference set** *if the value of each function is the same, viz.,*

$$f_{i_j}(x) = f_{i_k}(x), \qquad \textit{for any } i_j, i_k \in \mathcal{C}.$$

It can be proved easily that if a discrete Chebyshev problem is considered as a minimax problem, Definition 26 reduces to Definition 23. Note that the correspondence $f_{i+m}(x) = -f_i(x)$ is taken care of since we must have that all $\lambda_j > 0$.

As we have stated in § 3.1, under the classical Haar condition, the best linear Chebyshev approximation is a levelled reference function with the maximum deviation. This can be generalised to the nonlinear Chebyshev problem.

Consider a linear Chebyshev approximation problem under the Haar condition. Then, given any point cadre (equivalent to a reference in this case), there exists a unique levelled reference function defined on the point cadre and we say that the reference function $\phi(x, t)$ is *spanned* on the point cadre $\{t_0, \cdots, t_n\}$. For the nonlinear Chebyshev problem, however, whether a set of points forms a point cadre depends on the parameter x as well. Thus, the cadre itself becomes a local property. *Furthermore, it may or may not be possible to establish a levelled reference set upon the cadre.*

We have generalized the reference and reference function concepts to a general discrete Chebyshev problem and the minimax problem. Moreover, we believe that these concepts continue to characterise the best Chebyshev approximation.

For the linear Chebyshev approximation problem with the Haar condition, it is trivial to see that a reference set uniquely corresponds to a function whose errors alternate in signs on $n + 1$ points.

5. Optimality Conditions for a Local Minimum of a Minimax Problem.

In this section, we review the optimality conditions for minimax problems. We also establish more precise first order necessary conditions and hence recognise that the structure and characterization which are important to the best Chebyshev approximation carries over to a local minimum of a minimax problem.

THEOREM 27. *(First Order Necessary Conditions)*
If x^* *is a local minimizer of (1.2) then there exist multipliers* $\{\lambda_i\}$ *such that*

$$\sum_{i \in \mathcal{A}(x^*, 0)} \lambda_i \nabla f_i(x^*) = 0 \tag{5.1}$$

$$\sum_{i\in\mathcal{A}(x^*,0)} \lambda_i = 1 \tag{5.2}$$

$$\lambda_i \geq 0, \qquad i \in \mathcal{A}(x^*,0). \tag{5.3}$$

DEFINITION 28. *Define a function*

$$L(x,\lambda) = \sum_{j\in M} \lambda_j f_j(x), \tag{5.4}$$

This function is then called the **Lagrangian function** *for the minimax problem (1.2).*

At any local minimum of (1.2), there exists a set of multipliers $\{\lambda_j\}$ satisfying the above first order necessary conditions (5.1), (5.2) and (5.3). These $\{\lambda_j\}$ are called **Lagrangian multipliers**.

These necessary conditions could be more precisely stated in the following theorem.

THEOREM 29. *Suppose x^* is an optimal solution to a general minimax problem. Then, there exists a set of $l+1$ functions such that*
(1). $f_{i_j}(x^*) = \psi(x^*), \quad j = 0,1,\cdots,l;$
(2). $\sum_{j=0}^{l} \lambda_j \nabla f_{i_j}(x^*) = 0;$
(3). $\lambda_j > 0, \quad j = 0,1,\cdots,l;$
(4). $rank([\nabla f_{i_0}(x^*),\cdots,\nabla f_{i_l}(x^*)]) = l.$

A similar theorem has been stated in [22] for linear Chebyshev approximation. A proof of Theorem 29 can be found in [15] or [29].

DEFINITION 30. *A point x^* is called a* **stationary point** *of a minimax problem if and only if there exists a set of $l+1$ functions such that the statements (1)-(4) of Theorem 29 are satisfied.*

In terms of the reference set, the necessary optimality conditions for a local minimum is summarised as follows.

THEOREM 31. *Suppose x^* is a local minimum for a minimax problem (1.2). Then, there exists a set of $l+1$ functions $\{f_{i_0}(x),\cdots,f_{i_l}(x)\}$ which is a levelled reference set at x^* on the cadre $C = \{\nabla f_{i_0}(x^*),\cdots,\nabla f_{i_l}(x^*)\}$ with the maximum deviation.*

Proof. This follows immediately from Definition 26, Lemma 20 and Theorem 29. ■

At an optimum, the cadre multipliers associated with the cadre $\{\nabla f_{i_0}(x^*),\cdots,\nabla f_{i_l}(x^*)\}$, $f_{i_j}(x^*) = \psi(x^*)$, are identical to the non-zero Lagrangian multipliers.

This theorem illustrates the importance of the levelled reference set for the computation of a local minimum for Chebyshev problems.

Although it is true that Theorem 31 is equivalent to the first order necessary conditions, it is a characterisation of the local solution from another angle. Through this characterisation and the experience with the linear problems, we realise the key properties an efficient algorithm should attempt to construct, namely the reference set. Hence, as has been done for linear Chebyshev problems, we hope to design an efficient algorithm for the nonlinear problem.

In terms of continuous Chebyshev approximations, a reference function is an approximation function which satisfies the first order necessary conditions on a cadre without requiring the function values on the reference to be equal to the maximum deviation. A levelled reference function is a reference function whose function values on the reference are equivalent. A solution or a best approximation function is a levelled reference function with the maximum error achieved on the reference.

Therefore, a natural approach to finding a solution is as follows. Firstly, a cadre is located. Then, a reference set is constructed based on the cadre. Thirdly, a reference set is levelled.

We believe the extensions are particularly important from the computational point of view. In § 7, we will show how to exploit these characterisations but we first give a definition of degeneracy.

6. Degenerate Chebyshev Problems. For Chebyshev problems, degeneracy is a common occurrence.

We consider the general minimax problems of the form

$$\min_{x \in \mathcal{D}} \max_{t \in \mathcal{T}} f(x,t) \tag{6.1}$$

where $\mathcal{D} \subset \Re^n$ and $\mathcal{T} \subset \Re$. Note that $\mathcal{T}$ can be either connected or not.

DEFINITION 32. *A minimax problem (6.1) is termed* **nondegenerate** *at a point x_0 if and only if either the set of all active function gradients form a cadre at x_0 or the gradients of the active functions $\{\nabla f(x_0, t_i), i \in \mathcal{A}(x_0, 0)\}$ are linearly independent.*

If, at a local minimum x^*, the minimax problem is nondegenerate, then the cadre uniquely determines a levelled reference function. Otherwise, a cadre does not completely determine the structure of the solution, although the characterisation is still on a cadre.

The definition of degeneracy is for any minimax problem, continuous or discrete, linear or nonlinear.

In [22], the *singular* continuous linear Chebyshev problem has been studied. A Chebyshev problem is called *singular* if, at the optimum, the set of extreme points has cardinality less than $n+1$. We point out that the notion of singular is essentially different from our concept of degenerate.

For degenerate problems, there may be several cadres included in the set of extremal points at the optimal solution. Some of them might not satisfy the optimality conditions, but at least one of them does.

7. Computational Exploitation of the Structure and Characterization. We have established, algebraically, the structure and characterization for the solution of a minimax problem. We now investigate the possibility of computationally exploiting these properties.

We consider algorithms which always decrease the maximum function $\psi(x)$.

We will show that, for a discrete Chebyshev problem, it is possible to exploit the above properties in the sense that we can attempt to construct a levelled reference set and decrease the maximum function at the same time.

It is assumed that a discrete nonlinear Chebyshev problem

$$\min_{x \in \Re^n} \max_{i \in m} |f_i(x)|$$

is considered as the minimax problem

$$\min_{x \in \Re^n} \max_{i \in M} f_i(x), \tag{7.1}$$

where

$$m = \{1, \cdots, m\},$$

$$M = \{1, 2, \cdots, m, m+1, \cdots, 2m\},$$

$$f_{i+m}(x) = -f_i(x), \qquad i = 1, \cdots, m,$$

although, from a computational point of view, this is done implicitly rather than explicitly. The reference set is defined according to Definition 26.

7.1. The Importance of Exploitation. For the linear Chebyshev problem (with the Haar condition satisfied) and the varisolvent nonlinear Chebyshev approximation problem, a certain number of extreme points together with the alternating sign property uniquely determine a solution.

For function approximation problems or its discretised version, finding approximating functions with the structure of the best approximation is important. For other problems such as random problems, the importance is less obvious because two approximations with similar structure might be quite different since the continuity is lost.

As an example, let us consider the second Remez algorithm. Assuming the classical Haar condition, this algorithm for continuous linear Chebyshe approximation is a quadratically convergent method and usually finds a solution quickly (see, e.g. [8], page 97). We point out however, that the fast convergence of the Remez algorithm does not only come from the final iterations but is also a consequence of its global performance. At each iteration, the algorithm finds a set of $n+1$ points attaining the local minimum and local maximum residuals alternatively. These $n+1$ points always constitute a point cadre and moreover, the corresponding functions form a reference set. Hence the alternating sign property holds. Then a levelled reference function is found by requiring

$$\sum_{i=1}^{n} x_i \phi_i(t_{i_j}) + (-1)^j \xi = y(t_{i_j}),\ j = 0, 1, \cdots, n, \tag{7.2}$$

and x is in fact the Chebyshev solution on the reference $\{t_{i_j}\}_{i=0}^{n}$.

It is easy to observe that the structure of the characterisation of the best approximation has been exploited in the second Remez algorithm.

The general Remez algorithm is very expensive because it requires finding $n+1$ points upon which the errors achieve the local maximum and local minimum residuals alternately. In addition, it is also difficult to extend the general Remez algorithm to solve nonlinear Chebyshev problems (see, for example, [31], page 199). For linear Chebyshev problems, the number of points an approximating function can uniquely interpolate is always n. For continuous nonlinear Chebyshev approximations, however, the degree of the approximating function $\phi(x, t)$ varies with the parameter x. Since the Remez algorithm requires the degree of the function $\phi(x, t)$ to be known in advance, for nonlinear Chebyshev problems, it could be applied in a straightforward manner only when the approximating function $\phi(x, t)$ is a unisolvent function (which always has degree n).

In spite of the difficulties and the cost of the Remez algorithm, the fast convergence of the second Remez algorithm for continuous linear approximating functions or even for the unisolvent family does demonstrate the importance of exploiting the structure of the best approximations in algorithms for continuous Chebyshev problems.

Our experience with the linear Chebyshev problem [5] has also led us to believe that, for Chebyshev problems, it is important for an efficient algorithm to utilise, whenever possible, the structure and characterisation of the solution. In particular, constructing a reference function and its levelling are both important steps of the solution finding process.

7.2. The Working Set. If we assume any stationary point is a nondegenerate local minimum, finding a local minimum of the Chebyshev problem is equivalent to locating a levelled reference set including all the active functions.

DEFINITION 33. *A* **working set** $\mathcal{W} = \{i_0, i_1, \cdots, i_l\}$ *at a given point x is an index set of the functions which includes those of all the ϵ-active functions and is such that the following Jacobian matrix*

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

evaluated at x, is of full rank.

Thus, in the nondegenerate case, a point x^* is a stationary point if and only if there exists a working set at x^* which forms a levelled reference set.

Ideally, the best working set, $\mathcal{W}$, approximating the structure at optimality is some reference set which includes all the current active functions. This corresponds to solving the problem

$$\min_{\lambda} \| \sum_{i \in M} \lambda_i \nabla f_i \|_2$$

subject to

$$\sum_{i \in M} \lambda_i = 1, \quad \lambda_i \geq 0, \quad \forall i \in M,$$

$$\lambda_i > 0, \quad \forall i \in \mathcal{A}(x^k, 0).$$

If the minimum value is zero, $\mathcal{W} = \{ \ i \mid \lambda_i > 0 \ \}$ will be the desired working set. Unfortunately, it is relatively expensive to solve this problem and the existence of such a set is in general not guaranteed.

By Definition 26, a reference set is a function set, the gradients of which form a cadre and the signs of the functions are the same. We assume, for now, that we have ways of constructing the working set and we are able to identify whether a cadre has been located. Thus, assume that we have located a working set whose gradients form a cadre. Now we are going to see if we can construct a levelled reference set based on this cadre.

7.3. Levelling a Reference Set. Assume $\mathcal{W} = \{i_0, i_1, \cdots, i_l\}$ is a working set and all the functions in $\mathcal{W}$ form a reference set, not levelled. Then, moving along the direction v which is a solution to

$$f_{i_j}(x) + \nabla f_{i_j}^T v + \xi = 0 \tag{7.3}$$

levels all the functions in $\mathcal{W}$, up to the first order. Equivalently, the direction v can be defined by

$$v = -A(A^T A)^{-1} \Phi(x), \tag{7.4}$$

where

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

$$\Phi(x) = \begin{bmatrix} f_{i_0}(x) - f_{i_1}(x) \\ f_{i_0}(x) - f_{i_2}(x) \\ \vdots \\ f_{i_0}(x) - f_{i_l}(x) \end{bmatrix},$$

and $i_0 \in \mathcal{A}(x, 0)$. We call v a *vertical direction.* If a unit step along v is taken,

$$\Phi(x) + A^T v = 0,$$

then the functions in $\mathcal{W}$ would all have the same value, up to first order.

In this section, we will prove that the vertical direction obtained from a reference set is always a descent direction. First, we describe the following result.

LEMMA 34. *Let K be the convex hull of $\{a_i\}_{i=0}^{l}$. The system of linear inequalities:*

$$a_i^T x < 0, \qquad i = 0, \cdots, l$$

is inconsistent if and only if $0 \in K$.

Proof. This follows from Gordan's theorem (see page 31, [17]). ∎

Essentially, the lemma says that if zero is in the convex hull of a vector set $\{a_i\}_{i=0}^{l}$, there exists no direction which decreases or increases all the linear functions $\{a_i^T x,\ i \in K\}$ simultaneously.

Assume $\mathcal{W} = \{\mu, i_1, \cdots, i_l\}$. Thus $l \leq n$. It is clear that if $\mathcal{W}$ corresponds to a reference set, then $0 \in \mathrm{Conv}(\nabla f_\mu, \nabla f_{i_1}, \cdots, \nabla f_{i_l})$. On the other hand, if $0 \in \mathrm{Conv}(\nabla f_\mu, \nabla f_{i_1}, \cdots, \nabla f_{i_l})$ and $\{\nabla f_\mu, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}$ is a cadre and the functions in $\mathcal{W}$ have the same sign, then $\mathcal{W}$ is a reference set.

The importance of the lemma can be appreciated as follows. Suppose $\mathcal{W}$ is a working set corresponding to a reference set. If we level the functions in $\mathcal{W}$, we cannot do this by increasing or decreasing all the functions in $\mathcal{W}$ together. Thus, the only way in which these functions can be levelled is by decreasing the functions that are maximum and increasing some functions that are not.

Now we obtain the following very important property of the Chebyshev problem that justifies moving along the vertical direction when a reference set has been located.

LEMMA 35. *Suppose the functions in $\mathcal{W}$ form a reference set which includes all the current active functions. Then, the vertical direction defined from $\mathcal{W}$ is a descent direction for all the active functions if the reference set is not levelled.*

Proof. Assume $\mathcal{W} = \{\mu, i_1, \cdots, i_l\}$, where $f_\mu(x)$ attains the maximum deviation at x. Therefore,

$$f_i(x) \leq f_\mu(x), \qquad i \in \mathcal{W}.$$

Furthermore, because $\mathcal{W}$ is not a levelled reference set, there exists at least one ν such that

$$f_{i_\nu}(x) < f_\mu(x), \qquad i_\nu \in \mathcal{W}.$$

Since $\mathcal{W}$ is a reference set, there exist $\{\lambda_j\}_{j=0}^{l}$ satisfying

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

where

$$\lambda_j > 0, \qquad j = 0, \cdots, l, \quad \text{and} \quad \sum_{j=1}^{l} \lambda_j + \lambda_0 = 1.$$

Therefore

$$\nabla f_\mu = \sum_{j=1}^{l} \lambda_j (\nabla f_\mu - \nabla f_{i_j}).$$

By definition of the vertical direction (7.4),

$$\begin{aligned} \nabla {f_\mu}^T v &= \sum_{j=1}^{l} \lambda_j (\nabla f_\mu - \nabla f_{i_j})^T v \\ &= -\sum_{j=1}^{l} \lambda_j (f_\mu - f_{i_j}) \\ &< 0. \end{aligned}$$

Hence v is a descent direction for $f_\mu(x)$.

For any $f_k = f_\mu$, $k \in \mathcal{W}$ by hypothesis, and

$$(\nabla f_\mu - \nabla f_{i_j})^T v = 0, \qquad \text{where } k = i_j.$$

Thus

$$\nabla {f_{i_j}}^T v = \nabla {f_\mu}^T v < 0$$

and v is a descent direction for all the active functions. ∎

The fact that the vertical direction which levels functions in a reference set is also a descent direction is an important property, exploitation of which is desirable in an efficient algorithm.

We conclude this section by stating the following corollary of Lemma 35.

COROLLARY 36. *The direction which levels a reference set $\mathcal{W}$*

$$f_\mu - f_{i_j} + (\nabla f_\mu - \nabla f_{i_j})^T d = 0$$

always decreases the maximum deviation $\Delta = \max_{j \in \mathcal{W}} f_j(x)$ of this reference set.

7.4. Construction of Reference Sets. If the working set $\mathcal{W}$ corresponds to a cadre $\mathcal{C}$ but not a reference set, we want to be able to construct a reference set.

Assume $\mathcal{W} = \{i_0, i_1, \cdots, i_l\}$. By Definition 33, $\mathcal{A}(x, \epsilon) \subseteq \mathcal{W}$. If $\mathcal{W}$ is a cadre but yet not a reference set, by Definition 26, this means that either a function value or a multiplier has the incorrect sign, or both.

If $\lambda_j > 0$, $j = 0, \cdots, l$, and some function $f_{i_\nu}(x)$ has the wrong sign, then, as we will see below, (Lemma 38) simply continuing to level the functions will lead to a reference set.

Assume $\mathcal{W} = \{i_0, i_1, \cdots, i_l\}$ and $\mathcal{C} = \{\nabla f_{i_0}, \cdots, \nabla f_{i_l}\}$ is a cadre. There exists $\{\lambda_j\}$ such that

$$\sum_{j=0}^{l} \lambda_j \nabla f_{i_j} = 0,, \quad \lambda_j \neq 0, \quad j = 0, \cdots, l,$$

and

$$\sum_{j=0}^{l} \lambda_j = \begin{cases} 1 & \text{if } \sum_{j=0}^{l} \lambda_j \neq 0; \\ 0 & \text{otherwise.} \end{cases}$$

Suppose $\lambda_k < 0$. This suggests that $f_{i_k}(x)$ should be exchanged with $-f_{i_k}(x)$. It is important to note that the new set $\hat{\mathcal{W}}$ thus obtained is not a reference set because the function values $\{f_{i_j}(x),\ i_j \in \hat{\mathcal{W}}\}$ do not have the same signs (cf. Definition 26).

If the working set $\mathcal{W} = \{\mu, i_1, \cdots, i_l\}$ is a non-reference-set cadre and $\{\lambda_0, \cdots, \lambda_l\}$ are the cadre multipliers, we define a new set $\hat{\mathcal{W}}$ as follows:

$$(7.5)\qquad \begin{cases} \text{Let } \hat{\mathcal{W}} \leftarrow \emptyset. & \text{For } j = 0, \cdots, l \text{ do} & \\ & \hat{\mathcal{W}} \leftarrow \hat{\mathcal{W}} \cup \{i_j\} & \text{if } \lambda_j > 0, \\ & \hat{\mathcal{W}} \leftarrow \hat{\mathcal{W}} \cup \{i_j + m\} & \text{if } \lambda_j < 0 \text{ and } i_j \leq m, \\ & \hat{\mathcal{W}} \leftarrow \hat{\mathcal{W}} \cup \{i_j - m\} & \text{if } \lambda_j < 0 \text{ and } i_j > m. \end{cases}$$

It is clear that $\hat{\mathcal{C}} = \{\ \nabla f_i \mid i \in \hat{\mathcal{W}}\}$ is also a cadre.

Now we demonstrate that, when $\mathcal{W}$ is not a cadre, levelling the functions in $\hat{\mathcal{W}}$ (up to first order) is constructive.

LEMMA 37. *Suppose (7.1) is a discrete linear Chebyshev problem. Suppose* $\{a_{i_0}, \cdots, a_{i_l}\}$ *is a cadre, where* $a_{i_j} = \nabla f_{i_j}$ *and* $\mathcal{W} = \{i_0, \cdots, i_l\}$ *includes the indices of all active functions. Assume* $\hat{\mathcal{W}}$ *is obtained from (7.5). Then, after levelling the functions in* $\hat{\mathcal{W}}$*, i.e.,*

$$a_{k_j}^T x + \xi = b_{k_j}, \qquad k_j \in \hat{\mathcal{W}}$$

(or equivalently $a_{i_j}^T x + \sigma_j \xi = b_{i_j},\ i_j \in \mathcal{W},$ *where* $\sigma_j = sgn(\lambda_j)$ *) the functions in* $\hat{\mathcal{W}}$ *form a levelled reference set.*

Proof. Assume $\mathcal{W} = \{i_0, i_1, \cdots, i_l\}$. From (7.5), $\hat{\mathcal{W}} = \{k_0, k_1, \cdots, k_l\}$ where

$$k_j = \begin{cases} i_j & \text{if } \lambda_j > 0 \\ i_j + m & \text{if } \lambda_j < 0 \text{ and } i_j \leq m \\ i_j - m & \text{if } \lambda_j < 0 \text{ and } i_j > m \end{cases}.$$

By the definition of k_j, there exist positive multipliers $\{\theta_i = \sigma_i \lambda_i > 0, \quad i = 0, \cdots, l\}$ such that

$$(7.6)\qquad \sum_{j=0}^{l} \theta_j a_{k_j} = 0.$$

Initially, $\{f_{k_0}(x), \cdots, f_{k_l}(x)\}$ is not a reference set because the condition $f_{k_j}(x)\psi(x) > 0$ for all $k_j, j = 0, \cdots l$ is violated. Consider the linear system of equations

$$a_{k_j}^T x + \xi = b_{k_j}, \qquad k_j \in \hat{\mathcal{W}}. \tag{7.7}$$

Since $a_{k_j} = \sigma_j a_{i_j}$, it is clear the transpose of the coefficient matrix

$$\bar{A}^T = \begin{bmatrix} a_{i_0} & \cdots & a_{i_l} \\ \sigma_0 & \cdots & \sigma_l \end{bmatrix}$$

has the same rank as that of

$$\hat{A}^T = \begin{bmatrix} a_{k_0} & \cdots & a_{k_l} \\ 1 & \cdots & 1 \end{bmatrix}.$$

Furthermore, from (7.6), $\{\nabla f_{k_0}, \cdots, \nabla f_{k_l}\}$ is a cadre with the cadre multipliers summing to 1. Hence, the solution θ_j to

$$\sum_{j=0}^{l} \theta_j a_{k_j} = 0$$

is unique up to a scaling. Hence, there does not exist θ_j with some $\theta_j \neq 0$ such that

$$\sum_{j=0}^{l} \theta_j a_{k_j} = 0, \quad \text{where} \quad \sum_{j=0}^{l} \theta_j = 0.$$

Thus, $\hat{A}$, and hence $\bar{A}$, is of full rank $l+1$. Since there are $l+1$ equations, we know the above linear system has at least one solution and the functions in $\hat{\mathcal{W}}$ can be levelled.

Assume $\hat{x}$ is a solution to (7.7). After the functions in $\hat{\mathcal{W}}$ are levelled, the residuals $f_{k_j}(\hat{x}) = a_{k_j}^T \hat{x} - b_{k_j}$ have the same sign as ξ. Thus, $\hat{\mathcal{W}}$ is a reference set at $\hat{x}$. This completes the proof. ∎

The lemma is true only in the linear case. For a nonlinear discrete Chebyshev problem, since the gradients will be changed, the characteristic relation (7.6) does not necessarily hold. However, for a nonlinear problem, it is still a constructive step towards finding a reference set since it could be considered that the characteristic relation holds approximately, at least in the neighbourhood of small changes.

From $\mathcal{W}$, a new set $\hat{\mathcal{W}}$ which would be a levelled reference set if the function values were equal, is defined by (7.5). Thus, we attempt to construct the reference set by levelling the functions in $\hat{\mathcal{W}}$ via satisfying

$$f_{k_j}(x) + \xi = 0, \qquad k_j \in \hat{\mathcal{W}}.$$

Let $\sigma_j = \text{sgn}(\lambda_j)$, for $j = 0, \cdots, l$. From the definition of $\hat{\mathcal{W}}$, the above is equal to

$$\sigma_j f_{i_j}(x) + \xi = 0, \qquad i_j \in \mathcal{W}.$$

Thus, the direction v defined by

$$(7.8) \qquad [\nabla f_\mu - \sigma_0\sigma_j \nabla f_{i_j}]^T v = -(f_\mu - \sigma_0\sigma_j f_{i_j}), \qquad i_j \in \mathcal{W},\ i_j \neq i_0,$$

attempts to level the functions in $\hat{\mathcal{W}}$. We refer to $f_\mu(x)$ as a *representative function.*

Equivalently, we can write (7.8) as

$$(7.9) \qquad \begin{cases} \hat{A}v = -\hat{\Phi} \\ \hat{A} = [\nabla f_\mu - \sigma_0\sigma_1 \nabla f_{i_1}, \cdots, \nabla f_\mu - \sigma_0\sigma_l \nabla f_{i_l}] \\ \hat{\Phi} = [f_\mu - \sigma_0\sigma_1 f_{i_1}, \cdots, f_\mu - \sigma_0\sigma_l f_{i_l}]^T \end{cases} .$$

Next, we will see that this direction, v, is a descent direction for the maximum function $\psi(x)$.

LEMMA 38. *Suppose $\mathcal{C} = \{\nabla f_\mu, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}$ is a non-reference-set cadre with cadre multipliers $\{\lambda_j\}$ summing to one and $f_\mu(x)$ achieves the current maximum deviation for (7.1). Then, the vertical direction defined on $\mathcal{W} = \{\mu, i_1, \cdots, i_l\}$ by (7.9) decreases all the active functions, assuming $\mathcal{A}(x,0) \subseteq \mathcal{W}$.*

Proof. Suppose $f_\mu(x) = \psi(x)$. By assumption, $\mathcal{C} = \{\nabla f_\mu, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}$ is a cadre and there exist cadre multipliers $\{\lambda_j \neq 0\}$ such that:

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

Denote

$$\sigma_j = \mathrm{sgn}(\lambda_j), \qquad j = 0, 1, \cdots, l.$$

Consequently

$$(7.10) \qquad \lambda_0 \nabla f_\mu + \sum_{j=1}^{l} \sigma_0\sigma_j\lambda_j(\sigma_0\sigma_j \nabla f_{i_j}) = 0.$$

Denote

$$s \;=\; \lambda_0 + \sum_{j=1}^{l} \sigma_0\sigma_j\lambda_j.$$

Then $\mathrm{sgn}(s) = \sigma_0$. Substituting for λ_0, dividing both sides of (7.10) by s and denoting

$$\theta_j = \frac{\sigma_0\sigma_j\lambda_j}{s} > 0, \qquad j = 0, \cdots, l,$$

we obtain that

$$(7.11) \qquad \sum_{j=0}^{l} \theta_j = 1, \quad \text{and} \quad \nabla f_\mu = \sum_{j=1}^{l} \theta_j(\nabla f_\mu - \sigma_0\sigma_j \nabla f_{i_j}).$$

Using (7.8), we have

$$\nabla f_\mu{}^T v = -\sum_{j=1}^{l} \theta_j (f_\mu - \sigma_0 \sigma_j f_{i_j}). \tag{7.12}$$

We now prove $\nabla f_\mu^T v < 0$ by considering the following two cases:

Case–I $f_{i_j} = -f_\mu$, for some $i_j \in \mathcal{W}$.

In this case, define

$$k_j = \begin{cases} i_j + m & \text{if } i_j \le m, \\ i_j - m & \text{otherwise.} \end{cases}$$

Thus

$$k_j \in \mathcal{A}(x, 0) \subseteq \{\mu, i_1, \cdots, i_l\}.$$

Since $\nabla f_{k_j} = -\nabla f_{i_j}$, $\{\nabla f_{k_j}, \nabla f_{i_j}\}$ is a cadre. Hence $l = 1$ (in (7.11)) and

$$\mathcal{W} = \{\mu, i_1\}, \quad \mu = k_1 \quad \text{and} \quad \sigma_0 = \sigma_1 = 1.$$

Hence

$$\begin{aligned} \nabla f_\mu^T v &= \frac{1}{2}(\nabla f_\mu - \nabla f_{i_1})^T v \\ &= -\frac{1}{2}(f_\mu - f_{i_1}) \\ &= -f_\mu < 0. \end{aligned}$$

Case–II $f_{i_j} > -f_\mu$, for all $i_j \in \mathcal{W}$.

It is clear, from (7.12), that $\nabla f_\mu{}^T v < 0$ unless

$$f_\mu = \sigma_0 \sigma_j f_{i_j}, \qquad j = 1, \cdots, l.$$

Since $f_\mu(x) > 0$ and $-f_\mu < f_{i_j} \le f_\mu$, we know $\nabla f_\mu^T v < 0$, because, otherwise, $\sigma_0\sigma_j > 0, j = 1, \cdots, l$ and $f_\mu = f_{i_j}, j = 1, \cdots, l$. This means $\mathcal{C}$ is a levelled reference set, contradicting our hypothesis on $\mathcal{C}$.

Note that $f_{i_j} < -f_\mu$ is impossible by definition of f_μ and the fact that both f_{i_j} and $-f_{i_j}$ are indeed amongst the $f_i's$.

By the definition of the vertical direction (7.8) and (7.12), for any $i_\nu \in \mathcal{W}$,

$$\sigma_0 \sigma_\nu \nabla f_{i_\nu}{}^T v = f_\mu - \sigma_0 \sigma_\nu f_{i_\nu} - \sum_{j=1}^{l} \theta_j (f_\mu - \sigma_0 \sigma_j f_{i_j}).$$

For any ν such that $f_{i_\nu} = \triangle$, $i_\nu \in \mathcal{W}$. If $\sigma_\nu \sigma_0 = 1$, it follows that $\hat{\Phi}_\nu = 0$ and we have

$$\nabla f_{i_\nu}{}^T v = \nabla f_\mu{}^T < 0.$$

If $\sigma_\nu \sigma_0 = -1$,

$$\begin{aligned}
-\nabla f_{i_\nu}{}^T &= f_\mu - \sigma_0 \sigma_\nu f_{i_\nu} - \sum_{j=1}^{l} \theta_j (f_\mu - \sigma_0 \sigma_j f_{i_j}) \\
&> 2\Delta(1 - \sum_{j=1}^{l} \theta_j) \\
&= 2\Delta\theta_0 \\
&> 0.
\end{aligned}$$

In conclusion, all the active functions will be decreased. This completes the proof. ∎

Assume $\mathcal{W} = \{i_0, \cdots, i_l\}$ is the working set and $\mathcal{C} = \{\nabla f_{i_0}, \cdots, \nabla f_{i_l}\}$ is a cadre. Associated with $\mathcal{C}$, there exist unique cadre multipliers $\{\lambda_i\}$ such that

$$\sum_{j=0}^{l} \lambda_j \nabla f_{i_j} = 0. \tag{7.13}$$

We now establish an auxiliary lemma.

LEMMA 39. *Assume* $\{\lambda_j\}_{j=0}^{l}$ *is a set of real numbers which are not all zero and* $\sum_{j=0}^{l} \lambda_j$ *equals either* 1 *or* 0. *Define*

$$\theta_j = \frac{\sigma_0 \sigma_j \lambda_j}{s} > 0, \qquad j = 0, \cdots, l, \tag{7.14}$$

and

$$s = \lambda_0 + \sum_{j=1}^{l} \sigma_0 \sigma_j \lambda_j. \tag{7.15}$$

Then,

$$1 - 2 \sum_{\sigma_0 \sigma_j < 0} \theta_j = \begin{cases} 0 & \text{if } \sum_{j=0}^{l} \lambda_j = 0; \\ \frac{1}{s} & \text{if } \sum_{j=0}^{l} \lambda_j = 1. \end{cases} \tag{7.16}$$

Proof. By assumption,

$$\lambda_0 + \sum_{i=1}^{l} \lambda_i = \delta, \quad \text{where } \delta = 0 \text{ or } 1.$$

Define

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

This means

$$\sigma_0(\sigma_0\lambda_0 + \sigma_0 \sum_{j=1}^{l} \lambda_i) = \delta, \quad \text{or}$$

$$\sigma_0(\sigma_0\lambda_0 + \sum_{\sigma_0\sigma_i>0, i\neq 0} |\lambda_i| + \sigma_0 \sum_{\sigma_0\sigma_i<0} \lambda_i) = \delta.$$

Since

$$\begin{aligned} -\sum_{\sigma_0\sigma_i<0} |\lambda_i| &= \sum_{\sigma_0\sigma_i<0} \sigma_0\sigma_i\sigma_i\lambda_i \\ &= \sum_{\sigma_0\sigma_i<0} \sigma_0\lambda_i, \end{aligned}$$

we have that

$$\sigma_0(\sigma_0\lambda_0 + \sum_{i=1}^{l} |\lambda_i| + 2\sigma_0 \sum_{\sigma_0\sigma_i<0} \lambda_i) = \delta. \tag{7.17}$$

From (7.15),

$$s = \lambda_0 + \sigma_0 \sum_{i=1}^{l} |\lambda_i|$$

which with (7.17) gives

$$\sigma_0(\sigma_0 s - 2\sigma_0 \sum_{\sigma_0\sigma_i<0} \sigma_0\sigma_i\lambda_i) = \delta.$$

Dividing both sides by s we have

$$(1 - 2\sum_{\sigma_0\sigma_i<0} \frac{\sigma_0\sigma_i\lambda_i}{s}) = \frac{\delta}{s}.$$

Thus,

$$1 - 2\sum_{\sigma_0\sigma_j<0} \theta_j = \begin{cases} 0 & \text{if } \sum_{j=0}^{l} \lambda_j = 0; \\ \frac{1}{s} & \text{if } \sum_{j=0}^{l} \lambda_j = 1 \end{cases}$$

holds. This completes the proof. ∎

LEMMA 40. *Suppose $\mathcal{W} = \{\mu, i_1, \cdots, i_l\}$ consists only of indices of the active functions. Assume further that $\mathcal{C} = \{\nabla f_\mu, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}$ is a cadre. Assume the vertical direction v is determined from $\mathcal{W}$ as in (7.9). Then:*

1. *all the active functions with negative multipliers will be decreased more rapidly than all the other active functions if the cadre multipliers sum to one, i.e. $\sum_{j=0}^{l} \lambda_j = 1$;*

2. all the active functions are decreased equally (up to first order) if the cadre multipliers sum to zero, i.e., $\sum_{j=0}^{l} \lambda_j = 0$.

Proof. By assumption, $\mathcal{C} = \{\nabla f_\mu, \nabla f_{i_1}, \cdots, \nabla f_{i_l}\}$ is a cadre. Let $\triangle = \psi(x)$ and $\sigma_j = \mathrm{sgn}(\lambda_j)$.

As for (7.12), there exists multipliers $\{\theta_i\}$ such that

$$\nabla f_\mu{}^T = -\sum_{j=1}^{l} \theta_j (f_\mu - \sigma_0 \sigma_j f_{i_j}), \text{ where } \sum_{j=0}^{l} \theta_j = 1, \quad \theta_j > 0, \quad j = 0, \cdots, l.$$

Using the definition of v in (7.8), we have

$$\sigma_0 \sigma_\nu \nabla f_{i_\nu}{}^T v = (f_\mu - \sigma_0 \sigma_\nu f_{i_\nu}) + \nabla f_\mu{}^T v.$$

Thus, for any $\sigma_0 \sigma_\nu = 1$, we have

$$\nabla f_{i_\nu}{}^T v = \nabla f_\mu{}^T v.$$

For any $\sigma_0 \sigma_\nu = -1$, we have,

$$\nabla f_{i_\nu}{}^T v = -(f_\mu + f_{i_\nu}) - \nabla f_\mu{}^T v.$$

Therefore

$$\begin{aligned}
(7.18) \qquad & -(f_\mu + f_{i_\nu}) - 2\nabla f_\mu{}^T v \\
= \; & -2\triangle + 2\sum_{j=1}^{l} \theta_j (f_\mu - \sigma_0 \sigma_j f_{i_j}) \\
= \; & -2\triangle(1 - \sum_{j=1}^{l} \theta_j (1 - \sigma_0 \sigma_j)) \; (\text{ note } f_\mu = f_{i_j} = \triangle\,) \\
= \; & -2\triangle(1 - 2\sum_{\sigma_0 \sigma_j < 0} \theta_j) \\
= \; & \begin{cases} 0 & \text{if } \sum_{j=0}^{l} \lambda_j = 0; \\ -\frac{2\triangle}{s} & \text{if } \sum_{j=0}^{l} \lambda_j = 1. \end{cases} \qquad (\text{ by (7.16) })
\end{aligned}$$

Thus,

$$\nabla f_{i_\nu}{}^T v = \nabla f_\mu{}^T v \quad \text{if } \textstyle\sum_{j=0}^{l} \lambda_j = 0,$$

Furthermore, since s has the same sign as σ_0, we have, when $\sum_{j=0}^{l} \lambda_j = 1$,

$$\begin{cases} \nabla f_{i_\nu}{}^T v < \nabla f_\mu{}^T v, & \text{for any } \nu,\ \sigma_0 \sigma_\nu < 0,\ \sigma_0 > 0 \\ \nabla f_{i_\nu}{}^T v > \nabla f_\mu{}^T v, & \text{for any } \nu,\ \sigma_0 \sigma_\nu < 0,\ \sigma_0 < 0 \end{cases}$$

i.e., all the active functions with negative multiplers are decreased faster than those with positive multipliers. This completes the proof. ∎

Thus, if $\mathcal{C}$ is a cadre with the cadre multipliers satisfying $\sum_{j=0}^{l} \lambda_j = 0,\ \lambda_j \neq 0$, v decreases all the active functions equally.

If, on the other hand, a cadre is composed of the active functions and $\sum_{j=0}^{l} \lambda_j = 1$, all the active functions with negative cadre multipliers will actually be decreased faster than the other active functions. This amounts to dropping activities in terms of the underlying mathematical programming problem.

Dropping active constraints is an important part of active set methods for general nonlinear programming problems. Whenever a minimum on a subspace determined by the active constraints is reached and one of the multipliers is negative, a direction is usually chosen to make such a constraint strictly inactive, i.e., to drop this constraint. If more than one constraint has a negative multiplier, it is usually not possible for an algorithm to delete all the corresponding constraints, because of the difficulty of finding a feasible descent direction. Therefore, for most algorithms, instead of checking the possibility of all the constraints with negative multipliers being removed simultaneously, only a single constraint is deleted. By contrast, we are able to exploit multiple droppings.

It is interesting to note that, moving along the vertical direction defined by (7.8) will drop the functions with negative multipliers independently of the index associated with the representative function. There is no need to choose another representative function even if the multiplier of the current representative function is negative. Therefore, the only requirement for a representative function is that it is a function with maximum deviation. If more than one such function exists, the choice is arbitrary.

We have looked at the vertical direction defined in (7.9) from two different points of view. The motivation for the vertical direction is to construct a (levelled) reference set, assuming a cadre has been found.

8. Summary. It is well known that a best linear Chebyshev approximation corresponds to a characteristic structure. Algorithms for computation of the linear solution have been successfully able to make use of this structure. It is not so well recognised that a solution of a nonlinear Chebyshev problem also possesses a rich structure and characterisation can be computationally exploited.

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

Nonlinear Chebyshev approximation theory indicates that, theoretically, for certain classes of nonlinear problems at least, useful characterisations still exist.

In this paper, we have generalized the basic concepts which are important in characterising the best linear Chebyshev approximation to nonlinear Chebyshev problems. These characterizations are particularly important computationally.

We have established the structure and characterisation of a solution to nonlinear discrete Chebyshev problems which is a generalisation of the characterisation for a best linear Chebyshe approximation. These generalisations are motivated by the linear theory and our experience with Chebyshev problems. We emphasize the importance of exploiting these properties. Moreover, we have provided with ways of making use of the characterisations when we have located a cadre, namely to construct a levelled reference set from a reference set or a reference set from a cadre.

In [10], we present an efficient algorithm which exploits the theory of this paper.

REFERENCES

[1] D. H. ANDERSON AND M. R. OSBORNE, *Discrete, nonlinear approximations in polyhedral norms: a Levenberg-like algorithm*, Numerische Mathematik, 28 (1977), pp. 167–170.

[2] J. W. BANDLER AND C. CHARALAMBOUS, *Practical least pth optimization of networks*, IEEE Trans. Micro. Theory and Techniques, MTT-20 (1972), pp. 834–840.

[3] 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.

[4] R. H. BARTELS, A. R. CONN, AND C. CHARALAMBOUS, *On Cline's direct method for solving overdetermined linear systems in the l_∞ sense*, SIAM J. Numer. Anal., 15 (1978), pp. 255–270.

[5] 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., (to appear).

[6] M. BRANNIGAN, *Discrete Chebyshev approximation with linear constraints*, SIAM J. Numer. Anal., 22 (1985), pp. 1–14.

[7] C. CHARALAMBOUS AND A. R. CONN, *An efficient method to solve the minimax problem directly*, SIAM J. Numer. Anal., 15 (1978), pp. 162–187.

[8] E. W. CHENEY, *Introduction to Approximation Theory*, Mc Graw-Hill, 1966.

[9] A. R. CONN, *An efficient second order method to solve the (constrained) minimax problem*, Tech. Rep. CORR 79-5, Dept. of Combinatorics and Optimization, University of Waterloo, Ontario, Canada, 1979.

[10] 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.

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

[12] R. FLETCHER, *A model algorithm for composite nondifferentiable optimization problems*, Math. Prog. Study, 17 (1982), pp. 67–76.

[13] J. HALD AND K. MADSEN, *Combined LP and Quasi-Newton methods for minimax optimization*, Math. Prog., 20 (1981), pp. 49–62.

[14] S. P. HAN, *Variable metric methods for minimizing a class of nondifferentiable functions*, Math. Prog., 20 (1981), pp. 1–13.

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

[16] K. MADSEN, *Minimax solution of nonlinear equations without calculating derivatives*, Mathematical Programming Study, 3 (1975), pp. 110–126.

[17] O. L. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, 1969.

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

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

[20] W. MURRAY AND M. L. OVERTON, *A projected Lagrangian algorithm for nonlinear minimax optimization*, SIAM. J. Sci. Stat. Comput., 1 (1980), pp. 345–370.

[21] M. R. OSBORNE AND G. A. WATSON, *An algorithm for minimax approximation in the nonlinear case*, Computing J., 12 (1968), pp. 63–68.

[22] ——, *A note on singular minimax approximation problem*, J. Math. Anal. and Appl., 25 (1969), pp. 692–709.

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

[24] J. R. RICE, *The Approximation of Functions 1, Linear Theory*, Addison-Wesley, 1964.

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

[26] ——, *The Approximation of Functions 2, Nonlinear Theory*, Addison-Wesley, 1969.

[27] E. L. STIEFEL, *Numerical methods of Tchebycheff approximation*, in Proceedings of a symposium, math. research center, Univ. of Wisconsin Press, 1959.

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

[29] G. A. Watson, *Chebyshev approximation with applicatons to the numerical solution of differntial equations*, PhD thesis, Australian National University, 1969.

[30] G. A. Watson, *The minimax solution of an overdetermined system of non-linear equations*, J. Inst. Math. Appl., 23 (1979), pp. 167–180.

[31] G. A. Watson, *Approximation Theory and Numerical Methods*, John Wiley, 1980.

[32] R. S. Womersley and R. Fletcher, *An algorithm for composite nonsmooth optimization problems*, J. Opt. Theory and Applis., 48 (1986), pp. 493–523.