MONTANA STATE UNIVERSITY

Applications of approximation theory to initial value problems
by Kenneth Leroy Wiggins

A thesis submitted in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY in Mathematics
Montana State University
© Copyright by Kenneth Leroy Wiggins (1974)

Abstract:
Define F[x](t) = f(t, x(t), x'(t), x(g(t)), x'(g(t))), and L[x] (t) = x''(t) - F[x] (t). Consider the IVP (*) L[x] = 0, x(0) = C0, x'(0) = C1 on I = [-α,α].

Set pkl(t) = C0 + c1 and, for n≥2, choose p''kn as the polynomial of degree at most k-2 that solves the minimization problem (MP) inf ||p - F[pkn-1]|| , P □ Qk-2 where Qk-2 is the set of polynomials of degree at most k-2 and ||•|| = supI |•|. Then pkn(t) = c0 + c1t + ∫0t(t-s)p''kn(s)ds.

There is a positive number α0 such that for a≤α0 the sequence {Pkn}n=1∞ has a cluster point pk. Furthermore5 if Pk is the set of all polynomials of degree at most k satisfying the initial conditions of (*), then there is a positive number α1≤α0 such that α≤α1 implies that pk is the unique polynomial of degree k or less that solves the non- linear MP inf ||L[p]|| . The sequence {pk}k=1∞ converges p□Pk uniformly to the solution of (*). The second algorithm of Remes may be employed to compute the sequence (pk). The above analysis extends the work of M. S. Henry, G. D. Allinger, and D. E. Olson.

APPLICATIONS OF APPROXIMATION THEORY TO INITIAL VALUE PROBLEMS

by

KENNETH LEROY WIGGINS

A thesis submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Mathematics

Approved:

Robert D. Engle
Head, Major Department

Myron S Henry
Chairman, Examining Committee

Henry L Parsons
Graduate Dean

MONTANA STATE UNIVERSITY
Bozeman, Montana

August, 1974

ACKNOWLEDGEMENT

I especially want to thank my advisor, Dr. Myron S. Henry, for his able guidance, patience and encouragement throughout my graduate school experience. I have benefited immensely from my association with him.

Thanks are due to my wife, JoAnn, for her enthusiastic encouragement and her constant support.

The careful proof reading and helpful comments by Dr. David V.V. Wend are most appreciated. Appreciation is also expressed to Dr. John A. Brown and Mr. Robert J. Dillon for their help with the programming aspect of this study and to Linda Mosness for her efficient typing of this manuscript.

TABLE OF CONTENTS

CHAPTER		PAGE
INTRODUCTION		1
I.	APPROXIMATION THEORY	5
1.1	Introduction	5
1.2	Interpolation	5
1.3	Convexity	8
1.4	Tchebycheff Approximation	15
II.	SECOND ORDER NONLINEAR ORDINARY DIFFERENTIAL EQUATIONS	42
2.1	A Specific Example Comparing the BAS to the SAS	43
2.2	Existence of a SAS	54
2.3	A Constructive Existence Theorem for Solutions of the IVP (2.0.1)	57
2.4	Comparison of the SAS to the BAS	65
2.5	Error Estimation	77
2.6	A Numerical Example	82
III.	LINEAR SECOND ORDER FUNCTIONAL DIFFERENTIAL EQUATIONS	85
3.1	Existence of a SAS	86
3.2	Computing the BAS	87
3.3	A Theorem Concerning Solutions and BAS's of the IVP (3.0.1)	91
3.4	A Numerical Example	94
IV.	NONLINEAR SECOND ORDER FUNCTIONAL DIFFERENTIAL EQUATIONS	96
4.1	Existence of a SAS	97
4.2	A Constructive Existence Theorem for Solutions of the IVP (4.0.1)	101
4.3	Comparison of the SAS to the BAS	110
4.4	Error Estimation	119
4.5	A Numerical Example	124

CHAPTER PAGE

V. NONLINEAR n-th ORDER FUNCTIONAL DIFFERENTIAL EQUATIONS 126

5.1 Theorems Concerning SAS's and BAS's 128
5.2 Conclusion 130

BIBLIOGRAPHY 131

APPENDIX I - 132

APPENDIX II - 135

LIST OF TABLES

TABLE PAGE

1 50

2 54

LIST OF FIGURES

FIGURES PAGE

1 48

2 52

3 53

ABSTRACT

Define $F[x](t) = f(t, x(t), x'(t), x(g(t)), x'(g(t)))$, and $L[x](t) = x''(t) - F[x](t)$. Consider the IVP (*) $L[x] = 0$; $x(0) = c_0$, $x'(0) = c_1$ on $I = [-\alpha,\alpha]$. Set $p_{k1}(t) = c_0 + c_1 t$ and, for $n \geq 2$, choose p''_{kn} as the polynomial of degree at most k-2 that solves the minimization problem (MP) $\inf_{p \in Q_{k-2}} \|p - F[p_{kn-1}]\|$, where Q_{k-2} is the set of polynomials of degree at most k-2 and $\|\cdot\| = \sup_I |\cdot|$. Then $p_{kn}(t) = c_0 + c_1 t + \int_0^t (t-s)p''_{kn}(s)ds$. There is a positive number α_0 such that for $\alpha \leq \alpha_0$ the sequence $\{p_{kn}\}_{n=1}^{\infty}$ has a cluster point p_k. Furthermore, if $\mathcal{P}_k$ is the set of all polynomials of degree at most k satisfying the initial conditions of (*), then there is a positive number $\alpha_1 \leq \alpha_0$ such that $\alpha \leq \alpha_1$ implies that p_k is the unique polynomial of degree k or less that solves the nonlinear MP $\inf_{p \in \mathcal{P}_k} \|L[p]\|$. The sequence $\{p_k\}_{k=1}^{\infty}$ converges uniformly to the solution of (*). The second algorithm of Remes may be employed to compute the sequence $\{p_k\}$. The above analysis extends the work of M. S. Henry, G. D. Allinger, and D. E. Olson.

Introduction

Consider the initial value problem (IVP)

$$(*) \qquad x''(t) = f(t, x(t), x'(t), x(g(t)), x'(g(t)))$$

$$x(0) = c_0,\ x'(0) = c_1,\ t \in I_\alpha = [-\alpha, \alpha],$$

where $f \in C(I_\alpha \times R^4)$, $g \in C(I_\alpha)$ and $g(I_\alpha) \subseteq I_\alpha$.

It is known that under certain conditions, the IVP (*) has a unique solution [1, 7, 10]. For certain functions $g(t)$, numerical methods [4] may be employed to find approximate solutions on discrete subsets of I_α. These numerical techniques are generalizations of the Euler Method and they require g to satisfy $g(t) \geq t$ for $t < 0$ and $g(t) \leq t$ for $t \geq 0$.

The focus of this investigation is to find approximate solutions of (*) from the set

$$\mathbf{P}_k = \{c_0 + c_1 t + a_2 t^2 + \ldots + a_k t^k : (a_2, a_3, \ldots, a_k) \in R^{k-1}\}.$$

The approximation is in the context of the Tchebycheff norm defined by $\|h\|_\alpha = \max_{I_\alpha} |h(t)|$ for each $h \in C(I_\alpha)$. For notational convenience, define

$F[u](t) = f(t, u(t), u'(t), u(g(t)), u'(g(t)))$ for each $u \in C^1(I_\alpha)$. A polynomial p_k^* which solves the minimization

problem $\inf_{p \in \mathcal{P}_k} \|p'' - F[p]\|_\alpha$ is called a <u>best simultaneous approximate solution</u> (BAS) of the IVP (*) from $\mathcal{P}_k$. The word "simultaneous" arises from the fact that x is replaced by p in the right and left sides of (*) simultaneously.

G. D. Allinger and M. S. Henry [1, 6] have shown that under certain conditions, for each k, a BAS p_k^* exists, the IVP (*) has a unique solution y, and

$$\lim_{k \to \infty} \|p_k^{*(i)} - y^{(i)}\|_\alpha = 0, \quad (i = 0,1,2).$$

The primary shortcoming of the BAS is that for $k > 2$, it is very difficult to compute. This is due to the fact that the minimization problem which defines the BAS is often highly nonlinear in the coefficients $a_2, a_3, \ldots, a_k$. To circumvent these computational difficulties, S. E. Weinstein and D. E. Olson [8] introduced the following linearization scheme. Choose $p_{k1}(t) = c_0 + c_1 t$ or some other appropriate initial guess, and given p_{kn}, choose p''_{kn+1} to solve the minimization problem $\inf_{p \in Q_{k-2}} \|p - F[p_{kn}]\|_\alpha$ where Q_{k-2} is the set of polynomials of degree k - 2 or less. Since this step involves best approximating the known continuous function $F[p_{kn}]$ by a polynomial of degree at most k - 2, the second algorithm of Remes may be used. Set

$$p_{kn+1}(t) = c_0 + c_1 t + \int_0^t (t-s)p''_{kn+1}(s)ds.$$

This iteration procedure yields a sequence $\{p_{kn}\}_{n=1}^{\infty}$ of elements of $\mathcal{P}_k$. If the sequence $\{p_{kn}\}_{n=1}^{\infty}$ has a cluster point p_k, then p_k is called a substitute simultaneous approximate solution (SAS) of the IVP (*) from $\mathcal{P}_k$.

The following definition and example point out the most important difference between this study and that of Weinstein and Olson. Let $F(A,t)$ be a function which depends on the parameter vector $A = (a_1, a_2, \ldots, a_m) \in R^m$. Then $F(A,t)$ is said to have Property Z of degree n in $[a,b]$ if $A_1 \in R^m$, $A_2 \in R^m$ and $A_1 \neq A_2$ imply that $F(A_1,t) - F(A_2,t)$ has at most $n - 1$ zeros in $[a,b]$.

Define $H(A,t) = 2a_2 + 6a_3 t + \ldots + k(k-1)a_k t^{k-2}$

$$- f(t, c_0+c_1 t+a_2 t^2+\ldots+a_k t^k,$$

$$c_1+2a_2 t+\ldots+ka_k t^{k-1},$$

$$c_0+c_1 g(t)+a_2 g^2(t)+\ldots+a_k g^k(t),$$

$$c_1+2a_2 g(t)+\ldots+ka_k g^{k-1}(t)),$$

for $A = (a_2, a_3, \ldots, a_k) \in R^{k-1}$.

For a SAS p_k to be a BAS, Olson and Weinstein require $H(A,t)$ to have Property Z of degree $k - 1$ in $[-\alpha, \alpha]$. Property Z is usually very difficult to check; and, as the following example shows, $H(A,t)$ does not always have Property Z.

Consider the IVP

$$x''(t) = x^2(t) + 1 \qquad t \in [-\alpha, \alpha]$$

$$x(0) = x'(0) = 0,$$

and take $\mathcal{P}_2 = \{at^2 : a \in R\}$ as the approximating class. Then $H(A,t) = H(a,t) = 2a - a^2t^4 - 1$ for each $a \in R$. For a sufficiently large number a, the expression $H(a,t) - H(0,t) = 2a - a^2t^4$ has a zero in $[-\alpha, \alpha]$ and consequently, $H(A,t)$ does not have Property Z of degree 1 in $[-\alpha, \alpha]$ for any $\alpha > 0$. The analysis given in this study, which does not use Property Z, shows that if $\alpha \leq .5$, a SAS exists. If $\alpha \leq .14$, the SAS is also a BAS. Convergence theorems and error estimates are given which show that a SAS p_k is a "good" approximate solution regardless of whether or not p_k is a BAS.

CHAPTER I

Approximation Theory

1.1 Introduction.

The results contained in this chapter, unless otherwise mentioned, may be found in Cheney [2]. We first state the fundamental existence theorem.

Existence Theorem. A finite dimensional linear subspace of a normed linear space contains at least one point of minimum distance from a fixed point.

The main objective of this chapter is to review some of the classical theory of Tchebycheff approximation; however, first we need to consider interpolation and convexity.

1.2 Interpolation.

Let $x_0, x_1, \ldots, x_n$ be distinct points or nodes in $[a,b]$ and let $f \in C[a,b]$. One method of approximating f is that of finding a function that agrees with f on the points $x_0, x_1, \ldots, x_n$. This process is called interpolation. We state two theorems concerning polynomial interpolation.

1.2.1 Theorem. Given $x_0, x_1, \ldots, x_n$, distinct points in $[a,b]$, and values $y_0, y_1, \ldots y_n$, there exists a unique polynomial $p(x)$ of degree n or less satisfying $p(x_i) = y_i \quad (i = 0,1,\ldots,n)$.

Proof.

$$\text{Set } \ell_0(x) = \frac{(x-x_1)(x-x_2)\ldots(x-x_n)}{(x_0-x_1)(x_0-x_2)\ldots(x_0-x_n)},$$

$$\ell_n(x) = \frac{(x-x_0)(x-x_1)\ldots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\ldots(x_n-x_{n-1})},$$

and for $i = 1, 2, \ldots, n-1$, define

$$\ell_i(x) = \frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}.$$

Then $p(x) = \Sigma_{i=0}^{n} y_i \ell_i(x)$ is a polynomial of degree at most n, and $p(x_i) = y_i \quad (i = 0,1,\ldots,n)$.

For uniqueness, we note that the difference of two such polynomials has a zero at each x_i $(i = 0,1,\ldots,n)$ and thus must be identically zero.

1.2.2 Theorem. If $f \in C^{n+1}[a,b]$, and if $p(x)$ is the unique polynomial of degree at most n that agrees with f at the n+1 nodes, $x_0, x_1, \ldots, x_n$ contained in $[a,b]$, then

$||f-p|| \leq \frac{1}{(n+1)!}||f^{[n+1]}||\,||W||$ where $||\cdot||$ is the Tchebycheff norm defined by $||g|| = \max_{[a,b]} |g(t)|$ for each $g \in C[a,b]$ and $W(x) = \prod_{i=0}^{n} (x - x_i)$.

Proof.

Fix $x \in [a,b]$. If x is not one of the nodes, then $W(x) \neq 0$, and we may set

$$\lambda(x) = \frac{f(x) - p(x)}{W(x)}$$

and

$$\phi(t) = f(t) - p(t) - \lambda(x)W(t)$$

for each $t \in [a,b]$. We note that $f \in C^{n+1}[a,b]$ implies that $\phi \in C^{n+1}[a,b]$. Now $\phi(x_i) = 0$ $(i = 0,1,\ldots,n)$ and $\phi(x) = 0$, so $\phi(t)$ vanishes at $n+2$ distinct points in $[a,b]$, and by Rolle's Theorem, $\phi'(t)$ vanishes at $n+1$ distinct points. Continuing this argument, we see that $\phi^{[n+1]}(t)$ vanishes at least once on $[a,b]$. Say $\phi^{[n+1]}(\xi) = 0$. Then we have that

$$0 = \phi^{[n+1]}(\xi) = f^{[n+1]}(\xi) - p^{[n+1]}(\xi) - \lambda(x)W^{[n+1]}(\xi)$$

$$= f^{[n+1]}(\xi) - \lambda(x)\,(n+1)!$$

so

$$\lambda(x) = \frac{f^{[n+1]}(\xi)}{(n+1)!}$$

or

$$f(x) - p(x) = \frac{f^{[n+1]}(\xi)}{(n+1)!} \; W(x)$$

and

$$|f(x) - p(x)| \leq \frac{1}{(n+1)!} \; ||f^{[n+1]}|| \; ||W|| \; .$$

Finally, the continuity of f and p imply

$$||f - p|| \leq \frac{1}{(n+1)!} \; ||f^{[n+1]}|| \; ||W|| \; .$$

1.3 Convexity.

The concepts of convexity and convex hulls will be useful in the characterization of best approximations for continuous functions in the Tchebycheff norm.

1.3.1 Definition. Let K be a subset of a normed linear space X. The set K is said to be convex if whenever $x \in K$, $y \in K$ and $\theta \in [0,1]$ it is also true that $\theta x + (1-\theta)y \in K$.

By an induction argument, it can be shown that if K is convex, $f_i \in K$, $\theta_i > 0$ $(i = 1,2,\ldots,n)$ and $\Sigma_{i=1}^{n} \theta_i = 1$, then $\Sigma_{i=1}^{n} \theta_i f_i \in K$. The case for n=1 is clear, so we proceed to the induction step. Suppose the above statement

holds for n=k and consider $\Sigma_{i=1}^{k+1} \theta_i f_i$ where $f_i \in K$, $\theta_i \geq 0$ $(i = 1,\ldots,k+1)$ and $\Sigma_{i=1}^{k+1} \theta_i = 1$. If $\Sigma_{i=1}^{k} \theta_i = 0$, the result is clear, so we consider the case where $\lambda = \Sigma_{i=1}^{k} \theta_i > 0$ then $\Sigma_{i=1}^{k} \frac{\theta_i}{\lambda} f_i \in K$ by the induction hypothesis and $\Sigma_{i=1}^{k+1} \theta_i f_i = \lambda \Sigma_{i=1}^{k} \frac{\theta_i}{\lambda} f_i + \theta_{k+1} f_{k+1} \in K$ since K is convex and $\lambda + \theta_{k+1} = \Sigma_{i=1}^{k+1} \theta_i = 1$.

1.3.2 Definition. For any set A in a linear space, the convex hull of A is the set $\mathcal{K}(A) = \{\Sigma_{i=1}^{n} \theta_i f_i : n$ is a positive integer, $\Sigma_{i=1}^{n} \theta_i = 1$, $\theta_i \geq 0$ and $f_i \in A$ $(i = 1,2,\ldots,n)\}$.

The linear combinations of points of A which make up $\mathcal{K}(A)$ are called convex linear combinations.

1.3.3 Theorem of Caratheodory. If $A \subseteq R^n$, n dimensional Euclidean space, then each point in $\mathcal{K}(A)$ may be expressed as a convex linear combination of at most n+1 elements of A.

Proof.

Let $g \in \mathcal{K}(A)$ and consider the set

$K = \{k : \exists \theta_i \in [0,1]$, $f_i \in A$ $(i = 1,2,\ldots,k)$ such that

$$g = \Sigma_{i=1}^{k} \theta_i f_i\}.$$

Since K is a nonempty set of positive integers, there is a

smallest element say $k_0 \in K$. Thus there are elements $f_i \in A$ and $\theta_i \in [0,1]$ with $\Sigma_{i=1}^{k_0} \theta_i = 1$ and $g = \Sigma_{i=1}^{k_0} \theta_i f_i$. Since k_0 is minimal, we know that $\theta_i > 0$ $(i = 1,2,\ldots,k_0)$. The vectors $g_i = f_i - g$ $(i = 1,2,\ldots,k_0)$ are linearly dependent since $\Sigma_{i=1}^{k_0} \theta_i g_i = 0$. We now suppose by way of contradiction that $k_0 > n+1$. Then the vectors g_i $(i = 2,3,\ldots,k_0)$ are dependent since each $g_i \in R^n$. Choose α_i $(i = 2,3,\ldots,k_0)$ not all zero satisfying $\Sigma_{i=2}^{k_0} \alpha_i g_i = 0$. Define $\alpha_1 = 0$; then for all λ

$$\Sigma_{i=1}^{k_0} (\theta_i - \lambda\alpha_i)g_i = \Sigma_{i=1}^{k_0} \theta_i g_i - \lambda \Sigma_{i=2}^{k_0} \alpha_i g_i = 0.$$

We now choose λ so that $|\lambda|$ is as small as possible under the condition that at least one of the coefficients $\theta_i - \lambda\alpha_i$ vanishes. For each i with $\alpha_i \neq 0$, we have

$$\theta_i - \lambda\alpha_i \geq \theta_i - \frac{\theta_i}{\alpha_i} \alpha_i$$

by the minimality of λ, so $\theta_i - \lambda\alpha_i \geq 0$. We can also assert that not all these coefficients vanish since $\theta_1 - \lambda\alpha_1 = \theta_1 > 0$; however, at least one term does drop out of the sum

$$\Sigma_{i=1}^{k_0} (\theta_i - \lambda\alpha_i)g_i.$$

The equation

$$0 = \Sigma_{i=1}^{k_0} (\theta_i - \lambda\alpha_i)g_i = \Sigma_{i=1}^{k_0}(\theta_i - \lambda\alpha_i)(f_i - g)$$

implies that

$$g \Sigma_{i=1}^{k_0} (\theta_i - \lambda\alpha_i) = \Sigma_{i=1}^{k_0} (\theta_1 - \lambda\alpha_1)f_i$$

or

$$g = \Sigma_{i=1}^{k_0} \delta(\theta_i - \lambda\alpha_i)f_i \text{ where } \delta = 1/\Sigma_{i=1}^{k_0}(\theta_i - \lambda\alpha_i).$$

Since one of the coefficients $\theta_i - \lambda\alpha_i$ vanishes, g can be written as a convex linear combination of at most k_0-1 elements of A. This contradicts the minimality of k_0, so $k_0 \leq n+1$.

1.3.4 Theorem. Every closed convex subset of Euclidean n space possesses a unique point of minimum norm.

Proof.

Let K be such a set and define $p = \inf_{x \in K} \|x\|$.

There is a sequence $\{x_i\}_{i=1}^{\infty}$ with $x_i \in K$ and $\lim_{i \to \infty} \|x_i\| = p$.

By the parallelogram law

$$\|x_i - x_j\|^2 = 2|\|x_i\|^2 + 2|\|x_j\|^2 - \|x_i + x_j\|^2$$

$$= 2|\|x_i\|^2 + 2|\|x_j\|^2 - 4\|\tfrac{1}{2}x_i + \tfrac{1}{2}x_j\|^2.$$

Since K is convex, $\frac{1}{2}x_i + \frac{1}{2}x_j \in K$ so $\|\frac{1}{2}x_i + \frac{1}{2}x_j\| \geq p$ and

$$\|x_i - x_j\|^2 \leq 2|\|x_i\|^2 + 2|\|x_j\|^2 - 4p^2.$$

Now let $\epsilon > 0$ and choose M such that $i \geq M$ implies that $\|x_i\|^2 - p^2 < \epsilon^2/4$ and $j \geq M$ imply that $\|x_j\|^2 - p^2 < \epsilon^2/4$. Then $i \geq M$ and $j \geq M$ implies that

$$\|x_i - x_j\|^2 \leq 2|\|x_i\|^2 - 2p^2 + 2|\|x_j\|^2 - 2p^2 \leq \epsilon^2/2 + \epsilon^2/2 = \epsilon^2$$

and $\|x_i - x_j\| \leq \epsilon$. Thus the sequence $\{x_i\}_{i=1}^{\infty}$ is a Cauchy sequence and therefore converges to some point x, which is contained in K since K is closed. The inequality $|\,\|x_i\| - |\|x\|\,| \leq \|x_i - x\|$ shows that $\|x\| = p$. For the uniqueness, we assume there are two points x, y in K with $|\|x\| = \|y\| = p$. Then

$$\|x - y\|^2 = 2|\|x\|^2 + 2|\|y\|^2 - \|x + y\|^2$$

$$= 4p^2 - 4\|\tfrac{1}{2}x + \tfrac{1}{2}y\|^2$$

$$\leq 4p^2 - 4p^2$$

and we have that $x = y$.

1.3.5 Theorem on linear inequalities. Let U be a compact subset of R^n. A necessary and sufficient condition that the system of linear inequalities $\langle u,z\rangle > 0$ $(u \in U)$, with the usual inner product, be inconsistent is that zero be contained in the convex hull of U.

Proof.

For the sufficiency, we assume that $0 \in \mathcal{K}(U)$. Then $0 = \Sigma_{i=1}^{k}\theta_i u_i$ with $\Sigma_{i=1}^{k}\theta_i = 1$, $u_i \in U$ and $\theta_i \geq 0$ $(i = 1,2,\ldots,k)$. Thus for all z,

$$0 = \langle \Sigma_{i=1}^{k}\theta_i u_i,\ z\rangle = \Sigma_{i=1}^{k}\theta_i \langle u_i,\ z\rangle$$

and so $\langle u_i,\ z\rangle \leq 0$ for some i, and the system $\langle u,\ z\rangle > 0$ $(u \in U)$ is inconsistent.

For the necessity, we assume that $0 \notin \mathcal{K}(U)$. We first show that $\mathcal{K}(U)$ is closed. Let the sequence $\{x_k\}_{k=1}^{\infty}$ converge to x where $x_k \in \mathcal{K}(U)$.

By Caratheodory's theorem (1.3.3), each x_k may be written as a linear convex combination of n+1 or fewer points of U. That is, we may write

$x_k = \Sigma_{i=1}^{n+1}\theta_{ki}u_{ki}$ where $u_{ki} \in U$, $\theta_{ki} \geq 0$ $(i = 1,2,\ldots,n+1)$

and $\Sigma_{i=1}^{n+1} \theta_{ki} = 1$. The set $\Phi = \{(\theta_1,\theta_2,\ldots,\theta_{n+1}) : \theta_i \geq 0$ and $\Sigma_{i=1}^{n+1} \theta_i = 1\}$ is a closed bounded subset of R^{n+1} and hence is compact. Thus there is a subsequence of $\{(\theta_{k1},\theta_{k2},\ldots,\theta_{kn+1})\}_{k=1}^{\infty}$ which converges to $(\theta_1,\theta_2,\ldots,\theta_{n+1}) \in \Phi$. Since U^{n+1} is compact, the corresponding subsequence of $\{(u_{k1},u_{k2},\ldots,u_{kn+1})\}_{k=1}^{\infty}$ in turn has a subsequence $\{(u_{k(\ell)1},u_{k(\ell)2},\ldots,u_{k(\ell)n+1})\}_{\ell=1}^{\infty}$ which converges to $(u_1,u_2,\ldots,u_{n+1}) \in U^{n+1}$. Then, since $x_{k(\ell)} = \Sigma_{i=1}^{n+1} \theta_{k(\ell)i} u_{k(\ell)i}$ converges to $\Sigma_{i=1}^{n+1} \theta_i u_i$, $x = \Sigma_{i=1}^{n+1} \theta_i u_i$. Consequently $x \in \mathcal{K}(U)$ and thus $\mathcal{K}(U)$ is closed. Now by theorem (1.3.4) there is a point $z \in \mathcal{K}(U)$ such that $||z||$ is a minimum. Now let $u \in U$. Since $\mathcal{K}(U)$ is convex, $\theta \in [0,1]$ implies that $\theta u + (1 - \theta)z \in \mathcal{K}(U)$, and thus we have

$$0 \leq \|\theta u + (1 - \theta)z\|^2 - \|z\|^2$$

$$= \langle \theta(u - z) + z, \theta(u - z) + z \rangle - \|z\|^2$$

$$= \theta^2 \|u - z\|^2 + 2\theta \langle z, u - z \rangle .$$

For this inequality to hold for each $\theta \in [0,1]$, we must have $\langle z, u - z \rangle \geq 0$ or $\langle z, u \rangle \geq \langle z, z \rangle$; but $0 \notin \mathcal{K}(U)$ so $z \neq 0$

and we conclude that $\langle z, u \rangle > 0$ and z is a solution to the inequalities.

1.4 Tchebycheff Approximation.

Let X be a compact metric space. In this section we wish to review the theory of approximating functions in C(X) with the Tchebycheff or "max" norm. We want to choose our approximations out of families of functions more general than ordinary polynomials.

1.4.1 Definition. Let $\{g_1, g_2, \ldots, g_n\} \subseteq C(X)$. Linear combinations $\Sigma_{i=1}^{n} \alpha_i g_i$ will be termed generalized polynomials.

The existence theorem in Section 1.1 assures us that for each $f \in C(X)$, there is a generalized polynomial which is a best approximation for f. The next two theorems will characterize these best approximations.

1.4.2 Characterization Theorem. Let X be a compact metric space with $f \in C(X)$ and $\{g_1, g_2, \ldots, g_n\} \subseteq C(X)$. The coefficients $c_1, c_2, \ldots, c_n$ render the Tchebycheff norm of the error function $r = \Sigma_{i=1}^{n} c_i g_i - f$ a minimum if and only if the origin of n space lies in the convex hull of the set $K = \{r(x)\hat{x} : |r(x)| = ||r||\}$, where

$$\hat{x} = (g_1(x), g_2(x), \ldots, g_n(x)) \in R^n.$$

Proof.

The set $X_0 = \{x \in X: |r(x)| = ||r||\}$ is nonempty, since the continuous mapping $x \to |r(x)|$ assumes its supremum on the compact set X.

We now show the "if" part. Suppose $||r||$ is not a minimum. This means that there exists $d \in R^n$ satisfying

$$||\Sigma_{i=1}^n (c_i - d_i)g_i - f|| < ||\Sigma_{i=1}^n c_i g_i - f|| = ||r||.$$

If $x \in X_0$, then

$$|\Sigma_{i=1}^n (c_i - d_i)g_i(x) - f(x)| \leq ||\Sigma_{i=1}^n (c_i - d_i)g_i - f|| < ||r|| = |r(x)|$$

or

$$|r(x) - \Sigma_{i=1}^n d_i g_i(x)| < |r(x)|.$$

Thus,

$$[r(x) - \Sigma_{i=1}^n d_i g_i(x)]^2 < [r(x)]^2$$

or equivalently,

$$[r(x) - \langle d, \hat{x} \rangle]^2 < [r(x)]^2.$$

Multiplying the left side out, we have

$$-2r(x)\langle d, \hat{x} \rangle + (\langle d, \hat{x} \rangle)^2 < 0 \quad \forall\, x \in X_0.$$

For this inequality to hold, we must have

$$r(x) \langle d, \hat{x} \rangle > 0$$

and hence,

(1.4.3) $$\langle d, r(x)\, \hat{x} \rangle > 0 \quad \forall\, x \in X_0.$$

The continuity of $r(x)$ implies that X_0 is a closed subset of X and hence is compact. The continuity of $r(x)$ also implies the continuity of the mapping $x \to r(x)\hat{x}$, and thus we see that the set $K = \{ r(x)\hat{x} : x \in X_0 \}$ is the continuous image of a compact set and is therefore compact.

The above system (1.4.3) of inequalities may be written $\langle d, u \rangle > 0 \quad \forall\, u \in K$. Since K is compact, we may use the theorem on linear inequalities (1.3.5) to assert that zero is not contained in the convex hull of K, but this is a contradiction to our hypothesis.

Conversely, assume zero is not contained in the convex hull of K. Then by the theorem on linear inequalities (1.3.5) there exists a $d \in R^n$ such that

$$\langle d, u \rangle > 0 \quad \forall\, u \in K,$$

or equivalently,

(1.4.4) $$\langle d, r(x)\hat{x} \rangle > 0 \quad \forall\, x \in X_0.$$

The set X_0 is compact, and the mapping $\tau : X \to R$ defined by $x \to \langle d, r(x)\hat{x} \rangle$ is continuous since r is continuous and inner products are continuous. Thus

$$\min_{x \in X_0} \langle d, r(x)\hat{x} \rangle = \epsilon > 0.$$

Now set $X_1 = \{x \in X: \langle d, r(x)\hat{x} \rangle \leq \epsilon/2\}$, and assume for the time being that X_1 is nonempty. The continuity of τ implies that X_1 is a closed subset of X and is thus compact. We can then set

$$E = \max_{X_1} |r(x)| = |r(y)|$$

for some $y \in X_1$.

From the definition of X_0 and X_1, we see that $E < ||r||$. Our objective now is to find a number λ such that

$$||r - \Sigma_{i=1}^{n} \lambda \, d_i g_i|| < ||r|| .$$

Choose λ_1 so that

$$0 < \lambda_1 < \frac{||r|| - E}{||\Sigma \, d_i g_i||} .$$

We note that

$$r(x) \, \Sigma_{i=1}^{n} \, d_i g_i(x) = r(x) \langle d, \hat{x} \rangle = \langle d, r(x)\hat{x} \rangle > 0 \quad \forall \; x \in X_0$$

from the system (1.4.4) of inequalities and thus

$$|| \Sigma d_i g_i || > 0.$$

Then $\forall\ x \in X_1$

$$|r(x) - \Sigma_{i=1}^{n} \lambda_1 d_i g_i(x)| \leq |r(x)| + \lambda_1 |\Sigma d_i g_i(x)|$$

$$\leq E + \lambda_1 \, ||\Sigma d_i g_i||$$

$$< ||r||$$

by the definition of λ_1. Now choose λ_2 so that

$$0 < \lambda_2 < \frac{\epsilon}{||\Sigma d_i g_i||} .$$

Then $\forall\ x \notin X_1$

$$[r(x) - \Sigma\lambda_2 d_i g_i(x)]^2 = [r(x)]^2 - 2r(x)\Sigma\lambda_2 d_i g_i(x)$$

$$+ \lambda_2^2 [\Sigma d_i g_i(x)]^2$$

$$\leq [r(x)]^2 + \lambda_2[-2r(x)\Sigma d_i g_i(x) + \lambda_2 ||\Sigma d_i g_i||^2]$$

$$\leq [r(x)]^2 + \lambda_2[-\epsilon + \lambda_2 ||\Sigma d_i g_i||^2]$$

and thus by the definition of λ_2

$$[r(x) - \Sigma\lambda_2 d_i g_i(x)]^2 < [r(x)]^2 .$$

Now choose $\lambda = \min\{\lambda_1, \lambda_2\}$. Then

$$|r(x) - \Sigma\lambda d_i g_i(x)| = |\Sigma c_i g_i(x) - f(x) - \Sigma\lambda d_i g_i(x)|$$

$= |\Sigma(c_i - \lambda d_i) g_i(x) - f(x)| < |r(x)| \quad \forall\ x \in X$ and thus

$||\Sigma(c_i - \lambda d_i) g_i - f|| < ||r||$. But this is a contradiction to

our assumption that the coefficients $c_1, c_2, \ldots, c_n$ rendered the norm of r a minimum.

The last part of this proof covers the case in which X_1 is empty. That is, our contradiction comes from the inequality

$$[r(x) - \Sigma\lambda_2 d_i g_i(x)]^2 < [r(x)]^2 \quad \forall\, x \notin X_1.$$

By requiring the base functions $g_1, g_2, \ldots, g_n$ to satisfy an additional property, we can restate the characterization in a way which will be beneficial for the numerical determination of best approximations.

1.4.5 Definition. The system $\{g_1, g_2, \ldots, g_n\}$ satisfies the Haar condition on [a,b] if each determinant

$$D[x_1,x_2,\ldots,x_n] = \begin{vmatrix} g_1(x_1) & \cdots & g_n(x_1) \\ \vdots & & \vdots \\ g_1(x_n) & \cdots & g_n(x_n) \end{vmatrix}, \text{ where}$$

the points $x_1, x_2, \ldots, x_n$ are distinct elements of [a,b], is nonzero. A Cramer's rule argument shows that $\{g_1, g_2, \ldots, g_n\}$ satisfies the Haar condition if and only if each nontrivial linear combination $\Sigma_{i=1}^{n} \alpha_i g_i$ has at most n-1 zeros in [a,b].

The system $\{1, x, x^2, \ldots, x^{n-1}\}$ satisfies the Haar condition on any interval $[a,b]$, since each nontrivial linear combination $\alpha_1 + \Sigma_{i=2}^{n} \alpha_i x^{i-1}$ is a polynomial of degree at most n-1 and thus can have at most n-1 zeros. The system $\{x, x^2\}$ does not satisfy the Haar condition on $[0,1]$ since $x - x^2$ has two zeros in $[0,1]$. However, x and x^2 are linearly independent.

1.4.6 Lemma. Let $\{g_1, g_2, \ldots, g_n\} \subseteq C[a,b]$ satisfy the Haar condition. Let $a \leq x_0 < x_1 < \ldots < x_n \leq b$, and let $\lambda_0, \lambda_1, \ldots, \lambda_n$ be nonzero constants. Then $0 \in \mathcal{H}(\{\lambda_i \hat{x}_i : i = 0,1,\ldots,n\})$ if and only if the λ's alternate in sign, i.e. $\lambda_i \lambda_{i-1} < 0$ $(i = 1,2,\ldots,n)$, where

$$\hat{x}_i = (g_1(x_i), g_2(x_i), \ldots, g_n(x_i)).$$

Proof.

The Haar condition implies that all the determinants $D[t_1, t_2, \ldots, t_n]$ with $t_1 < t_2 < \ldots < t_n$ agree in sign. To see this, suppose that

$$D[t_1, t_2, \ldots, t_n] < 0 < D[s_1, s_2, \ldots, s_n].$$

The mapping $(x_1, x_2, \ldots, x_n) \to D[x_1, x_2, \ldots, x_n]$ defined on the connected set

$C = \{\alpha(t_1, t_2, \ldots, t_n) + (1 - \alpha)(s_1, s_2, \ldots, s_n) : 0 \leq \alpha \leq 1\}$ is continuous and thus the image of C is connected. This means there is a number $\alpha_0 \in (0,1)$ such that

$$D[\alpha_0(t_1, t_2, \ldots, t_n) + (1 - \alpha_0)(s_1, s_2, \ldots, s_n)] = 0.$$

From the Haar condition, it must be that $\alpha_0 t_i + (1 - \alpha_0)s_i = \alpha_0 t_j + (1 - \alpha_0)s_j$ for some distinct i and j. Then $\alpha_0(t_i - t_j) = (1 - \alpha_0)(s_j - s_i)$ implies that $t_i - t_j$ and $s_j - s_i$ have the same sign, but this contradicts the assumption that $t_1 < t_2 < \ldots < t_n$ and $s_1 < s_2 < \ldots < s_n$.

Now the origin lies in the convex hull of the points $\lambda_i \hat{x}_i$ $(i = 0, 1, 2, \ldots, n)$ if and only if the equation $\Sigma_{i=0}^{n} \theta_i \lambda_i \hat{x}_i = 0$ has a solution $(\theta_0, \theta_1, \ldots, \theta_n)$ when $\theta_i \geq 0$ $(i = 0, 1, \ldots, n)$ and not all the θ's are zero. We may assume $\Sigma_{i=0}^{n} \theta_i = 1$ since the above equation can be divided by the positive number $\Sigma_{i=0}^{n} \theta_i$. The Haar condition implies that each θ_i is positive, so we may write the above equation in the form

$$\hat{x}_0 = \Sigma_{i=1}^{n} \frac{-\theta_i \lambda_i}{\theta_0 \lambda_0} \hat{x}_i .$$

Now, solving by Cramer's rule, we obtain

$$\frac{-\theta_i\lambda_i}{\theta_0\lambda_0} = \frac{D[x_1,\ldots,x_{i-1},x_0,x_{i+1},\ldots,x_n]}{D[x_1,x_2,\ldots,x_n]}$$

$$= \frac{(-1)^{i-1}\, D[x_0,x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n]}{D[x_1,x_2,\ldots,x_n]}$$

since each column interchange changes the sign of the determinant in the numerator. Since the two determinants on the right agree in sign, it must be that

$$\operatorname{sgn}\left(\frac{-\theta_i\lambda_i}{\theta_0\lambda_0}\right) = (-1)^{i-1}$$

or $\operatorname{sgn}(\lambda_i) = (-1)^i \operatorname{sgn}(\lambda_0)$ and thus the λ's alternate in sign. The above steps are each reversible, so the proof of the lemma is complete.

1.4.7 Alternation Theorem. Let $\{g_1, g_2, \ldots, g_n\} \subseteq C[a,b]$ satisfy the Haar condition, and let X be a closed subset of $[a,b]$ containing at least n+1 points. Then $p = \Sigma_{i=1}^{n} c_i g_i$ is a best approximation on X to a function $f \in C[X]$ if and only if the error function $r = f - p$ exhibits on X at least n+1 alternations, i.e. there are n+1 points

$x_0 < x_1 < \cdots < x_n \in X \ni r(x_i) = -r(x_{i+1}) = \pm\, ||r||$ where $||r|| = \max_X |r(x)|$. The set $\{x_0,\ldots,x_n\}$ is sometimes called

an "extremal set".

Proof.

By the characterization theorem (1.4.2), $||r||$ is a minimum if and only if $0 \in \mathcal{H}(\{ r(x)\hat{x} : |r(x)| = ||r||\})$. Thus $0 = \Sigma_{i=0}^{k} \lambda_i r(x_i)\hat{x}_i$ where $\Sigma_{i=1}^{k} \lambda_i = 1$ and $\lambda_i > 0$.

By Caratheodory's theorem (1.3.3), we may assume $k \leq n$ and by the Haar condition $k \geq n$, thus $k = n$. If we now assume $x_0 < x_1 < \ldots < x_n$, then the lemma asserts that $0 = \Sigma_{i=0}^{n} \lambda_i r(x_i)\hat{x}_i$ if and only if the numbers $\lambda_i\ r(x_i)$ alternate in sign. Since $\lambda_i > 0$, the numbers $\lambda_i\ r(x_i)$ alternate in sign if and only if the numbers $r(x_i)$ alternate in sign. Since the x_i's were chosen so that $r(x_i) = ||r||$, the proof is complete.

Now suppose we are approximating the function $f \in C[a,b]$ by linear combinations of $g_1, g_2, \ldots, g_n$, and suppose further that, for some linear combination $p = \Sigma_{i=1}^{n} \alpha_i g_i$, the error function $r = f - p$ has the extremal set $\{x_0, x_1, \ldots, x_n\}$. The alternation theorem then asserts that if the system $\{g_1, g_2, \ldots, g_n\}$ satisfies the Haar condition, then p is a best approximation for f. However, from the proof of lemma (1.4.6), a sufficient condition for

p to be a best approximation for f is that all the determinants $D_0 = D[x_1, \ldots, x_n]$,

$$D_i = D[x_0, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n],$$

$(i = 1, 2, \ldots, n-1)$, $D_n = D[x_0, \ldots, x_{n-1}]$

agree in sign. Thus the Haar condition, which is sometimes very difficult to check, can be replaced by an easily checkable condition involving n+1 determinants. We restate these ideas as a corollary to the alternation theorem.

1.4.8 Corollary. Let $f \in C[a,b]$ and $p = \Sigma_{i=1}^{n} \alpha_i g_i$. Then p is a best approximation for f if the determinants D_i $(i = 0, 1, \ldots, n)$ agree in sign.

The alternation theorem will play a major role in the computation of best approximations.

Example.

Let $f(x) = x^2 \in C[-1,1]$ and let $\{g_1, g_2\} = \{1, x\}$. The best approximation to f by linear combinations of 1 and x is $p(x) = 1/2$ since the error function $r(x) = x^2 - 1/2$ has the extremal set $\{-1, 0, 1\}$. The norm of the error is $||r|| = 1/2$.

1.4.9 Theorem of de La Vallee Poussin. Let $f \in C[a,b]$ and $\{g_1, g_2, \ldots, g_n\}$ satisfy the Haar condition on the interval $[a,b]$ and let $g = \Sigma\alpha_i g_i$ be a generalized polynomial such that $f - g$ assumes alternately positive and negative values at $n+1$ consecutive points $x_i \in [a,b]$. Then $E(f) \geq \min_i |f(x_i) - g(x_i)|$ where $E(f)$ is the infimum of $\| f - p\|$ as p ranges over all generalized polynomials $\Sigma_{i=1}^{n} c_i g_i$.

Proof.

Suppose there is a generalized polynomial g_0 satisfying $\| f - g_0\| < \min_i |f(x_i) - g(x_i)|$. Then $g_0 - g = (g_0 - f) + (f - g)$ is alternately positive and negative at the points x_i $(i = 0, 1, \ldots, n)$. Thus $g_0 - g$ has n zeros in $[a,b]$ and by the Haar condition $g_0 \equiv g$. This contradicts the above inequality, so for each generalized polynomial p we have

$$\|f - p\| \geq \min_i |f(x_i) - g(x_i)|.$$

Thus $\inf_{p \in G} \|f - p\| \geq \min_i |f(x_i) - g(x_i)|$ where

$$G = \text{span}\{g_1, g_2, \ldots, g_n\}.$$

Hence

$$E(f) \geq \min |f(x_i) - g(x_i)|.$$

This concludes the proof of the theorem.

We now need an estimate on the size of $E_n(f)$ where $E_n(f) = \inf_{p \in Q_n} ||f - p||$ with Q_n denoting the set of all polynomials of degree n or less.

1.4.10 Definition. Let f be defined on [a,b]. The modulus of continuity of f(x) on [a,b], denoted by $\omega(\delta)$, is defined for $\delta > 0$ by

$$\omega(\delta) = \sup_{\substack{x_1, x_2 \in [a,b] \\ |x_1 - x_2| \leq \delta}} |f(x_1) - f(x_2)|$$

Note. The function $\omega(\delta)$ is nondecreasing.

We now state Jackson's theorem, which gives an upper bound on $E_n(f)$.

1.4.11 Jackson's Theorem. If $f \in C[a,b]$, then

$$E_n(f) \leq 6\omega\left(\frac{b-a}{2n}\right).$$

The proof of Jackson's theorem may be found in Rivlin [9].

We next consider the question of unicity of best approximations. That is, given

$$\{f, g_1, g_2, \ldots, g_n\} \subseteq C[a,b],$$

when is the best approximation to f by linear combination of the g's unique? The lack of unicity impedes the numerical computation of best approximations.

1.4.12 Strong Unicity Theorem. Let the system $\{g_1, \ldots, g_n\} \subseteq C[a,b]$ satisfy the Haar condition, and let $g^* = \Sigma\alpha_i g_i$ be a best approximation to a given continuous function f. There is a constant $\gamma > 0$ depending on f, such that for each $g = \Sigma\beta_i g_i$,

$$||f - g|| \geq ||f - g^*|| + \gamma\, ||g^* - g|| .$$

Proof.

Case 1. Suppose $||f - g^*|| = 0$. Then $f \equiv g^*$, $||g - g^*|| \leq ||g - f|| + ||f - g^*|| = ||g - f||$ and we may choose $\gamma = 1$.

Case 2. Suppose $||f - g^*|| > 0$. By the characterization theorem (1.4.2), the origin lies in the convex hull of $\{r(x)\hat{x}: |r(x)| = ||r||\}$, where $r = f - g^*$ and $\hat{x} = (g_1(x), g_2(x), \ldots, g_n(x))$.

Thus there are points $x_0, x_1, \ldots, x_k$ in $[a,b]$, and signs $\sigma_0, \sigma_1, \ldots, \sigma_k$ such that $r(x_i) = \sigma_i \|r\|$ $(i = 0, 1, \ldots, k)$ and $\bar{0} = \Sigma_{i=0}^{k} \theta_i \sigma_i \|r\| \hat{x}_i$ for some constants $\theta_i > 0$ with $\Sigma\theta_i = 1$. Now since $\|r\| > 0$, we have

$$\bar{0} = \Sigma_{i=0}^{k} \theta_i \sigma_i (g_1(x_i), \ldots, g_n(x_i)) \quad \text{or}$$

$$0 = \Sigma_{i=0}^{k} \theta_i \sigma_i g_j(x_i) \qquad (j = 1, 2, \ldots, n).$$

By Caratheodory's theorem (1.3.3), we may assume $k \leq n$, and by the Haar condition $k \geq n$ so $k = n$. Then for each $q = \Sigma a_i g_i$

$$(1.4.13) \qquad 0 = \Sigma_{i=0}^{n} \theta_i \sigma_i q(x_i).$$

By the Haar condition, not all the numbers $q(x_i)$ are zero, thus not all the numbers $\sigma_i q(x_i)$ are zero. Now $\theta_i > 0$ and equation (1.4.13) imply that $\sigma_i q(x_i) > 0$ for some index i and thus $\max_i \sigma_i q(x_i) > 0$. Since the mappings $q \to \sigma_i q(x_i)$ $(i = 1, 2, \ldots, n)$ are each continuous, the mapping $q \to \max_i \sigma_i q(x_i)$ is also continuous and hence assumes its infimum on each compact set. Since $\max_i \sigma_i q(x_i) > 0$, we conclude that $\inf_{\|q\| = 1} \max_i \sigma_i q(x_i) = \gamma > 0$.

Let $g = \Sigma_{i=1}^{n} \beta_i g_i$ be an arbitrary linear combination of the g's. If $g = g^*$, the conclusion of the theorem is trivial. If $g \neq g^*$, then

$$q = \frac{g^* - g}{||g^* - g||}$$

has norm 1 and thus $\max_i \sigma_i q(x_i) \geq \gamma$, and there is an index i such that

$$\sigma_i q(x_i) \geq \gamma.$$

Now $||f - g|| \geq |(f - g)(x_i)|$

$$\geq \sigma_i (f - g)(x_i)$$

$$= \sigma_i (f - g^*)(x_i) + \sigma_i (g^* - g)(x_i)$$

$$= ||f - g^*|| + \sigma_i (g^* - g)(x_i)$$

$$\geq ||f - g^*|| + \gamma ||g^* - g|| ,$$

and the proof is complete.

We will need more detailed information concerning the size of γ. It is clear that once a γ is found, any smaller γ will suffice. The following construction of a suitable γ is due to A. K. Cline [3].

Let $\{g_1, g_2, \ldots, g_n\} \subseteq C[a,b]$ satisfy the Haar condition and let $G = \text{span}\{g_1, g_2, \ldots, g_n\}$ and $G_1 = \{g \in G: \|g\| = 1\}$. Let $f \in C[a,b]$, let g^* be the best approximation to f out of G and let E denote the point set $\{x_0, x_1, \ldots, x_n\}$ from the proof of the strong unicity theorem (1.4.12). Then $\gamma = \min_{g \in G_1} \max_{x \in E} (\text{sgn}(f(x)-g^*(x))g(x)$.

In lemma (1.4.17) below, we will show that γ may be determined by taking minima on a set smaller than G. For $g \in G$, set

$$\gamma(g) = \max_{x \in E} [\text{sgn}(f(x) - g^*(x))g(x)]$$

and

$$G_1^* = \{g \in G_1: \gamma(g) = \text{sgn}(f(x) - g^*(x))g(x)$$

$$\text{for at least n values } x_i \text{ of E}\}.$$

1.4.14 Note: If $f \notin G$, then $\gamma(g)$ is positive for each nontrivial g contained in G.

Proof.

Suppose there is an element $g \in G$ such that $\gamma(g) \leq 0$. Then since $f(x_i) - g^*(x_i)$ assumes alternately positive and negative values on E, it must be that g also assumes alter-

nately positive and negative values on E. The Haar condition implies that $g \equiv 0$.

1.4.15 Lemma. If $f \notin G$, then G_1^* contains n+1 elements.

Proof.

With $E = \{x_0, x_1, \ldots, x_n\}$, the proof of the strong unicity theorem shows that there are n+1 positive scalars θ_i $(i = 0, 1, \ldots, n)$ satisfying

$$(1.4.16) \qquad 0 = \Sigma_{j=0}^{n} \theta_j \sigma_j g(x_j) \quad \text{for each } g \in G$$

where $\sigma_j = \operatorname{sgn}[f(x_j) - g^*(x_j)]$. For $i = 0, 1, \ldots, n$, define $p_i \in G$ to be the solution of the interpolation problem

$$p_i(x_j) = \sigma_j \quad (j = 0, 1, \ldots, n;\ j \neq i).$$

The Haar condition on $\{g_1, g_2, \ldots, g_n\}$ insures that each p_i can be uniquely determined. Then

$$\sigma_i p_i(x_i) = \frac{-1}{\theta_i} \sum_{\substack{j=1 \\ j\neq i}}^{n} \theta_j \sigma_j p_i(x_j) = \frac{-1}{\theta_i} \sum_{\substack{j=1 \\ j\neq 1}}^{n} \theta_i < 0$$

since $\sigma_j p_i(x_j) = 1$ for $i \neq j$. Thus

$$\gamma(p_i) = \max_{x \in E} [\operatorname{sgn}(f(x) - g^*(x))]p_i(x)$$

$$= 1 = \operatorname{sgn}[f(x) - g^*(x)]p_i(x)$$

for $x \in \{x_i : i = 0, 1, \ldots, n;\ i \neq j\}$. We have $p_i/\|p_i\| \in G_1^*$ for $i = 0, 1, \ldots, n$ and G_1^* has at least $n+1$ elements.

We now show that G_1^* has exactly $n+1$ elements. Let $p \in G_1^*$. From equation (1.4.16) we have $\operatorname{sgn} p(x_i) = -\sigma_i$ for some index i, and then, since $p \in G_1^*$, $\gamma(p) = \operatorname{sgn}[f(x_j) - g^*(x_j)]p(x_j)$ $(j = 0, 1, \ldots, n;\ j \neq i)$. Solving for $p(x_j)$, we obtain the equation

$$p(x_j) = \gamma(p)\sigma_j \quad (j = 0, 1, \ldots, n;\ j \neq i).$$

Since p and p_i are each elements of G, and the system $\{g_1, g_2, \ldots, g_n\}$ satisfies the Haar condition, p and p_i are uniquely determined by their values on the points x_j $(j = 0, 1, \ldots, n;\ j \neq i)$. Consequently, $p(x) \equiv \gamma(p)p_i(x)$; but $\|p\| = |\gamma(p)|\,\|p_i\|$ and $\|p\| = \|p_i\| = 1$ imply that $|\gamma(p)| = \gamma(p) = 1$. This means that $p(x) \equiv p_i(x)$, and there are exactly $n+1$ elements of G_1^*.

1.4.17 Lemma. The quantity $\gamma = \min_{g \in G_1} \gamma(g)$ is also given by $\gamma = \min_{g \in G_1^*} \gamma(g)$.

Proof.

We first show that if $g \in G_1 \sim G_1^*$, then there is an element $g' \in G_1 \ni \gamma(g') < \gamma(g)$. Define

$$J = \{x \in E: \gamma(g) = \operatorname{sgn}[f(x) - g^*(x)]g(x)\}.$$

Case 1. Suppose there is an $x^* \notin J$ with $|g(x^*)| = \|g\| = 1$. Since g is not contained in G_1^*, J contains at most n-1 points, so we may choose $h \in G$ to solve the interpolation problem

$$h(x) = 0 \quad \forall\, x \in J \quad \text{and}$$

$$h(x^*) = g(x^*).$$

Now set $u = g + \lambda h$. For $x \in J$,

$$\operatorname{sgn}[f(x) - g^*(x)]u(x) = \operatorname{sgn}[f(x) - g^*(x)](g(x) + \lambda h(x))$$

$$= \operatorname{sgn}[f(x) - g^*(x)]g(x)$$

$$= \gamma(g), \text{ and for } x \in E \sim J$$

$$\operatorname{sgn}[f(x) - g^*(x)]u(x) = \operatorname{sgn}[f(x) - g^*(x)]g(x)$$

$$+ \lambda \operatorname{sgn}[f(x) - g^*(x)]h(x)$$

$$\leq \operatorname{sgn}[f(x) - g^*(x)]g(x) + |\lambda| \, \|h\|$$

$$\leq \gamma(g) \quad \text{for } \lambda \text{ sufficiently small since } x \notin J.$$

We now have that $\forall\, x \in E$