OPTIMAL STARTING APPROXIMATIONS FOR ITERATIVE SCHEMES FROM CLASSES OF RATIONAL FUNCTIONS

Thesis for the Degree of Ph. D.

MICHIGAN STATE UNIVERSITY
JOHN BYRON GIBSON

1971

This is to certify that the

thesis entitled

Optimal Starting Approximations for
Iterative Schemes from Classes
of Rational Functions

presented by

John Byron Gibson

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Mathematics

Gerald D. Taylor
Major professor

Date 8-4-71

O-7639

ABSTRACT

OPTIMAL STARTING APPROXIMATIONS
FOR ITERATIVE SCHEMES
FROM CLASSES OF RATIONAL FUNCTIONS

By

John Byron Gibson

Given a compact subset X of the closed interval $[a,b]$, let $C(X)$ denote the space of all continuous real-valued functions defined on X, normed by $\|f\| = \max_{x \in X} \{|w(x)f(x)|\}$ where $w \in C(X)$ and $w > 0$ on X. In the case $w \equiv 1$ the norm is the usual Chebyshev norm and is denoted by $\|\cdot\|_\infty$. Let K be a convex subset of $C(X)$ and Φ a continuous mapping of K into $C(X)$. We shall then be interested in the problem of approximating $g \in \Phi(K)$ by elements of $\Phi(M)$ where M is a fixed subset of K. If we impose certain restrictions on Φ, K and M, a Chebyshev type theory can be developed for this nonlinear approximation problem which is similar to the standard Chebyshev theory. The theory obtained has application to many iterative processes that can be used to approximate functions such as $\exp(x)$ and $x^{\frac{1}{n}}$, $n = 2,3,\ldots$ and consequently we make the following definition:

<u>Definition</u>: An element $p \in M$ is called an optimal starting approximation (or best starting approximation) for g, with respect to Φ and M, if $\|g - \Phi(p)\| \le \|g - \Phi(q)\|$ for all $q \in M$.

In this paper we investigate the above problem for M a subset of K consisting of certain classes of rational functions. In Chapter I the aforementioned Chebyshev theory is developed. Classically, one would apply the characterizing alternation theorem we obtain to develop a Remez exchange algorithm for the computation of optimal starting approximations. However, if the operator Φ satisfies certain properties the optimal starting approximation for a function f can be calculated as a constant multiple of the best relative approximation to f in the case of relative error or as a translate of the best uniform approximation to f in the case of uniform error.

In Chapter II the theory is applied to the Newton operator and optimal starting approximations are calculated for the functions x^{α}, $\alpha \in (0,1)$, $\exp(x)$ and $\ln x$ in the manner described above. A number of additional iterative schemes are presented in Chapter III and analyzed for optimal starting approximations for the functions $x^{\frac{1}{n}}$, $n = 2,3,\ldots$ and $\exp(x)$.

Recently D. Moursund has introduced the concept of a generalized weight function which is a more general means of measuring error than that described earlier. Chapter IV is concerned with a development of the Chebyshev theory, analogous to that in Chapter I, for this type of weight function.

Finally, in Chapter V an improved Newton iteration scheme for approximating $\exp(x)$ is presented and is shown to be optimal in the sense of this paper.

OPTIMAL STARTING APPROXIMATIONS FOR ITERATIVE SCHEMES FROM CLASSES OF RATIONAL FUNCTIONS

By

John Byron Gibson

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Mathematics

1971

ACKNOWLEDGEMENTS

I am deeply indebted to my major professor Gerald D. Taylor for his many suggestions and guidance in the preparation of this paper. His enthusiasm for and insights into the area of approximation theory have inspired me greatly.

I would also like to thank Dr. Edwin Kaufman for his many helpful comments.

TABLE OF CONTENTS

INTRODUCTION

In [15] D.G. Moursund examined the problem of approximating $\sqrt{x}$ over an interval $[a,b]$ $(a > 0)$ by applying the Newton-Raphson iteration scheme to classes of polynomial and rational approximants. For polynomial approximation the problem may be formulated in the following way: set $K = \{f \in C[a,b]: f > 0 \text{ on } [a,b]\}$, $M = P_n \cap K$ where P_n consists of all polynomials of degree less than or equal to n, and define $\Phi: K \to C[a,b]$ via $\Phi(h)(x) = \frac{1}{2}\left(h(x) + \frac{x}{h(x)}\right)$. Here Φ denotes a single Newton iteration for calculating $\sqrt{x}$ starting with $h(x)$. Then we are interested in approximating $\sqrt{x}$ with the class $\{\frac{1}{2}\left(p(x) + \frac{x}{p(x)}\right)\}$, for $p \in M$, using the relative error (the notion of relative error will be defined later). Moursund was able to show that there exists a unique solution p^* to $\inf\left\{\left\|\frac{\sqrt{x} - \Phi(p)(x)}{\sqrt{x}}\right\|_\infty : p \in M\right\}$ and moreover that p^* is also the unique solution to $\inf\left\{\left\|\frac{\sqrt{x} - \Phi^m(p)(x)}{\sqrt{x}}\right\|_\infty : p \in M\right\}$, $m = 2,3,\ldots,$ where $\Phi^k(h) = \Phi(\Phi^{k-1}(h))$. More recently, P.H. Sterbenz and C.T. Fike [24] and R.F. King and D.L. Phillips [7] both showed (independently) that p^* is a multiple of $\bar{p}$, the best relative approximation to $\sqrt{x}$ on $[a,b]$.

The above problem may be posed in a more general setting. Given a compact subset X of the closed interval $[a,b]$, let $C(X)$ denote the space of all continuous real-valued functions defined on X, normed by $\|f\| = \max_{x \in X} \{|w(x)f(x)|\}$ where $w \in C(X)$ and $w > 0$

on X. In the case $w \equiv 1$ the norm is the usual Chebyshev norm and is denoted by $\|\cdot\|_{\infty}$. Let K be a convex subset of $C(X)$ and Φ a continuous mapping of K into $C(X)$. We shall then be interested in the problem of approximating $g \in \Phi(K)$ by elements of $\Phi(M)$ where M is a fixed subset of K.

G. Meinardus and G.D. Taylor [11] have considered the above more general problem for M a subset of K consisting of members of a Haar subspace of $C[a,b]$. (The notion of a Haar subspace will be defined in the paper.) By imposing additional restrictions on Φ, K and M they were able to develop a theory for the above nonlinear approximation problem which is analogous to the classical Chebyshev theory. Moreover they were able to preserve the behavior of the problem considered by Moursund and to obtain results similar to those obtained by Sterbenz-Fike and King-Phillips in the more general setting. Because of the application of this theory to iterative processes we make the following definition:

Definition: $p \in M$ is an optimal starting approximation (or best starting approximation) for g, with respect to Φ and M, if $\|g - \Phi(p)\| \leq \|g - \Phi(q)\|$ for all $q \in M$.

In this paper we shall investigate the above problem for M a subset of K consisting of certain classes of rational functions. In Chapter I we develop a Chebyshev type theory for this highly nonlinear approximation problem and in Chapter II we apply the theory to the Newton operator. A number of iterative schemes are given in Chapter III and analyzed for optimal starting approximations.

The concept of a generalized weight function was introduced by D. Moursund in [13] and [14]. In Chapter IV we show that the

Chebyshev type theory can also be developed for this more general means of measuring error.

Finally, in Chapter V we present an improved Newton iteration scheme for approximating e^x which is shown to be optimal in the sense of this paper.

CHAPTER I

GENERAL THEORY FOR OPTIMAL STARTING APPROXIMATIONS

Section 1: Definitions and Basic Results

In our development we wish to make use of the results and techniques of the classical Chebyshev rational approximation. Consequently we introduce the following definitions and results from the paper of Meinardus and Taylor [11]. Throughout this section we shall assume that X, K, M and Φ are as described in the introduction.

Definition 1.1: The operator Φ is called pointwise strictly monotone at $f \in K$ if for each $h,k \in K$ we have $|\Phi(h)(x_0) - \Phi(f)(x_0)| < |\Phi(k)(x_0) - \Phi(f)(x_0)|$ for each $x_0 \in X$ where either $k(x_0) < h(x_0) \leq f(x_0)$ or $f(x_0) \leq h(x_0) < k(x_0)$.

Lemma 1.2: Let $\Phi: K \to C(X)$ be pointwise strictly monotone at $f \in K$. If $k \in K$ and at $x_0 \in X$, $k(x_0) \neq f(x_0)$ then $\Phi(k)(x_0) \neq \Phi(f)(x_0)$.

Proof: Let $h \equiv f$ in Definition 1.1.∎

Definition 1.3: The operator Φ is said to be pointwise fixed at $f \in K$ if $h \in K$ with $h(x_0) = f(x_0)$ for $x_0 \in X$ implies $\Phi(h)(x_0) = \Phi(f)(x_0)$.

In general the notions of pointwise strictly monotone and pointwise fixed are independent. Indeed, define $K \subseteq C[0,1]$ by

$K = \{f(x) = ax^2 : 0 \le a \le 1\}$. If $\Phi_1: K \to C[0,1]$ is defined by $\Phi_1(ax^2) = x^2$ then Φ_1 is pointwise fixed at x^2 yet not pointwise strictly monotone at x^2. Now, if $\Phi_2: K \to C[0,1]$ is defined by $\Phi_2(ax^2) = x^2 + (1-a)$ then Φ_2 is continuous and pointwise strictly monotone at $f(x) = x^2$. Indeed, if $ax_0^2 < bx_0^2 \le x_0^2$ then $x_0 \ne 0$ and $a < b \le 1$ so $|\Phi_2(ax^2)(x_0) - x_0^2| = 1 - a > 1 - b = |\Phi_2(bx^2)(x_0) - x_0^2|$ since $\Phi_2(x^2) = x^2$. Each $ax^2 \in K$ has the property that $ax^2 = x^2$ at $x = 0$; but $\Phi_2(ax^2)(0) \ne 0 = \Phi_2(x^2)(0)$ if $a \ne 1$. Hence Φ_2 being pointwise strictly monotone at x^2 does not imply Φ_2 is pointwise fixed at x^2.

Next we shall show that the composition of two continuous operators possessing the above properties is again such an operator provided of course that domains and ranges mesh correctly.

<u>Lemma 1.4</u>: Let $\Phi: K \to C(X)$ and $\Psi: L \to C(X)$ be continuous operators with $\Phi(K) \subseteq L$, Φ pointwise strictly monotone and pointwise fixed at $f \in K$ and Ψ pointwise strictly monotone and pointwise fixed at $\Phi(f) \in L$. Then $\Psi\Phi: K \to C(X)$ is a continuous pointwise strictly monotone operator at f which is also pointwise fixed at f.

<u>Proof</u>: It is clear that $\Psi\Phi: K \to C(X)$ is continuous and pointwise fixed at $f \in K$. To establish the pointwise strict montonicity of $\Psi\Phi$ at f let us first show that if $h,k \in K$ with $k(x_0) < h(x_0) \le f(x_0)$ or $f(x_0) \le h(x_0) < k(x_0)$ then either $\Phi(k)(x_0) < \Phi(h)(x_0) \le \Phi(f)(x_0)$ or $\Phi(f)(x_0) \le \Phi(h)(x_0) < \Phi(k)(x_0)$. Since Φ is pointwise strictly monotone at f we have that $|\Phi(k)(x_0) - \Phi(f)(x_0)| > |\Phi(h)(x_0) - \Phi(f)(x_0)|$. Suppose that $\Phi(k)(x_0) < \Phi(f)(x_0) < \Phi(h)(x_0)$ or $\Phi(h)(x_0) < \Phi(f)(x_0) < \Phi(k)(x_0)$. Then for $0 \le \alpha \le 1$, the functions

$\ell_\alpha(x) = \alpha h(x) + (1-\alpha)k(x)$ are members of K (since K is convex) and furthermore $\Phi(\ell_\alpha)(x_0)$ is a continuous function of α for each fixed $x_0 \in X$. Thus, by the intermediate value theorem, there exists $\alpha_0 \in (0,1)$ such that $\Phi(\ell_{\alpha_0})(x_0) = \Phi(f)(x_0)$. Therefore, in view of Lemma 1.2, $\ell_\alpha(x_0) = f(x_0)$. This is a contradiction. Consequently we have either $\Phi(k)(x_0) < \Phi(h)(x_0) \le \Phi(f)(x_0)$ or $\Phi(f)(x_0) \le \Phi(h)(x_0) < \Phi(k)(x_0)$. An application of the pointwise strict monotonicity of Ψ at $\Phi(f)$ yields the desired result. ∎

In order to be able to iterate with the operator Φ we must necessarily demand that $\Phi(K) \subseteq K$. With this restriction we are able to relate the notions of pointwise strictly monotone and pointwise fixed according to the following lemma:

Lemma 1.5: If $\Phi\colon K \to K$ is a continuous operator on the convex set K which is pointwise strictly monotone at $f \in K$ then Φ is pointwise fixed at f.

Proof: Suppose $h \in K$ and $h(x_0) = f(x_0)$ for some $x_0 \in X$. Clearly, if $k(x_0) = f(x_0)$ for all $k \in K$ then $\Phi(h)(x_0) = \Phi(f)(x_0)$ and so Φ is pointwise fixed at f. If this is not the case then there exists $k \in K$ such that $k(x_0) \neq f(x_0)$. Set $k_\alpha = \alpha f + (1-\alpha)k$ and note that $k_\alpha(x_0) \neq f(x_0)$ for $\alpha \in (0,1)$. Now $\|k_\alpha - f\| \to 0$ as $\alpha \to 1$ and so $\|\Phi(k_\alpha) - \Phi(f)\| \to 0$ as $\alpha \to 1$. Thus $\Phi(k_\alpha)(x_0) \to \Phi(f)(x_0)$ since norm convergence is stronger than pointwise convergence. Now Φ being pointwise strictly monotone at f implies $|\Phi(k_\alpha)(x_0) - \Phi(f)(x_0)| > |\Phi(h)(x_0) - \Phi(f)(x_0)|$ for all $\alpha \in (0,1)$ and so $|\Phi(h)(x_0) - \Phi(f)(x_0)| = 0$. Hence $\Phi(h)(x_0) = \Phi(f)(x_0)$ and so Φ is pointwise fixed at f. ∎

For the sake of any iteration processes we may be interested in, we state the following corollary to Lemma 1.5:

Corollary 1.6: If $\Phi: K \to K$ is continuous, $\Phi(f) = f$ for some $f \in K$ and Φ is pointwise strictly monotone at $f \in K$ then $\Phi^m: K \to K$ defined inductively by $\Phi^m(h) = \Phi(\Phi^{m-1}(h))$, $m = 2,3,\ldots$, satisfies $\Phi^m(f) = f$ and Φ^m is pointwise strictly monotone at f.

Finally we shall state a general existence theorem for best approximation in our setting. The theorem is very general and for specific operators is sometimes difficult to check.

Theorem 1.7: Let K be a convex subset of $C(X)$, $M \subseteq K$ and $\Phi: K \to C(X)$ be continuous. Then corresponding to $g \in \Phi(K)$ there exists $r^* \in M$ minimizing $\|g - \Phi(r)\|$, $r \in M$, provided there exists a compact subset M_1 of M and a positive constant η such that $s \in M \sim M_1$ implies $\inf \{\|g - \Phi(r)\| : r \in M\} + \eta \le \|g - \Phi(s)\|$.

Proof: A continuous function on a compact set assumes its minimum. ▮

Section 2: Existence, Uniqueness, and Characterization Theorems

In this section we shall specialize M and K and develop an alternation theory for characterizing best approximations similar to the standard Chebyshev theory. From this development we are able, when we have existence, to conclude uniqueness using standard Chebyshev theory arguments. Before proceeding to the first characterization theorem we must establish some additional notation.

Let X be a compact subset of the closed interval $[a,b]$ containing at least $m + n + 2$ points where m and n are fixed for the discussion at hand. Let P_n denote the set of all polynomials p with $\partial p \le n$ (∂p denotes the degree of p) and set

$$R_m^n[a,b] = \{\frac{p}{q} : p \in P_n,\ q \in P_m,\ (p,q) = 1,\ q > 0 \text{ on } [a,b]\}$$

where $(p,q) = 1$ denotes the fact that the polynomials p and q are relatively prime.

The following well-known lemma from Rice [20] will be used extensively:

Lemma 1.8: Given $r = \frac{p}{q} \in R_m^n[a,b]$, $\tau > 0$ and any set

$$\{x_i : i = 1,\dots,r,\ r < 1 + \max\{m + \partial p,\ n + \partial q\},\ x_i < x_{i+1}\}$$

of points in X, there is a rational function $r_\varepsilon \in R_m^n[a,b]$ such that

(i) $\|r_\varepsilon - r\|_{L^\infty[a,b]} < \tau$,

and (ii) $\operatorname{sgn}(r(x) - r_\varepsilon(x)) = (-1)^{i+1}$, $x \in (x_i, x_{i+1})$, $i = 0,1,\dots,r$

where $x_0 = a$ if $x_1 \neq a$ and $x_{r+1} = b$ if $x_r \neq b$.

Proof: Set $\psi(x) = \prod_{i=1}^{r} (x_i - x)$. Now $p(x)$ and $q(x)$ are relatively prime and so there are polynomials $p_1(x) \in P_n$ and $q_1(x) \in P_m$ such that $\psi(x) = p(x)q_1(x) - p_1(x)q(x)$. The proof of this fact depends on the Euclidean Algorithm and can be found in Rivlin [21]. Let

$$r_\varepsilon(x) = \frac{p(x) - \varepsilon p_1(x)}{q(x) - \varepsilon q_1(x)}$$

where $\varepsilon > 0$. Then

$$r(x) - r_\varepsilon(x) = \frac{-\varepsilon\psi(x)}{q(x)(q(x) - \varepsilon q_1(x))} \quad . \tag{1.1}$$

Since $q(x) > 0$ on $[a,b]$, ε may be chosen sufficiently small, say $0 < \varepsilon \le \varepsilon_0$, so that the denominator of (1.1) is positive. Now, given $\tau > 0$, select ε such that $0 < \varepsilon \le \varepsilon_0$ and $|\varepsilon| \left\|\frac{\psi}{q(q-\varepsilon q_1)}\right\|_\infty < \tau$.

Then $\|r - r_\varepsilon\| < \tau$ and, since (ii) is clear, the result is established.∎

<u>Theorem 1.9</u>: Let $\Phi: K \to C(X)$ be a continuous operator where K is a convex subset of $C(X)$. Let $M = K \cap R_m^n[a,b]$ be a nonempty relatively open subset of $R_m^n[a,b]$. Finally, assume that Φ is pointwise strictly monotone and pointwise fixed at $f \in K \sim M$. Then $r \in M$ is the best starting approximation for $\Phi(f)$ if and only if there exist points $\{x_i\}_{i=1}^N \subseteq X$, $N = 2 + \max\{m + \partial p, n + \partial q\}$, for which

(i) $x_1 < x_2 < \ldots < x_N$,

(ii) $|w(x_i)(\Phi(f)(x_i) - \Phi(r)(x_i))| = \|\Phi(f) - \Phi(r)\|$, $i = 1,\ldots,N$

and

(iii) $\operatorname{sgn}(f(x_i) - r(x_i)) = (-1)^{i+1}\operatorname{sgn}(f(x_1) - r(x_1))$, $i = 1,\ldots,N$.

<u>Proof</u>: (Sufficiency). Since $f \notin M$ we know that there exists a point $x_0 \in X$ for which $f(x_0) \neq r(x_0)$. Thus $\|\Phi(f) - \Phi(r)\| \neq 0$. Suppose that $|w(x_i)(\Phi(f)(x_i) - \Phi(r)(x_i))| = \|\Phi(f) - \Phi(r)\|$ and $\operatorname{sgn}(f(x_i) - r(x_i)) = (-1)^{i+1}\operatorname{sgn}(f(x_1) - r(x_1))$, $i = 1,\ldots,N$. Then $\Phi(f)(x_i) \neq \Phi(r)(x_i)$. Let $r_1 \in M$ be such that $\|\Phi(f) - \Phi(r_1)\| \leq \|\Phi(f) - \Phi(r)\|$. At x_i, $i = 1,\ldots,N$, $|w(x_i)(\Phi(f)(x_i) - \Phi(r_1)(x_i))| \leq |w(x_i)(\Phi(f)(x_i) - \Phi(r)(x_i))|$. Now at each x_i either $r(x_i) > f(x_i)$ or $r(x_i) < f(x_i)$ since Φ is pointwise fixed at f. If $r(x_i) > f(x_i)$ then $r(x_i) \geq r_1(x_i)$ for otherwise $r_1(x_i) > r(x_i) > f(x_i)$, which implies $|w(x_i)(\Phi(f)(x_i) - \Phi(r_1)(x_i))| > |w(x_i)(\Phi(f)(x_i) - \Phi(r)(x_i))|$ which is a contradiction. Similarly $r(x_i) < f(x_i)$ implies $r(x_i) \leq r_1(x_i)$. Now $r(x_i) > f(x_i)$ implies $r(x_i) \geq r_1(x_i)$ and $r(x_{i+1}) < f(x_{i+1})$ implies $r(x_{i+1}) \leq r_1(x_{i+1})$. Therefore $(r-r_1)(x_i) \geq 0$ and $(r-r_1)(x_{i+1}) \leq 0$ (or conversely)

for $i = 1,\ldots,N-1$. Set $s = r_1 - r = (f-r) - (f-r_1)$. If $s(x_i) \neq 0$ then $\operatorname{sgn} s(x_i) = \operatorname{sgn}(f(x_i) - r(x_i))$ (since $\operatorname{sgn}(f(x_i) - r(x_i)) = \operatorname{sgn}(r_1(x_i) - r(x_i))$, as noted above). If $s(x_i) \neq 0$ and $s(x_{i+1}) = s(x_{i+2}) = \ldots = s(x_{i+j-1}) = 0$, while $s(x_{i+j}) \neq 0$, then since $x_1,\ldots,x_N$ is an alternating set for $f-r$, $(-1)^k[f(x_k) - r(x_k)]$ is of one sign for $k = 1,2,\ldots,N$ and so $\operatorname{sgn} s(x_i) = (-1)^j \operatorname{sgn} s(x_{i+j})$. Therefore, if j is odd, $s(x)$ has an odd number of zeros (counting multiple zeros according to multiplicity) in $[x_i, x_{i+j}]$, while if j is even $s(x)$ has an even number of zeros in $[x_i, x_{i+j}]$. Since, by definition, s has at least $j-1$ zeros in $[x_i, x_{i+j}]$, we can therefore conclude that s has at least j zeros in $[x_i, x_{i+j}]$. Thus, s has at least $N-1$ zeros in $[a,b]$. But the zeros of s are the zeros of its numerator $p_1 q - q_1 p$, which is a polynomial of degree at most $\max\{\partial p_1 + \partial q, \partial q_1 + \partial p\} \leq N-2$. Hence $s \equiv 0$ and so $r \equiv r_1$.

(Necessity). Suppose that there exists $\{x_i\}_{i=1}^{N'} \subseteq X$, $N' < N$, N' maximal on which $|w(x_i)(\Phi(f)(x_i) - \Phi(r)(x_i))| = \|\Phi(f) - \Phi(r)\|$ and $\operatorname{sgn}(f(x_i) - r(x_i)) = (-1)^{i+1}\operatorname{sgn}(f(x_1) - r(x_1))$. First let us dispose of the case where the error curve is constant and $f-r$ has constant sign on X. (A technical difficulty which results in the general case requires us to consider this separately.) Without loss of generality we shall assume $w(x)(\Phi(f)(x) - \Phi(r)(x)) \equiv \|\Phi(f) - \Phi(r)\|$ and $f(x) - r(x) > 0$ for all $x \in X$. Then there exists $\epsilon_1 > 0$ such that $f(x) > r_1(x) > r(x)$ where $r_1(x) = \dfrac{p(x) + \epsilon_1}{q(x)}$. Also, since M is relatively open in $R_m^n[a,b]$, there exists $\epsilon_2 > 0$ with $0 < \epsilon_2 \leq \epsilon_1$ such that $r_2(x) = \dfrac{p(x) + \epsilon_2}{q(x)} \in M$ and $f(x) > r_2(x) > r(x)$. This implies $\|\Phi(f) - \Phi(r_2)\| < \|\Phi(f) - \Phi(r)\|$, which is a contradiction.

Now let $I_1, I_2, \ldots, I_{N'}$ be a collection of relatively open intervals in $[a,b]$ such that $x_i \in I_i$, $\bar{I}_i \cap \bar{I}_j = \phi$ for $i \neq j$, all extreme points $= \{x \in X : |w(x)(\Phi(f)(x) - \Phi(r)(x))| = \|\Phi(f) - \Phi(r)\|\} \subseteq \bigcup_{i=1}^{N'} I_i$ and for each extreme point in I_i the function $f-r$ has constant sign. Let

$$Y = X \cap \left(\bigcap_{i=1}^{N'} \tilde{I}_i \right)$$

where $\tilde{I}_i$ denotes the complement of I_i with respect to $[a,b]$. Y is a compact subset of X and $|w(x)(\Phi(f)(x) - \Phi(r)(x))| < \|\Phi(f) - \Phi(r)\|$ for all $x \in Y$. By continuity there exists $\rho > 0$ for which $\max_{x \in Y} |w(x)(\Phi(f)(x) - \Phi(r)(x))| \leq \|\Phi(f) - \Phi(r)\| - \rho$. Next, let

$$W_i = \{x \in X \cap \bar{I}_i : |w(x)(\Phi(f)(x) - \Phi(r)(x))| \geq \frac{\|\Phi(f) - \Phi(r)\|}{2}$$

$$\text{and } \operatorname{sgn}(f(x) - r(x)) = \operatorname{sgn}(f(x_i) - r(x_i))\}$$

where $\bar{I}_i$ denotes the closure of I_i in $[a,b]$. $W = \bigcup_{i=1}^{N'} W_i$ is a compact subset of X and so by continuity there exists an $\eta > 0$ such that $|f(x) - r(x)| \geq \eta$ on W. Set

$$Z_i = \{x \in X \cap \bar{I}_i : |w(x)(\Phi(f)(x) - \Phi(r)(x))| \geq \frac{\|\Phi(f) - \Phi(r)\|}{2}$$

$$\text{and } \operatorname{sgn}(f(x) - r(x)) \neq \operatorname{sgn}(f(x_i) - r(x_i))\}$$

and let $Z = \bigcup_{i=1}^{N'} Z_i$. Observe that $|w(x)(\Phi(f)(x) - \Phi(r)(x))| < \|\Phi(f) - \Phi(r)\|$ for all $x \in Z$ by the construction of the intervals $\{\bar{I}_i\}$. Finally, set

$$U_i = \{x \in X \cap \overline{I}_i : |w(x)(\Phi(f)(x) - \Phi(r)(x))| \le \frac{\|\Phi(f) - \Phi(r)\|}{2}\}$$

and let $U = \bigcup_{i=1}^{N'} U_i$. Then, by continuity, there exists $\delta > 0$, $\delta \le \rho$, such that

$$\max_{x \in Z \cup U} |w(x)(\Phi(f)(x) - \Phi(r)(x))| \le \|\Phi(f) - \Phi(r)\| - \delta .$$

Given $\tau > 0$ we can, using Lemma 1.8, select $r_\epsilon \in R_m^n[a,b]$ such that

$$\begin{aligned} &\text{(i)} \quad \|r_\epsilon - r\|_{L^\infty[a,b]} < \tau , \\ \text{and} \quad &\text{(ii)} \quad \operatorname{sgn}(r_\epsilon(x) - r(x)) = \operatorname{sgn}(f(x_i) - r(x_i)) \quad \text{for all } x \in I_i, \\ &\qquad i = 1,\ldots,N' . \end{aligned} \tag{1.2}$$

By the continuity of Φ we can select $\epsilon_1 > 0$ such that for $0 < \epsilon \le \epsilon_1$, r_ϵ satisfies (1.2) and

$$\max_{x \in Y \cup Z \cup U} |w(x)(\Phi(f)(x) - \Phi(r_\epsilon)(x))| \le \|\Phi(f) - \Phi(r)\| - \frac{\delta}{2} .$$

Next, by continuity of f and r, we can select ϵ_2, $0 < \epsilon_2 \le \epsilon_1$, such that for $0 < \epsilon \le \epsilon_2$, r_ϵ lies strictly between $f(x)$ and $r(x)$ on $W = \bigcup_{i=1}^{N'} W_i$. Hence by the strict monotonicity of Φ at f we have that

$$\max_{x \in W} |w(x)(\Phi(f)(x) - \Phi(r_\epsilon)(x))| < \|\Phi(f) - \Phi(r)\|.$$

Thus for ϵ with $0 < \epsilon \le \epsilon_2$, r_ϵ is such that $\|\Phi(f) - \Phi(r_\epsilon)\| < \|\Phi(f) - \Phi(r)\|$. Finally, since M is relatively open in $R_m^n[a,b]$, we can select ϵ_3 with $0 < \epsilon_3 \le \epsilon_2$ so that $r_{\epsilon_3} \in M$ and $\|\Phi(f) - \Phi(r_{\epsilon_3})\| < \|\Phi(f) - \Phi(r)\|$. Thus r_{ϵ_3} is a better starting approximation than r which is a contradiction and this concludes the proof of the theorem. ▮

Corollary 1.10: If $\Phi(f)$ has a best starting approximation under the setting of the above theorem than it is unique.

Proof: Uniqueness of best approximation is established in a manner analogous to that in which sufficiency was verified in Theorem 1.9. The technique of proof may be found in Rivlin [21, p. 123]. ▮

Next we turn our attention to the question of existence of best approximation. J.R. Rice [20] points out that the standard argument used to establish existence for classical rational approximation may fail if X is not an interval. Since we wish to employ the classical techniques we must also require that X be an interval. Before stating the existence theorem we prove the following lemmas:

Lemma 1.11: Let K be a convex subset of $C[a,b]$ and $\Phi: K \to C[a,b]$ be a continuous operator which is pointwise strictly monotone and pointwise fixed at $f \in K \sim M$ where $M = K \cap R_m^n[a,b]$. Further assume that for each $x \in [a,b]$ there exists a sequence $\{r_i\} \subseteq M$ with $\{r_i(x)\}$ unbounded from above and below and such that $|\Phi(r_i)(x)| \to \infty$ as $i \to \infty$. Then given $P > 0$ there exists a number $N(P)$ (independent of x) such that $|w(x)(\Phi(f)(x) - \Phi(r)(x))| \le P$ implies $|r(x)| \le N(P)$ where $r \in M$ and w is any continuous positive weight function on $[a,b]$.

Remark 1: It may well happen that all $r \in M$ are uniformly bounded from above or below on $[a,b]$. In this case we will already have the desired upper or lower uniform bound on $r(x)$ (for all $r \in M$). The above lemma is easily modified to cope with the aforementioned situations.

Proof: Let $t \in [a,b]$ be arbitrary but fixed. Then there exists $r_t, s_t \in M$ with $r_t(t) > f(t)$ and $s_t(t) < f(t)$ such that $|w(t)(\Phi(f)(t) - \Phi(r_t)(t))| > P + 1$ and $|w(t)(\Phi(f)(t) - \Phi(s_t)(t))| > P + 1$. By continuity there exists $\delta(t) > 0$ such that $|x-t| < \delta(t)$ and $x \in [a,b]$ imply $|w(x)(\Phi(f)(x) - \Phi(r_t)(x))| > P$ and $|w(x)(\Phi(f)(x) - \Phi(s_t)(x))| > P$. The neighborhoods $(t-\delta(t), t + \delta(t))$ form an open cover of $[a,b]$ and so by the compactness of $[a,b]$ there exists a finite subcover $\{(t_i-\delta(t_i), t_i + \delta(t_i))\}_{i=1}^{k}$. Set $N(P) = \max_{1\le i\le k} \{\|r_{t_i}\|, \|s_{t_i}\|\}$ to obtain the desired result.∎

The next lemma is analogous to a result found in Rice [20, p. 75].

Lemma 1.12: Let $r(x) = \dfrac{a_0 + a_1x +\ldots+ a_nx^n}{b_0 + b_1x +\ldots+ b_mx^m} \in R_m^n[a,b]$, $0 \le a < b$ where $b_0 + b_1x +\ldots+ b_mx^m \neq 0$ for all $x \in [a,b]$ and $\sum_{i=0}^{m} b_i^2 = 1$. If $\max_{x\in[a,b]} |r(x)| \le S$ then there exists $N(S)$ such that $\max_i |a_i| \le N(S)$, $\max_i |b_i| \le N(S)$.

Proof: $$|r(x)| \ge \frac{|\sum_{i=0}^{n} a_ix^i|}{\max_{x\in[a,b]} |\sum_{i=0}^{m} b_ix^i|} \ge \frac{|\sum_{i=0}^{n} a_ix^i|}{(m+1)B} \quad \text{where } B = \max\{b^m,1\}.$$

Hence $(m+1)BS \ge (m+1)B \max_{x\in[a,b]} |r(x)| \ge \max_{x\in[a,b]} |\sum_{i=0}^{n} a_ix^i|$.

Thus the polynomial $\sum_{i=0}^{n} a_ix^i$ is uniformly bounded on $[a,b]$ and so, by Markoff's Inequality, [1, p. 91], also the parameters a_i. That is, there exists $N_1(S)$ such that $\max_i |a_i| \le N_1(S)$. By the normalization of the denominator, $\max_i |b_i| \le 1$. Setting $N(S) = \max\{1,N_1(S)\}$ we have the desired result.∎

<u>Theorem 1.13</u>: Let $\Phi: C[a,b] \to C[a,b]$ be a continuous operator which is pointwise strictly monotone at $f \in C[a,b] \sim M$ and pointwise fixed at each $k \in C[a,b]$ where $M = R_m^n[a,b]$ $(m,n$ fixed). Assume that for each $x \in [a,b]$, $|\Phi(r_i)(x)| \to \infty$ whenever $\{r_i\} \subseteq M$ and $r_i(x) \to \pm\infty$. Then there exists a best starting approximation for $\Phi(f)$ from M.

<u>Proof</u>: Let $\{r_i\} \subseteq M$ be such that $\lim_{i\to\infty} \|\Phi(f) - \Phi(r_i)\| = \inf_{r\in M} \|\Phi(f) - \Phi(r)\|$. Let us first normalize the functions $r_i(x) = \frac{p_i(x)}{q_i(x)}$ so that $\sum_{j=0}^{m} b_{ij}^2 = 1$ where $q_i(x) = \sum_{j=0}^{m} b_{ij}x^j$. Clearly there exists $P > 0$ such that $\|\Phi(f) - \Phi(r_i)\| \le P$ for all i and so by Lemma 1.11 there exists $N(P)$ such that $\|r_i\| \le N(P)$ for all i. According to Lemma 1.12 the coefficients of the elements of $\{r_i\}$ are uniformly bounded. Therefore there exists a subsequence of $\{r_i(x)\}$ (which we shall also denote by $\{r_i(x)\}$) such that $\lim_{i\to\infty} r_i(x) = \bar{r}(x)$ for all $x \in [a,b]$ except for zeros of the denominator of $\bar{r}(x)$. If $\{x_i: i = 1,\ldots,k\}$ denotes the zeros of the denominator of $\bar{r}$ then it is clear that $k \le m$. Using the uniform boundedness of the sequence $\{r_i\}$, it is straightforward to show that each x_i is a removable singularity of $\bar{r}$. Upon cancellation, we obtain a rational function $r^* \in R_m^n[a,b]$ which we claim is the desired solution. First, using the pointwise fixed property of Φ it is easy to show that if $\{k_i\} \subseteq C[a,b]$, $k \in C[a,b]$, $x_0 \in [a,b]$ and $\lim_{i\to\infty} k_i(x_0) = k(x_0)$ then $\lim_{i\to\infty} \Phi(k_i)(x_0) = \Phi(k)(x_0)$. Let $\epsilon > 0$ be given. Then there exists $N = N(\epsilon)$ such that $\|\Phi(f) - \Phi(r_i)\| \le I + \epsilon$ for all $i \ge N$ where $I = \inf_{r\in M} \|\Phi(f) - \Phi(r)\|$.

For any $x_0 \in [a,b] \sim \bigcup_{i=1}^{k} \{x_i\}$, $|w(x_0)(\Phi(f)(x_0) - \Phi(r_i)(x_0))| \le I + \epsilon$ for $i \ge N$ and so $|w(x_0)(\Phi(f)(x_0) - \Phi(r^*)(x_0))| = \lim_{i\to\infty} |w(x_0)(\Phi(f)(x_0) - \Phi(r_i)(x_0))| \le I + \epsilon$. Thus $\|\Phi(f) - \Phi(r^*)\| = \max_{x\in[a,b]} |w(x)(\Phi(f)(x) - \Phi(r^*)(x))| = \sup_{\substack{x\in[a,b] \\ x \ne x_i}} |w(x)(\Phi(f)(x) - \Phi(r^*)(x))| \le I + \epsilon$. Hence $\|\Phi(f) - \Phi(r^*)\| = \inf_{r\in M} \|\Phi(f) - \Phi(r)\|$. ∎

In the next theorem we wish to consider a characterization of the best starting approximation for $\Phi(f)$ from a family of functions having restricted ranges. Much of the general theory for restricted range approximation by both polynomial and rational functions can be found in Taylor [25] and [26], Schumaker and Taylor [23], and Loeb, Moursund and Taylor [10]. It is interesting to note that the problem of approximation with restricted range also encompasses one-sided approximation [4]. The theory of uniform approximation with restricted ranges is of particular significance for the problem of obtaining starting values for various iterative schemes. This will become evident later on.

For our considerations let $\ell(x), u(x) \in C(X)$ with $\ell(x) < u(x)$ for all $x \in X$. K. Taylor [29] has pointed out that this inequality restriction on ℓ and u can be relaxed. Define $K = \{f \in C(X): \ell(x) \le f(x) \le u(x)\}$ and set $M = K \cap R_m^n[a,b]$ which we assume is nonempty.

<u>Theorem 1.14</u>: Let $\Phi: K \to C(X)$ be a continuous operator which is pointwise strictly monotone and pointwise fixed at $f \in K \sim M$ where K and M are as defined above. Then $r = \frac{p}{q} \in M$ is a best starting approximation for $\Phi(f)$ if and only if there exists $\{x_i\}_{i=1}^{N} \subseteq X$,

$N = 2 + \max\{n + \partial q,\ m + \partial p\}$, for which

(i) $x_1 < x_2 < \ldots < x_N$

(ii) $|w(x_i)(\Phi(f)(x_i) - \Phi(r)(x_i))| = \|\Phi(f) - \Phi(r)\|$,

$r(x_i) = u(x_i)$, or $r(x_i) = \ell(x_i)$

(iii) $\operatorname{sgn}^*(f(x_i) - r(x_i)) = (-1)^{i+1}\operatorname{sgn}^*(f(x_1) - r(x_1))$

where
$$\operatorname{sgn}^*(f(x) - r(x)) = \begin{cases} \operatorname{sgn}(f(x) - r(x)) & \text{if } r(x) \neq \ell(x) \text{ and } r(x) \neq u(x) \\ +1 & \text{if } r(x) = \ell(x) \\ -1 & \text{if } r(x) = u(x)\ . \end{cases}$$

Proof: (Sufficiency). Suppose $r \in M$ possesses the properties (i)-(iii) on the set $\{x_i\}_{i=1}^N$. Let $r_1 \in M$ be such that $\|\Phi(f) - \Phi(r_1)\| \le \|\Phi(f) - \Phi(r)\|$. If $r(x_i) = \ell(x_i)$ then $r(x_i) \le r_1(x_i)$ since $r_1 \in M$ and if $r(x_i) = u(x_i)$ then $r(x_i) \ge r_1(x_i)$. If $|w(x_i)(\Phi(f)(x_i) - \Phi(r)(x_i))| = \|\Phi(f) - \Phi(r)\|$ then $|w(x_i)(\Phi(f)(x_i) - \Phi(r)(x_i))| \ge |w(x_i)(\Phi(f)(x_i) - \Phi(r_1)(x_i))|$. Since $f \notin M$ we have that $\|\Phi(f) - \Phi(r)\| > 0$ so that $\operatorname{sgn}^*(f(x_i) - r(x_i)) \neq 0$. If $\operatorname{sgn}^*(f(x_i) - r(x_i)) = +1$ then $f(x_i) > r(x_i)$ and by the pointwise strict monotonicity of Φ at f we must have $r_1(x_i) \ge r(x_i)$. Likewise, if $\operatorname{sgn}^*(f(x_i) - r(x_i)) = -1$ then $r_1(x_i) \le r(x_i)$. Hence we obtain the desired result as in the proof of Theorem 1.9.

(Necessity). Suppose $r \in M$ is a best starting approximation for $\Phi(f)$ and that $r(x)$ has the desired behavior on a set of points $\{x_i\}_{i=1}^{N'} \subseteq X$, $N' < N$, where N' is maximal and $x_1 < x_2 < \ldots < x_{N'}$. First we shall consider the case where $w(x)(\Phi(f)(x) - \Phi(r)(x)) \equiv \lambda$, $|\lambda| > 0$, for all $x \in X$ and f-r has constant sign. A technical difficulty which results in the general case necessitates that we consider this situation separately. Without loss of generality assume that $f(x) - r(x) < 0$ for all $x \in X$. Then there exists $\epsilon > 0$ such that $f(x) < r^*(x) < r(x)$ where $r^*(x) = \frac{p(x) - \epsilon}{q(x)}$ and so $\|\Phi(f) - \Phi(r^*)\| < \|\Phi(f) - \Phi(r)\|$, which is a contradiction.

Let $I_1, I_2, \ldots, I_{N'}$ be a collection of relatively open intervals in $[a,b]$ such that $x_i \in I_i$, $\bar{I}_i \cap \bar{I}_j = \emptyset$ for $i \neq j$, all extreme points $= \{x \in X: |w(x)(\Phi(f)(x) - \Phi(r)(x))| = \|\Phi(f) - \Phi(r)\|,\ r(x) = \ell(x), \text{ or } r(x) = u(x)\} \subseteq \bigcup_{i=1}^{N'} I_i$ and for each extreme point in each I_i the function $\operatorname{sgn}^*(f(x) - r(x))$ has the same value. Given $\tau > 0$ we can, according to Lemma 1.8, select $r_\epsilon \in R_m^n[a,b]$ such that

$$\text{(i)} \quad \|r_\epsilon - r\|_{L^\infty[a,b]} < \tau , \tag{1.3}$$

and

$$\text{(ii)} \quad \operatorname{sgn}(r_\epsilon(x) - r(x)) = \operatorname{sgn}^*(f(x_i) - r(x_i)) \quad \text{for all } x \in I_i,\ i = 1,\ldots,N' .$$

We shall now show that there exists $\epsilon > 0$ such that $r_\epsilon \in M$ and $\|\Phi(f) - \Phi(r_\epsilon)\| < \|\Phi(f) - \Phi(r)\|$. Let $Y = X \cap (\bigcup_{i=1}^{N'} \tilde{I}_i)$. Y is a compact subset of X and $|w(x)(\Phi(f)(x) - \Phi(r)(x))| < \|\Phi(f) - \Phi(r)\|$ for all $x \in Y$. Hence, by the continuity of Φ, we can select $\epsilon_1 > 0$ such that for $0 < \epsilon \le \epsilon_1$, $r_\epsilon(x)$ satisfies (1.3) and $\max_{x \in Y} |w(x)(\Phi(f)(x) - \Phi(r_\epsilon)(x))| < \|\Phi(f) - \Phi(r)\|$. In addition $r(x) \neq u(x)$ and $r(x) \neq \ell(x)$ for all $x \in Y$ so there exists ϵ_2 with $0 < \epsilon_2 \le \epsilon_1$ such that $0 < \epsilon \le \epsilon_2$ implies $\ell(x) \le r_\epsilon(x) \le u(x)$ for $x \in Y$. Let

$$W_i = \{x \in X \cap \bar{I}_i : |w(x)(\Phi(f)(x) - \Phi(r)(x))| \ge \frac{\|\Phi(f) - \Phi(r)\|}{2} \text{ and } \operatorname{sgn}(f(x) - r(x)) = \operatorname{sgn}^*(f(x_i) - r(x_i))\}$$

and set

$$V_i = \{x \in X \cap \bar{I}_i : r(x) = u(x) \text{ or } r(x) = \ell(x)\},\ i = 1,2,\ldots,N'.$$

With these definitions we see that all extreme points contained in I_i are included in $W_i \cup V_i$. Using the strict monotonicity of Φ at f and the continuity of Φ there exists ϵ_3, with $0 < \epsilon_3 \le \epsilon_2$, such that for $0 < \epsilon \le \epsilon_3$, $r_\epsilon(x)$ satisfies (1.3),

$$\max_{x \in W_i \cup V_i} |w(x)(\Phi(f)(x) - \Phi(r_\epsilon)(x))| < \|\Phi(f) - \Phi(r)\| \text{ and}$$
$$\ell(x) \le r_\epsilon(x) \le u(x).$$

Do this for $i = 1,2,\ldots,N'$ and let ϵ_4 be such that $0 < \epsilon_4 \le \epsilon_3$ and for $0 < \epsilon \le \epsilon_4$, $r_\epsilon(x)$ satisfies (1.3),

$$\max_{x \in W \cup V} |w(x)(\Phi(f)(x) - \Phi(r_\epsilon)(x))| < \|\Phi(f) - \Phi(r)\| \text{ and}$$
$$\ell(x) \le r_\epsilon(x) \le u(x)$$

where $W = \bigcup_{i=1}^{N'} W_i$ and $V = \bigcup_{i=1}^{N'} V_i$. Set

$$Z_i = \{x \in X \cap \bar{I}_i : |w(x)(\Phi(f)(x) - \Phi(r)(x))| \ge \frac{\|\Phi(f) - \Phi(r)\|}{2}$$
$$\text{and } \operatorname{sgn}(f(x) - r(x)) \ne \operatorname{sgn}^*(f(x_i) - r(x_i))\}$$

and let $Z = \bigcup_{i=1}^{N'} Z_i$. By the construction of the intervals $\{I_i\}$ we have that $r(x) \ne u(x)$, $r(x) \ne \ell(x)$ and $|w(x)(\Phi(f)(x) - \Phi(r)(x))| < \|\Phi(f) - \Phi(r)\|$ for all $x \in Z$. Finally, let

$$U_i = \{x \in X \cap \bar{I}_i : |w(x)(\Phi(f)(x) - \Phi(r)(x))| \le \frac{\|\Phi(f) - \Phi(r)\|}{2}\}$$

and set $U = \bigcup_{i=1}^{N'} U_i$. By continuity there exists ϵ_5 with $0 < \epsilon_5 \le \epsilon_4$ so that for $x \in Z \cup U$ and $0 < \epsilon \le \epsilon_5$, $r_\epsilon(x)$ satisfies (1.3),

$|w(x)(\Phi(f)(x) - \Phi(r_\epsilon)(x))| < \|\Phi(f) - \Phi(r)\|$ and $\ell(x) \le r_\epsilon(x) \le u(x)$.

Combining the above results we have that for $0 < \epsilon \le \epsilon_5$, $r_\epsilon \in M$ and $\|\Phi(f) - \Phi(r_\epsilon)\| < \|\Phi(f) - \Phi(r)\|$. ∎

Corollary 1.15: Under the conditions of the above theorem the best starting approximation for $\Phi(f)$ is unique.

Proof: Uniqueness is derived from the above characterization theorem in a manner analogous to that in which Corollary 1.10 was established. ∎

Remark 1: The question of existence in the above setting is still unanswered. Thereom 1.13 is not applicable here since it requires that the domain of Φ be $C[a,b]$.

In 1955 S. Paszkowski (see [2]) considered the problem of approximating a given continuous function on an interval $[a,b]$ by elements of an n-dimensional Haar subspace (to be defined) which were also required to interpolate the given function at certain prescribed points in $[a,b]$ called nodes. He considered, using a classical approach, the questions of existence, uniqueness and characterization of best approximation. In 1968 Deutsch [2] extended and continued the study of Paszkowski. Using duality theory, Deutsch was able to recover the characterization theorem of Paszkowski and to actually formulate exchange algorithms for computation. In 1969, Loeb, Moursund, Schumaker and Taylor [9] generalized the above work of Paszkowski and Deutsch to the following problem: let X be a compact subset of $[a,b]$ and suppose $\{x_i\}_{i=1}^{p} \subseteq X$ satisfies $x_1 < x_2 < \ldots < x_p$. Let $\{m_i\}_{i=1}^{p}$ be a set of positive integers with $m = \sum_{i=1}^{p} m_i < n$ and assume X contains at least $n - m + p + 1$ points. Suppose further that H is an n-dimensional extended Haar system of order ν, $\nu = \max_i (m_i) + 1$, (which we shall define) and set

$$M = \{p \in H : p^{(j)}(x_i) = a_{ij},\ 1 \le i \le p \text{ and } 0 \le j \le m_i - 1\},$$

where $\{a_{ij}\}$ is a set of m real numbers. Then the problem considered was to find best approximations to elements of

$$K = \{f \in C(X) : f(x_i) = a_{i0},\ 1 \le i \le p\}$$

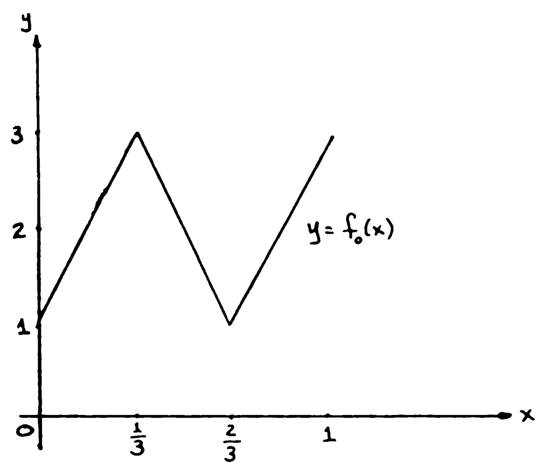
by elements of M. The error of approximation was measured by a generalized weight function, a notion to be discussed in Chapter IV. We shall refer to the elements of M as p-point osculating Haar functions. A. Perrie [19] has examined a problem analogous to the above by replacing the extended Haar system by classes of rational functions. The approximants used by Perrie are referred to as p-point osculating rational functions.

We shall now examine a problem analogous to that considered by Perrie for our operator setting. Let $X \subseteq [a,b]$ be compact, $\{y_i\}_{i=1}^{p}$ a fixed set of p points in X where $y_1 < y_2 <\ldots< y_p$ and $\{m_i\}_{i=1}^{p}$ a fixed set of positive integers with $m^* = \sum_{i=1}^{p} m_i <$ $n + 1$ (n as in R_m^n). We shall assume that X contains at least $n + m - m^* + p + 2$ points. Furthermore, let $\{a_i\}_{i=1}^{p}$ be any fixed set of p real numbers and define $K = \{f \in C(X) : f(y_i) = a_i,$ $i = 1,\ldots,p\}$. Let a_{ij}, $i = 1,\ldots,p$, $j = 0,1,\ldots,m_i-1$ be a second set of real numbers where $a_{i0} = a_i$, $i = 1,2,\ldots,p$ and set $M = \{r \in R_m^n[a,b] : r^{(j)}(y_i) = a_{ij},\ 0 \le j \le m_i-1,\ 1 \le i \le p\}$. Following the discussion in Perrie, we shall refer to the elements of M as p-point osculating rational functions. As before we are interested in approximating $g \in \Phi(K)$ by elements of $\Phi(M)$ where $\Phi: K \to C(X)$

is continuous.

In the case of ordinary rational approximation we know that for each $f \in C[a,b]$ there exists a best approximation from $R_m^n[a,b]$. However, in the case of interpolating rational functions we can no longer insure existence as is demonstrated by the following example due to H.L. Loeb [8]: Define $f_0 \in C[0,1]$ as follows:

$$f_0(x) = \begin{cases} 1 + 6x, & x \in [0, \frac{1}{3}] \\ 3 - 6(x - \frac{1}{3}), & x \in [\frac{1}{3}, \frac{2}{3}] \\ 1 + 6(x - \frac{2}{3}), & x \in [\frac{2}{3}, 1] \end{cases}$$



Here $a = 0$, $b = 1$, $m = n = 1$, $w(x) \equiv 1$. Let $K = \{f \in C[0,1] : f(0) = 1\}$. Obviously $f_0(x) \in K$. Define the operator $\Phi: K \to C[0,1]$ via $\Phi(f) = f$. It is clear that Φ is pointwise strictly monotone and pointwise fixed at $f_0 \in K$. By the usual alternation theory for rational approximation $g(x) \equiv 2$ is a best approximation to $f_0(x)$ from $R_1^1[0,1]$. Notice that $2 \notin M = \{r \in R_1^1[0,1] : r(0) = 1\}$.

<u>Claim</u>: There exists no $r \in M$ such that $\|\Phi(f_0) - \Phi(r)\| = \|r - f_0\| = \text{dist}(f_0, M) = \text{dist}(\Phi(f_0), \Phi(M))$.

Proof: For each $n = 1,2,\ldots,$ define $r_n(x) = \dfrac{2nx+1}{nx+1}$, $x \in [0,1]$. Clearly $r_n \in M$ for all n since $r_n(0) = 1$. Also $r_n'(x) = \dfrac{n}{(nx+1)^2} > 0$ and $r_n(1) = \dfrac{2n+1}{n+1} < 2$ for all n. Hence $1 \le r_n(x) \le 2$ for all n and for all $x \in [0,1]$. Furthermore $f_0(1) - r_n(1) > 1$ and $\max\{|f_0(x) - r_n(x)| : x \in [0, \frac{1}{6}]\} \le 1$ for all n. Therefore $\|f_0 - r_n\| = \max\{|f_0(x) - r_n(x)| : x \in [\frac{1}{6}, 1]\}$. Now $\{r_n\}$ approaches 2 uniformly on $[\frac{1}{6}, 1]$ and so $\lim_{n\to\infty} \|f_0 - r_n\| = \lim_{n\to\infty} \max\{|f_0(x) - r_n(x)| : x \in [\frac{1}{6}, 1]\} = \max\{|f_0(x) - 2| : x \in [\frac{1}{6}, 1]\} = \|f - 2\|$. Since $M \subseteq R_1^1[0,1]$, $\|f_0 - 2\| \le \inf\{\|f_0 - r\| : r \in M\}$. Hence $\inf\{\|f_0 - r\| : r \in M\} = \|f_0 - 2\|$. But, by uniqueness in $R_1^1[0,1]$, there exists no $r^* \in M$ such that $\|f_0 - r^*\| = \|f_0 - 2\| = \inf\{\|f_0 - r\| : r \in M\}$ and hence the claim. ∎

In order to establish the characterization theorem for the above setting we need to introduce the following definitions and concepts. First we wish to define the notion of an n-dimensional extended Haar subspace of order ν of $C[a,b]$. In doing this, we shall follow the notation and development given in [5]. For completeness we first define what is meant by a Haar subspace of $C[a,b]$.

Definition 1.16: A finite-dimensional subspace H of $C[a,b]$ is called a Haar subspace if each nonzero function in H possesses at most $n-1$ zeros, where $n = \dim H$.

Let $\{u_i(t)\}_{i=1}^n$ be a family of functions in $C[a,b]$ (with each sufficiently differentiable so that what follows is well-defined) and $a \le t_1 \le t_2 \le \ldots \le t_n \le b$ a given set of points. Define

$$U^*\begin{pmatrix}1,\dots,n\\ t_1,\dots,t_n\end{pmatrix}=\begin{vmatrix}\tilde u_1(t_1) & \tilde u_1(t_2) & \dots & \tilde u_1(t_n)\\ \tilde u_2(t_1) & \tilde u_2(t_2) & \dots & \tilde u_2(t_n)\\ \vdots & \vdots & & \vdots\\ \tilde u_n(t_1) & \tilde u_n(t_2) & \dots & \tilde u_n(t_n)\end{vmatrix}$$

where for each fixed i:

$$\tilde u_i(t_j)=\begin{cases}u_i(t_j) & \text{if } t_{j-1}<t_j\ ,\\ u_i^{(r)}(t_j) & \text{if } t_{j-r}=t_{j-r+1}=\dots=t_j,\ 1\le i\le n.\end{cases}$$

Definition 1.17: The functions $u_1,\dots,u_n$ will be called an extended Chebyshev system of order ν on $[a,b]$ provided $u_i\in C^{\nu-1}[a,b]$, $i=1,\dots,n$, and $U^*\begin{pmatrix}1,\dots,n\\ t_1,\dots,t_n\end{pmatrix}>0$ for all choices $t_1\le t_2\le\dots\le t_n$, $t_i\in[a,b]$, where equality occurs in groups of at most ν consecutive values of t_i.

With the above we are able to define the notion of an extended Haar subspace.

Definition 1.18: An n-dimensional subspace H of $C[a,b]$ is said to be an extended Haar subspace of order ν, $\nu\le n$, if there exists a basis for H which is an extended Chebyshev system of order ν on $[a,b]$.

Remark 1: It is straightforward to show that H is an extended Haar subspace of order ν if and only if each nonzero element in H possesses at most $n-1$ zeros in $[a,b]$ counting multiple zeros in the following manner:

(1) if $f^{(j)}(z)=0$, $j=0,\dots,\nu-1$, we say that z is a zero of multiplicity ν;

(2) otherwise we say that z is a zero of multiplicity $m = \min\{j: f^{(j)}(z) \neq 0, j = 0,\ldots,\nu-1\}$.

For a fixed $r = \frac{p}{q} \in R_m^n[a,b]$, we shall write $P_n + rP_m$ to denote the subspace $\{p + rq : p \in P_n, q \in P_m\}$ of $C[a,b]$. It is well-known (see [19]) that $P_n + rP_m$ is an extended Haar subspace of order $\nu = \dim(P_n + rP_m) = 1 + \max\{n + \partial q, m + \partial p\}$. Define

$$S(r) = \{h \in P_n + rP_m : h^{(j)}(y_i) = 0, \ i = 1,\ldots,p, \ j = 0,\ldots,m_i-1\}.$$

It is clear that $S(r)$ is a subspace of $P_n + rP_m$. Let us now restrict P_n, P_m and r to X and view $S(r)$ as a subspace of $C(X)$. The following lemma, whose proof is modeled after an analogous result from the work of Perrie (see [19, p. 13]), yields the exact dimension of the subspace $S(r)$ of $C(X)$.

<u>Lemma 1.19</u>: $\dim S(r) = 1 + \max\{n + \partial q, m + \partial p\} - m^* \ (= d)$.

<u>Proof</u>: Let $x_1,\ldots,x_d$ be d distinct points in X which are distinct from the interpolating points $y_1,\ldots,y_p$. For each i, $i = 1,\ldots,d$, choose $h_i \in P_n + rP_m$ such that $h_i \in S(r)$ and $h_i(x_j) = \delta_{ij}$, $i,j = 1,2,\ldots,d$. The set $\{h_i\}_{i=1}^d$ is clearly linearly independent. For $h \in S(r)$, $g = \sum_{i=1}^{d} h(x_i)h_i$ is an element of $S(r)$ and $h-g$ possesses at least $1 + \max\{n + \partial q, m + \partial p\}$ zeros in X. But the numerator of $h-g$ is an element of $P_{\max\{n+\partial q, m+\partial p\}}$ and so $h \equiv g$. Thus $\{h_1,\ldots,h_d\}$ is a basis for $S(r)$. ∎

If $\{g_1,\ldots,g_d\}$ represents a basis for $S(r)$ then for each $x \in X$ we shall write $\hat{x}$ to denote the vector $(g_1(x),\ldots,g_d(x)) \in E^d$ (E^1 = real numbers). If Y is a subset of X and ψ is a real-valued function defined on Y, then $H\{\psi(y)\hat{y} : y \in Y\}$ denotes the convex hull of the vectors $\psi(y)\hat{y} \in E^d$, for $y \in Y$.

We conclude our preparatory discussion with an important result due to Salzer [22].

<u>Lemma 1.20</u>: Let $f \in C^s(X)$ where $s = \max_{1 \le i \le p} \{m_i - 1\}$. At points where $q(y_i) \neq 0$, the system

$$\left(\frac{p}{q}\right)^{(j)}(y_i) = f^{(j)}(y_i),\ j = 0,1,\ldots,m_i-1,\ i = 1,\ldots,p,$$

is equivalent to

$$p^{(j)}(y_i) = (fq)^{(j)}(y_i),\ j = 0,1,\ldots,m_i-1,\ i = 1,\ldots,p.$$

Finally we are able to establish the following characterization theorem.

<u>Theorem 1.21</u>: Let $\Phi: K \to C(X)$ be a continuous operator which is pointwise strictly monotone and pointwise fixed at $f \in K \sim M$ where K and M have been previously defined. Then the following four statements are equivalent:

(i) $r^* = \frac{p^*}{q^*} \in M$ is a best starting approximation for $\Phi(f)$ from M.

(ii) $0 \in H\{\sigma(x)\hat{x} : x \in X(r^*)\}$ where $\sigma(x) = \mathrm{sgn}(f(x) - r^*(x))$, $X(r^*) = \{x \in X : |w(x)(\Phi(f)(x) - \Phi(r^*)(x))| = \|\Phi(f) - \Phi(r^*)\|\}$, $\{g_1,\ldots,g_d\}$ is a basis for $S(r^*)$ and $\hat{x} = (g_1(x),\ldots,g_d(x))$ where $d = \max\{n + \partial q^*, m + \partial p^*\} + 1 - m^*$.

(iii) There exist $d + 1$ consecutive points $x_1 < x_2 < \ldots < x_{d+1}$ in $X \sim \bigcup_{i=1}^{p} \{y_i\}$ such that

(a) $|w(x_i)(\Phi(f)(x_i) - \Phi(r^*)(x_i))| = \|\Phi(f) - \Phi(r^*)\|$, $i = 1,2,\ldots,d+1$,

(b) $\operatorname{sgn}\{(f(x_i) - r^*(x_i))\Delta_i\} = (-1)^{i+1}\operatorname{sgn}\{(f(x_1) - r^*(x_1))\Delta_1\}$, $i = 1,\ldots,d+1$, where

$$\Delta_i = \begin{vmatrix} g_1(x_1) & g_1(x_2) & \cdots & g_1(x_{i-1}) & g_1(x_{i+1}) & \cdots & g_1(x_{d+1}) \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ g_d(x_1) & g_d(x_2) & \cdots & g_d(x_{i-1}) & g_d(x_{i+1}) & \cdots & g_d(x_{d+1}) \end{vmatrix}$$

(iv) There exist $d + 1$ consecutive points $x_i \in X \sim \bigcup_{i=1}^{p} \{y_i\}$ such that

(a) $|w(x_i)(\Phi(f)(x_i) - \Phi(r^*)(x_i))| = \|\Phi(f) - \Phi(r^*)\|$, $i = 1,2,\ldots,d+1$,

(b) $\operatorname{sgn}\{[f(x_i) - r^*(x_i)]\Pi(x_i)\} = (-1)^{i+1}\operatorname{sgn}\{[f(x_i) - r^*(x_i)]\Pi(x_i)\}$ for $i = 1,2,\ldots,d+1$, where $\Pi(t) = (y_1 - t)^{m_1}\cdots(y_p - t)^{m_p}$ if $p \neq 0$ and $\Pi(t) \equiv 1$ if $p = 0$.

Proof: (i) implies (ii):

Suppose $0 \notin H\{\sigma(x)\hat{x} : x \in X(r^*)\}$. By the Theorem on Linear Inequalities [1] there exists $h \in S(r^*)$ such that $\sigma(x)h(x) > 0$ for all $x \in X(r^*)$. Set $h = p - r^*q$ where $p \in P_n$ and $q \in P_m$. We then have that $h^{(j)}(y_i) = 0$, $i = 1,\ldots, p$, $j = 0,1,\ldots,m_i-1$. Set $r_\lambda = \dfrac{p^* - \lambda p}{q^* - \lambda q}$. Now there exists $\lambda_0 > 0$ such that $q^*(x) - \lambda q(x) > 0$ for $|\lambda| \le \lambda_0$ and for all $x \in X$ since $q^*(x)$ is positive and $q(x)$ is bounded on the compact set X. For $|\lambda| \le \lambda_0$, we have $r_\lambda \in M$ since $r^* - r_\lambda = \dfrac{p^*}{q^*} - \dfrac{p^* - \lambda p}{q^* - \lambda q} = \dfrac{\lambda q^*(p - r^* q)}{q^*(q^* - \lambda q)} = \dfrac{\lambda h}{q^* - \lambda q}$, and so $h^{(j)}(y_i) = 0$, $i = 1,\ldots,p$, $j = 0,1,\ldots,m_i-1$, implies that $r_\lambda^{(j)}(y_i) = r^{*(j)}(y_i) = a_{ij}$, $i = 1,2,\ldots,p$, $j = 0,1,\ldots,m_i-1$. Since

$X(r^*)$ is compact, the number $\delta = \min\{\sigma(x)h(x) : x \in X(r^*)\} > 0$. Set $E = \|\Phi(f) - \Phi(r^*)\|$ and $X_1 = \{x \in X : \sigma(x)h(x) > \frac{1}{2}\delta$ and $|w(x)(\Phi(f)(x) - \Phi(r^*)(x))| > \frac{E}{2}\}$. Note that $X(r^*) \subseteq X_1$ and X_1 is open; hence $X_2 = X \sim X_1$ is a closed subset of X and thus compact. By the continuity of Φ and the compactness of X_2 there exists λ_1 such that $0 < \lambda_1 \leq \lambda_0$ and $|\lambda| \leq \lambda_1$ implies $|w(x)(\Phi(f)(x) - \Phi(r_\lambda)(x))| < E$ for all $x \in X_2$ since $|w(x)(\Phi(f)(x) - \Phi(r^*)(x))| < E$ for all $x \in X_2$. Set

$$\mu = \inf\{|f(x) - r^*(x)| : |w(x)(\Phi(f)(x) - \Phi(r^*)(x))| \geq \frac{E}{2} \text{ and } x \in X\}.$$

Note that $\mu > 0$. Now choose λ_2 such that $0 < \lambda_2 \leq \lambda_1$ and $|\lambda| \leq \lambda_2$ implies $|\lambda| \left\|\frac{h}{q^*-\lambda q}\right\| \leq \frac{\mu}{2}$. For $x \in X_1$, $|r^*(x) - r_\lambda(x)| \leq \|r^* - r_\lambda\| = |\lambda| \left\|\frac{h}{q^*-\lambda q}\right\| \leq \frac{\mu}{2} < |f(x) - r^*(x)|$. Consequently $(f - r^*)(x)$ and $(f - r_\lambda)(x)$ have the same signum function for $x \in X_1$. For $\lambda < 0$ with $|\lambda| \leq \lambda_2$ we have $\sigma(x)(r^*(x) - r_\lambda(x)) = \frac{\lambda\sigma(x)h(x)}{q^*(x)-\lambda q(x)} < 0$ for all $x \in X_1$. Therefore $\sigma(x)(r^*(x) - r_\lambda(x)) < 0$ for all $x \in X_1$ and so $\operatorname{sgn}(f(x) - r^*(x)) = -\operatorname{sgn}(r^*(x) - r_\lambda(x))$ for all $x \in X_1$. Hence for $f(x) > r^*(x)$ we have $f(x) > r_\lambda(x) > r^*(x)$ and for $f(x) < r^*(x)$ we have $f(x) < r_\lambda(x) < r^*(x)$. Applying the pointwise strict monotonicity of Φ at f we find that $|w(x)(\Phi(f)(x) - \Phi(r_\lambda)(x))| < |w(x)(\Phi(f)(x) - \Phi(r^*)(x))| \leq \|\Phi(f) - \Phi(r^*)\| = E$ for all $x \in X_1$ and λ as described above. Assembling the above results we have $|w(x)(\Phi(f)(x) - \Phi(r_\lambda)(x))| < E$ for all $x \in X$ (λ as above) and so $\|\Phi(f) - \Phi(r_\lambda)\| < \|\Phi(f) - \Phi(r^*)\|$. Therefore r_λ is a better starting approximation than r^*.

(ii) implies (iii):

Suppose $0 \in H\{\sigma(x)\hat{x} : x \in X(r^*)\}$. Then by Caratheodory's Theorem [1, p.17] there exist d_1 ($d_1 \leq d+1$) positive scalars λ_i with $\sum_{i=1}^{d_1} \lambda_i = 1$ and d_1 consecutive points $x_1, x_2, \ldots, x_{d_1} \in X(r^*)$ such that

$$\sum_{i=1}^{d_1} \lambda_i \sigma(x_i) g_j(x_i) = 0 \quad \text{for} \quad j = 1,2,\ldots,d. \tag{1.4}$$

We claim that $d_1 = d+1$. Suppose $d_1 \leq d$. Then $\det[g_j(\bar{x}_i)] = 0$, $i = 1,\ldots,d$, $j = 1,\ldots,d$ where $\bar{x}_i = x_i$, $i = 1,\ldots,d_1$ and $\{\bar{x}_i\}_{i=d_1+1}^{d} \cap (\{x_i\}_{i=1}^{d_1} \cup \{y_i\}_{i=1}^{p}) = \emptyset$. This then implies that there exists a d-tuple $\alpha = (\alpha_1,\ldots,\alpha_d)$, $\alpha \neq (0,0,\ldots,0)$ such that $\sum_{j=1}^{d} \alpha_j g_j(\bar{x}_i) = 0$, $i = 1,\ldots,d$. Set $g = \sum_{j=1}^{d} \alpha_j g_j$. Then $g \in S(r^*)$ and so g has, in addition to simple zeros at the $\bar{x}_i$, $i = 1,\ldots,d$, zeros of order m_i at the points y_i, $i = 1,\ldots,p$. In total, g possesses at least $d + m^* = \max\{n + \partial q^*, m + \partial p^*\} + 1 - m^* + m^* = \max\{n + \partial q^*, m + \partial p^*\} + 1$ zeros. But the numerator of g is an element of $P_{\max\{n+\partial q^*, m+\partial p^*\}}$ and so $g \equiv 0$. This is a contradiction and so $d_1 = d + 1$. Rewriting (1.4) as

$$\sum_{i=2}^{d+1} \lambda_i \sigma(x_i) g_j(x_i) = -\lambda_1 \sigma(x_1) g_j(x_1), \; j = 1,2,\ldots,d$$

and solving by Cramer's rule, we obtain

$$\lambda_i \sigma(x_i) = \frac{(-1)^{i+1} \lambda_1 \sigma(x_1) \Delta_i}{\Delta_1}, \; i = 2,\ldots,d+1 .$$

We now claim that $\Delta_i \neq 0$ for all i. On the contrary suppose that $\Delta_i = 0$ for some i. Define h_i by

$$h_i(s) = \begin{vmatrix} g_1(s) & g_1(x_2) & \cdots & g_1(x_{i-1}) & g_1(x_{i+1}) & \cdots & g_1(x_{d+1}) \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ g_d(s) & g_d(x_2) & \cdots & g_d(x_{i-1}) & g_d(x_{i+1}) & \cdots & g_d(x_{d+1}) \end{vmatrix}$$

Then $h_i(x_1) = h_i(x_2) = \ldots = h_i(x_{i-1}) = h_i(x_{i+1}) = \ldots = h_i(x_{d+1}) = 0$ and since $h_i \in S(r^*)$ we have that h_i possesses at least $d + m^*$ zeros. But $h_i \not\equiv 0$ since $P_n + r^*P_m$ is an extended Haar subspace of order $1 + \max\{n + \partial q^*, m + \partial p^*\}$. Hence we have a contradiction. Now $\lambda_i > 0$ and $\sigma(x_i) = \operatorname{sgn}(f(x_i) - r^*(x_i))$ and so the result follows easily.

(iii) implies (iv):

Let $\{z_1,\ldots,z_{d-1}\}$ be an arbitrary set of $d-1$ consecutive points in $X \sim \bigcup_{i=1}^{p} \{y_i\}$ and define

$$g(s) = \begin{vmatrix} g_1(s) & g_1(z_1) & \cdots & g_1(z_{d-1}) \\ \vdots & \vdots & & \vdots \\ g_d(s) & g_d(z_1) & \cdots & g_d(z_{d-1}) \end{vmatrix}$$

Note that $g \in S(r^*)$ since g is a linear combination of the basis elements of $S(r^*)$. Furthermore, $g(s)$ has a nonzero coefficient by the remark at the end of (ii) implies (iii). By construction, $g(s)$ has $d-1 + m^* = \max\{n + \partial q^*, m + \partial p^*\}$ zeros counting multiplicities. These consist of a zero of order m_i at each y_i and a simple zero at each z_i (the zeros at z_i are simple since $g \not\equiv 0$). Hence g changes sign at each z_i and changes sign at y_i if and only if m_i is odd. Set $z_i = x_{i+2}$, $i = 1,\ldots,d-1$, for the choice of x_i's in part (iii). Then

$$g(x_1) = \begin{vmatrix} g_1(x_1) & g_1(z_1) & \cdots & g_1(z_{d-1}) \\ \vdots & \vdots & & \vdots \\ g_d(x_1) & g_d(z_1) & \cdots & g_d(z_{d-1}) \end{vmatrix} = \begin{vmatrix} g_1(x_1) & g_1(x_3) & \cdots & g_1(x_{d+1}) \\ \vdots & \vdots & & \vdots \\ g_d(x_1) & g_d(x_3) & \cdots & g_d(x_{d+1}) \end{vmatrix} = \Delta_2 .$$

Similarly $g(x_2) = \Delta_1$. Now if $y_{i_1},\ldots,y_{i_q} \in (x_1,x_2)$, then

$\operatorname{sgn} \Delta_2 = \operatorname{sgn} g(x_1) = (-1)^{m_{i_1}+\ldots+m_{i_q}} \operatorname{sgn} g(x_2) = (-1)^{m_{i_1}+\ldots+m_{i_q}} \operatorname{sgn} \Delta_1$

$= \operatorname{sgn} \dfrac{\Pi(x_1)}{\Pi(x_2)} \operatorname{sgn} \Delta_1$. Thus $\operatorname{sgn}\{[f(x_2) - r^*(x_2)]\Pi(x_2)\} =$

$\operatorname{sgn}\{[f(x_2) - r^*(x_2)]\}\operatorname{sgn} \Pi(x_2) = \operatorname{sgn}\{[f(x_2) - r^*(x_2)]\} \dfrac{\operatorname{sgn} \Delta_2}{\operatorname{sgn} \Delta_1} \operatorname{sgn} \Pi(x_1) =$

$\operatorname{sgn}\{[f(x_2) - r^*(x_2)]\Delta_2\} \dfrac{\operatorname{sgn} \Pi(x_1)}{\operatorname{sgn} \Delta_1} = -\operatorname{sgn}\{[f(x_1) - r^*(x_1)]\Delta_1\} \dfrac{\operatorname{sgn} \Pi(x_1)}{\operatorname{sgn} \Delta_1} =$

$-\operatorname{sgn}\{[f(x_1) - r^*(x_1)]\Pi(x_1)\}$.

Continuing in this manner, we obtain the desired result.

(iv) implies (i):

Suppose there exists $r \in M$ such that $\|\Phi(f) - \Phi(r)\| \le \|\Phi(f) - \Phi(r^*)\|$. Now $r^*(x_i) < f(x_i)$ implies $r^*(x_i) \le r(x_i)$ and similarly $r^*(x_i) > f(x_i)$ implies $r^*(x_i) \ge r(x_i)$. Thus, using (iv), we have that $[r^*(x_i) - r(x_i)]\Pi(x_i) \ge 0$ and $[r^*(x_{i+1}) - r(x_{i+1})]\Pi(x_{i+1}) \le 0$ (or conversely) for $i = 1,\ldots,d$. Set $s = r - r^*$. If $s(x_i) \ne 0$ and $s(x_{i+1}) =\ldots= s(x_{i+j-1}) = 0$, while $s(x_{i+j}) \ne 0$, then since $(-1)^k[(r^*(x_k) - f(x_k))\Pi(x_k)]$ is of one sign for $k = 1,\ldots,d+1$ we have that $\operatorname{sgn}\{[s(x_i)]\Pi(x_i)\} = \operatorname{sgn}\{[f(x_i) - r^*(x_i)]\Pi(x_i)\} = (-1)^j\operatorname{sgn}\{[f(x_{i+j}) - r^*(x_{i+j})]\Pi(x_{i+j})\} = (-1)^j\operatorname{sgn}\{[s(x_{i+j})]\Pi(x_{i+j})\}$. Now we have two cases to consider.

First assume that the sum of the m_k's such that $y_k \in (x_i, x_{i+j})$ is even. Then $\Pi(x_i)\Pi(x_{i+j}) > 0$ and $\operatorname{sgn} s(x_i) = (-1)^j \operatorname{sgn} s(x_{i+j})$. Thus if j is odd, $s(x)$ has an odd number of zeros (counting multiple zeros according to multiplicity) in $[x_i, x_{i+j}]$, while if j is even $s(x)$ has an even number of zeros in $[x_i, x_{i+j}]$. Since, by definition, s has at least $j-1 + \Sigma m_k$ (sum over $y_k \in (x_i, x_{i+j})$) zeros in $[x_i, x_{i+j}]$, we can therefore conclude that s has at least $j + \Sigma m_k$ zeros in $[x_i, x_{i+j}]$. Counting zeros, we find that s has at least $d + m^* = 1 + \max\{n + \partial q^*, m + \partial p^*\}$ zeros in $[a,b]$. But the zeros of s are the zeros of its numerator $p^* q - q^* p$, which is a polynomial of degree at most $\max\{n + \partial q^*, m + \partial p^*\}$. Hence $s \equiv 0$. Similarly, we obtain the same conclusion if the sum of the m_k's, such that $y_k \in (x_i, x_{i+j})$, is odd. Thus $r \equiv r^*$. ∎

In order to establish uniqueness of best approximation in the above setting we require the following lemma:

<u>Lemma 1.22</u>: Let $r^* \in M$ be a best starting approximation to $f \in K \sim M$ where K and M are as previously defined and $\Phi: K \to C(X)$ is pointwise strictly monotone and pointwise fixed at f. Then $h \in S(r^*)$ and $h(x)\sigma(x) \geq 0$ for all $x \in X(r^*)$ imply $h \equiv 0$.

<u>Proof</u>: Let $h \in S(r^*)$ with $h(x)\sigma(x) \geq 0$ for all $x \in X(r^*)$ and suppose $h \not\equiv 0$. Let $z_1, z_2, \ldots, z_t$ be the zeros of $h(x)$ which are in $X(r^*)$. Notice that $z_i \neq y_j$ for any pair i,j since $f(y_i) = r^*(y_i)$ and $\|\Phi(f) - \Phi(r^*)\| > 0$. Furthermore $t < d = \max\{n + \partial q^*, m + \partial p^*\} + 1 - m^*$ since if $t \geq d$ then h possesses at least $1 + \max\{n + \partial q^*, m + \partial p^*\}$ zeros which is impossible since the numerator of h is an element of $P_{\max\{n+\partial q^*, m+\partial p^*\}}$. From the standard interpolation theory we know that there exists $g \in S(r^*)$

such that $g(z_i) = \sigma(z_i)$, $i = 1,2,\ldots,t$. Since $\sigma(x)h(x) \geq 0$ for all $x \in X(r^*)$, we claim that there exists $\lambda > 0$ such that $\sigma(x)(h(x) + \lambda g(x)) > 0$ for all $x \in X(r^*)$. For each $i = 1,\ldots,t$, choose an open interval, $I(z_i)$, about z_i such that $\sigma(x)g(x) > 0$ for all $x \in I(z_i)$ and $I(z_i) \cap I(z_j) = \emptyset$ for $i \neq j$. This we are able to do since $\sigma(z_i)g(z_i) = (\sigma(z_i))^2 > 0$ and $\sigma(x)g(x)$ is continuous at $x = z_i$. Set $Y = X(r^*) \cap \left(\bigcap_{i=1}^{t} \{\widetilde{I(z_i)}\}\right)$. Y is closed in $[a,b]$, hence compact, and so $\sigma(x)h(x) > 0$ for all $x \in Y$. Let $m = \inf\{\sigma(x)h(x) : x \in Y\}$ and $M = \inf\{\sigma(x)g(x) : x \in Y\}$. Note that $m > 0$ and $M > -\infty$. If $M \geq 0$, let $\lambda > 0$ be arbitrary; if $M < 0$, choose $\lambda > 0$ such that $m + \lambda M > 0$. Then $\sigma(x)(h(x) + \lambda g(x)) > 0$ for all $x \in X(r^*)$. Since $S(r^*)$ is a subspace of $P_n + r^*P_m$ and $h,g \in S(r^*)$ we have that $h + \lambda g \in S(r^*)$. Thus, by the Theorem on Linear Inequalities, $0 \notin H\{\sigma(x)\hat{x} : x \in X(r^*)\}$ and so r^* is not a best approximation to $f \in K \sim M$. This is a contradiction. ∎

<u>Theorem 1.23</u>: If $r^* \in M$ is a best starting approximation to $\Phi(f)$, then r^* is unique.

<u>Proof</u>: Suppose $r \in M$ is such that $\|\Phi(f) - \Phi(r)\| = \|\Phi(f) - \Phi(r^*)\|$. Then if $\sigma(x) = \operatorname{sgn}(f(x) - r^*(x))$ we have for $x \in X(r^*)$: (i) if $f(x) > r^*(x)$ then $r(x) \geq r^*(x)$ and (ii) if $f(x) < r^*(x)$ then $r(x) \leq r^*(x)$. Hence we conclude that $\sigma(x)(r(x) - r^*(x)) \geq 0$ for all $x \in X(r^*)$. Consequently $\sigma(x)\left(\frac{p(x)}{q(x)} - \frac{p^*(x)}{q^*(x)}\right) \geq 0$, so $\sigma(x)(p(x) - q(x)r^*(x)) \geq 0$ since $q(x) > 0$ on $[a,b]$. By Lemma 1.20 we have that $p-r^*q \in S(r^*)$. Therefore Lemma 1.22 gives $p(x) - r^*(x)q(x) \equiv 0$ and so $r \equiv r^*$. ∎

Finally we shall give a generalization of a theorem of de la Vallée Poussin, for the problem at hand, which is often useful in estimating the deviation of the best starting approximation to $\Phi(f)$.

Theorem 1.24: Let $f \in K \sim M$ and let $r^* = \frac{p^*}{q^*} \in M$ be the best starting approximation to $\Phi(f)$ from M. If $r \in M$, $\{x_i\}_{i=1}^{d+1}$ a set of $d+1$ consecutive points in $X \sim \bigcup_{i=1}^{p} \{y_i\}$, and $\{\lambda_i\}_{i=1}^{d+1}$ is a set of positive numbers such that

$$\text{(i)} \quad |w(x_i)(\Phi(f)(x_i) - \Phi(r)(x_i))| = \lambda_i,$$

$$\text{(ii)} \quad \operatorname{sgn}\{[f(x_i) - r(x_i)]\Pi(x_i)\} = (-1)^{i+1}\operatorname{sgn}\{[f(x_1) - r(x_1)]\Pi(x_1)\},$$

then $\min_i \lambda_i \leq \|\Phi(f) - \Phi(r^*)\|$.

Proof: If $r \equiv r^*$ the result is clear and so we suppose $r \not\equiv r^*$. Assume $\min_i \lambda_i > \|\Phi(f) - \Phi(r^*)\|$. Therefore $\lambda_i > \|\Phi(f) - \Phi(r^*)\|$, $i = 1,\dots,d+1$ (where $d = \max\{n + \partial q^*, m + \partial p^*\} + 1 - m^*$) and so $|w(x_i)(\Phi(f)(x_i) - \Phi(r)(x_i))| > |w(x_i)(\Phi(f)(x_i) - \Phi(r^*)(x_i))|$, $i = 1,2,\dots,d+1$. Now proceeding as in the proof that (iv) implies (i) in Theorem 1.21 we obtain the fact that $r \equiv r^*$ which is a contradiction. Hence $\min_i \lambda_i \leq \|\Phi(f) - \Phi(r^*)\|$. ∎

Section 3: Computation of an Optimal Starting Approximation

In this section we shall examine the problem of computation of a best starting approximation for an operator Φ. The characterization theorems we have developed, especially Theorem 1.9, are of particular importance in establishing computational procedures. From the classical theory an optimal starting approximation would be computed using a modified Remes algorithm which involves solving a nonlinear system of equations with Newton's method of higher order.

However, if the operator Φ satisfies certain conditions a best starting approximation can be more easily obtained. Moreover, the best starting approximation may be independent of the number of applications of Φ provided this iteration is well-defined. We now wish to consider sufficient conditions on Φ for which this behavior occurs. The following definitions and results are due to Meinardus and Taylor [11].

<u>Definition 1.25</u>: Let $\Phi: K \to C(X)$ be a continuous operator. We say that Φ possesses Property I at $f \in K$ provided for each $r \in K$ and $x,y \in X$, $\frac{r(x)}{f(x)} = \frac{r(y)}{f(y)}$ implies $\frac{\Phi(r)(x)}{f(x)} = \frac{\Phi(r)(y)}{f(y)}$, and $\frac{r(y)}{f(y)} < \frac{r(x)}{f(x)} \leq 1$ or $\frac{r(y)}{f(y)} > \frac{r(x)}{f(x)} \geq 1$ implies $\left|1 - \frac{\Phi(r)(x)}{f(x)}\right| < \left|1 - \frac{\Phi(r)(y)}{f(y)}\right|$.

<u>Definition 1.26</u>: Let $\Phi: K \to C(X)$ be a continuous operator. Φ is said to be one-sided at f provided either $\Phi(k) \geq \Phi(f)$ for all $k \in K$ or $\Phi(k) \leq \Phi(f)$ for all $k \in K$.

<u>Theorem 1.27</u>: Let K be a convex subset of $C(X)$ and $\Phi: K \to C(X)$ be pointwise strictly monotone, pointwise fixed and possess Property I at $f \in K$ where $\Phi(f) = f$ and $f > 0$. Norm $C(X)$ by $\|h\| = \|\frac{h}{f}\|_\infty$ $(h \in C(X))$. Let $r = \frac{p}{q} \in R_m^n[a,b]$ $(m,n,a,b$ fixed) be the best relative approximation to f from $R_m^n[a,b]$ with deviation λ; that is, $\|\frac{f-r}{f}\|_\infty = \inf\limits_{s \in R_m^n[a,b]} \|\frac{f-s}{f}\|_\infty = \lambda$. If $M = K \cap R_m^n[a,b]$ is nonempty and relatively open in $R_m^n[a,b]$ and $\delta r \in M$ for $\delta \in \left[\frac{1}{1+\lambda}, \frac{1}{1-\lambda}\right]$ then there exists $\delta_0 \in \left(\frac{1}{1+\lambda}, \frac{1}{1-\lambda}\right)$ for which $\delta_0 r$ is the best starting approximation for f (with respect to Φ).

Proof: First we observe that $\lambda < 1$ since arbitrarily small positive functions exist in $R_m^n[a,b]$. From the standard Chebyshev theory we know that there exists $N = 2 + \max\{n + \partial q, m + \partial p\}$ consecutive points, $x_1 < x_2 < \ldots < x_N$, in X such that

$$\text{(i)} \quad \left|\frac{f(x_i) - r(x_i)}{f(x_i)}\right| = \left\|\frac{f-r}{f}\right\|_\infty, \quad i = 1,\ldots,N,$$

$$\text{(ii)} \quad \operatorname{sgn}(f(x_i) - r(x_i)) = (-1)^{i+1}\operatorname{sgn}(f(x_1) - r(x_1)), \quad i = 1,\ldots,N.$$

Without loss of generality we may assume that $r(x_1) > f(x_1)$. It is easily seen, using (i) above, that $\frac{1}{1+\lambda} r(x_1) = f(x_1)$ and $\frac{1}{1-\lambda} r(x_2) = f(x_2)$ and in general $1 - \lambda \le \frac{r(x)}{f(x)} \le 1 + \lambda$. Hence, in view of the above results, $\frac{r(x_1)}{f(x_1)} \ge \frac{r(x)}{f(x)} \ge \frac{r(x_2)}{f(x_2)}$ and so $\frac{\gamma r(x_1)}{f(x_1)} \ge \frac{\gamma r(x)}{f(x)} \ge \frac{\gamma r(x_2)}{f(x_2)}$, where $\gamma \in \left[\frac{1}{1+\lambda}, \frac{1}{1-\lambda}\right]$, for all $x \in X$. By Property I,

$$\left|1 - \frac{\Phi(\gamma r)(x)}{f(x)}\right| \le \max\left(\left|1 - \frac{\Phi(\gamma r)(x_1)}{f(x_1)}\right|, \left|1 - \frac{\Phi(\gamma r)(x_2)}{f(x_2)}\right|\right).$$

Also, $\frac{f(x_{i+2}) - r(x_{i+2})}{f(x_{i+2})} = \frac{f(x_i) - r(x_i)}{f(x_i)}$, $i = 1,2,\ldots,N-2$, implies

$\frac{r(x_{i+2})}{f(x_{i+2})} = \frac{r(x_i)}{f(x_i)}$, $i = 1,\ldots,N-2$ (using Property (ii)). Thus, for $\gamma \in \left[\frac{1}{1+\lambda}, \frac{1}{1-\lambda}\right]$, $\left|1 - \frac{\Phi(\gamma r)(x_{i+2})}{f(x_{i+2})}\right| = \left|1 - \frac{\Phi(\gamma r)(x_i)}{f(x_i)}\right|$ since Φ possesses Property I at f. Using the pointwise strict monotonicity of Φ at f, we find that $\left|1 - \frac{\Phi(\gamma r)(x_1)}{f(x_1)}\right|$ decreases to 0 as γ decreases to $\frac{1}{1+\lambda}$ and $\left|1 - \frac{\Phi(\gamma r)(x_2)}{f(x_2)}\right|$ increases from 0 as γ

decreases to $\frac{1}{1+\lambda}$. Similarly, $\left|1 - \frac{\Phi(\gamma r)(x_1)}{f(x_1)}\right|$ increases from 0

γ increases to $\frac{1}{1-\lambda}$ and $\left|1 - \frac{\Phi(\gamma r)(x_2)}{f(x_2)}\right|$ decreases to 0 as γ

increases to $\frac{1}{1-\lambda}$. Since both $\left|1 - \frac{\Phi(\gamma r)(x_1)}{f(x_1)}\right|$ and $\left|1 - \frac{\Phi(\gamma r)(x_2)}{f(x_2)}\right|$

are continuous functions of γ, there exists $\gamma_0 \in \left(\frac{1}{1+\lambda}, \frac{1}{1-\lambda}\right)$ such

that $\left|1 - \frac{\Phi(\gamma_0 r)(x_1)}{f(x_1)}\right| = \left|1 - \frac{\Phi(\gamma_0 r)(x_2)}{f(x_2)}\right|$. Hence $\left|\frac{\Phi(f)(x_i) - \Phi(\gamma_0 r)(x_i)}{f(x_i)}\right|$

$= \left|1 - \frac{\Phi(\gamma_0 r)(x_i)}{f(x_i)}\right| = \max_{x \in X} \left|\frac{\Phi(f)(x) - \Phi(\gamma_0 r)(x)}{f(x)}\right| = \left\|\frac{\Phi(f) - \Phi(\gamma_0 r)}{f}\right\|_\infty$

$= \|\Phi(f) - \Phi(\gamma_0 r)\|$. Finally, it is clear from the choice of γ that $\operatorname{sgn}(f(x_i) - \gamma r(x_i)) = \operatorname{sgn}(f(x_i) - r(x_i))$. Thus, by Theorem 1.9, $\gamma_0 r$ is the best starting approximation for f. ∎

Definition 1.25 and Theorem 1.27 of Meinardus and Taylor have rather natural analogs in the setting of uniform approximation ($w(x) \equiv 1$).

<u>Definition 1.28</u>: The operator Φ is said to possess Property J at $f \in K$ if for each $r \in K$ and $x,y \in X$, $f(x) - r(x) = f(y) - r(y)$ implies $\Phi(f)(x) - \Phi(r)(x) = \Phi(f)(y) - \Phi(r)(y)$, and $0 \le f(x) - r(x) < f(y) - r(y)$ or $0 \ge f(x) - r(x) > f(y) - r(y)$ implies $|\Phi(f)(y) - \Phi(r)(y)| > |\Phi(f)(x) - \Phi(r)(x)|$.

<u>Theorem 1.29</u>: Let $\Phi\colon K \to C(X)$, K be a convex subset of $C(X)$, and $M = K \cap R_m^n[a,b]$ be a relatively open subset of $R_m^n[a,b]$ with $m \le n$. Assume Φ is pointwise strictly monotone, pointwise fixed and possesses Property J at $f \in K \sim M$. If $r \in R_m^n[a,b]$ is the best uniform approximation to f from $R_m^n[a,b]$ with deviation λ

and if $r + c \in M$ for $c \in [-\lambda,\lambda]$ then there exists $c_0 \in (-\lambda,\lambda)$ for which $r + c_0$ is the best starting approximation to $\Phi(f)$ from M.

Proof: Since r is the best uniform approximation to f from $R_m^n[a,b]$, there exist N points, $x_1 <\ldots< x_N$, in X, $N = 2 + \max\{m + \partial p, n + \partial q\}$, such that

(i) $|f(x_i) - r(x_i)| = \|f - r\|$, $i = 1,\ldots,N$

(ii) $\operatorname{sgn}(f(x_i) - r(x_i)) = (-1)^{i+1}\operatorname{sgn}(f(x_1 - r(x_1))$, $i = 1,\ldots,N$.

Now $f(x_i) - r(x_i) = f(x_{i+2}) - r(x_{i+2})$, $i = 1,2,\ldots,N-2$ and so by Property J $|\Phi(f)(x_i) - \Phi(r)(x_i)| = |\Phi(f)(x_{i+2}) - \Phi(r)(x_{i+2})|$, $i = 1,\ldots,N-2$. Also by Property J we see that for any $c \in [-\lambda,\lambda]$,

$$|\Phi(f)(x) - \Phi(r+c)(x)| \leq \max(|\Phi(f)(x_1) - \Phi(r+c)(x_1)|, |\Phi(f)(x_2)-\Phi(r+c)(x_2)|).$$

Note that for each r and x, $\Phi(r+c)(x)$ is a continuous function of c. Without loss of generality assume $f(x_1) > r(x_1)$. Using the fact that Φ is pointwise strictly monotone and pointwise fixed at $f \in K$ we have that $|\Phi(f)(x_1) - \Phi(r+c)(x_1)|$ is a strictly decreasing function of c which tends to 0 as c tends to λ and $|\Phi(f)(x_2) - \Phi(r+c)(x_2)|$ is a strictly increasing function of c which tends to zero as c tends to $-\lambda$. Since both $|\Phi(f)(x_1) - \Phi(r+c)(x_1)|$ and $|\Phi(f)(x_2) - \Phi(r+c)(x_2)|$ are continuous functions of c there exists $c_0 \in (-\lambda,\lambda)$ for which $|\Phi(f)(x_1) - \Phi(r+c_0)(x_1)| = |\Phi(f)(x_2) - \Phi(r+c_0)(x_2)|$. Thus $|\Phi(f)(x_i) - \Phi(r+c_0)(x_i)| = \|\Phi(f) - \Phi(r+c_0)\|$, $i = 1,2,\ldots,N$. Furthermore, since $c_0 \in (-\lambda,\lambda)$, $\operatorname{sgn}(f(x_i) - (r+c_0)(x_i)) = (-1)^{i+1}\operatorname{sgn}(f(x_{i+1}) - (r+c_0)(x_{i+1}))$ and so appealing to Theorem 1.9

we have the desired result. ▮

<u>Theorem 1.30</u>: Let $\Phi: K \to K$ satisfy the following properties:

(i) Φ is continuous

(ii) $\Phi(f) = f$

(iii) Φ is pointwise strictly monotone at $f \in K \sim M$

(iv) Φ is a one-sided operator at f

(v) Φ possesses Property I (Property J) at f.

Then $\Phi^m = \Phi(\Phi^{m-1})$, $m = 2,3,\ldots$ has all the same properties as Φ and, moreover, the best starting approximation for $\Phi(f)$ is also the best starting approximation for $\Phi^m(f)$, $m = 2,3,\ldots$.

<u>Proof</u>: The first part of the conclusion follows from Corollary 1.6 and a simple inductive application of Property I (Property J). For the second part let us suppose, without loss of generality, that $\Phi(k) \geq \Phi(f) = f$ for all $k \in K$ and that we are dealing with relative approximation. Now if $r = \frac{p}{q}$ is a best starting approximation for $\Phi(f)$ then there exist N points $\{x_i\}_{i=1}^N \subseteq X$, $N = 2 + \max\{n + \partial q, m + \partial p\}$, such that $x_1 < x_2 < \ldots < x_N$ and

$$\text{(i)}\quad \left|\frac{\Phi(r)(x_i) - f(x_i)}{f(x_i)}\right| = \left\|\frac{\Phi(r) - f}{f}\right\|_\infty, \quad i = 1,\ldots,N$$

$$\text{(ii)}\quad \operatorname{sgn}(f(x_i) - r(x_i)) = (-1)^{i+1}\operatorname{sgn}(f(x_1) - r(x_1)), \quad i = 1,2,\ldots,N.$$

Now by applying Property I to these points m times it is clear that r is also the best starting approximation for $\Phi^m(f) = f$. ▮

In attempting to construct a theorem analogous to Theorem 1.27 for the case of restricted range approximation certain difficulties arise. The multiplicative shift of the best relative approximation to f from $R_m^n[a,b]$ may well remove us from the approximating class.

However in one particular setting a result can be obtained. If $\ell(x), u(x) \in C(X)$ denote the upper and lower constraining curves for the restricted range problem, take $\ell(x) = f(x)$ or $u(x) = f(x)$ where $f(x)$ is the function we desire to approximate. Without loss of generality take $\ell(x) = f(x)$ and set $K = \{h \in C(X): f(x) \leq h(x) \leq u(x)\}$. This situation will be described as modified one-sided approximation from above since the usual one-sided approximation would have $u(x) \equiv \infty$. In this setting we are able to state the following analog of Theorem 1.27.

Theorem 1.31: Let $\Phi: K \to C(X)$, $K = \{h \in C(X): f(x) \leq h(x) \leq u(x)\}$ where $u(x) \in C(X)$, $f > 0$ on X, and $C(X)$ is normed by $\|h\| = \|\frac{h}{f}\|_\infty$. Further suppose that $\Phi(f) = f$, Φ is pointwise strictly monotone, pointwise fixed, and possesses Property I at f. Let $r \in M$ be the best relative approximation to f from M where $M = K \cap R^n_m[a,b]$. Then r is the optimal starting approximation for f from M (with respect to Φ).

Proof: The proof follows directly from Theorem 1.14 and the fact that Φ possesses Property I at f. ∎

Remark 1: In applying the above theory one may find it more efficient (in terms of the computer) to compose different operators (each of which defines an iteration procedure) rather than iterating with a given operator. In view of this we state the following: suppose $\Phi: K \to C(X)$ possesses Property I at $f \in K$, Φ is one-sided at f and $\psi: L \to C(X)$ possesses Property I at $\Phi(f)$ where $\Phi(K) \subseteq L$. If $\psi(f) = \Phi(f) = f$ then $\Psi\Phi$ possesses Property I at $f \in K$. With this observation and the results in Section 1 all of the theorems of this section may be restated for the composition operator $\Psi\Phi$ or more generally for the composition of operators $\Psi_1,\ldots,\Psi_k$, provided the domains and ranges mesh correctly and each of the operators Ψ_i possesses the appropriate properties.

CHAPTER II

APPLICATION TO THE NEWTON OPERATOR AND ASSOCIATED RESULTS

Section 1: Introduction and General Theory

In this chapter we shall apply the theory of Chapter I for a special choice of Φ, namely the operator associated with the well-known Newton iteration scheme. First set

$$S = \{f \in C^2(0,\infty): f', f'' \neq 0 \text{ for all } x > 0 \text{ and image } f = (0,\infty)\}.$$

The choice of $(0,\infty)$ is quite arbitrary; actually, one could select $(-\infty,\infty)$ or some other region on which to work. Note that domain $f^{-1} = (0,\infty)$ since image $f = (0,\infty)$. For a fixed $x \in [a,b] \subseteq (0,\infty)$ we can solve $f^{-1}(y) - x = 0$ by Newton's method to obtain the value of f at x where $f \in S$. In particular, if $y_0(x) = y(x)$ is the initial guess to $f(x)$ then inductively we define $y_n(x)$ by

$$y_n(x) = y_{n-1}(x) - \{f^{-1}(y_{n-1}(x)) - x\}\{f'[f^{-1}(y_{n-1}(x))]\}, \quad n = 1,2,\ldots \tag{2.1}$$

Of course we must insure that $y_{n-1}(x) > 0$, $n = 1,2,\ldots$ so that f^{-1} is defined. The sequence (2.1), which represents the Newton iteration scheme for determining the unique zero of the equation $f^{-1}(y) - x = 0$, converges quadratically to $f(x)$ provided the initial guess $y_0(x)$ is sufficiently good. As mentioned, the above consideration is for a fixed value of x. Our goal is to be able to calculate $f(x)$ for all $x \in [a,b]$ on a computer. To accomplish this we shall select a class of functions M (specifically $M \subseteq R_m^n[a,b]$ for our purposes) which are easily programmed and then select a member of M as the initial guess. Precisely, we wish to find an element r of M for which

$$\left\|\frac{y_{n,r}(x) - f(x)}{f(x)}\right\|_\infty \leq \left\|\frac{y_{n,s}(x) - f(x)}{f(x)}\right\|_\infty$$

for all $s \in M$ where $y_{n,s}(x)$ denotes the n-th Newton iterate at x with initial guess $s(x)$. Numerically the above method would be applied to functions f for which f^{-1} is easily evaluated such as $f(x) = x^{\frac{1}{n}}$, $n = 2,3,\ldots$. This particular problem was examined by D.G. Moursund and G.D. Taylor [16] and is a generalization of the subroutine used to calculate $\sqrt{x}$ on a computer such as the CDC-3600. We shall now show that the theory we have developed in Chapter I is also applicable to this problem as well as others.

The first problem confronting us is that of determining a suitable convex subset K of $C[a,b]$ such that the operator N_f defined by

$$N_f(h)(x) = h(x) - \{f^{-1}(h(x)) - x\}\{f'[f^{-1}(h(x))]\}$$

maps K to K (into or onto). It is imperative that we examine two cases:

Case 1: Fix $f \in S$ and assume that either $f' > 0$ and $f'' < 0$ on $(0,\infty)$ or $f' < 0$ and $f'' < 0$ on $(0,\infty)$. Let

$$K = \{h \in C[a,b]: h(x) > 0 \text{ for all } x \in [a,b]\}.$$

Setting $N(x,y) = y - \{f^{-1}(y) - x\}\{f'[f^{-1}(y)]\}$ and calculating $\frac{\partial N(x,y)}{\partial y}$, it is easily seen that N_f is pointwise strictly monotone at f and one-sided from above at f. It is also clear that $N_f: K \to K$ is continuous and $N_f(f) = f$. Observe that the class of functions $f(x) = x^{\frac{1}{n}}$, $n = 2,3,\ldots$ are suitable choices for f.

Case 2: Fix $f \in S$ where either $f' > 0$ and $f'' > 0$ or $f' < 0$ and $f'' > 0$ holds on $(0,\infty)$. The choice for K in this setting is somewhat more difficult since N_f is a one-sided operator from below and consequently large values of $h(x_0)$, for fixed $x_0 \in [a,b]$, may yield negative values for $N_f(h)(x_0)$. Since we wish to iterate with N_f and we have assumed the image of f is $(0,\infty)$, we must bound the members of K from above. Specifically, for fixed $f \in S$ as above, we wish to find, if possible, a function $\phi(x) \in C[a,b]$ such that if

$$K = \{h \in C[a,b]: 0 < h(x) < \phi(x) \text{ for all } x \in [a,b]\}$$

then $N_f: K \to K$. If we set $N(x,y) = y - \{f^{-1}(y) - x\}\{f'[f^{-1}(y)]\}$ (as above) and calculate $\frac{\partial N(x,y)}{\partial y}$, it is clear that for each fixed x $N(x,y)$ is a strictly increasing function of y for $0 < y < f(x)$ and a strictly decreasing function of y for $y > f(x)$. Moreover, $N(x,y) > 0$ for $0 < y < f(x)$. Let us now assume that for each fixed $x \in [a,b]$, $\lim_{y \to +\infty} N(x,y) = -\infty$. This guarantees us that for each $x \in [a,b]$ there exists $y_x > 0$ such that $N(x,y_x) = 0$ (clearly $y_x > f(x)$ for each $x \in [a,b]$). To construct the function $\phi(x)$ mentioned above we shall make use of the Implicit Function Theorem. First we have that $N(x,y)$ is defined on

$[a,b] \times E^1_+ \subseteq E^2$ where $E^1_+ = \{s \in E^1 : s > 0\}$. Let $x_0 \in [a,b]$ be arbitrary but fixed and let $y_0 \in E^1_+$ be such that $N(x_0,y_0) = 0$. Then the following holds:

(i) $[a,b] \times [y: y > f(x_0)\} = \eta_{(x_0,y_0)}$ is a neighborhood of (x_0,y_0) in $[a,b] \times E^1_+$, $N(x,y)$ is continuous on $\eta_{(x_0,y_0)}$ and $\frac{\partial N(x,y)}{\partial y} = -\frac{\{f^{-1}(y) - x\}f''[f^{-1}(y)]}{f'[f^{-1}(y)]}$ is continuous on $\eta_{(x_0,y_0)}$ by the assumptions on f.

(ii) $N(x_0,y_0) = 0$

(iii) $\left.\frac{\partial N(x,y)}{\partial y}\right|_{(x_0,y_0)} \neq 0$.

Hence, by the Implicit Function Theorem, there exist neighborhoods η_{x_0} of x_0, η_{y_0} of y_0 and a function $\phi_0(x)$ defined on η_{x_0} such that

(iv) for all $x \in \eta_{x_0}$, $\phi_0(x) \in \eta_{y_0}$, $(x,\phi_0(x)) \in \eta_{(x_0,y_0)}$ and $N(x,\phi_0(x)) = 0$.

(v) $\phi_0(x)$ is uniquely determined by (iv)

(vi) ϕ_0 is continuous on η_{x_0}

(vii) $\phi_0(x_0) = y_0$.

Now consider $\{\eta_x : x \in [a,b]\}$. This is an open cover of $[a,b]$, which is compact, and so can be reduced to a finite subcover $\eta_{x_1},\ldots,\eta_{x_n}$. For each $i = 1,2,\ldots,n$ we have $\phi_i(x)$ defined on η_{x_i} and continuous there. Suppose $\eta_{x_i} \cap \eta_{x_j} \neq \phi$ for $i \neq j$. We wish to show that $\phi_i(x) = \phi_j(x)$ for $x \in \eta_{x_i} \cap \eta_{x_j}$. Let $z \in \eta_{x_i} \cap \eta_{x_j}$. Then $N(z,\phi_i(z)) = N(z,\phi_j(z)) = 0$ and $\phi_i(x_i) = y_i > f(x_i)$.

Claim: $\phi_i(z) > f(z)$.

Assume to the contrary. Then $\phi_i(z) < f(z)$ since $\phi_i(z) = f(z)$ implies $0 = N(z,\phi_i(z)) = N(z,f(z)) = f(z) > 0$. Define $g_i(x) = \phi_i(x) - f(x)$. Then g_i is continuous on η_{x_i} with $g_i(x_i) > 0$, $g_i(z) < 0$. Hence there exists $\xi \in \eta_{x_i}$ such that $g_i(\xi) = 0$. Therefore $\phi_i(\xi) = f(\xi)$ and so $0 = N(\xi,\phi_i(\xi)) = N(\xi,f(\xi)) > 0$. This is a contradiction and thus the claim. Similarly $\phi_j(z) > f(z)$. Since N is decreasing for $y > f(z)$ and $N(z,\phi_i(z)) = N(z,\phi_j(z)) = 0$ we have that $\phi_i(z) = \phi_j(z)$. Consequently we are able to define a function ϕ on $[a,b]$ via $\phi(x) = \phi_j(x)$ if $x \in \eta_{x_j}$, $j = 1,\dots,n$. The function ϕ is continuous on $[a,b]$ and $\phi(x) > f(x) > 0$ for all $x \in [a,b]$. Set

$$K = \{h \in C[a,b]: 0 < h(x) < \phi(x) \text{ for all } x \in [a,b]\}.$$

Then if $h \in K$, $N_f(h)(x) \le f(x) < \phi(x)$ for all $x \in [a,b]$ and from the previous discussion $N_f(h)(x) > 0$ for all $x \in [a,b]$. Hence, for the above choice of K, $N_f: K \to K$ and so iterating with N_f via $N_f^m = N_f(N_f^{m-1})$, $m = 2,3,\dots$ is well-defined. As in Case 1 we know that N_f is continuous, $N_f(f) = f$ and N_f is pointwise strictly monotone at f. Moreover, N_f is one-sided from below at f.

Remark 1: For the function $f(x) = e^x$ (which falls in Case 2), the function $\phi(x) = e^{1+x}$ satisfies $N(x,e^{1+x}) = 0$ for all $x \in [a,b]$ where $N(x,y)$ is defined by $N(x,y) = y(1 + x - \ln y)$. Hence we set

$$K = \{h \in C[a,b]: 0 < h(x) < e^{1+x} \text{ for all } x \in [a,b]\}.$$

With the preceding discussion in mind we can state the following analog of Theorem 1.9 for both of these cases (with the respective K's defined above).

<u>Theorem 2.1</u>: Let $f \in S$, K be as above and $M = K \cap R^n_m[a,b]$ (fixed m,n). Then $r_k = \frac{p_k}{q_k} \in M$ is the unique best starting approximation for f with respect to the k-th Newton iteration if and only if there exists $\{x_i\}_{i=1}^N \subseteq [a,b]$, $N = 2 + \max\{n + \partial q_k, m + \partial p_k\}$, with $x_1 < x_2 < \ldots < x_N$ for which

(i) $$\left|\frac{f(x_i) - N_f^k(r_k)(x_i)}{f(x_i)}\right| = \left\|\frac{f - N_f^k(r_k)}{f}\right\|_\infty, \quad i = 1,2,\ldots,N$$

(ii) $$\operatorname{sgn}(f(x_i) - r_k(x_i)) = (-1)^{i+1}\operatorname{sgn}(f(x_1) - r_k(x_1)),$$

$$i = 1,2,\ldots,N.$$

<u>Section 2</u>: Optimal Starting Approximations for Roots, Exp(x) and Ln(x)

In general the existence of a best starting approximation is a rather difficult problem which depends upon the particular function f and the interval $[a,b]$. For the special case that $f(x) = x^\alpha$, $\alpha \in (0,1)$ or $f(x) = e^x$ existence has been established and, moreover, the best starting approximation is independent of the number of iterations. For $f(x) = x^\alpha$, $\alpha \in (0,1)$, set $K = \{h \in C[a,b]: h(x) > 0$ for all $x \in [a,b]\}$ where $0 < a < b$. The operator N_f is defined by

$$N_f(h)(x) = \alpha\left[\left(\frac{1}{\alpha} - 1\right)h(x) + \frac{x}{h^{\frac{1}{\alpha}-1}(x)}\right], \text{ for } h \in K.$$

In the case of $f(x) = e^x$, set $K = \{h \in C[a,b]: 0 < h(x) < e^{1+x}$ for all $x \in [a,b]\}$ and define $N_f(h)(x) = h(x)(1 + x - \ln h(x))$ for $h \in K$. If $r \in M$ denotes the best relative approximation to

e^x from $R_m^n[a,b]$ with error λ, then we must apparently require that $\lambda < \frac{e-1}{e+1}$ in order to apply Theorem 1.27. However this requirement will be shown to be unnecessary. Before stating the next theorem due to Meinardus and Taylor [11], we introduce the following notation: for each $\alpha \in (0,1)$, let r_α denote the best relative approximation to x^α from $R_m^n[a,b]$ with deviation λ_α; that is,

$$\left\| \frac{x^\alpha - r_\alpha}{x^\alpha} \right\|_\infty = \inf_{s \in R_m^n[a,b]} \left\| \frac{x^\alpha - s}{x^\alpha} \right\|_\infty = \lambda_\alpha .$$

Let r denote the best relative approximation to $f(x) = e^x$ from $R_m^n[a,b]$ with deviation λ; that is, $\left\| \frac{e^x - r}{e^x} \right\|_\infty = \inf_{s \in R_m^n[a,b]} \left\| \frac{e^x - s}{e^x} \right\|_\infty = \lambda.$

Theorem 2.2: For $m = 1,2,\ldots$ the following is true:

(a) The best starting approximation for m Newton iterations for the calculation of x^α is $\gamma_\alpha r_\alpha$ where

$$\gamma_\alpha = \left[\frac{(1+\lambda_\alpha)^{\beta-1} - (1-\lambda_\alpha)^{\beta-1}}{2(\beta-1)\lambda_\alpha(1-\lambda_\alpha^2)^{\beta-1}} \right]^\alpha, \quad \beta = \frac{1}{\alpha} .$$

(b) The best starting approximation for m Newton iterations for the calculation of e^x is γr where

$$\gamma = \exp\left[\frac{1}{2\lambda}\left(2\lambda + (1-\lambda)\ln(1-\lambda) - (1+\lambda)\ln(1+\lambda)\right)\right] = e\left(\frac{1-\lambda}{1+\lambda}\right)^{\frac{1}{2\lambda}} \frac{1}{\sqrt{1-\lambda^2}} .$$

Proof: The proof of (a) appears in [28] and will be omitted. For (b) we shall apply Theorems 1.9 and 1.30. First we shall show that $N = N_{e^x}$ possesses Property I at $f(x) = e^x$ where $N(r)(x) = r(x)\left(1 - \ln \frac{r(x)}{e^x}\right)$. If $r \in K$ (as defined earlier), $y,z \in [a,b]$

and $\frac{r(y)}{e^y} = \frac{r(z)}{e^z}$ then we clearly have $\frac{N(r)(y)}{e^y} = \frac{N(r)(z)}{e^z}$. To obtain the second condition necessary for Property I we set $\varphi(t) = 1 - t + t\,\ell nt$ (for $t > 0$) and examine $\varphi'(t)$. Since $\varphi'(t) < 0$ for $0 < t < 1$ and $\varphi'(t) > 0$ for $t > 1$, $\varphi(1)$ is the only local minimum for $\varphi(t)$ when $t > 0$ and hence the second condition for Property I is also satisfied. Next we shall show that the best relative approximation, r, to e^x belongs to $K = \{h \in C[a,b]: 0 < h(x) < e^{1+x} \text{ for all } x \in [a,b]\}$. This follows from the fact that $(1-\lambda)e^x \le r(x) \le (1+\lambda)e^x$ and $0 < \lambda < 1$. Proceeding as in the proof of Theorem 1.27 we find that γr has an error curve of the type described in Theorem 1.9 where $\gamma = e\left(\frac{1-\lambda}{1+\lambda}\right)^{\frac{1}{2\lambda}} \frac{1}{\sqrt{1-\lambda^2}}$. The value of γ is found by solving the equation

$$\left| 1 - \frac{N(\gamma r)(x_1)}{e^{x_1}} \right| = \left| 1 - \frac{N(\gamma r)(x_2)}{e^{x_2}} \right|$$

where $r(x_1) = (1+\lambda)e^{x_1}$ and $r(x_2) = (1-\lambda)e^{x_2}$. Since $0 < \gamma < \frac{e}{2}$ we have that $\gamma r \in K$ and so by Theorem 1.9 γr is the best starting approximation for e^x with respect to N. Furthermore, by Theorem 1.30, γr is the best starting approximation for e^x with respect to N^m $(m = 1,2,\ldots)$. ▮

The aforementioned Property J is a rather natural analog of Property I for the case of uniform approximation. The Newton operators used for approximating the functions $f(x) = e^x$ and $f(x) = x^\alpha$, $\alpha \in (0,1)$ fail to possess Property J and hence Theorem 1.29 does not apply. However, the Newton operator for approximating

$f(x) = \ell n\, x$ does possess Property J as will be shown. Incidentally, the Newton operator for approximating $\ell n\, x$ fails to possess Property I. We now turn our attention to the operator $N \equiv N_{\ell n x}$. For $0 < a < b$, set $K = \{h \in C[a,b]: h(x) > \ell n\, x - 1$ for all $x \in [a,b]\}$ and define $N_{\ell n\, x}(h)(x) = h(x) - 1 + \frac{x}{e^{h(x)}} = h(x) - 1 + e^{[\ell n\, x - h(x)]}$ for $h \in K$. It is straightforward to show that $N \equiv N_{\ell n x} : K \to K$, is pointwise strictly monotone at $\ell n\, x$ and is one-sided from above. To prove that N possesses Property J at $\ell n\, x$ we define $\varphi(t) = (t-1) + e^{-t}$ and examine $\varphi'(t)$. Since

$$\varphi'(t) = 1 - e^{-t} \begin{cases} > 0 & \text{for } t > 0 \\ = 0 & \text{for } t = 0 \\ < 0 & \text{for } t < 0 \end{cases}, \text{ the desired result is obvious.}$$

With the above we can prove the following result:

Theorem 2.3: Let $N \equiv N_{\ell n\, x}$ denote the Newton operator for approximating $f(x) = \ell n\, x$, $K = \{h \in C[a,b]: h(x) > \ell n\, x - 1$ for all $x \in [a,b]\}$ where $0 < a < b$ and $M = K \cap R^n_m[a,b]$ where $m \le n$. The best starting approximation (from M) for k Newton iterations for the calculation of $\ell n\, x$ (in the uniform norm) is $r^* + c_o$ where r^* is the best uniform approximation to $\ell n\, x$ from $R^n_m[a,b]$ with deviation λ (we assume $0 < \lambda < 1$) and $c_0 = \ell n \frac{\sinh \lambda}{\lambda}$.

Proof: We have already noted that N is pointwise strictly monotone at $\ell n\, x$, pointwise fixed at $\ell n\, x$, possesses Property J at $\ell n\, x$, maps K to K, and is one-sided from above. Since $\lambda < 1$ we also have that $r^* \in M$. Proceeding as in Theorem 1.29 we find that $r^* + c_0$ has an error curve of the type described in Theorem 1.9 where $c_0 = \ell n \frac{\sinh \lambda}{\lambda}$. The value of c_0 is found by solving the equation

$$|N(\ell n\ x)(x_1) - N(r^* + c)(x_1)| = |N(\ell n\ x)(x_2) - N(r^* + c)(x_2)|$$

for c, where $f(x_1) - r^*(x_1) = \lambda$ and $f(x_2) - r^*(x_2) = -\lambda$. Since $\frac{\sinh \lambda}{\lambda} > 1$ we have that $c_0 > 0$ and so $r^* + c_0 \in M$. Thus, by Theorem 1.9, $r^* + c_0$ is the best starting approximation for $\ell n\ x$ with respect to N. Furthermore, by Theorem 1.30, $r^* + c_0$ is the best starting approximation for $\ell n\ x$ with respect to N^k, $k = 1,2,\ldots$. Hence we have the desired result.∎

<u>Section 3</u>: An Associated Operator

Let us once again turn our attention to the Newton operator $N \equiv N_{e^x}$ defined by $N(h)(x) = h(x)(1 + x - \ell n\ h(x))$ ($h \in K$ as previously defined) used for approximating $f(x) = e^x$. Iterating with this operator is computationally impractical since the time required for the evaluation of the function $\ell n\ x$ is approximately that for the evaluation of e^x (say on the CDC-3600). To improve the situation we would like to replace the logarithm function in the above formula by a rational function $r \in R_m^n$ with m,n sufficiently small as to make the rational function more computationally desirable. At present the value of this scheme is somewhat questionable. However, if we define

$$N_r(h)(x) = h(x)[1 + x - r(h(x))] ,$$

for h an element of some convex set K' with $N_r: K' \to K'$ and r to be determined then we can show that an optimal starting approximation $r_\nu(x)$ to e^x with respect to N_r^ν $(\nu = 1,2,\ldots)$ exists.

Let $x \in [a,b]$, $0 < a < b$. Observe that for $h(x) \le e^x$ $(h \in K' \subseteq K)$ we have $N(h)(x) \ge h(x)(1 + x - \ln e^x) = h(x)$. Since $e^{x+1} - e^x$ is a continuous function on the compact set $[a,b]$, there exists $\delta_0 > 0$ such that $e^{x+1} - e^x \ge \delta_0 > \frac{\delta_0}{2}$ for all $x \in [a,b]$. Furthermore, for fixed x, $N(x,y) = N(y)(x)$ is a continuous function of $y(x)$, is monotone decreasing for $y \ge e^x$ and $N(x,e^{x+1}) = 0$. Hence $N(x,e^{x+1} - \frac{\delta_0}{2}) = \epsilon_0(x) > 0$. Set $K' = \{h \in C[a,b]: \epsilon_0(x) \le h(x) \le e^{x+1} - \frac{\delta_0}{2}\}$. By compactness and continuity considerations, $\epsilon_0(x) > 0$ for all $x \in [a,b]$ and so there exists $\alpha > 0$ such that $\epsilon_0(x) \ge \alpha$. Similarly, there exists $\beta > 0$ for which $e^{x+1} - \frac{\delta_0}{2} \le \beta$. Thus we have $0 < \alpha < \beta$ and $\alpha \le \epsilon_0(x) \le h(x) \le e^{x+1} - \frac{\delta_0}{2} \le \beta$ for all $h \in K'$. Let $r(y)$ be the best one-sided uniform approximation to $\ln y$ from below from $R_m^n[\alpha,\beta]$ and suppose $\lambda = \|\ln y - r(y)\|_{L^\infty[\alpha,\beta]}$. Define the operator N_r by $N_r(h)(x) = h(x)(1 + x - r(h(x)))$ where $h \in K'$. It is clear that $N_r(h)(x) \ge N(h)(x)$ for all $h \in K'$ and $x \in [a,b]$ since $\alpha \le h(x) \le \beta$ for $x \in [a,b]$ and $r(y) \le \ln y$ for all $y \in [a,b]$.

Claim: If r defined above belongs to a sufficiently large class of rational functions (that is, $m + n$ is sufficiently large) then $N_r: K' \to K'$.

Proof: For $\epsilon_0(x) \le h(x) \le e^x$, $N_r(h)(x) \ge N(h)(x) \ge h(x) \ge \epsilon_0(x)$ and so $N_r(h)(x) \ge \epsilon_0(x)$ for $h \in K'$ (regardless of the class r belongs to). Now choose m,n sufficiently large, say $m = m_0$, $n = n_0$, such that $m_0 + n_0$ is minimal and $e^{x+1} - e^x \ge (1-\lambda)\frac{\delta_0}{2} + \lambda e^{b+1}$. This we can do since $e^{x+1} - e^x > \frac{\delta_0}{2}$ for all $x \in [a,b]$ and $\lambda \to 0$ as $m,n \to \infty$ (in fact as $n \to \infty$). We shall now consider N_r only where r is the best one-sided uniform approximation to $\ln y$ from below from $R_m^n[\alpha,\beta] \supseteq R_{m_0}^{n_0}[\alpha,\beta]$. For such r,

$$N_r(h)(x) - N(h)(x) = h(x)(\ln h(x) - r(h(x))) \le \lambda h(x) \le \lambda(e^{x+1} - \frac{\delta_0}{2}) \le$$

$$\lambda\, e^{b+1} - \lambda \frac{\delta_0}{2} .$$

Therefore $N_r(h)(x) \le N(h)(x) + \lambda\, e^{b+1} - \frac{\lambda\delta_0}{2} \le e^x + \lambda\, e^{b+1} - \frac{\lambda\delta_0}{2}$

$$\le e^{x+1} - (1-\lambda)\frac{\delta_0}{2} - \lambda\, e^{b+1} + \lambda\, e^{b+1} - \frac{\lambda\delta_0}{2} = e^{x+1} - \frac{\delta_0}{2} .$$

Hence $N_r: K' \to K'$ for r described above and so we are able to iterate with the operator N_r.

With the above definitions of δ_0, r, λ and K' we are able to state the following existence theorem for the operator N_r.

Theorem 2.4: Let K' and r be as above for the fixed interval $[a,b]$ and let $M = K' \cap R_m^n[a,b]$. (Here r is fixed and subject to the above restrictions and m,n are arbitrary but fixed.) Then for any arbitrary but fixed natural number k there exists $s^* \in M$ such that $\|w(e^x - N_r^k(s^*))\|_{L^\infty[a,b]} \le \|w(e^x - N_r^k(s))\|_{L^\infty[a,b]}$ for all $s \in M$ where w is any continuous positive weight function on $[a,b]$.

Proof: Select a sequence $\{s_n\}_{n=1}^\infty$, $s_n \in M$, such that

$$\lim_{n\to\infty} \|w(e^x - N_r^k(s_n))\|_\infty = \inf_{s\in M} \|w(e^x - N_r^k(s))\|_\infty \quad (\|\cdot\|_\infty = \|\cdot\|_{L^\infty[a,b]})$$

which exists by the definition of infimum. Note that $s_n \in M$ for all n implies $\alpha \le \epsilon_0(x) \le s_n(x) \le e^{x+1} - \frac{\delta_0}{2} \le \beta$ for all $x \in [a,b]$, $n = 1,2,\ldots$. Thus $\{s_n(x)\}_{n=1}^\infty$ is a uniformly bounded sequence of rational functions and hence, by Lemma 1.12, there exists a uniform bound on the coefficients of all the $s_n(x)$. Therefore there exists a subsequence of $\{s_n\}_{n=1}^{\infty}$ (which we shall also denote by $\{s_n\}_{n=1}^{\infty}$) such that $\lim_{n\to\infty} s_n(x) = \bar{s}(x)$ (a rational function) for all $x \in [a,b]$ except for points which are the zeros of the denominator of $\bar{s}(x)$. If $\{x_i: i = 1,\ldots,\ell\}$ denotes the zeros of the denominator of $\bar{s}$ then it is clear that $\ell \le m$. Using the uniform boundedness of the sequence $\{s_i\}$, it is straightforward to show that each x_i is a removable singularity of $\bar{s}$. Upon cancellation, we obtain a rational function $s^* \in R_m^n[a,b]$ which we claim is the solution we seek. Observe that $\epsilon_0(x) \le s^*(x) \le e^{x+1} - \frac{\delta_0}{2}$ and so $\epsilon_0(x) \le N_r^k(s^*)(x) \le e^{x+1} - \frac{\delta_0}{2}$. We also have that if $x_0 \in [a,b]$ and $\lim_{n\to\infty} s_n(x_0) = s^*(x_0)$ then $\lim_{n\to\infty} N_r(s_n)(x_0) = N_r(s^*)(x_0)$. Inductively we may conclude that if $x_0 \in [a,b]$ and $\lim_{n\to\infty} s_n(x_0) = s^*(x_0)$ then $\lim_{n\to\infty} N_r^k(s_n)(x_0) = N_r^k(s^*)(x_0)$. Let $\epsilon > 0$ be given. Then there exists $N = N(\epsilon)$ such that $\|w(e^x - N_r^k(s_n))\|_\infty \le I + \epsilon$ for all $n \ge N$ where $I = \inf_{s\in M} \|w(e^x - N_r^k(s))\|_\infty$. For any $x_0 \in [a,b] \sim \bigcup_{i=1}^{\ell} \{x_i\}$,

$|w(x_0)(e^{x_0} - N_r^k(s_n)(x_0))| \le I + \epsilon$ for $n \ge N$ and so

$$|w(x_0)(e^{x_0} - N_r^k(s^*)(x_0))| = \lim_{n\to\infty} |w(x_0)(e^{x_0} - N_r^k(s_n)(x_0))| \le I + \epsilon.$$

Thus $\|w(e^x - N_r^k(s^*))\|_\infty = \max_{x\in[a,b]} |w(x)(e^x - N_r^k(s^*)(x))| =$

$\sup_{\substack{x\in[a,b] \\ x\neq x_i}} |w(x)(e^x - N_r^k(s^*)(x))| \le I + \epsilon$. Hence $\|w(e^x - N_r^k(s^*))\|_\infty =$

$\inf_{s\in M} \|w(e^x - N_r^k(x))\|_\infty$. ∎

CHAPTER III

FURTHER EXAMPLES OF OPERATORS DEFINING ITERATIVE SCHEMES

Section 1: Definition and Basic Results

In Chapter II we were concerned exclusively with iteration using Newton's method. We shall now consider other iterative methods for the solution of equations and examine them for various properties which have been defined. First we introduce the important concept of order which affords us a means of classifying the iterative schemes which we shall discuss. The following definition is due to Traub [30]:

Definition 3.1: An iteration function $\phi_f: E^1 \to E^1$ defined by $\phi_f(x_k) = x_{k+1}$, $k = 0,1,2,\ldots$ for finding a root α of the equation $f^{-1}(y) - x = 0$ (fixed x) is said to be of order p (or have order of convergence p) if there exists a nonzero constant C such that $\dfrac{|\phi_f(x_k) - \alpha|}{|x_k - \alpha|^p} \to C$ as $k \to \infty$. The number C is called the asymptotic error constant.

Lemma 3.2: If the order of an iteration function exists, then it is unique.

Proof: Assume the iteration function ϕ_f has two orders, p_1 and p_2. Set $p_2 = p_1 + \delta$, $\delta > 0$. Then

$$\lim_{k\to\infty} \frac{|\phi_f(x_k) - \alpha|}{|x_k - \alpha|^{p_2}} = \lim_{k\to\infty} \frac{|\phi_f(x_k) - \alpha|}{|x_k - \alpha|^{p_1+\delta}} = C_2 ,$$

and $$\lim_{k\to\infty} \frac{|\phi_f(x_k) - \alpha|}{|x_k - \alpha|^{p_1}} = 0,$$ which is a contradiction. ∎

<u>Remark 1</u>: The iteration functions with which we are concerned are continuous one-point iteration functions; that is, each successive iterate x_{k+1} is obtained using only information from the previous iterate x_k. No old information is reused.

<u>Remark 2</u>: A detailed discussion of order of convergence and related topics may be found in Traub [30].

With each iteration function $\phi_f(x_k) = x_{k+1}$ we can associate an iteration operator $\Phi_f: S \to S \subseteq C[a,b]$ via $\Phi_f(s_k) = s_{k+1}$ for approximating the continuous function $f \in C[a,b]$. The subset S of $C[a,b]$ depends upon f and $[a,b]$. We now embark upon a study of iteration operators Φ_f associated with known one-point iteration functions ϕ_f for $f(x) = e^x$ and $f(x) = x^{\frac{1}{p}}$, $p > 1$, p an integer. The one-point iteration functions ϕ_f which we shall be concerned with are order-preserving Pade rational approximations to a certain class of iteration functions generated by inverse hyperosculatory interpolation at a single point. This material first appeared in a paper of Traub [31] and is presented in detail in Chapter 5 of his book [30]. The iteration functions mentioned are all of integral order and many are widely known and have other origins (suitable reference will be made if such is the case). Since linear convergence (convergence of order 1) is not very interesting nor practical we commence with a study of iteration functions of order 2. We shall refer to the schemes generated by iteration functions (operators) as iterative schemes.

Section 2: Iteration Functions of Order 2

1. First, we shall only mention that the Newton iteration function already discussed has order of convergence 2.

2. The only other iterative scheme of order 2 which seems to exist in the literature is Newton's method for computing roots using the function $g(y) = \frac{y^p - x}{y^p}$, $p > 1$, p an integer. Applying Newton's method to this function, we obtain the iteration function $\phi_f(x_n) = \frac{x_n}{p}((p+1) - \frac{x_n^p}{x})$. For $0 < a < b$, define $K = \{h(x) \in C[a,b]:$ $0 < h(x) < ((p+1)x)^{\frac{1}{p}}$ for all $x \in [a,b]\}$ and $\Phi_f(h)(x) = \frac{h(x)}{p}((p+1) - \frac{(h(x))^p}{x})$ for $h \in K$. With these definitions it is easily checked that Φ_f is continuous, $\Phi_f\colon K \to K$ and $\Phi_f(x^{\frac{1}{p}}) = x^{\frac{1}{p}}$. Set $\psi(t) = \frac{t}{p}((p+1) - t^p)$. Then $\psi'(t) = \frac{p+1}{p}(1-t^p)$ and so

$$\psi'(t) \begin{cases} < 0, & t > 1 \\ = 0, & t = 1 \\ > 0, & 0 < t < 1\,. \end{cases}$$

Therefore Φ_f is pointwise strictly monotone at $x^{\frac{1}{p}}$, $p = 2,3,\ldots$, possesses Property I at $x^{\frac{1}{p}}$ and is one-sided from below at $x^{\frac{1}{p}}$. (It should be remarked that the function $\psi(t)$ analyzed above actually only yields the fact that Φ_f possesses Property I. However, the argument to obtain Φ_f pointwise strictly monotone is entirely analogous and shall be omitted. This format will be followed throughout our analysis.) In order to apply Theorem 1.27 we must require that if r_p denotes the best relative approximation to $x^{\frac{1}{p}}$ from $R_m^n[a,b]$ (m,n arbitrary but fixed throughout this discussion) with deviation λ_p then $\delta r_p \in M$ for $\delta \in \left[\frac{1}{1+\lambda_p}, \frac{1}{1-\lambda_p}\right]$. To insure this we require that $\lambda_p \le \frac{(p+1)^{1/p} - 1}{(p+1)^{1/p} + 1}$ which is apparently a rather severe restriction on λ_p for large p (that is, r_p must be a very good approximation to $x^{1/p}$). If we so restrict λ_p, then Theorem 1.27 is applicable and there exists $\gamma_p \in \left(\frac{1}{1+\lambda_p}, \frac{1}{1-\lambda_p}\right)$ such that $\gamma_p r_p$ is the best starting approximation for $x^{1/p}$ with respect to Φ_f. Since iteration is well-defined and Φ_f is a one-sided operator we have that $\gamma_p r_p$ is the best starting approximation for $x^{1/p}$ with respect to Φ_f^k, $k = 1,2,\ldots$. Analyzing the proof of Theorem 1.27, it is clear that γ_p is the unique solution, in the above interval, to the equation

$$\frac{\Phi_f(\gamma_p r_p)(x_1)}{x_1^{1/p}} - 1 = \frac{\Phi_f(\gamma_p r_p)(x_2)}{x_2^{1/p}} - 1$$

where $r_p(x_1) = (1+\lambda_p)x_1^{1/p}$ and $r_p(x_2) = (1-\lambda_p)x_2^{1/p}$. Solving, we find that $\gamma_p = \left[\frac{2\lambda_p(p+1)}{(1+\lambda_p)^{p+1} - (1-\lambda_p)^{p+1}}\right]^{1/p}$.

<u>Section 3</u>: Iteration Functions of Order 3

1. The first iteration scheme of order 3 to be considered is attributed to Halley (see [30]) and, as pointed out by Traub, has the distinction of being one of the most frequently rediscovered iteration functions in the literature. A recent rediscovery is due to J.S. Frame [3]. If we wish to find a root of the equation $g(y) = f^{-1}(y) - x = 0$ (fixed x) then the iteration function ϕ_f is defined by

$$x_{n+1} = \phi_f(x_n) = x_n - \frac{g(x_n)}{g'(x_n) + \frac{g''(x_n)}{2}\left(\frac{-g(x_n)}{g'(x_n)}\right)} .$$

A. First, let us define $g(y) = \ln y - x$. Then $g(y) = 0$ is equivalent to $y = e^x$ (for fixed x). For this choice of g the Halley iteration function becomes

$$x_{n+1} = \phi_f(x_n) = x_n \left[\frac{2 + x - \ln x_n}{2 - x + \ln x_n}\right].$$

In order to approximate e^x on an interval $[a,b]$ (a,b arbitrary with $a < b$) let us put $K = \{h(x) \in C[a,b]: h(x) > e^{x-2}$ for all $x \in [a,b]\}$ and define the associated iteration operator

$\Phi_f(h)(x) = h(x) \left[\frac{2 + x - \ln h(x)}{2 - x + \ln h(x)}\right]$ for $h \in K$. With these definitions $\Phi_f: K \to C[a,b]$. If we wish to iterate with the operator Φ_f then we must further restrict K. We shall omit the restriction and remark that it is straightforward. Clearly Φ_f is continuous for $h \in K$, $\Phi_f(e^x) = e^x$ and Φ is pointwise fixed at $h \in K$. Setting $\psi(t) = t\left[\frac{2 - \ln t}{2 + \ln t}\right]$ for $t \in (e^{-2},\infty)$, we find that $\psi'(t) = -\frac{(\ln t)^2}{(2 + \ln t)^2} \le 0$. Thus $\psi(t)$ is decreasing for $t > e^{-2}$ and so Φ_f is clearly pointwise strictly monotone at e^x and possesses Property I at e^x. In order to apply Theorem 1.27 we must require that $\delta r \in M = K \cap R^n_m[a,b]$ for $\delta \in \left[\frac{1}{1+\lambda}, \frac{1}{1-\lambda}\right]$ where r denotes the best relative approximation to e^x from $R^n_m[a,b]$. To insure this we must demand that $\lambda < \frac{e^2 - 1}{e^2 + 1}$ which is a very mild restriction on λ since we always have $\lambda < 1$. Let us remark that it may indeed be the case that we need not restrict λ at all as was discovered earlier in the discussion of the Newton operator for approximating e^x. With the above restriction the best starting approximation for e^x with respect to Φ_f is a multiple, γ, of the best relative approximation to e^x and is given implicitly by

$$\frac{\Phi_f(\gamma r)(x_1)}{e^{x_1}} + \frac{\Phi_f(\gamma r)(x_2)}{e^{x_2}} = 2$$

where x_1 and x_2 are such that $r(x_1) = (1+\lambda)e^{x_1}$ and $r(x_2) = (1-\lambda)e^{x_2}$. This equation reduces to solving

$$\gamma(1+\lambda)\left[\frac{2 - \ln \gamma(1+\lambda)}{2 + \ln \gamma(1+\lambda)}\right] + \gamma(1-\lambda)\left[\frac{2 - \ln \gamma(1-\lambda)}{2 + \ln \gamma(1-\lambda)}\right] = 2$$

which would be solved numerically on a computer. It is important to note that in this case the best starting approximation need not be independent of the number of iterations of Φ_f since Φ_f is not a one-sided operator.

B. If we desire to find $\sqrt[p]{x}$ (for fixed x), $p \geq 2$, p an integer, then we are confronted with finding a zero of the equation $g(y) = y^p - x$. In this case the Halley iteration function is given by

$$x_{n+1} = \phi_f(x_n) = x_n \left[\frac{(p-1)x_n^p + (p+1)x}{(p+1)x_n^p + (p-1)x}\right].$$

For the operator setting let $K = \{h \in C[a,b]: h(x) > 0$ for all $x \in [a,b]\}$ where $0 < a < b$ and set $\Phi_f(h)(x) = h(x)\left[\frac{(p-1)(h(x))^p + (p+1)x}{(p+1)(h(x))^p + (p-1)x}\right]$ for $h \in K$. It is clear that Φ_f is continuous and $\Phi_f(x^{\frac{1}{p}}) = x^{\frac{1}{p}}$, $p = 2,3,\ldots$. Set $\psi(t) = t\left[\frac{(p-1)t^p + (p+1)}{(p+1)t^p + (p-1)}\right]$ where $t > 0$. Differentiating, we obtain $\psi'(t) = (p^2-1)\left[\frac{t^p - 1}{(p+1)t^p + (p-1)}\right]^2 \geq 0$ and hence $\psi(t)$ is increasing for $t > 0$. Consequently Φ_f is pointwise strictly monotone at $x^{\frac{1}{p}}$

and possesses Property I at $x^{\frac{1}{p}}$. Since $\Phi_f(h_n)(x_0) \to 0$ as $h_n(x_0) \to 0$ for each fixed $x_0 \in [a,b]$ and Φ_f is monotone we have that $\Phi_f: K \to K$. Thus, repeated application of Φ_f is well-defined. If $M = K \cap R_m^n[a,b]$ then $\delta r_p \in M$ for $\delta \in \left[\frac{1}{1+\lambda_p}, \frac{1}{1-\lambda_p}\right]$ where r_p denotes the best relative approximation to $x^{1/p}$ from $R_m^n[a,b]$ with deviation λ_p. Therefore, by Theorem 1.27, the best starting approximation for $x^{1/p}$, $p = 2,3,\ldots$ from M with respect to Φ_f is $\gamma_p r_p$ for γ_p the unique solution in $\left(\frac{1}{1+\lambda_p}, \frac{1}{1-\lambda_p}\right)$ to the equation

$$\frac{\Phi_f(\gamma_p r_p)(x_1)}{x_1^{1/p}} + \frac{\Phi_f(\gamma_p r_p)(x_2)}{x_2^{1/p}} = 2$$

where $r_p(x_1) = (1+\lambda_p)x_1^{1/p}$, $r_p(x_2) = (1-\lambda_p)x_2^{1/p}$. As in part A., the best starting approximation need not be independent of the number of iterations of Φ_f. In the present literature the above scheme for roots is often ascribed to Bailey (see [30]).

2. The second and final scheme of order 3 which evolves from the work of Traub is defined by the iteration function

$$x_{n+1} = \phi_f(x_n) = x_n - \frac{g(x_n)}{g'(x_n)} - \frac{g''(x_n)g^2(x_n)}{2g'^3(x_n)}$$

where we are again searching for a root of the equation $g(y) = f^{-1}(y) - x$.

A. Let $g(y) = \ell n\, y - x$. Then $g(y) = 0$ is equivalent to $y = e^x$ (fixed x). For this choice of g, ϕ_f is defined by

$$x_{n+1} = \phi_f(x_n) = \frac{x_n}{2}[1 + (\ell n\, x_n - x - 1)^2].$$

To approximate e^x on an interval $[a,b]$ $(a < b)$ set

$$K = \{h \in C[a,b]: h(x) > 0 \text{ for all } x \in [a,b]\}$$

and define $\Phi_f(h)(x) = \frac{h(x)}{2}[1 + (\ln h(x) - x - 1)^2]$ for $h \in K$. Clearly Φ_f is continuous, $\Phi_f: K \to K$ and $\Phi_f(e^x) = e^x$. Setting $\psi(t) = \frac{t}{2}[1 + (\ln t - 1)^2]$ for $t > 0$ and differentiating we obtain $\psi'(t) = \frac{1}{2}(\ln t)^2 \geq 0$. Hence $\psi(t)$ is increasing for $t > 0$ and so Φ_f is pointwise strictly monotone at e^x and possesses Property I at e^x. If $M = K \cap R_m^n[a,b]$ then $\delta r \in M$ for $\delta \in \left[\frac{1}{1+\lambda}, \frac{1}{1-\lambda}\right]$ where r denotes the best relative approximation to e^x from $R_m^n[a,b]$ with deviation λ. Thus the best starting approximation for e^x from M with respect to Φ_f is γr for some $\gamma \in \left(\frac{1}{1+\lambda}, \frac{1}{1-\lambda}\right)$ (γ is obtained by solving a certain transcendental equation as in the above work). Since the operator Φ_f is not one-sided, the best starting approximation need not be independent of the number of iterations.

B. For computing $x^{1/p}$, $p = 2,3,\ldots$ the iteration function is defined by

$$x_{n+1} = \phi_f(x_n) = \frac{x_n(1-p)}{2p^2}\left[\left(\frac{2p-1}{p-1} - \frac{x}{x_n^p}\right)^2 + \frac{p^2(1-2p)}{(1-p)^2}\right]$$

where $g(y) = y^p - x$. For $0 < a < b$, set $K = \{h \in C[a,b]: h(x) > 0 \text{ for all } x \in [a,b]\}$ and define $\Phi_f(h)(x) =$

$$\frac{1-p}{2p^2}h(x)\left[\left(\frac{2p-1}{p-1} - \frac{x}{(h(x))^p}\right)^2 + \frac{p^2(1-2p)}{(1-p)^2}\right]$$, for $h \in K$. With these definitions, $\Phi_f: K \to C[a,b]$ and so, as mentioned earlier, K needs to be further restricted if we wish to iterate with Φ_f. Clearly Φ_f is continuous and $\Phi_f(x^{1/p}) = x^{1/p}$. Setting

$\psi(t) = \frac{1-p}{2p^2} t \left[\left(\frac{2p-1}{p-1} - t^{-p} \right)^2 + \frac{p^2(1-2p)}{(1-p)^2} \right]$ for $t > 0$ and differentiating we find that $\psi'(t) = \frac{(2p-1)(p-1)}{2p^2} \left[1 - \frac{1}{t^p} \right]^2 > 0$. Thus Φ_f is pointwise strictly monotone at $x^{1/p}$ and possesses Property I at $x^{1/p}$. Hence we may apply Theorem 1.27 as before and shall omit the details. The lack of one-sidedness of Φ_f may again destroy independence of the best starting approximation upon the number of iterations.

Section 4: Iteration Functions of Order 4

Next we consider a class of iteration schemes with order of convergence 4. The first three schemes which we shall examine are derived from the previously mentioned work of Traub.

1. The first fourth-order iteration function is defined by

$$x_{n+1} = \phi_f(x_n) = x_n - \frac{g(x_n)}{g'(x_n)} \left[1 + \frac{g''(x_n)}{2g'(x_n)} \frac{g(x_n)}{g'(x_n)} + \left[2 \left(\frac{g''(x_n)}{2g'(x_n)} \right)^2 - \frac{g'''(x_n)}{6g'(x_n)} \right] \frac{g^2(x_n)}{[g'(x_n)]^2} \right].$$

A. For approximating e^x for some fixed x this iteration function becomes

$$x_{n+1} = \phi_f(x_n) = x_n[1 - (\ln x_n - x) + \tfrac{1}{2}(\ln x_n - x)^2 - \tfrac{1}{6}(\ln x_n - x)^3]$$

where $g(y) = \ln y - x$. To approximate e^x on the entire interval $[a,b]$ $(a < b)$ set $K' = \{h \in C[a,b]: h(x) > 0 \text{ for all } x \in [a,b]\}$ and define Φ_f by $\Phi_f(h)(x) = h(x)[1 - (\ln h(x) - x) + \frac{1}{2}(\ln h(x) - x)^2 - \frac{1}{6}(\ln h(x) - x)^3]$ for $h \in K'$. As before Φ_f is continuous and $\Phi_f(e^x) = e^x$. Set $\psi(t) = t[1 - (\ln t) + \frac{1}{2}(\ln t)^2 - \frac{1}{6}(\ln t)^3]$ for $t > 0$. Upon differentiation we obtain $\psi'(t) = -\frac{1}{6}(\ln t)^3$ and so it is clear that Φ_f is pointwise strictly monotone at e^x, possesses Property I at e^x and is one-sided from below at e^x. As with the Newton operator in approximating e^x, Φ_f being one-sided from below implies that large values of $h(x)$ (fixed x) yield negative values for $\Phi_f(h)(x)$. By the Implicit Function Theorem the equation $\Phi_f(h)(x) = 0$ defines $h(x) > e^x$ as a continuous function of x. In our case the desired $h(x)$ is obtained by solving

$$1 - (\ln h(x) - x) + \frac{1}{2}(\ln h(x) - x)^2 - \frac{1}{6}(\ln h(x) - x)^3 = 0 \quad \text{for } h(x).$$

If r_0 is the unique real zero of $1 - u + \frac{1}{2}u^2 - \frac{1}{6}u^3 = 0$ then $h(x) = e^{x+r_0}$. Note that $1 < r_0 < 2$. Setting $K = \{h \in C[a,b]: 0 < h(x) < e^{x+r_0} \text{ for all } x \in [a,b]\}$, we have that $\Phi_f: K \to K$. If r denotes the best relative approximation to e^x from $R_m^n[a,b]$ with deviation λ then the best starting approximation for Φ_f^k, $k = 1,2,\ldots$ is γr provided $\delta r \in M$ for $\delta \in \left[\frac{1}{1+\lambda}, \frac{1}{1-\lambda}\right]$. To insure this we require that $\delta r(x) \le \delta(1+\lambda)e^x \le \frac{1+\lambda}{1-\lambda} e^x < e^{x+r_0}$, or equivalently $\lambda < \frac{e^{r_0}-1}{e^{r_0}+1}$. With this restriction on λ the best starting approximation is γr where γ is computed by solving

$$\frac{\Phi_f(\gamma r)(x_1)}{e^{x_1}} = \frac{\Phi_f(\gamma r)(x_2)}{e^{x_2}} \quad \text{where } r(x_1) = (1+\lambda)e^{x_1} \text{ and } r(x_2) = (1-\lambda)e^{x_2}.$$

Upon simplification we find that γ is the unique solution in $\left(\frac{1}{1+\lambda}, \frac{1}{1-\lambda}\right)$ to the transcendental equation

$$(1+\lambda)[1- \ln \gamma(1+\lambda)+\frac{1}{2}(\ln \gamma(1+\lambda))^2-\frac{1}{6}(\ln \gamma(1+\lambda))^3] = (1-\lambda)[1-(\ln \gamma(1-\lambda)) + \frac{1}{2}(\ln \gamma(1-\lambda))^2 - \frac{1}{6}(\ln \gamma(1-\lambda))^3] .$$

B. For the approximation of $x^{1/p}$, $p = 2,3,\ldots$ the iteration function is given by

$$x_{n+1} = \phi_f(x_n) = -\frac{x_n}{p}\left[\left(\frac{x_n^p - x}{x_n^p}\right) + \frac{p-1}{2p}\left(\frac{x_n^p - x}{x_n^p}\right)^2 + \frac{(p-1)(2p-1)}{6p^2}\left(\frac{x_n^p - x}{x_n^p}\right)^3 - p\right]$$

where $g(y) = y^p - x$. For $0 < a < b$, set $K = \{h \in C[a,b]: h(x) > 0$ for all $x \in [a,b]\}$ and define the associated iteration operator Φ_f by

$$\Phi_f(h)(x) = -\frac{h(x)}{p}\left[\left(\frac{(h(x))^p - x}{(h(x))^p}\right) + \frac{p-1}{2p}\left(\frac{(h(x))^p - x}{(h(x))^p}\right)^2 + \frac{(p-1)(2p-1)}{6p^2}\left(\frac{(h(x))^p - x}{(h(x))^p}\right)^3 - p\right]$$

for $h \in K$. Φ_f is continuous and $\Phi_f(x^{1/p}) = x^{1/p}$. Set

$$\psi(t) = -\frac{t}{p}\left[\left(1 - (\tfrac{1}{t})^p\right) + \frac{p-1}{2p}\left(1 - (\tfrac{1}{t})^p\right)^2 + \frac{(p-1)(2p-1)}{6p^2}\left(1 - (\tfrac{1}{t})^p\right)^3 - p\right]$$

for $t > 0$.

Then $\psi'(t) = \frac{(p-1)(2p-1)(3p-1)}{6p^3}\left(\frac{t^p - 1}{t^p}\right)^3$ and so $\psi(t)$ is increasing for $t > 1$ and decreasing for $0 < t < 1$. Therefore Φ_f is pointwise strictly monotone at $x^{1/p}$, $p = 2,3,\ldots$, possesses Property I at $x^{1/p}$ and is one-sided from above. The one-sidedness of Φ_f yields $\Phi_f: K \to K$ and hence we are able to iterate. Setting $M = K \cap R_m^n[a,b]$, we obtain that $\delta r_p \in M$ for $\delta \in \left[\frac{1}{1+\lambda_p}, \frac{1}{1-\lambda_p}\right]$ where r_p denotes the best starting approximation to $x^{1/p}$ from $R_m^n[a,b]$ with deviation λ_p. Hence, according to Theorems 1.27 and 1.30, the best starting approximation for $x^{1/p}$, $p = 2,3,\ldots$ with respect to Φ_f^k, $k = 1,2,\ldots$ is $\gamma_p r_p$ for $\gamma_p \in \left(\frac{1}{1+\lambda_p}, \frac{1}{1-\lambda_p}\right)$ the unique solution to the equation

$$\frac{\Phi_f(\gamma_p r_p)(x_1)}{x_1^{1/p}} = \frac{\Phi_f(\gamma_p r_p)(x_2)}{x_2^{1/p}} \quad \text{where} \quad r_p(x_1) = (1+\lambda_p)x_1^{1/p},$$

$r_p(x_2) = (1-\lambda_p)x_2^{1/p}$.

In the next two iteration schemes derived from the work of Traub we shall only be concerned with approximating e^x. The results for the approximation of $x^{1/p}$, $p = 2,3,\ldots$ are analogous although the algebra involved is much more tedious.

2. The next iteration function of order 4 with which we shall be concerned is defined by

$$x_{n+1} = \phi_f(x_n) = x_n - \frac{\dfrac{g(x_n)}{g'(x_n)}}{1 - \dfrac{g''(x_n)g(x_n)}{2(g'(x_n))^2} + \left(\dfrac{g'''(x_n)}{6g'(x_n)} - \dfrac{(g''(x_n))^2}{4(g'(x_n))^2}\right)\dfrac{(g(x_n))^2}{(g'(x_n))^2}}.$$

For the approximation of e^x we have $g(y) = \ln y - x$ and the iteration function is given by $x_{n+1} = \phi_f(x_n) = x_n \left[1 - \dfrac{(\ln x_n - x)}{1 + \frac{1}{2}(\ln x_n - x) + \frac{1}{12}(\ln x_n - x)^2} \right]$. Set $K = \{h \in C[a,b]: h(x) > 0$ for all $x \in [a,b]\}$ where $a < b$ and define

$$\Phi_f(h)(x) = h(x)\left[1 - \frac{(\ln h(x) - x)}{1 + \frac{1}{2}(\ln h(x) - x) + \frac{1}{12}(\ln h(x) - x)^2}\right] \quad \text{for} \quad h \in K.$$

Then Φ_f is continuous, $\Phi_f: K \to K$ and $\Phi_f(e^x) = e^x$. Setting

$$\psi(t) = t\left[1 - \frac{\ln t}{1 + \frac{1}{2}\ln t + \frac{1}{12}(\ln t)^2}\right]$$

we find that $\psi'(t) = \dfrac{\frac{1}{144}(\ln t)^4}{(1 + \frac{1}{2}\ln t + \frac{1}{12}(\ln t)^2)^2} \geq 0$ for all $t > 0$.

Hence Φ_f is pointwise strictly monotone at e^x and possesses

Property I at e^x. The conditions of Theorem 1.27 are all satisfied and so we are able to conclude that the best starting approximation for e^x on $[a,b]$ (with respect to Φ_f) from $M = K \cap R_m^n[a,b]$ is a positive multiple (γ) of the best relative approximation to e^x on $[a,b]$ from $R_m^n[a,b]$. Let r denote the best relative approximation to e^x on $[a,b]$ with deviation λ. Then the value of γ we are seeking is obtained by solving the equation

$\frac{\Phi_f(\gamma r)(x_1)}{e^{x_1}} + \frac{\Phi_f(\gamma r)(x_2)}{e^{x_2}} = 2$, where $r(x_1) = (1+\lambda)e^{x_1}$ and $r(x_2) = (1-\lambda)e^{x_2}$, which would be effected numerically on a computer. Since Φ_f fails to be one-sided the best starting approximation need not be independent of the number of iterations.

3. The final iteration function with order of convergence 4 obtainable from Traub's work which we shall consider is defined by

$$x_{n+1}=\phi_f(x_n)=x_n - \frac{\frac{g(x_n)}{g'(x_n)}\left[\frac{g''(x_n)}{2g'(x_n)} + \left(\frac{g'''(x_n)}{6g'(x_n)} - \frac{(g''(x_n))^2}{4(g'(x_n))^2}\right)\frac{g(x_n)}{g'(x_n)}\right]}{\frac{g''(x_n)}{2g'(x_n)} - \frac{g(x_n)}{g'(x_n)}\left(\frac{(g''(x_n))^2}{2(g'(x_n))^2} - \frac{g'''(x_n)}{6g'(x_n)}\right)}.$$

For the approximation of e^x (fixed $x \in [a,b]$), $g(y) = \ln y - x$ and the iteration function is defined by

$$x_{n+1} = \phi_f(x_n) = x_n\left[1 + \frac{(\ln x_n - x)(\ln x_n - x - 6)}{2(\ln x_n - x + 3)}\right].$$

To approximate e^x on the entire interval $[a,b]$ set

$$K = \{h \in C[a,b]\colon h(x) > e^{x-3} \text{ for all } x \in [a,b]\}$$

and define the associated iteration operator $\Phi_f(h)(x) = h(x)\left[1 + \frac{(\ln h(x) - x)(\ln h(x) - x - 6)}{2(\ln h(x) - x + 3)}\right]$ where $h \in K$. It is clear that Φ_f is continuous and $\Phi_f(e^x) = e^x$. Furthermore we shall demonstrate that Φ_f is one-sided from above at e^x and consequently $\Phi_f(K) \subseteq K$. Set $\psi(t) = t\left[1 + \frac{(\ell n\, t)(\ell n\, t - 6)}{2(\ell n\, t + 3)}\right]$ for $t > \frac{1}{e^3}$. Differentiating, we obtain $\psi'(t) = \frac{(\ell n\, t)^3}{2(\ell n\, t + 3)^2}$ and so Φ_f is pointwise strictly monotone at e^x, possesses Property I at e^x and is one-sided from above at e^x. To be able to apply Theorem 1.27 we must further demand that $\lambda < \frac{e^3-1}{e^3+1}$ where λ denotes the error associated with r, the best relative approximation to e^x on $[a,b]$ from $R^n_m[a,b]$. As noted before, this restriction is extremely mild since we always have $\lambda < 1$. With this restriction Theorems 1.27 and 1.30 are applicable and hence the best starting approximation for e^x on $[a,b]$ with respect to Φ^k_f, $k = 1,2,\ldots$ is γr where γ is the unique solution to the equation $\frac{\Phi_f(\gamma r)(x_1)}{e^{x_1}} = \frac{\Phi_f(\gamma r)(x_2)}{e^{x_2}}$ in the interval $\left(\frac{1}{1+\lambda}, \frac{1}{1-\lambda}\right)$ (x_1, x_2 as before). For our definition of Φ_f this reduces to solving a third degree equation in $\ell n\, \gamma$.

4. Finally we shall examine an iteration function also of rational form which has order of convergence 4 and is due to Kiss (see [30]). As above we shall only consider the iteration function ϕ_f corresponding to $g(y) = \ell n\, y - x$. The iteration function of Kiss is defined by

$$x_{n+1} = \phi_f(x_n) = x_n - \frac{\frac{g(x_n)}{g'(x_n)}\left(1 - \frac{g''(x_n)}{2g'(x_n)}\frac{g(x_n)}{g'(x_n)}\right)}{1 - \frac{g''(x_n)}{g'(x_n)}\frac{g(x_n)}{g'(x_n)} + \frac{g'''(x_n)g^2(x_n)}{6g'(x_n)(g'(x_n))^2}} .$$

For the special choice $g(y) = \ell n\, y - x$ we have

$$x_{n+1} = \phi_f(x_n) = x_n \left[1 - \frac{(\ell n\, x_n - x)(1 + \frac{1}{2}(\ell n\, x_n - x))}{1+(\ell n\, x_n - x) + \frac{1}{3}(\ell n\, x_n - x)^2} \right].$$

To approximate e^x on the interval $[a,b]$ where $a < b$ set $K'' = \{h \in C[a,b]: h(x) > 0$ for all $x \in [a,b]\}$ and define Φ_f by

$$\Phi_f(h)(x) = h(x) \left[1 - \frac{(\ell n\, h(x) - x)(1 + \frac{1}{2}(\ell n\, h(x) - x))}{1 + (\ell n\, h(x) - x) + \frac{1}{3}(\ell n\, h(x) - x)^2} \right] \text{ for } h \in K''.$$

It is easily shown that $1 + (\ell n\, h(x) - x) + \frac{1}{3}(\ell n\, h(x) - x)^2 \neq 0$ for all $h \in K''$ and $x \in [a,b]$. Set $\psi(t) = t\left[1 - \frac{\ell n\, t(1 + \frac{1}{2}\ell n\, t)}{1 + \ell n\, t + \frac{1}{3}(\ell n\, t)^2}\right]$ for $t > 0$. Then $\psi'(t) = \frac{-\frac{1}{18}(\ell n\, t)^3(\ell n\, e^3 t)}{(1 + (\ell n\, t) + \frac{1}{3}(\ell n\, t)^2)^2}$. Note that $\psi'(t) = 0$ for $t = 1$ and for $t = e^{-3}$. To insure that Φ_f will be pointwise strictly monotone at e^x and possess Property I we shall demand that $t > e^{-3}$. If we set $K' = \{h \in C[a,b]: h(x) > e^{x-3}$ for all $x \in [a,b]\}$, we have that $\Phi_f: K' \to C[a,b]$, Φ_f is continuous, $\Phi_f(e^x) = e^x$, Φ_f is pointwise strictly monotone at e^x, possesses Property I at e^x and is one-sided from below. For each fixed $x \in [a,b]$, $\Phi_f(h)(x)$ is decreasing for $h(x) > e^x$ and increasing for $h(x) < e^x$ and $\Phi_f(h)(x) = 0$ at $h_1(x) = e^{x-\sqrt{6}}$, $h_2(x) = e^{x+\sqrt{6}}$. If we set

$$K = \{h \in C[a,b]: e^{x-2} < h(x) < e^{x+2} \text{ for all } x \in [a,b]\}$$

then it is easy to show that $\Phi_f: K \to K$ and the aforementioned properties of Φ_f are preserved. In order to apply Theorem 1.27 we must require that $\delta r \in M = K \cap R^n_m[a,b]$ for $\delta \in \left[\frac{1}{1+\lambda}, \frac{1}{1-\lambda}\right]$ where r denotes the best relative approximation to e^x from $R^n_m[a,b]$ with deviation λ. This requirement is equivalent to

$\lambda < \dfrac{e^2 - 1}{e^2 + 1}$ which is again not a severe restriction. Applying Theorems 1.27 and 1.30 (subject to the above restriction) we know that there exists a unique $\gamma \in \left(\frac{1}{1+\lambda}, \frac{1}{1-\lambda}\right)$ such that γr is the best starting approximating to e^x on $[a,b]$ from M (with respect to Φ_f^k, $k = 1,2,3,\ldots$). The value of γ is obtained by solving the equation $\dfrac{\Phi_f(\gamma r)(x_1)}{e^{x_1}} = \dfrac{\Phi_f(\gamma r)(x_2)}{e^{x_2}}$ where x_1 and x_2 are such that $r(x_1) = (1+\lambda)e^{x_1}$ and $r(x_2) = (1-\lambda)e^{x_2}$. For the given operator Φ_f this equation becomes

$$(1+\lambda)\left[\frac{6 - (\ln \gamma(1+\lambda))^2}{3 + 3\ln \gamma(1+\lambda) + (\ln \gamma(1+\lambda))^2}\right] =$$

$$(1-\lambda)\left[\frac{6 - (\ln \gamma(1-\lambda))^2}{3 + 3\ln \gamma(1-\lambda) + (\ln \gamma(1-\lambda))^2}\right]$$

and we are confronted with solving a certain quartic equation in $\ln \gamma$.

Section 5: A Sequence of Schemes for Approximating $\sqrt{x}$

In this section we shall consider a sequence of iterative schemes for approximating $\sqrt{x}$ which were introduced by G. Merz [12] and subsequently analyzed by Meinardus and Taylor [11]. The iteration function(s) of Merz are defined by

$$x_{n+1} = \phi_k(x_n) = \sqrt{x}\,\frac{(x_n + \sqrt{x})^k + (x_n - \sqrt{x})^k}{(x_n + \sqrt{x})^k - (x_n - \sqrt{x})^k}$$

for each fixed integer $k \geq 2$ (ϕ_k would be written ${}_k\phi_f$ in the earlier notation, where $f(x) = \sqrt{x}$). Let us note that, for each k, ϕ_k possesses a purely algebraic expression. Merz has shown that for each fixed integer $k \geq 2$ the above sequence $\{x_n\}$ converges to $\sqrt{x}$ starting with any initial guess $x_0 > 0$ and furthermore the order of convergence is k. Observe that for $k = 2$, $\phi_2(y) = \frac{1}{2}(y + \frac{x}{y})$ which is the well-known Newton iteration scheme.

It is interesting to note that $\phi_k(\phi_m(y)) = \phi_{k \cdot m}(y)$ and so one has the choice of repeated iterations with a lower order formula or fewer iterations with a higher order formula to obtain the same final iterate.

Since we wish to approximate $\sqrt{x}$ on an interval $[a,b]$, $0 < a < b$, we consider the associated iteration operator(s) as before. Set

$$K = \{h \in C[a,b]: h(x) > 0 \text{ for all } x \in [a,b]\}$$

and define $\Phi_k(h)(x) = \sqrt{x}\,\dfrac{(h(x) + \sqrt{x})^k + (h(x) - \sqrt{x})^k}{(h(x) + \sqrt{x})^k - (h(x) - \sqrt{x})^k}$ for $h \in K$.

It is easily seen that $\Phi_k: K \to K$, Φ_k is continuous and $\Phi_k(\sqrt{x}) = \sqrt{x}$ for each $k \geq 2$. Set $\psi_k(t) = \dfrac{(1+t)^k + (1-t)^k}{(1+t)^k - (1-t)^k}$ for $t > 0$.

Upon computing $\psi_k'(t)$ it becomes clear that for k even, Φ_k is pointwise strictly monotone at $\sqrt{x}$, one-sided from above at $\sqrt{x}$ and possesses Property I at $\sqrt{x}$. Likewise, for k odd Φ_k is pointwise strictly monotone at $\sqrt{x}$ and possesses Property I at $\sqrt{x}$. Hence, by Lemma 1.5 and Theorem 1.27 we have that for each k the best starting approximation for $\sqrt{x}$ on $[a,b]$ from $M = K \cap R_m^n[a,b]$ (with respect to Φ_k) is a multiple of the best relative approximation to $\sqrt{x}$ on $[a,b]$ from $R_m^n[a,b]$. Furthermore, if k is even we have that $\Phi_{k^m} = \Phi_k^m = \Phi_k(\Phi_k^{m-1})$, $m = 2,3,\ldots$ have the same starting value by Theorem 1.30. If r denotes the best relative approximation to $\sqrt{x}$ on $[a,b]$ with deviation λ, then γr is the optimal starting approxition for Φ_{k^m} (k even), $m = 1,2,\ldots$ where $\gamma \in \left(\frac{1}{1+\lambda}, \frac{1}{1-\lambda}\right)$ is the unique solution to the equation $\frac{\Phi_k(\gamma r)(x_1)}{\sqrt{x_1}} - 1 = \frac{\Phi_k(\gamma r)(x_2)}{\sqrt{x_2}} - 1$ where x_1 and x_2 are such that $r(x_1) = (1+\lambda)\sqrt{x_1}$ and $r(x_2) = (1-\lambda)\sqrt{x_2}$. Using the fact that k is even and $\gamma \in \left(\frac{1}{1+\lambda}, \frac{1}{1-\lambda}\right)$, we obtain $\gamma = \left(\frac{1}{1-\lambda^2}\right)^{\frac{1}{2}}$. Hence $\left(\frac{1}{1-\lambda^2}\right)^{\frac{1}{2}} r$ is the best starting approximation for every even ordered scheme. For the case $k = 2$ and $R_m^n[a,b]$ replaced by $P_n[a,b]$ this result was earlier obtained (independently) by P.H. Sterbenz and C.T. Fike [24] and R.F. King and D.L. Phillips [7]. Additional background concerning these and associated results appears in Taylor [27].

In the case that k is odd, Φ_k is pointwise strictly monotone at $\sqrt{x}$ and possesses Property I at $\sqrt{x}$ but is no longer one-sided. Hence we may no longer conclude that the best starting approximation for $\sqrt{x}$ is independent of the number of iterations. Computing the best starting approximation for a single application of Φ_k (according to Theorem 1.27), we obtain γr where γ satisfies the equation $\frac{\Phi_k(\gamma r)(x_1)}{\sqrt{x_1}} - 1 = 1 - \frac{\Phi_k(\gamma r)(x_2)}{\sqrt{x_2}}$ and r, x_1, x_2 are as before. This reduces to solving

$$\left(\frac{1 - \lambda\gamma + \gamma}{1 + \lambda\gamma - \gamma}\right)^k + \left(\frac{1 + \lambda\gamma + \gamma}{1 - \lambda\gamma - \gamma}\right)^k = 2$$

and the unique solution $\gamma \in \left(\frac{1}{1+\lambda}, \frac{1}{1-\lambda}\right)$ must (apparently) be found using numerical methods on a computer. Note that γ, and hence the best starting approximation, depends on k which is not the case for k even. Since $\Phi_{k^m} = \Phi_k^m$, Φ_k^m possesses Property I for $m = 1,2,\ldots$. Thus Theorem 1.27 is applicable for any number of iterations of Φ_k. For k odd, the best starting approximation for Φ_k^m, $m = 1,2,\ldots$ is $\gamma_m r$ where γ_m is the only solution of

$$\left(\frac{1 - \lambda\gamma_m + \gamma_m}{1 + \lambda\gamma_m - \gamma_m}\right)^{k^m} + \left(\frac{1 + \lambda\gamma_m + \gamma_m}{1 - \lambda\gamma_m - \gamma_m}\right)^{k^m} = 2 \quad \text{in the interval}$$

$$\left(\frac{1}{1+\lambda}, \frac{1}{1-\lambda}\right).$$

Remark 1: In this chapter we have been concerned with approximation using the relative error. The iteration functions could also be analyzed in the uniform approximation setting. In this case we would be checking for Property J. However in view of some earlier remarks we should probably examine the function $g(y) = e^y - x$; that is, consider an approximation of the function $f(x) = \ell n\ x$ on some interval $[a,b]$ $(0 < a < b)$.

CHAPTER IV

OPTIMAL STARTING APPROXIMATIONS FOR A GENERALIZED WEIGHT FUNCTION

Section 1: Introduction and Examples

In [13], [14] D.G. Moursund introduced the concept of uniform approximation using a generalized weight function. The original development here was concerned with polynomial approximation. Applications pertaining to the notion of a generalized weight function and concerning both polynomial and rational approximation appear in [15], [16] and [26].

Definition 4.1: Let X be a compact metric space. A real-valued function $W(x,y)$ defined on $X \times (-\infty,\infty)$ is called a generalized weight function if:

(i) $W(x,y)$ is continuous on $X \times (-\infty,\infty)$

(ii) $\operatorname{Sgn} W(x,y) = \operatorname{sgn} y$ for all $x \in X$ (in particular, $W(x,0) = 0$).

(iii) For each $x \in X$, $W(x,y)$ is strictly monotone increasing in y with $\lim_{|y|\to\infty} |W(x,y)| = \infty$.

The problem which we shall consider is the following: given $X \subseteq [a,b]$, $f \in K \subseteq C(X)$ and $M = K \cap R_m^n[a,b]$ (m,n fixed), find $r^* \in M$ such that $\|W[x,\Phi(f)(x) - \Phi(r^*)(x)]\| = \inf_{r \in M} \|W[x,\Phi(f)(x) - \Phi(r)(x)]\|$ where $\Phi: K \to C(X)$ is continuous and $\|W[x,\Phi(f)(x) - \Phi(r)(x)]\| = \max_{x \in X} |W[x,\Phi(f)(x) - \Phi(r)(x)]|$.

Before proceeding to the basic results we shall give some examples of generalized weight functions (here we take Φ to be the identity operator).

Example 1: $W(x,y) \equiv y$ corresponds to ordinary uniform approximation.

Example 2: $W(x,y) \equiv \frac{y}{f(x)}$ corresponds to ordinary relative error approximation provided $f \neq 0$ on X (f is the function being approximated).

Let X be compact, $X \subseteq [a,b]$, $K \subseteq C(X)$ and $\Phi: K \to C(X)$ a continuous operator which is pointwise strictly monotone and pointwise fixed at $f \in K$.

Definition 4.2: The operator Φ is said to be sign-preserving if for each $k \in K$, $\text{sgn}(f(x) - k(x)) = \text{sgn}(\Phi(f)(x) - \Phi(k)(x))$ for all $x \in X$.

Definition 4.3: The operator Φ is said to be sign-reversing if for each $k \in K$, $\text{sgn}(f(x) - k(x)) = -\text{sgn}(\Phi(f)(x) - \Phi(k)(x))$ for all $x \in X$.

Section 2: Existence, Uniqueness, and Characterization Theorems

In this section we shall be concerned with existence, uniqueness, and characterization of optimal starting approximations using a generalized weight function.

Theorem 4.4: Let K be a convex subset of $C(X)$ and $\Phi: K \to C(X)$ a continuous operator which is pointwise strictly monotone and pointwise fixed at $f \in K \sim M$. Suppose further that Φ is sign-preserving, sign-reversing or a one-sided operator and let $W(x,y)$ be a generalized weight function on X. Then, for $M = K \cap R_m^n[a,b]$ relatively open in $R_m^n[a,b]$, $r \in M$ is a best starting approximation for $\Phi(f)$ relative to $W(x,y)$ if and only if there exists $\{x_i\}_{i=1}^N \subseteq X$, $N = 2 + \max\{n + \partial q, m + \partial p\}$, for which

(i) $x_1 < x_2 < \ldots < x_N$,

(ii) $|W[x_i, \Phi(f)(x_i) - \Phi(r)(x_i)]| = \|W[x, \Phi(f)(x) - \Phi(r)(x)]\|$

(iii) $\operatorname{sgn}(f(x_i) - r(x_i)) = (-1)^{i+1}\operatorname{sgn}(f(x_1) - r(x_1))$,

$i = 1,2,\ldots,N$.

Proof: We shall prove the theorem for Φ a one-sided operator from below and for Φ sign-preserving and remark that the other cases follow in a similar manner.

Case 1: $\Phi(f) \geq \Phi(h)$ for all $h \in K$.

(Sufficiency) Suppose there exists $\{x_i\}_{i=1}^N \subseteq X$ such that $|W[x_i, \Phi(f)(x_i) - \Phi(r)(x_i)]| = \|W[x, \Phi(f)(x) - \Phi(r)(x)]\|$ and $\operatorname{sgn}(f(x_i) - r(x_i)) = (-1)^{i+1}\operatorname{sgn}(f(x_1) - r(x_1))$. Note that $f \notin M$ implies that there exists $x_0 \in X$ such that $f(x_0) \neq r(x_0)$. Thus $\Phi(f)(x_0) \neq \Phi(r)(x_0)$ and so $|W[x_0, \Phi(f)(x_0) - \Phi(r)(x_0)]| \neq 0$ which implies $\|W[x, \Phi(f)(x) - \Phi(r)(x)]\| \neq 0$. Let $r^* \in M$ be such that $\|W[x, \Phi(f)(x) - \Phi(r^*)(x)]\| \leq \|W[x, \Phi(f)(x) - \Phi(r)(x)]\|$. Since $\operatorname{sgn} W[x, \Phi(f)(x) - \Phi(r)(x)] = \operatorname{sgn}(\Phi(f)(x) - \Phi(r)(x)) = 1$ or 0 for all $x \in X$ and all $r \in M$ we have $W[x_i, \Phi(f)(x_i) - \Phi(r^*)(x_i)] = |W[x_i, \Phi(f)(x_i) - \Phi(r^*)(x_i)]| \leq |W[x_i, \Phi(f)(x_i) - \Phi(r)(x_i)]| = W[x_i, \Phi(f)(x_i) - \Phi(r)(x_i)]$. Hence $\Phi(f)(x_i) - \Phi(r^*)(x_i) \leq \Phi(f)(x_i) - \Phi(r)(x_i)$ since $W(x,y)$ is strictly monotone increasing in y for each fixed $x \in X$. If $r(x_i) > f(x_i)$ then $r(x_i) \geq r^*(x_i)$. Otherwise $r^*(x_i) > r(x_i) > f(x_i)$ and so

$\Phi(f)(x_i) - \Phi(r^*)(x_i) = |\Phi(f)(x_i) - \Phi(r^*)(x_i)| > |\Phi(f)(x_i) - \Phi(r)(x_i)| =$ $\Phi(f)(x_i) - \Phi(r)(x_i)$ which is a contradiction. Similarly $r(x_i) < f(x_i)$ implies $r(x_i) \le r^*(x_i)$. Thus, by the standard argument, $r - r^*$ possesses at least $N-1$ zeros and so $r \equiv r^*$. Hence we have thé desired result.

(Necessity) Assume there exists $\{x_i\}_{i=1}^{N'} \subseteq X$, N' maximal, $N' < N$ on which $|W[x_i,\Phi(f)(x_i) - \Phi(r)(x_i)]| = \|W[x,\Phi(f)(x) - \Phi(r)(x)]\|$ and $\operatorname{sgn}(f(x_i) - r(x_i)) = (-1)^{i+1}\operatorname{sgn}(f(x_1) - r(x_1))$. About each x_i construct a relatively open interval I_i such that all extreme points $= \{x \in X: |W[x,\Phi(f)(x) - \Phi(r)(x)]| = \|W[x,\Phi(f)(x) - \Phi(r)(x)]\|\} \subseteq \bigcup_{i=1}^{N'} I_i$, $\overline{I}_j \cap \overline{I}_k = \phi$ for $j \neq k$ and for all extreme points in each I_i the function $f-r$ has constant sign. In what follows, we make explicit use of the fact that $|W[x,\Phi(f)(x) - \Phi(r)(x)]| =$ $W[x,\Phi(f)(x) - \Phi(r)(x)]$. Let

$$Y = X \cap \left(\bigcap_{i=1}^{N'} \tilde{I}_i \right)$$

where $\tilde{I}_i$ denotes the complement of I_i with respect to $[a,b]$. Y is a compact subset of X and $W[x,\Phi(f)(x) - \Phi(r)(x)] <$ $\|W[x,\Phi(f)(x) - \Phi(r)(x)]\|$ for all $x \in Y$. Hence by the continuity of W (as a function of x) there exists $\rho > 0$ such that

$$\max_{x \in Y} W[x,\Phi(f)(x) - \Phi(r)(x)] \le \|W[x,\Phi(f)(x) - \Phi(r)(x)]\| - \rho.$$

For each i, let

$$V_i = \{x \in X \cap \overline{I}_i: W[x,\Phi(f)(x) - \Phi(r)(x)] \ge \frac{\|W[x,\Phi(f)(x) - \Phi(r)(x)]\|}{2}$$
$$\text{and } \operatorname{sgn}(f(x) - r(x)) = \operatorname{sgn}(f(x_i) - r(x_i))\}.$$

Now $V = \bigcup_{i=1}^{N'} V_i$ is a compact subset of X and, since $f(x) - r(x) \neq 0$ on V, there exists $\eta > 0$ such that $|f(x) - r(x)| \geq \eta$ on V. Let

$$Z_i = \{x \in X \cap \bar{I}_i : W[x,\Phi(f)(x) - \Phi(r)(x)] \geq \frac{\|W[x,\Phi(f)(x) - \Phi(r)(x)]\|}{2}$$

$$\text{and } \operatorname{sgn}(f(x) - r(x)) \neq \operatorname{sgn}(f(x_i) - r(x_i))\} .$$

Observe that $W[x,\Phi(f)(x) - \Phi(r)(x)] < \|W[x,\Phi(f)(x) - \Phi(r)(x)]\|$ for all $x \in Z = \bigcup_{i=1}^{N'} Z_i$ by the construction of the intervals $\{\bar{I}_i\}$. Finally, let

$$U_i = \{x \in X \cap \bar{I}_i : W[x,\Phi(f)(x) - \Phi(r)(x)] \leq \frac{\|W[x,\Phi(f)(x) - \Phi(r)(x)]\|}{2}\} ,$$

and set $U = \bigcup_{i=1}^{N'} U_i$. Then by continuity there exists $\delta > 0$, $\delta \leq \rho$, such that

$$\max_{x \in Z \cup U} W[x,\Phi(f)(x) - \Phi(r)(x)] \leq \|W[x,\Phi(f)(x) - \Phi(r)(x)]\| - \delta .$$

Given $\tau > 0$ we can, using Lemma 1.8, select $r \in R_m^n[a,b]$ such that (1.2) holds. By the continuity of Φ and W we can select $\epsilon_1 > 0$ such that for $0 < \epsilon \leq \epsilon_1$, r_ϵ satisfies (1.2) and

$$\max_{x \in Y \cup Z \cup U} W[x,\Phi(f)(x) - \Phi(r_\epsilon)(x)] < \|W[x,\Phi(f)(x) - \Phi(r)(x)]\| - \frac{\delta}{2} .$$

Next, by continuity of f and r, we can select ϵ_2, $0 < \epsilon_2 \leq \epsilon_1$, such that for $0 < \epsilon \leq \epsilon_2$, r_ϵ lies strictly between $f(x)$ and $r(x)$ on $V = \bigcup_{i=1}^{N'} V_i$. Then, since Φ is one-sided from below at f, we have $\Phi(f)(x) - \Phi(r_\epsilon)(x) < \Phi(f)(x) - \Phi(r)(x)$ on $V = \bigcup_{i=1}^{N'} V_i$ where $0 < \epsilon \leq \epsilon_2$. Hence $W[x,\Phi(f)(x) - \Phi(r_\epsilon)(x)] < W[x,\Phi(f)(x) - \Phi(r)(x)] \leq \|W[x,\Phi(f)(x) - \Phi(r)(x)]\|$ on V if $0 < \epsilon \leq \epsilon_2$. Therefore $\max_{x \in V} W[x,\Phi(f)(x) - \Phi(r_\epsilon)(x)] < \|W[x,\Phi(f)(x) - \Phi(r)(x)]\|$. Consequently for ϵ such that $0 < \epsilon \leq \epsilon_2$,

r_ϵ is such that

$$\|W[x,\Phi(f)(x) - \Phi(r_\epsilon)(x)]\| < \|W[x,\Phi(f)(x) - \Phi(r)(x)]\|.$$

Finally, since M is relatively open in $R_m^n[a,b]$, we can select ϵ_3 with $0 < \epsilon_3 \le \epsilon_2$ so that $r_{\epsilon_3} \in M$ and $\|W[x,\Phi(r)(x) - \Phi(r_{\epsilon_3})(x)]\| < \|W[x,\Phi(f)(x) - \Phi(r)(x)]\|$ and hence the result.

Case 2: Φ is sign-preserving.

(Sufficiency) Suppose that there exists $\{x_i\}_{i=1}^N \subseteq X$ such that $|W[x_i,\Phi(f)(x_i) - \Phi(r)(x_i)]| = \|W[x,\Phi(f)(x) - \Phi(r)(x)]\|$ and $\text{sgn}(f(x_i) - r(x_i)) = (-1)^{i+1}\text{sgn}(f(x_1) - r(x_1))$. Let $r^* \in M$ be such that $\|W[x,\Phi(f)(x) - \Phi(r^*)(x)]\| \le \|W[x,\Phi(f)(x) - \Phi(r)(x)]\|$. Suppose $r(x_i) < f(x_i)$. Then $r(x_i) \le r^*(x_i)$ for otherwise $r^*(x_i) < r(x_i) < f(x_i)$ and so $\Phi(f)(x_i) - \Phi(r)(x_i) < \Phi(f)(x_i) - \Phi(r^*)(x_i)$ which implies $|W[x,\Phi(f)(x_i) - \Phi(r^*)(x_i)]| = W[x_i,\Phi(f)(x_i) - \Phi(r^*)(x_i)] > W[x_i,\Phi(f)(x_i) - \Phi(r)(x_i)] = |W[x_i,\Phi(f)(x_i) - \Phi(r)(x_i)]|$, which is a contradiction. Similarly $r(x_i) > f(x_i)$ implies $r(x_i) \ge r^*(x_i)$. Thus, as in Case 1 above, $r \equiv r^*$ and we have the desired result.

(Necessity) Assume there exists $\{x_i\}_{i=1}^{N'} \subseteq X$, $N' < N$, N' maximal on which $|W[x_i,\Phi(f)(x_i) - \Phi(r)(x_i)]| = \|W[x,\Phi(f)(x) - \Phi(r)(x)]\|$ and $\text{sgn}(f(x_i) - r(x_i)) = (-1)^{i+1}\text{sgn}(f(x_1) - r(x_1))$. About each x_i construct a relatively open interval I_i such that all extreme points $= \{x \in X: |W[x,\Phi(f)(x) - \Phi(r)(x)]| = \|W[x,\Phi(f)(x) - \Phi(r)(x)]\|\} \subseteq \bigcup_{i=1}^{N'} I_i$, $\bar{I}_j \cap \bar{I}_k = \emptyset$ for $j \ne k$ and for all extreme points in each I_i the function $f-r$ has constant sign. Construct the sets Y, V, Z and U as in the proof for Case 1. Using the same arguments as in the above proof we can find $\delta > 0$ such that

$$\max_{x\in Y\cup Z\cup U} |W[x,\Phi(f)(x) - \Phi(r)(x)]| \le \|W[x,\Phi(f)(x) - \Phi(r)(x)]\| - \delta .$$

By Lemma 1.8 we can select $r_\epsilon \in R_m^n[a,b]$ satisfying (1.2). By the continuity of Φ and W we can select $\epsilon_1 > 0$ such that for $0 < \epsilon \le \epsilon_1$, r_ϵ satisfies (1.2) and

$$\max_{x\in Y\cup Z\cup U} |W[x,\Phi(f)(x) - \Phi(r_\epsilon)(x)]| \le \|W[x,\Phi(f)(x) - \Phi(r)(x)]\| - \frac{\delta}{2} .$$

Next, by continuity of f and r, we can select ϵ_2, $0 < \epsilon_2 \le \epsilon_1$, such that for $0 < \epsilon \le \epsilon_2$, r_ϵ lies strictly between $f(x)$ and $r(x)$ on $V = \bigcup_{i=1}^{N'} V_i$. If $f(x_i) < r(x_i)$ on V_i then $f(x) < r_\epsilon(x) < r(x)$ for all $x \in V_i$ and so $\Phi(f)(x) - \Phi(r)(x) < \Phi(f)(x) - \Phi(r_\epsilon)(x) < 0$ $(0 < \epsilon \le \epsilon_2)$ which implies $|W[x,\Phi(f)(x) - \Phi(r_\epsilon)(x)]| = -W[x,\Phi(f)(x) - \Phi(r_\epsilon)(x)] < -W[x,\Phi(f)(x) - \Phi(r)(x)] = |W[x,\Phi(f)(x) - \Phi(r)(x)]|$ and so $|W[x,\Phi(f)(x) - \Phi(r_\epsilon)(x)]| < \|W[x,\Phi(f)(x) - \Phi(r)(x)]\|$ on V_i. Similarly, if $f(x_i) > r(x_i)$ on V_i we obtain $|W[x,\Phi(f)(x) - \Phi(r_\epsilon)(x)]| < \|W[x,\Phi(f)(x) - \Phi(r)(x)]\|$ for $0 < \epsilon \le \epsilon_2$. Thus $\max_{x\in V} |W[x,\Phi(f)(x) - \Phi(r_\epsilon)(x)]| < \|W[x,\Phi(f)(x) - \Phi(r)(x)]\|$. Consequently, for ϵ subject to $0 < \epsilon \le \epsilon_2$, r_ϵ is such that $\|W[x,\Phi(f)(x) - \Phi(r_\epsilon)(x)]\| < \|W[x,\Phi(f)(x) - \Phi(r)(x)]\|$. Since M is relatively open in $R_m^n[a,b]$ we may choose ϵ_3 with $0 < \epsilon_3 \le \epsilon_2$ such that $r_{\epsilon_3} \in M$ and $\|W[x,\Phi(f)(x) - \Phi(r_{\epsilon_3})(x)]\| < \|W[x,\Phi(f)(x) - \Phi(r)(x)]\|$. Thus we have the desired result. ▮

<u>Corollary 4.5</u>: If $\Phi(f)$ has a best starting approximation under the setting of the above theorem then it is unique.

<u>Proof</u>: Using the properties of the generalized weight function, uniqueness is established in the same way that sufficiency was established in Theorem 4.4. ▮

As noted earlier we can only guarantee existence of a best starting approximation in the case where $X = [a,b]$ and $K = C[a,b]$.

Theorem 4.6: Let $\Phi: C[a,b] \to C[a,b]$ be a continuous operator which is pointwise strictly monotone at $f \in C[a,b] \sim M$ and pointwise fixed at each $k \in C[a,b]$ where $M = R_m^n[a,b]$ (m,n fixed). Further assume that for each $x \in [a,b]$, $|\Phi(r_i)(x)| \to \infty$ whenever $\{r_i\} \subseteq M$ and $r_i(x) \to \pm\infty$. Then there exists a best starting approximation for $\Phi(f)$, relative to the generalized weight function $W(x,y)$, from the class $\Phi(M)$.

Proof: The result follows at once by combining Lemma 1 of [17] with Lemma 1.11. ▮

Remark 1: Similar results corresponding to restricted range approximation and osculatory rational interpolation for the operator setting may also be obtained in the generalized weight function setting.

Remark 2: Theorem 4.4 yields a generalization of the work which we have done in Chapter I. If we choose $W(x,y) = w(x)\cdot y$ ($w(x)$ a positive, continuous weight function on X) then the problem of determining $r \in M$ such that

$$\|W[x,\Phi(f)(x) - \Phi(r)(x)]\| = \inf_{s\in M} \|W[x,\Phi(f)(x) - \Phi(s)(x)]\| \tag{4.1}$$

reduces to determining $r \in M$ such that $\|\Phi(f) - \Phi(r)\| = \inf_{s\in M} \|\Phi(f) - \Phi(s)\|$ where $\|h\| = \max_{x\in X} |w(x)h(x)|$, which is the problem studied earlier. A question which arises is the following: can the problem (4.1) be cast in terms of another generalized weight function $\tilde{W}$ for which the solution is already known? In the problem (4.1) we are approximating $\Phi(f)$ be elements of $\Phi(M)$ with respect to the generalized weight function W. Suppose we could find another generalized weight function $\tilde{W}$ such that the problem (4.1) would reduce to finding $r \in M$ such that $\|\tilde{W}[x,g(x) - r(x)]\| = \inf_{s \in M} \|\tilde{W}[x,g(x) - s(x)]\|$ where $\Phi(f) = g$. This problem has already been studied by Moursund and Taylor [17].

CHAPTER V

AN IMPROVED NEWTON ITERATION SCHEME FOR APPROXIMATING EXP(X) WHICH IS OPTIMAL

Section 1: Introduction

In a recent paper I. Ninomiya [18] suggested a modification of the Newton-Raphson method for calculating square roots which reduced the error at each step of the algorithm. This method was generalized by R.F. King [6] to a modified Newton method for calculating integral roots. Both Ninomiya and King show that the improved Newton iteration converges by a factor of approximately 2^{2^n-1} faster than the usual Newton iteration (comparing the n-th iterate of each). R.F. King notes that in actual practice one would not use this scheme for calculating a p-th root on a machine which has p as its floating base since in this case an exponential add order is much faster than a multiplication, and it would usually be better to use Newton's method with an extra iteration (if necessary).

G.D. Taylor [27] has shown that the suggested modification of Newton's method above for integral roots is optimal in the sense of our work. We shall now show that an analogous modification of Newton's method for $f(x) = e^x$ yields an improved iteration scheme for approximating this function.

Let $[a,b]$ be a fixed interval and K a convex subset of $C[a,b]$. Recall that if $N(h)(x) = h(x)(1 + x - \ln h(x))$ $(h \in K)$

is the Newton operator used to approximate e^x on the interval $[a,b]$ then $N(h)(x) \le e^x$ for all $h \in K$ and $x \in [a,b]$; that is, N is one-sided from below. With this in mind we are able to construct an improved Newton iteration procedure for computing e^x which is optimal and which differs from the usual Newton method by a multiplicative factor at each step. At each step of the algorithm the multiplicative factor has the effect of "translating" the previous Newton iterate upward sufficiently to halve the relative error.

Section 2: The Improved Newton Iteration Scheme

Let $[a,b]$ be a fixed interval as above and define K by $K = \{h \in C[a,b]: 0 < h(x) < e^{1+x}$ for all $x \in [a,b]\}$. For fixed nonnegative integers m and $\bar{m}$ let $R_m^{\bar{m}}$ denote the usual class of rational functions defined on the interval $[a,b]$ (See Chapter I). Set $M = K \cap R_m^{\bar{m}}$ (assume $M \neq \phi$) and define $N(h)(x) = h(x)(1 + x - \ln h(x))$ for $h \in K$. This is, of course, one Newton iteration for calculating e^x starting with the initial guess $h(x)$. Set $E_+^n = \{(c_1,\ldots,c_n) : (c_1,\ldots,c_n) \in E^n$ and $c_i > 0,\ i = 1,\ldots,n\}$ where E^n denotes Euclidean n-space. For $(c_1,\ldots,c_n,h) \in E_+^n \times K$ formally define the sequence

$$\begin{aligned} y_0(x) &= h(x) \\ y_k(x) &= c_k\, N(y_{k-1})(x),\ k = 1,2,\ldots \end{aligned} \tag{5.1}$$

which may have no meaning for some choices of $c_1,\ldots,c_n$. Define $S_n = \{(c_1,\ldots,c_n,r) : c_i \in E_+^1,\ r \in M$ and $y_{n-1} \in K\}$; that is, $(c_1,\ldots,c_n,r) \in S_n$ if and only if $y_1 = c_1N(r) \in K$, $y_2 = c_2N(y_1) = c_2N(c_1N(r)) \in K,\ldots,\ y_{n-1} = c_{n-1}N(y_{n-2}) \in K$. Thus S_n consists of all tuples $(c_1,\ldots,c_n,r)$ for which $r \in M$ and the iteration above is well-defined. Now define $T_n: S_n \to C^+[a,b] = \{f \in C[a,b]: f(x) > 0$ for all $x \in [a,b]\}$ by $T_n(c_1,\ldots,c_n,r)(x) \equiv y_n(x)$ ($y_n(x)$ given by (5.1)). Observe that $(c_1,\ldots,c_{n-1},c_n,r) \in S_n$ implies that $(c_1,\ldots,c_{n-1},c,r) \in S_n$ for any $c \in E_+^1$, $(c_1,\ldots,c_k,r) \in S_k$, $k = 1,2,\ldots,n$ and $T_n(c_1,\ldots,c_{n-1},c_n,r) = c_n T_n(c_1,\ldots,c_{n-1},1,r) = c_n N(T_{n-1}(c_1,\ldots,c_{n-1},r))$. We shall show that there exists a unique best approximation to e^x in the relative norm from $T_n(S_n)$; that is, there exists $(c_1^{(n)},\ldots,c_n^{(n)},r^*) \in S_n$ for which

$$\left\| \frac{e^x - T_n(c_1^{(n)},\ldots,c_n^{(n)},r^*)(x)}{e^x} \right\|_\infty \le \left\| \frac{e^x - T_n(c_1,\ldots,c_n,r)(x)}{e^x} \right\|_\infty \tag{5.2}$$

for all $(c_1,\ldots,c_n,r) \in S_n$ with equality holding in the above if and only if $c_i = c_i^{(n)}$, $i = 1,2,\ldots,n$ and $r \equiv r^*$.

Let $\bar{r}(x)$ denote the best relative approximation to e^x on $[a,b]$ from $\bar{R}_m^m$ with deviation λ_0; that is,

$$\left\| \frac{\bar{r}(x) - e^x}{e^x} \right\|_\infty = \inf_{r \in \bar{R}_m^m} \left\| \frac{r(x) - e^x}{e^x} \right\|_\infty = \lambda_0 . \tag{5.3}$$

Clearly $0 < \lambda_0 < 1$ so that $\bar{r} \in M$. The unique best approximation $T_n(c_1^{(n)},\ldots,c_n^{(n)},r^*)$ to e^x from $T_n(S_n)$ is given by the following set of recursive relations:

$$\beta_k = \exp\left[\frac{1}{2\lambda_k}\left(2\lambda_k+(1-\lambda_k)\ln(1-\lambda_k)-(1+\lambda_k)\ln(1+\lambda_k)\right)\right]$$
$$= e\left(\frac{1-\lambda_k}{1+\lambda_k}\right)^{\frac{1}{2\lambda_k}} \frac{1}{\sqrt{1-\lambda_k^2}},\quad k = 0,1,\ldots \tag{5.4}$$

$$\lambda_k = \frac{1 - \beta_{k-1}(1 - \lambda_{k-1})[1 - \ln \beta_{k-1}(1 - \lambda_{k-1})]}{1 + \beta_{k-1}(1 - \lambda_{k-1})[1 - \ln \beta_{k-1}(1 - \lambda_{k-1})]}, \quad k = 1,2,\ldots \tag{5.5}$$

$$d_k = \frac{2}{\beta_{k-1}(1 - \lambda_{k-1})[1 - \ln \beta_{k-1}(1 - \lambda_{k-1})] + 1}, \quad k = 1,2,\ldots \tag{5.6}$$

$$c_k^{(n)} = \beta_k d_k, \quad k = 1,2,\ldots,n-1 \tag{5.7}$$

$$c_n^{(n)} = d_n \tag{5.8}$$

and

$$r^*(x) = \beta_0 \bar{r}(x) . \tag{5.9}$$

In passing from n to $n+1$ the tuples $(c_1^{(n)},\ldots,c_n^{(n)},r^*)$ and $(c_1^{(n+1)},\ldots,c_{n+1}^{(n+1)},r^*)$ are related as follows:

$$\begin{aligned} c_i^{(n)} &= c_i^{(n+1)}, \quad i = 1,\ldots,n-1 , \\ c_n^{(n+1)} &= \beta_n c_n^{(n)} \end{aligned} \tag{5.10}$$

and

$$c_{n+1}^{(n+1)} N\{\beta_n(T_n(c_1^{(n)},\ldots,c_n^{(n)},r^*))\} = T_{n+1}(c_1^{(n+1)},\ldots,c_{n+1}^{(n+1)},r^*) .$$

Note that r^* is the best starting approximation for calculating e^x with the usual Newton iteration; that is,

$$\left\|\frac{N^n(r^*)(x) - e^x}{e^x}\right\|_\infty \le \left\|\frac{N^n(r)(x) - e^x}{e^x}\right\|_\infty \tag{5.11}$$

for all $r \in M$ with equality holding if and only if $r \equiv r^*$, where $N(r)(x) = r(x)(1 + x - \ln r(x))$ and $N^n(r)(x) = N(N^{n-1}(r))(x)$, $n = 2,3,\ldots$ Finally we have that β_k is such that

$$\beta_k(1-\lambda_k)[1-\ln \beta_k(1-\lambda_k)] = \beta_k(1+\lambda_k)[1-\ln \beta_k(1+\lambda_k)], \quad k = 0,1,2,\ldots \tag{5.12}$$

Lemma 5.1: Let N_x be defined by $N_x(t) = t(1 + x - \ln t)$ for $t > 0$. Let x,y be two given real numbers and let θ,φ be such that $0 < \theta < 1 < \varphi$. Then, for $\frac{1}{\theta} < t < \frac{e}{\theta}$ we have

$$\frac{N_y(\varphi e^y)}{e^y} > \frac{N_y(t\varphi e^y)}{e^y} \cdot \frac{e^x}{N_x(t\theta e^x)}$$

and for $0 < t < \frac{1}{\varphi}$ we have

$$\frac{N_x(\theta e^x)}{e^x} > \frac{N_x(t\theta e^x)}{e^x} \cdot \frac{e^y}{N_y(t\varphi e^y)} .$$

Proof: Set $f(t) = \frac{N_y(t\varphi e^y)}{e^y} \cdot \frac{e^x}{N_x(t\theta e^x)} - \frac{N_y(\varphi e^y)}{e^y}$

$$= \frac{\varphi}{\theta}\left[\frac{1 - \ln t\varphi}{1 - \ln t\theta}\right] - \varphi[1 - \ln \varphi] .$$

Now consider the function $h(t) = t(1 - \ln t)$ for $t > 0$. Upon differentiation, we obtain

$$h'(t) < 0 \quad \text{if} \quad t > 1, \; h'(t) > 0 \quad \text{if} \quad t < 1. \tag{5.13}$$

Thus since $\frac{\varphi}{\theta} > \varphi > 1$ it is clear that $f(\frac{1}{\theta}) < 0$. We also have that $f'(t) = \frac{\varphi}{\theta}\left[\frac{\ln t\theta - \ln t\varphi}{t(1 - \ln t\theta)^2}\right] < 0$ if $t \neq \frac{e}{\theta}$. Hence the first result.

Similarly, set

$$g(t) = \frac{N_x(t\theta e^x)}{e^x} \cdot \frac{e^y}{N_y(t\varphi e^y)} - \frac{N_x(\theta e^x)}{e^x}$$

$$= \frac{\theta}{\varphi}\left[\frac{1 - \ln t\theta}{1 - \ln t\varphi}\right] - \theta(1 - \ln \theta) .$$

Using (5.13) and the fact that $\frac{\theta}{\varphi} < \theta < 1$ it is clear that $g(\frac{1}{\varphi}) < 0$.

Now $g'(t) = \frac{\theta}{\varphi}\left[\frac{\ell n\, t\varphi - \ell n\, t\theta}{t(1 - \ell n\, t\varphi)^2}\right] > 0$ for $0 < t < \frac{1}{\varphi}$. Thus the second result. ▌

<u>Theorem 5.2</u>: $(c_1^{(n)},\ldots,c_n^{(n)},r^*)$ for $c_i^{(n)}$, $i = 1,\ldots,n$ and r^* given by (5.7)-(5.9) is an element of S_n and $T_n(c_1^{(n)},\ldots,c_n^{(n)},r^*)$ is the unique best relative approximation to e^x from $T_n(S_n)$ for each $n = 1,2,\ldots$. Furthermore

$$\left\|\frac{T_n(c_1^{(n)},\ldots,c_n^{(n)},r^*)(x) - e^x}{e^x}\right\|_\infty = \lambda_n , \qquad (5.14)$$

$T_n(c_1^{(n+1)},\ldots,c_n^{(n+1)},r^*)$ intersects e^x in $[a,b]$ and

$$\left\|\frac{N(T_n(c_1^{(n+1)},\ldots,c_n^{(n+1)},r^*))(x) - e^x}{e^x}\right\|_\infty \le \left\|\frac{N(T_n(c_1,\ldots,c_n,r))(x) - e^x}{e^x}\right\|_\infty \qquad (5.15)$$

for all $(c_1,\ldots,c_n,r) \in S_n$ with equality if and only if $c_i = c_i^{(n+1)}$, $i = 1,\ldots,n$ and $r \equiv r^*$. This last inequality states that $T_n(c_1^{(n+1)},\ldots,c_n^{(n+1)},r^*)$ is the unique best relative starting approximation to e^x for the Newton operator N, from the class $T_n(S_n)$.

<u>Proof</u>: We shall proceed by induction on n. For convenience we shall use the following notation: for $(c_1,\ldots,c_n,r) \in S_n$ set

$$M(r) = \left\{x: \frac{r(x) - e^x}{e^x} = \max_{y\in[a,b]} \frac{r(y) - e^y}{e^y}\right\}$$

$$m(r) = \left\{x: \frac{r(x) - e^x}{e^x} = \min_{y\in[a,b]} \frac{r(y) - e^y}{e^y}\right\}$$

$$M(c_1,\ldots,c_n,r) = \left\{ x : \frac{T_n(c_1,\ldots,c_n,r)(x) - e^x}{e^x} = \max_{y\in[a,b]} \frac{T_n(c_1,\ldots,c_n,r)(y) - e^y}{e^y} \right\}$$

$$m(c_1,\ldots,c_n,r) = \left\{ x : \frac{T_n(c_1,\ldots,c_n,r)(x) - e^x}{e^x} = \min_{y\in[a,b]} \frac{T_n(c_1,\ldots,c_n,r)(y) - e^y}{e^y} \right\}.$$

Observe that $M(c_1,\ldots,c_n,r) = M(c_1,\ldots,c_{n-1},1,r)$ and that $m(c_1,\ldots,c_n,r) = m(c_1,\ldots,c_{n-1},1,r)$.

To establish the case $n = 1$ we shall establish various conditions that a best relative approximation from $T_1(S_1)$ must satisfy and then show that $T_1(c_1^{(1)},r^*) \equiv T_1(d_1,r^*)$ is strictly minimal within this subset of $T_1(S_1)$. First $r^* \in M$ since $0 < \bar{r}(x) \le (1+\lambda_0)e^x$, $r^*(x) = \beta_0\bar{r}(x)$ and $\beta_0(1+\lambda_0) < e$ as is easily checked. Thus $(c_1^{(1)},r^*) \in S_1$ as $c_1^{(1)}$ is clearly positive. Now if $T_1(c,r)$ is to be a minimal solution then c must clearly be given by

$$c = \frac{2}{\left[\max_{x\in[a,b]} \frac{T_1(1,r)(x)}{e^x} + \min_{x\in[a,b]} \frac{T_1(1,r)(x)}{e^x}\right]} \quad . \qquad (5.16)$$

Since $T_1(1,r)(x) = N(r)(x)$ we have that $1 \ge \max_{x\in[a,b]} \frac{T_1(1,r)(x)}{e^x} > \min_{x\in[a,b]} \frac{T_1(1,r)(x)}{e^x}$ with equality holding in the above if and only if r intersects e^x. With the above selection for c, we have

$c \max_{x\in[a,b]} \frac{T_1(1,r)(x)}{e^x} - 1 = 1 - c \min_{x\in[a,b]} \frac{T_1(1,r)(x)}{e^x}$. Thus

$$\max_{x\in[a,b]}\left(\frac{e^x - T_1(c,r)(x)}{e^x}\right) = 1 - \min_{x\in[a,b]} \frac{T_1(c,r)(x)}{e^x} = 1 -$$

$$c \min_{x\in[a,b]} \frac{T_1(1,r)(x)}{e^x} = c \max_{x\in[a,b]} \frac{T_1(1,r)(x)}{e^x} - 1 = \max_{x\in[a,b]}\left(\frac{T_1(c,r)(x)-e^x}{e^x}\right).$$

Now let $c_1 \in E_+^1$ be such that $c_1 > c$. Then for $x \in M(1,r)$ we have $\frac{T_1(c_1,r)(x) - e^x}{e^x} = \frac{c_1T_1(1,r)(x) - e^x}{e^x} > \left\|\frac{T_1(c,r)(x) - e^x}{e^x}\right\|_\infty$. Likewise, for $c_1 \in E_+^1$ and $c_1 < c$ it is clear that $T_1(c,r)$ is a strictly better relative approximation to e^x than $T_1(c_1,r)$. Hence if $T_1(c,r)$ is to be a minimal solution then c must be as in (5.16).

Next suppose $r \in M$ intersects e^x and let

$$c = \frac{2}{\left[1 + \min\limits_{x\in[a,b]} \frac{T_1(1,r)(x)}{e^x}\right]}$$

where we shall assume that $r \not\equiv r^*$.

Since r^* is the best starting approximation for the Newton operator to e^x we have (using $T_1(1,r) \equiv N(r)$)

$$\max_{x\in[a,b]} \frac{e^x - T_1(1,r)(x)}{e^x} > \max_{x\in[a,b]} \frac{e^x - T_1(1,r^*)(x)}{e^x}$$

or equivalently,

$$\min_{x\in[a,b]} \frac{T_1(1,r)(x)}{e^x} < \min_{x\in[a,b]} \frac{T_1(1,r^*)(x)}{e^x}. \tag{5.17}$$

Let us now show that r^* intersects e^x. Since $\frac{\bar{r}(x)}{e^x}$ has a maximum value of $(1 + \lambda_0)$ and a minimum value of $(1 - \lambda_0)$, we must show that $\beta_0(1 - \lambda_0) < 1 < \beta_0(1 + \lambda_0)$. We shall verify the second inequality and the first follows by an analogous argument.

Claim: $\beta_0(1+\lambda_0) > 1$.

Proof: Recall that $\beta_0 = e\left(\frac{1-\lambda_0}{1+\lambda_0}\right)^{\frac{1}{2\lambda_0}} \frac{1}{\sqrt{1-\lambda_0^2}}$ and so $\beta_0(1+\lambda_0) =$
$e\left(\frac{1-\lambda_0}{1+\lambda_0}\right)^{\frac{1}{2\lambda_0}-\frac{1}{2}}$. Taking logarithms we find that $\beta_0(1+\lambda_0) > 1$ is equivalent to $\ell n\left(\frac{1-\lambda_0}{1+\lambda_0}\right) > \frac{2\lambda_0}{\lambda_0-1}$. Consider the function $f(x) = \ell n\left(\frac{1-x}{1+x}\right) - \frac{2x}{x-1}$ for $0 < x < 1$. Upon differentiation we find that $f'(x) = \frac{4x}{(x-1)^2(x+1)} > 0$ for $0 < x < 1$ which implies f is increasing for $0 < x < 1$. Now $f(0) = 0$ implies $f(\lambda_0) > 0$ and so the desired result. Notice that the above argument is independent of the index of β and λ and thus we may conclude that $\beta_k(1-\lambda_k) < 1 < \beta_k(1+\lambda_k)$, $k = 0,1,2,\ldots$, a fact we shall make use of later.

Hence r^* intersects e^x and so by (5.17)

$$c^* = \frac{2}{\left[1 + \min\limits_{x\in[a,b]} \frac{T_1(1,r^*)(x)}{e^x}\right]} < c . \tag{5.18}$$

Let $x_1 \in M(1,r)$ and $y_1 \in M(1,r^*)$. Then $\left\|\frac{T_1(c,r)(x) - e^x}{e^x}\right\|_\infty =$

$$\frac{c\,T_1(1,r)(x_1) - e^{x_1}}{e^{x_1}} = \frac{c\,T_1(1,r^*)(y_1) - e^{y_1}}{e^{y_1}} > \left\|\frac{T_1(c_1^*,r^*)(x) - e^x}{e^x}\right\|_\infty .$$

This shows that if $r \in M$ intersects e^x and $r \not\equiv r^*$, then $T_1(c,r)$ is not a best relative approximation to e^x from $T_1(S_1)$. Let us now show that c^* of (5.18) is equal to $c_1^{(1)}$ of (5.8). First

$$\min_{x\in[a,b]} \frac{T_1(1,r^*)(x)}{e^x} = \min_{x\in[a,b]} \frac{r^*(x)}{e^x}(1 + x - \ell n\, r^*(x)) =$$

$\min\limits_{x\in[a,b]} \frac{r^*(x)}{e^x}\left(1 - \ln \frac{r^*(x)}{e^x}\right)$. Since $\beta_0(1 - \lambda_0) \le \frac{r^*(x)}{e^x} \le \beta_0(1 + \lambda_0)$

it is easily shown that the minimum desired is $\beta_0(1-\lambda_0)[1-\ln \beta_0(1-\lambda_0)] =$ $\beta_0(1+\lambda_0)[1-\ln \beta_0(1+\lambda_0)]$. Therefore $c^* = \frac{2}{\left[1 + \min\limits_{x\in[a,b]} \frac{T_1(1,r^*)(x)}{e^x}\right]} =$

$\frac{2}{\beta_0(1-\lambda_0)[1-\ln \beta_0(1-\lambda_0)] + 1} = c_1^{(1)}$ as required.

Now it still remains to consider the case where $r \in M$ does not intersect e^x on $[a,b]$. First we shall examine the case where $r(x) > e^x$ for all $x \in [a,b]$. Set

$$\min_{x\in[a,b]} \frac{r(x) - e^x}{e^x} = \alpha > 0 . \tag{5.19}$$

Using Property I, it is easily shown that $m(r) = M(1,r)$ and $M(r) = m(1,r)$. Furthermore it is evident that $r(x) = (1 + \alpha)e^x$ on $m(r)$. Now let $\hat{r}$ denote the best one-sided relative approximation to $f(x) = (1 + \alpha)e^x$ on $[a,b]$ from above from the class $R_m^{\bar{m}}$. Clearly $\hat{r} \in M$ if there exists $r \in M$ for which $r(x) \ge (1 + \alpha)e^x$ on $[a,b]$. Thus for every $r \in M$ satisfying $r(x) \ge (1 + \alpha)e^x$ for all $x \in [a,b]$ we have

$$\left\|\frac{\hat{r}(x) - (1 + \alpha)e^x}{(1 + \alpha)e^x}\right\|_\infty \le \left\|\frac{r(x) - (1 + \alpha)e^x}{(1 + \alpha)e^x}\right\|_\infty$$

with equality if and only if $r \equiv \hat{r}$. The function $\hat{r}(x) = \frac{\hat{p}(x)}{\hat{q}(x)}$ is uniquely characterized by a set of points $\{x_i\}_{i=1}^d \subseteq [a,b]$ satisfying $a \le x_1 <\ldots< x_d \le b$ where $d = \bar{m} + m + 2 - \delta$, $\delta = \min(\bar{m} - \partial\hat{p},\ m - \partial\hat{q})$ and on which $E(x) = \frac{\hat{r}(x) - (1 + \alpha)e^x}{(1 + \alpha)e^x}$ is alternately zero and $\|E(x)\|_\infty$ (or alternately equal to $\|E(x)\|_\infty$

and zero).

We shall now show that $(\hat{c},\hat{r})$ where $\hat{r}$ is the best one-sided relative approximation to $(1+\alpha)e^x$ is a better solution than any (c,r) with $r \in M$ satisfying (5.19), where $\hat{c}$ and c are the constants given by (5.16) corresponding to $\hat{r}$ and r respectively. By the uniqueness of the best one-sided relative approximation $\hat{r}$ to $(1+\alpha)e^x$, we have that

$$\max_{x\in[a,b]} \frac{r(x) - (1+\alpha)e^x}{(1+\alpha)e^x} > \max_{x\in[a,b]} \frac{\hat{r}(x) - (1+\alpha)e^x}{(1+\alpha)e^x} .$$

Select $x_0, x_1 \in [a,b]$ such that $\frac{r(x_0)}{e^{x_0}} = \max_{x\in[a,b]} \frac{r(x)}{e^x}$, $\frac{\hat{r}(x_1)}{e^{x_1}} = \max_{x\in[a,b]} \frac{\hat{r}(x)}{e^x}$. Then $\frac{r(x_0)}{e^{x_0}} > \frac{\hat{r}(x_1)}{e^{x_1}} > 1$ and so $\frac{N(r)(x_0)}{e^{x_0}} < \frac{N(\hat{r})(x_1)}{e^{x_1}}$ since the function

$f(t) = t(1 - \ln t)$ is increasing for $t < 1$, decreasing for $t > 1$. (5.20)

Now, since $x_0 \in M(r) = m(1,r)$, $x_1 \in M(\hat{r}) = m(1,\hat{r})$,

$$\min_{x\in[a,b]} \frac{T_1(1,r)(x)}{e^x} < \min_{x\in[a,b]} \frac{T_1(1,\hat{r})(x)}{e^x} .$$

We now claim that $\max_{x\in[a,b]} \frac{T_1(1,r)(x)}{e^x} = \max_{x\in[a,b]} \frac{T_1(1,\hat{r})(x)}{e^x}$. We can see this as follows. Let $x_1 \in m(r)$. Then $r(x_1) = (1+\alpha)e^{x_1}$. Select $x_0 \in [a,b]$ such that $\hat{r}(x_0) = (1+\alpha)e^{x_0}$. Then $\frac{r(x_1)}{e^{x_1}} = \frac{\hat{r}(x_0)}{e^{x_0}}$ and so $\frac{T_1(1,r)(x_1)}{e^{x_1}} = \frac{T_1(1,\hat{r})(x_0)}{e^{x_0}}$. Therefore,

$$\max_{x\in[a,b]} \frac{T_1(1,r)(x)}{e^x} = \frac{T_1(1,r)(x_1)}{e^{x_1}} = \frac{T_1(1,\hat{r})(x_0)}{e^{x_0}} = \max_{x\in[a,b]} \frac{T_1(1,\hat{r})(x)}{e^x}$$

since $m(r) = M(1,r)$. Thus $\hat{c} < c$ and so

$$\left\|\frac{e^x - T_1(c,r)(x)}{e^x}\right\|_\infty = c \max_{x\in[a,b]} \frac{T_1(1,r)(x)}{e^x} - 1 > \hat{c} \max_{x\in[a,b]} \frac{T_1(1,r)(x)}{e^x} - 1$$

$$= \hat{c} \max_{x\in[a,b]} \frac{T_1(1,\hat{r})(x)}{e^x} - 1 = \max_{x\in[a,b]} \frac{T_1(\hat{c},\hat{r})(x)}{e^x} - 1 = \left\|\frac{e^x - T_1(\hat{c},\hat{r})(x)}{e^x}\right\|_\infty .$$

Thus the only possibility for a minimal solution in this case is $T_1(\hat{c},\hat{r})$.

Now $\bar{r}$, the best relative approximation to e^x, has a characterizing sequence $a \le x_1 < x_2 < \dots < x_d \le b$, $d = \bar{m} + m + 2 - \delta$, $\delta = \min(\bar{m} - \partial\bar{p}, m - \partial\bar{q})$ where $\bar{r} = \frac{\bar{p}}{\bar{q}}$, on which

$$\left|\frac{\bar{r}(x_i) - e^{x_i}}{e^{x_i}}\right| = \left\|\frac{\bar{r}(x) - e^x}{e^x}\right\|_\infty = \lambda_0, \ i = 1,\dots,d \ \text{ and}$$

$$\operatorname{sgn} \frac{\bar{r}(x_i) - e^{x_i}}{e^{x_i}} = (-1)^{i+1} \operatorname{sgn} \frac{\bar{r}(x_1) - e^{x_1}}{e^{x_1}}, \ i = 1,\dots, d. \ \text{Set}$$

$\lambda_1 = \frac{1+\alpha}{1-\lambda_0}$. Then $\lambda_1\bar{r}$ turns out to be the best one-sided relative approximation to $(1+\alpha)e^x$ on $[a,b]$ from above as $\{x_i\}_{i=1}^d$ is a characterizing sequence for $\lambda_1\bar{r}$ which satisfies the conditions of a best one-sided approximation. By uniqueness, $\hat{r} \equiv \lambda_1 \bar{r} \equiv \frac{1+\alpha}{1-\lambda_0}\frac{r^*}{\beta_0} \equiv \lambda r^*$ where $\lambda = \frac{1+\alpha}{\beta_0(1-\lambda_0)}$. Note that $\frac{1}{\beta_0(1-\lambda_0)} < \lambda < \frac{e}{\beta_0(1-\lambda_0)}$, $\hat{r}(x) > e^x$ for all $x \in [a,b]$ and r^* intersects e^x.

We shall now make use of Lemma 5.1 to show that $T_1(c_1^{(1)},r^*)$ is a (strictly) better relative approximation to e^x than

$T_1(\hat{c},\hat{r}) \equiv T_1(\hat{c},\lambda r^*)$. Clearly $M(r^*) = M(\lambda r^*)$ and $m(r^*) = m(\lambda r^*)$. Let $x_1 \in m(r^*)$ and $x_2 \in M(r^*)$. Then

$$r^*(x_1) = \beta_0(1-\lambda_0)e^{x_1} \qquad \lambda r^*(x_1) = \lambda\beta_0(1-\lambda_0)e^{x_1}$$

$$r^*(x_2) = \beta_0(1+\lambda_0)e^{x_2} \qquad \lambda r^*(x_2) = \lambda\beta_0(1+\lambda_0)e^{x_2} .$$

With the notation in the Lemma, set $x = x_1$, $y = x_2$, $\varphi = \beta_0(1+\lambda_0)$, $\theta = \beta_0(1-\lambda_0)$, $t = \lambda$. Then

$$\min_{x\in[a,b]} \frac{T_1(1,r^*)(x)}{e^x} = \frac{T_1(1,r^*)(x_2)}{e^{x_2}} = \frac{N_{x_2}(\beta_0(1+\lambda_0)e^{x_2})}{e^{x_2}}$$

$$> \frac{N_{x_2}(\lambda\beta_0(1+\lambda_0)e^{x_2})}{e^{x_2}} \cdot \frac{e^{x_1}}{N_{x_1}(\lambda\beta_0(1-\lambda_0)e^{x_1})} = \frac{T_1(1,\lambda r^*)(x_2)}{e^{x_2}} \cdot \frac{e^{x_1}}{T_1(1,\lambda r^*)(x_1)} \tag{5.21}$$

$$= \min_{x\in[a,b]}\left\{\frac{T_1(1,\lambda r^*)(x)}{e^x}\right\}\left\{\max_{x\in[a,b]} \frac{T_1(1,\lambda r^*)(x)}{e^x}\right\}^{-1}$$

since $m(1,\lambda r^*) = M(\lambda r^*)$ and $M(1,\lambda r^*) = m(\lambda r^*)$ as $\lambda r^*(x) > e^x$ for all $x \in [a,b]$. Hence, by (5.21),

$$\left\|\frac{T_1(\hat{c},\lambda r^*)(x) - e^x}{e^x}\right\|_\infty = 1 - \frac{T_1(\hat{c},\lambda r^*)(x_2)}{e^{x_2}} = 1 - \frac{\hat{c}\, T_1(1,\lambda r^*)(x_2)}{e^{x_2}}$$

$$= 1 - \frac{2\,\dfrac{T_1(1,\lambda r^*)(x_2)}{e^{x_2}}}{\left[\dfrac{T_1(1,\lambda r^*)(x_1)}{e^{x_1}} + \dfrac{T_1(1,\lambda r^*)(x_2)}{e^{x_2}}\right]}$$

$$= 1 - \frac{2}{\left[1 + \dfrac{T_1(1,\lambda r^*)(x_1)}{e^{x_1}} \cdot \dfrac{e^{x_2}}{T_1(1,\lambda r^*)(x_2)}\right]}$$

$$> 1 - \frac{2}{\left[1 + \frac{e^{x_2}}{T_1(1,r^*)(x_2)}\right]} = 1 - \frac{T_1(c_1^{(1)},r^*)(x_2)}{e^{x_2}} = \left\|\frac{T_1(c_1^{(1)},r^*)(x)-e^x}{e^x}\right\|_\infty$$

thus showing that $T_1(c_1^{(1)},r^*)$ is a better relative approximation to e^x than $T_1(\hat{c},\hat{r})$. A similar type of argument applies to the case where $r \in M$ satisfies $r(x) < e^x$ for all $x \in [a,b]$ and we shall omit the details. This case will be considered in the induction step.

Collecting the above results, we find that we have established the following: $T_1(c_1^{(1)},r^*)$ is the unique best relative approximation to e^x from $T_1(S_1)$. To complete the case $n = 1$ we must establish (5.14) and (5.15). For (5.14), let $x_1 \in m(1,r^*)$. Then

$$\left\|\frac{T_1(c_1^{(1)},r^*)(x) - e^x}{e^x}\right\|_\infty = \frac{e^{x_1} - c_1^{(1)}T_1(1,r^*)(x_1)}{e^{x_1}} = 1 - \frac{c_1^{(1)}T_1(1,r^*)(x_1)}{e^{x_1}}$$

$$= 1 - \left\{\frac{2}{\beta_0(1-\lambda_0)[1-\ln\beta_0(1-\lambda_0)] + 1}\right\}\left\{\beta_0(1-\lambda_0)[1 - \ln\beta_0(1-\lambda_0)]\right\}$$

$$= \frac{1 - \beta_0(1-\lambda_0)[1 - \ln\beta_0(1-\lambda_0)]}{1 + \beta_0(1-\lambda_0)[1 - \ln\beta_0(1-\lambda_0)]} = \lambda_1 .$$

Next we must show that $T_1(c_1^{(2)},r^*)$ intersects e^x in $[a,b]$.

From an earlier result, $\max_{x\in[a,b]}\left(\frac{e^x - T_1(c_1^{(1)},r^*)(x)}{e^x}\right) = \max_{x\in[a,b]}\left(\frac{T_1(c_1^{(1)},r^*)(x) - e^x}{e^x}\right) = \lambda_1$. Now $c_1^{(2)} = \beta_1 c_1^{(1)}$, $\beta_1(1-\lambda_1) < 1 < \beta_1(1+\lambda_1)$ and $(1-\lambda_1) \le \frac{T_1(c_1^{(1)},r^*)(x)}{e^x} \le (1+\lambda_1)$ imply that

$\beta_1(1-\lambda_1) \le \dfrac{T_1(c_1^{(2)},r^*)(x)}{e^x} \le \beta_1(1+\lambda_1)$. Since $\dfrac{T_1(c_1^{(2)},r^*)(x)}{e^x}$ is a continuous function of x and assumes the values $\beta_1(1-\lambda_1)$ and $\beta_1(1+\lambda_1)$ it assumes all values between. Hence there exists $x \in [a,b]$ such that $\dfrac{T_1(c_1^{(2)},r^*)(x)}{e^x} = 1$, and so $T_1(c_1^{(2)},r^*)$ intersects e^x in $[a,b]$. Finally, we shall establish the fact that

$$\left\|\frac{N(T_1\ (c_1^{(2)},r^*))(x) - e^x}{e^x}\right\|_\infty \le \left\|\frac{N(T_1(c,r))(x) - e^x}{e^x}\right\|_\infty \quad \text{for all}$$

$(c,r) \in S_1$ with equality if and only if $c = c_1^{(2)}$ and $r \equiv r^*$. Since $(1-\lambda_1)e^y \le T_1(c_1^{(1)},r^*)(y) \le (1+\lambda_1)e^y$ for all $y \in [a,b]$ we have that $\beta_1(1-\lambda_1)e^y \le T_1(c_1^{(2)},r^*)(y) \le \beta_1(1+\lambda_1)e^y$ for all $y \in [a,b]$. Recall that $\dfrac{N_y(\beta_1(1-\lambda_1)e^y)}{e^y} = \dfrac{N_x(\beta_1(1+\lambda_1)e^x)}{e^x}$ for each pair of points $x,y \in [a,b]$. Let $(c,r) \in S_1$ be such that

$$\left\|\frac{N(T_1(c,r))(x) - e^x}{e^x}\right\|_\infty \le \left\|\frac{N(T_1(c_1^{(2)},r^*))(x) - e^x}{e^x}\right\|_\infty .$$

This implies $\beta_1(1-\lambda_1)e^y \le T_1(c,r)(y) \le \beta_1(1+\lambda_1)e^y$. Therefore

$$(1-\lambda_1)e^y \le T_1(\frac{c}{\beta_1},\ r)(y) \le (1+\lambda_1)e^y$$

and so

$$-\lambda_1 \le \frac{T_1(\frac{c}{\beta_1},\ r)(y) - e^y}{e^y} \le \lambda_1 \ .$$

This allows us to conclude that $T_1(\frac{c}{\beta_1},\ r)$ is as good a relative approximation to e^x as $T_1(c_1^{(1)},r^*)$. By the uniqueness of $T_1(c_1^{(1)},r^*)$ we have

$$c_1^{(1)} = \frac{c}{\beta_1},\ r \equiv r^*;\ \text{that is, } c = \beta_1 c_1^{(1)} = c_1^{(2)},\ r \equiv r^*.$$

Before we proceed to the induction step let us establish the following result for which we shall have need.

Claim: $\beta_k < \frac{e}{2}$ for $k = 0,1,2,\ldots$.

Proof: Since $\beta_k = e\left(\frac{1-\lambda_k}{1+\lambda_k}\right)^{\frac{1}{2\lambda_k}} \frac{1}{\sqrt{1-\lambda_k^2}}$ it suffices to show that $\left(\frac{1-\lambda_k}{1+\lambda_k}\right)^{\frac{1}{2\lambda_k}} \frac{1}{\sqrt{1-\lambda_k^2}} < \frac{1}{2}$, or equivalently $\ell n \frac{1-\lambda_k}{1+\lambda_k} < \frac{2\lambda_k}{1-\lambda_k}\left(\ell n \frac{1+\lambda_k}{2}\right)$.

It is clear that $0 < \lambda_k < 1$. Set $f(x) = \frac{2x}{1-x}\ell n \frac{1+x}{2} - \ell n \frac{1-x}{1+x}$ for $0 < x < 1$. Upon simplification, $f(x) = \frac{1+x}{1-x}\ell n \frac{1+x}{2} - \ell n(1-x) + \ell n\, 2$. Now $f(0) = 0$ and $f'(x) = \frac{2}{1-x}\left[1 + \frac{1}{1-x}\ell n \frac{1+x}{2}\right]$. It is easily shown that $1 + \frac{1}{1-x}\ell n \frac{1+x}{2} > 0$ for $0 < x < 1$ and so $f'(x) > 0$ for $0 < x < 1$. Hence the desired result follows.

Now assume the theorem is true for $k = n$. First let us show that $(c_1^{(n+1)},\ldots,c_{n+1}^{(n+1)},r^*) \in S_{n+1}$, or equivalently $y_n = c_n^{(n+1)}N(y_{n-1}) \in K$ since $r^* \in M$ from before. By assumption $(c_1^{(n)},\ldots,c_n^{(n)},r^*) \in S_n$ and so $y_{n-1} \in K$. Thus $N(y_{n-1})$ is well-defined. Now $c_n^{(n+1)} = \beta_n c_n^{(n)}$ and so $y_n = \beta_n c_n^{(n)} N(y_{n-1})$. Clearly $0 < c_n^{(n)} < 2$ and $0 < N(y_{n-1}) \le e^x$ since N is one-sided from below. Thus, by the claim, $\beta_n c_n^{(n)} < \frac{e}{2} \cdot 2 = e$ and so $0 < y_n = \beta_n c_n^{(n)} N(y_{n-1}) \le \beta_n c_n^{(n)} e^x < e^{1+x}$ and so $y_n \in K$. We shall now show that $T_{n+1}(c_1^{(n+1)},\ldots,c_{n+1}^{(n+1)},r^*)$ for $c_1^{(n+1)},\ldots,c_{n+1}^{(n+1)}$ and r^* given by (5.7)-(5.9) is the unique best relative approximation to e^x from $T_{n+1}(S_{n+1})$. As in the case $n = 1$ we shall show that for any $(c_1,\ldots,c_{n+1},r) \in S_{n+1}$, where we allow c_{n+1} to vary,

$T_{n+1}(c_1,\ldots,c_{n+1},r)(x)$ is closest to e^x in the relative norm only when

$$c_{n+1} = \frac{2}{\left[\max_{x\in[a,b]} \frac{T_{n+1}(c_1,\ldots,c_n,1,r)(x)}{e^x} + \min_{x\in[a,b]} \frac{T_{n+1}(c_1,\ldots,c_n,1,r)(x)}{e^x}\right]} \tag{5.22}$$

where $\max_{x\in[a,b]} \frac{T_{n+1}(c_1,\ldots,c_n,1,r)(x)}{e^x} \le 1$ with equality only at those points y where $T_n(c_1,\ldots,c_n,r)(y) = e^y$. This is true since $T_{n+1}(c_1,\ldots,c_n,1,r) = N(T_n(c_1,\ldots,c_n,r))$. For the above choice of c_{n+1}, $x_1 \in M(c_1,\ldots,c_n,1,r)$ and $x_2 \in m(c_1,\ldots,c_n,1,r)$ we have

$$\left\| \frac{e^y - T_{n+1}(c_1,\ldots,c_n,c_{n+1},r)(y)}{e^y} \right\|_\infty = \frac{T_{n+1}(c_1,\ldots,c_n,c_{n+1},r)(x_1) - e^{x_1}}{e^{x_1}}$$

$$= \frac{e^{x_2} - T_{n+1}(c_1,\ldots,c_n,c_{n+1},r)(x_2)}{e^{x_2}} .$$

Now if $\hat{c}_{n+1} > c_{n+1}$ then for $x_1 \in M(c_1,\ldots,c_n,1,r)$

$$\frac{T_{n+1}(c_1,\ldots,c_n,\hat{c}_{n+1},r)(x_1) - e^{x_1}}{e^{x_1}} > \frac{T_{n+1}(c_1,\ldots,c_n,c_{n+1},r)(x_1) - e^{x_1}}{e^{x_1}}$$

$$= \left\| \frac{T_{n+1}(c_1,\ldots,c_n,c_{n+1},r)(x) - e^x}{e^x} \right\|_\infty .$$

Similarly, if $\hat{c}_{n+1} < c_{n+1}$, then for $x_2 \in m(c_1,\ldots,c_n,1,r)$

$$\frac{e^{x_2} - T_{n+1}(c_1,\ldots,c_n,\hat{c}_{n+1},r)(x_2)}{e^{x_2}} > \left\| \frac{T_{n+1}(c_1,\ldots,c_n,c_{n+1},r)(x) - e^x}{e^x} \right\|_\infty$$

and thus the desired result. Next we must show that c_{n+1} of (5.22)

gives the correct value of $c_{n+1}^{(n+1)}$ corresponding to $c_1^{(n+1)},\dots,c_n^{(n+1)}$ and r^*. By the induction assumption $\left\|\frac{T_n(c_1^{(n)},\dots,c_n^{(n)},r^*)(x) - e^x}{e^x}\right\|_\infty = \lambda_n$. Furthermore, the minimum of $\frac{T_{n+1}(c_1^{(n+1)},\dots,c_n^{(n+1)},1,r^*)(x)}{e^x}$ occurs on the set $M(c_1^{(n+1)},\dots,c_n^{(n+1)},r^*) \cup m(c_1^{(n+1)},\dots,c_n^{(n+1)},r^*)$ and is given by $\beta_n(1-\lambda_n)[1 - \ln \beta_n(1-\lambda_n)]$. Hence $c_{n+1}^{(n+1)} = \frac{2}{1+\beta_n(1-\lambda_n)[1-\ln \beta_n(1-\lambda_n)]}$, as desired. Next let $(c_1,\dots,c_{n+1},r) \in S_{n+1}$ where c_{n+1} satisfies (5.22) and suppose $T_n(c_1,\dots,c_n,r)$ intersects e^x on $[a,b]$. Then by the induction assumption we know that

$$\left\|\frac{N(T_n(c_1^{(n+1)},\dots,c_n^{(n+1)},r^*))(x) - e^x}{e^x}\right\|_\infty \le \left\|\frac{N(T_n(c_1,\dots,c_n,r))(x) - e^x}{e^x}\right\|_\infty$$

for all $(c_1,\dots,c_n,r) \in S_n$ with equality only if $c_i^{(n+1)} = c_i$, $i = 1,\dots,n$ and $r \equiv r^*$. Suppose that $(c_1,\dots,c_n,r) \neq (c_1^{(n+1)},\dots,c_n^{(n+1)},r^*)$. Then

$$\min_{x\in[a,b]} \frac{T_{n+1}(c_1,\dots,c_n,1,r)(x)}{e^x} < \min_{x\in[a,b]} \frac{T_{n+1}(c_1^{(n+1)},\dots,c_n^{(n+1)},1,r^*)(x)}{e^x}$$

and so $c_{n+1} > c_{n+1}^{(n+1)}$. Thus, for $x_1 \in M(c_1,\dots,c_{n+1},r)$ and $x_2 \in M(c_1^{(n+1)},\dots,c_{n+1}^{(n+1)},r^*)$

$$\left\|\frac{T_{n+1}(c_1,\dots,c_{n+1},r)(x)-e^x}{e^x}\right\|_\infty = \frac{c_{n+1}N(T_n(c_1,\dots,c_n,r))(x_1) - e^{x_1}}{e^{x_1}}$$

$$= \frac{c_{n+1}N(T_n(c_1^{(n+1)},\dots,c_n^{(n+1)},r^*))(x_2)-e^{x_2}}{e^{x_2}}$$

$$> \frac{c_{n+1}^{(n+1)} N(T_n(c_1^{(n+1)},\ldots,c_n^{(n+1)},r^*))(x_2) - e^{x_2}}{e^{x_2}}$$

$$= \left\| \frac{T_{n+1}(c_1^{(n+1)},\ldots,c_{n+1}^{(n+1)},r^*)(x) - e^x}{e^{x_2}} \right\|_\infty .$$

Finally we consider the case where $(c_1,\ldots,c_{n+1},r) \in S_{n+1}$ for which $T_n(c_1,\ldots,c_n,r)(x)$ does not intersect e^x on $[a,b]$. Let us first consider the case where

$$\min_{x\in[a,b]} \frac{e^x - T_n(c_1,\ldots,c_n,r)(x)}{e^x} = \alpha > 0 ;$$

that is, $e^x > T_n(c_1,\ldots,c_n,r)(x)$ for all $x \in [a,b]$. Set

$$\max_{x\in[a,b]} \frac{e^x - T_n(c_1,\ldots,c_n,r)(x)}{e^x} = \beta .$$

Since $T_n(S_n) \subseteq C^+[a,b]$ we have that $\beta < 1$ and so $\alpha < \beta < 1$.

Set $\lambda = \dfrac{1-\alpha}{\beta_n(1+\lambda_n)}$. First $\max\limits_{x\in[a,b]} \dfrac{T_n(c_1,\ldots,c_n,r)(x)}{e^x} = 1 - \alpha.$

Now $\max\limits_{x\in[a,b]} \dfrac{\lambda T_n(c_1^{(n+1)},\ldots,c_n^{(n+1)},r^*)(x)}{e^x} =$

$\lambda \max\limits_{x\in[a,b]} \dfrac{T_n(c_1^{(n+1)},\ldots,c_n^{(n+1)},r^*)(x)}{e^x} =$

$\lambda \max\limits_{x\in[a,b]} \dfrac{T_n(c_1^{(n)},\ldots,c_{n-1}^{(n)},\beta_n c_n^{(n)},r^*)(x)}{e^x} =$

$\lambda\beta_n \max\limits_{x\in[a,b]} \dfrac{T_n(c_1^{(n)},\ldots,c_n^{(n)},r^*)(x)}{e^x} = \lambda\beta_n(1+\lambda_n) = 1 - \alpha.$ Therefore

$$\max_{x\in[a,b]} \frac{\lambda T_n(c_1^{(n+1)},\ldots,c_n^{(n+1)},r^*)(x)}{e^x} =$$

$$\max_{x\in[a,b]} \frac{T_n(c_1,\ldots,c_n,r)(x)}{e^x} = 1 - \alpha. \tag{5.23}$$

By (5.15),

$$\min_{x\in[a,b]} \frac{T_n(c_1^{(n+1)},\ldots,c_{n-1}^{(n+1)},\lambda c_n^{(n+1)},r^*)(x)}{e^x} >$$

$$\min_{x\in[a,b]} \frac{T_n(c_1,\ldots,c_n,r)(x)}{e^x}, \tag{5.24}$$

for otherwise $T_n(c_1,\ldots,c_{n-1},\frac{c_n}{\lambda},r)(x)$ is a better relative starting approximation to e^x for the Newton operator than is $T_n(c_1^{(n+1)},\ldots,c_n^{(n+1)},r^*)(x)$. Here we have used the fact that

$$m(c_1,\ldots,c_n,r) = m(c_1,\ldots,c_n,1,r),\ M(c_1,\ldots,c_n,r) = M(c_1,\ldots,c_n,1,r) \tag{5.25}$$

which is established using Property I. Choose $x_0,x_1 \in [a,b]$ such that $x_0 \in M(c_1^{(n+1)},\ldots,\lambda c_n^{(n+1)},r^*)$ and $x_1 \in M(c_1,\ldots,c_n,r)$. Then, by (5.23),

$$\frac{T_n(c_1^{(n+1)},\ldots,\lambda c_n^{(n+1)},r^*)(x_0)}{e^{x_0}} = \frac{T_n(c_1,\ldots,c_n,r)(x_1)}{e^{x_1}}$$

and so, by (5.25),

$$\max_{x\in[a,b]} \frac{T_{n+1}(c_1^{(n+1)},\ldots,c_n^{(n+1)},\lambda c_n^{(n+1)},1,r^*)(x)}{e^x} =$$

$$\frac{T_{n+1}(c_1^{(n+1)},\ldots,c_n^{(n+1)},\lambda c_n^{(n+1)},1,r^*)(x_0)}{e^{x_0}} = \frac{T_{n+1}(c_1,\ldots,c_n,1,r)(x_1)}{e^{x_1}} =$$

$$\max_{x\in[a,b]} \frac{T_{n+1}(c_1,\ldots,c_n,1,r)(x)}{e^x}.$$

Choose $y_0, y_1 \in [a,b]$ such that $y_0 \in m(c_1^{(n+1)},\ldots,\lambda c_n^{(n+1)},r^*)$ and $y_1 \in m(c_1,\ldots,c_n,r)$. Then

$$1 > \frac{T_n(c_1^{(n+1)},\ldots,\lambda c_n^{(n+1)},r^*)(y_0)}{e^{y_0}} > \frac{T_n(c_1,\ldots,c_n,r)(y_1)}{e^{y_1}}$$

and so

$$\min_{x\in[a,b]} \frac{T_{n+1}(c_1,\ldots,c_n,1,r)(x)}{e^x} = \frac{T_{n+1}(c_1,\ldots,c_n,1,r)(y_1)}{e^{y_1}}$$

$$< \frac{T_{n+1}(c_1^{(n+1)},\ldots,\lambda c_n^{(n+1)},1,r^*)(y_0)}{e^{y_0}} =$$

$$\min_{x\in[a,b]} \frac{T_{n+1}(c_1^{(n+1)},\ldots,\lambda c_n^{(n+1)},1,r^*)(x)}{e^x}$$

where we have made use of (5.13) and (5.25). Hence $\hat{c}_{n+1} < c_{n+1}$ where $\hat{c}_{n+1}$ corresponds to $(c_1^{(n+1)},\ldots,c_{n-1}^{(n+1)},\lambda c_n^{(n+1)},r^*)$ and c_{n+1} to $(c_1,\ldots,c_n,r)$. Thus

$$\left\| \frac{e^x - T_{n+1}(c_1^{(n+1)},\ldots,c_{n-1}^{(n+1)},\lambda c_n^{(n+1)},\hat{c}_{n+1},r^*)(x)}{e^x} \right\|_\infty$$

$$= \hat{c}_{n+1} \max_{x\in[a,b]} \frac{T_{n+1}(c_1^{(n+1)},\ldots,c_{n-1}^{(n+1)},\lambda c_n^{(n+1)},1,r^*)(x)}{e^x} - 1$$

$$= \hat{c}_{n+1} \max_{x\in[a,b]} \frac{T_{n+1}(c_1,\ldots,c_n,1,r)(x)}{e^x} - 1$$

$$< c_{n+1} \max_{x\in[a,b]} \frac{T_{n+1}(c_1,\ldots,c_n,1,r)(x)}{e^x} - 1$$

$$= \left\| \frac{e^x - T_{n+1}(c_1,\dots,c_{n+1},r)(x)}{e^x} \right\|_\infty .$$

Thus, we must show that $T_{n+1}(c_1^{(n+1)},\dots,c_{n+1}^{(n+1)},r^*)(x)$ is a (strictly) better relative approximation to e^x than $T_{n+1}(c_1^{(n+1)},\dots,c_{n-1}^{(n+1)}, \lambda c_n^{(n+1)},\hat{c}_{n+1},r^*)(x)$. Here we shall make use of the second part of Lemma 5.1. Set $x = x_1$, $y = y_1$, $\theta = \beta_n(1-\lambda_n)$, $\varphi = \beta_n(1+\lambda_n)$ and $\lambda = t$ where $\lambda = \dfrac{1-\alpha}{\beta_n(1+\lambda_n)}$. Observe that $0 < \lambda < \dfrac{1}{\varphi}$. In order to apply the lemma let us make the following observations. For $x_1 \in M(c_1^{(n+1)},\dots,c_n^{(n+1)},r^*)$ and $x_2 \in m(c_1^{(n+1)},\dots,c_n^{(n+1)},r^*)$,

$$\frac{T_n(c_1^{(n+1)},\dots,c_n^{(n+1)},r^*)(x_1)}{e^{x_1}} = \frac{1-\alpha}{\lambda} = \beta_n(1+\lambda_n)$$

and so $T_n(c_1^{(n+1)},\dots,c_n^{(n+1)},r^*)(x_1) = \beta_n(1+\lambda_n)e^{x_1}$. Also

$$\left\| \frac{T_n(c_1^{(n)},\dots,c_n^{(n)},r^*)(x) - e^x}{e^x} \right\|_\infty = \lambda_n \quad \text{implies}$$

$$\left\| \frac{T_n(c_1^{(n+1)},\dots,c_{n-1}^{(n+1)}, \frac{c_n^{(n+1)}}{\beta_n},r^*)(x) - e^x}{e^x} \right\|_\infty = \lambda_n \quad \text{which in turn yields}$$

$T_n(c_1^{(n+1)},\dots,c_n^{(n+1)},r^*)(x_2) = \beta_n(1-\lambda_n)e^{x_2}$. Applying the lemma,

$$\frac{N(\beta_n(1-\lambda_n)e^{x_2})}{e^{x_2}} > \frac{N(\lambda\beta_n(1-\lambda_n)e^{x_2})}{e^{x_2}} \cdot \frac{e^{x_1}}{N(\lambda\beta_n(1+\lambda_n)e^{x_1})}$$

or equivalently,

$$\frac{N(T_n(c_1^{(n+1)},\ldots,c_n^{(n+1)},r^*))(x_2)}{e^{x_2}} > \frac{N(\lambda T_n(c_1^{(n+1)},\ldots,c_n^{(n+1)},r^*))(x_2)}{e^{x_2}} \cdot \frac{e^{x_1}}{N(\lambda T_n(c_1^{(n+1)},\ldots,c_n^{(n+1)},r^*))(x_1)} .$$

Now

$$\left\|\frac{T_{n+1}(c_1^{(n+1)},\ldots,c_{n-1}^{(n+1)},\lambda c_n^{(n+1)},\hat{c}_{n+1},r^*)(x) - e^x}{e^x}\right\|_\infty$$

$$= 1 - \frac{\hat{c}_{n+1}T_{n+1}(c_1^{(n+1)},\ldots,c_{n-1}^{(n+1)},\lambda c_n^{(n+1)},1,r^*)(x_2)}{e^{x_2}}$$

$$= 1 - \frac{2}{\left[\frac{N(\lambda T_n(c_1^{(n+1)},\ldots,c_n^{(n+1)},r^*))(x_1)}{e^{x_1}} \cdot \frac{e^{x_2}}{N(\lambda T_n(c_1^{(n+1)},\ldots,c_n^{(n+1)},r^*))(x_2)} + 1\right]}$$

$$> 1 - \frac{2}{\left[1 + \frac{e^{x_2}}{N(T_n(c_1^{(n+1)},\ldots,c_n^{(n+1)},r^*))(x_2)}\right]}$$

$$= 1 - \frac{T_{n+1}(c_1^{(n+1)},\ldots,c_{n+1}^{(n+1)},r^*)(x_2)}{e^{x_2}}$$

$$= \left\|\frac{T_{n+1}(c_1^{(n+1)},\ldots,c_{n+1}^{(n+1)},r^*)(x) - e^x}{e^x}\right\|_\infty ,$$

showing that $T_{n+1}(c_1^{(n+1)},\ldots,c_{n+1}^{(n+1)},r^*)(x)$ is a (strictly) better relative approximation to e^x than $T_{n+1}(c_1^{(n+1)},\ldots,c_{n-1}^{(n+1)},\lambda c_n^{(n+1)},\hat{c}_{n+1},r^*)(x)$. Similarly, the case where $T_n(c_1,\ldots,c_n,r)(x) > e^x$ leads to the same result.

Gathering the above results, we have that $T_{n+1}(c_1^{(n+1)},\dots,c_{n+1}^{(n+1)},r^*)$ is the unique best relative approximation to e^x on $[a,b]$ from $T_{n+1}(S_{n+1})$. To complete the induction we must verify the following:

$$\left\| \frac{T_{n+1}(c_1^{(n+1)},\dots,c_{n+1}^{(n+1)},r^*)(x) - e^x}{e^x} \right\|_\infty = \lambda_{n+1}\,, \tag{5.26}$$

$$T_{n+1}(c_1^{(n+2)},\dots,c_{n+1}^{(n+2)},r^*)(x) \quad \text{intersects} \quad e^x \quad \text{on} \quad [a,b] \tag{5.27}$$

and

$$\left\| \frac{N(T_{n+1}(c_1^{(n+2)},\dots,c_{n+1}^{(n+2)},r^*))(x) - e^x}{e^x} \right\|_\infty \le \left\| \frac{N(T_{n+1}(c_1,\dots,c_{n+1},r))(x) - e^x}{e^x} \right\|_\infty \tag{5.28}$$

for each $(c_1,\dots,c_{n+1},r) \in S_{n+1}$ with equality if and only if $c_i = c_i^{(n+2)}$, $i = 1,\dots,n+1$ and $r \equiv r^*$.

To establish (5.26) observe that for $x_1 \in m(c_1^{(n+1)},\dots,c_n^{(n+1)},r^*)$

$$\left\| \frac{T_{n+1}(c_1^{(n+1)},\dots,c_{n+1}^{(n+1)},r^*)(x) - e^x}{e^x} \right\|_\infty =$$

$$= \frac{e^{x_1} - c_{n+1}^{(n+1)} T_{n+1}(c_1^{(n+1)},\dots,c_n^{(n+1)},1,r^*)(x_1)}{e^{x_1}}$$

$$= 1 - \left\{ \frac{2}{\beta_n(1-\lambda_n)[1-\ell n\ \beta_n(1-\lambda_n)]+1} \right\} \left\{ \beta_n(1-\lambda_n)[1-\ell n\ \beta_n(1-\lambda_n)] \right\}$$

$$= \frac{1 - \beta_n(1-\lambda_n)[1 - \ln \beta_n(1-\lambda_n)]}{1 + \beta_n(1-\lambda_n)[1 - \ln \beta_n(1-\lambda_n)]} = \lambda_{n+1} .$$

Similarly, $\left\| \frac{T_{n+1}(c_1^{(n+1)},\ldots,c_{n+1}^{(n+1)},r^*)(x) - e^x}{e^x} \right\|_\infty = \lambda_{n+1}$ for $x \in M(c_1^{(n+1)},\ldots,c_n^{(n+1)},r^*)$. By the induction assumption $T_n(c_1^{(n+1)},\ldots,c_n^{(n+1)},r^*)(x)$ intersects e^x on $[a,b]$. Now

$$(1-\lambda_{n+1}) \le \frac{T_{n+1}(c_1^{(n+1)},\ldots,c_{n+1}^{(n+1)},r^*)(x)}{e^x} \le (1+\lambda_{n+1}) \quad \text{and so}$$

$$\beta_{n+1}(1-\lambda_{n+1}) \le \frac{T_{n+1}(c_1^{(n+2)},\ldots,c_{n+1}^{(n+2)},r^*)(x)}{e^x} \le \beta_{n+1}(1+\lambda_{n+1}).$$

Since $\frac{T_{n+1}(c_1^{(n+2)},\ldots,c_{n+1}^{(n+2)},r^*)(x)}{e^x}$ is a continuous function of x it assumes all values between. Thus there exists $x_0 \in [a,b]$ such that $T_{n+1}(c_1^{(n+1)},\ldots,c_{n+1}^{(n+2)},r^*)(x_0) = e^{x_0}$ and so (5.27) is established. Finally (5.28) follows as in the case $n = 1$. Indeed, by (5.26)

$$(1-\lambda_{n+1})e^y \le T_{n+1}(c_1^{(n+1)},\ldots,c_{n+1}^{(n+1)},r^*)(y) \le (1+\lambda_{n+1})e^y$$

for all $y \in [a,b]$. Consequently

$$\beta_{n+1}(1-\lambda_{n+1})e^y \le T_{n+1}(c_1^{(n+2)},\ldots,c_{n+1}^{(n+2)},r^*)(y) \le \beta_{n+1}(1+\lambda_{n+1})e^y.$$

Recall that

$$\frac{N_y(\beta_{n+1}(1-\lambda_{n+1})e^y)}{e^y} = \frac{N_x(\beta_{n+1}(1+\lambda_{n+1})e^x)}{e^x}$$

for each pair of points $x,y \in [a,b]$. Let $(c_1,\ldots,c_{n+1},r) \in S_{n+1}$

satisfy

$$\left\| \frac{N(T_{n+1}(c_1,\ldots,c_{n+1},r))(x) - e^x}{e^x} \right\|_\infty \leq$$

$$\left\| \frac{N(T_{n+1}(c_1^{(n+2)},\ldots,c_{n+1}^{(n+2)},r^*))(x) - e^x}{e^x} \right\|_\infty .$$

This can only happen if

$$\beta_{n+1}(1-\lambda_{n+1})e^y \leq T_{n+1}(c_1,\ldots,c_{n+1},r)(y) \leq \beta_{n+1}(1+\lambda_{n+1})e^y .$$

Thus

$$-\lambda_{n+1} \leq \frac{T_{n+1}(c_1,\ldots,c_n, \frac{c_{n+1}}{\beta_{n+1}},r)(y) - e^y}{e^y} \leq \lambda_{n+1}$$

and so $T_{n+1}(c_1,\ldots,c_n, \frac{c_{n+1}}{\beta_{n+1}},r)$ is as good a relative approximation to e^x as $T_{n+1}(c_1^{(n+1)},\ldots,c_{n+1}^{(n+1)},r^*)$. By uniqueness of $T_{n+1}(c_1^{(n+1)},\ldots,c_{n+1}^{(n+1)},r^*)$ we have that $c_i = c_i^{(n+2)}$, $i = 1,2,\ldots,n+1$ and $r \equiv r^*$. This completes the proof of the theorem.

Section 3: Execution of the Scheme

In this section we shall show how this iterative scheme would be programmed. Suppose we wish to calculate e^x for $x \in [a,b]$ within a tolerance of $\epsilon > 0$. First we calculate λ_0 and $\bar{r}$ (where m and $\bar{m}$ are specified) with:

$$\lambda_0 = \left\| \frac{\bar{r}(x) - e^x}{e^x} \right\|_\infty = \inf_{r \in R_m^{\bar{m}}} \left\| \frac{r(x) - e^x}{e^x} \right\|_\infty$$

and then calculate the sequences $\{\beta_k\}_{k=0}^{\infty}$ and $\{\lambda_k\}_{k=0}^{\infty}$, comparing the value of λ_k with ε at each step:

$$\beta_k = e\left(\frac{1-\lambda_k}{1+\lambda_k}\right)^{\frac{1}{2\lambda_k}} \frac{1}{\sqrt{1-\lambda_k^2}} \ , \quad k = 0,1,2,\ldots$$

$$\lambda_k = \frac{1 - \beta_{k-1}(1-\lambda_{k-1})[1 - \ln \beta_{k-1}(1-\lambda_{k-1})]}{1 + \beta_{k-1}(1-\lambda_{k-1})[1 + \ln \beta_{k-1}(1-\lambda_{k-1})]} \ .$$

Let n be the first integer for which $\varepsilon > \lambda_n$. For this n calculate $c_k^{(n)}$, $k = 1,2,\ldots,n$ using

$$d_k = \frac{2}{\beta_{k-1}(1-\lambda_{k-1})[1 - \ln \beta_{k-1}(1-\lambda_{k-1})] + 1} \ , \quad k = 1,\ldots,n$$

$$c_k^{(n)} = \beta_k d_k \ , \quad k = 1,2,\ldots,n-1$$

$$c_n^{(n)} = d_n$$

and $$r^*(x) = \beta_0 \bar{r}(x) \ .$$

We then store the values $c_k^{(n)}$, $k = 1,\ldots,n$. The algorithm is:

$$y_0(x) = r^*(x)$$

$$y_k(x) = c_k^{(n)} y_{k-1}(x)(1 \pm x - \ln y_{k-1}(x)), \quad k = 1,2,\ldots,n \ .$$

<u>Remark 1</u>: We would like to point out that Theorem 5.2 is of theoretical interest since it is one of the few cases where we can find a best approximation from an algebraic combination of approximants.

BIBLIOGRAPHY

BIBLIOGRAPHY

1. E.W. Cheney, Introduction to Approximation Theory, McGraw-Hill, Inc., New York, 1966.

2. F. Deutsch, Uniform Approximation with Interpolatory Constraints, J. Math. Anal. Appl. 24 (1968), pp. 62-79.

3. J.S. Frame, A Variation of Newton's Method, American Mathematical Monthly, Vol. 51 (1944), pp. 36-38.

4. W. Kammerer, Optimal Approximations of Functions; One-sided Approximation and Extrema Preserving Approximations, Doctoral Thesis, University of Wisconsin, 1959.

5. S. Karlin and W.J. Studden, Tchebyscheff Systems: With Applications in Analysis and Statistics, Interscience Publishers, New York, 1966.

6. R.F. King, Improved Newton Iteration for Integral Roots, Math. Comp. 25 (1971), pp. 299-304.

7. R.F. King and D.L. Phillips, The Logarithmic Error and Newton's Method for the Square Root, Comm. ACM 12 (1969), pp. 87-88.

8. H.L. Loeb, Un Contre-exemple a un Resultat de M. Claude Gilormini, C.R. Acad. Sci. Paris, Ser. A 266 (1968), pp. A237-A238.

9. H.L. Loeb, D.G. Moursund, L.L. Schumaker, G.D. Taylor, Uniform Generalized Weight Function Polynomial Approximation With Interpolation, Siam J. Numer. Anal. 6 (1969), pp. 284-293.

10. H.L. Loeb, D.G. Moursund and G.D. Taylor, Uniform Rational Weighted Approximations Having Restricted Ranges, J. Approx. Theory 1 (1968), pp. 401-411.

11. G. Meinardus and G.D. Taylor, Optimal Starting Approximations for Iterative Schemes (to appear).

12. G. Merz, Padésche Näherungsbrüche und Iterationsverfahren höherer Ordnung, Doctoral Thesis, 1967.

13. D.G. Moursund, Chebyshev Approximation Using a Generalized Weight Function, Siam J. Numer. Anal. 3 (1966), pp. 435-450.

14. D.G. Moursund, Computational Aspects of Chebyshev Approximation Using a Generalized Weight Function, Siam J. Numer. Anal. 5 (1968), pp. 126-137.

15. ________, Optimal Starting Values for the Newton-Raphson Calculation of $\sqrt{x}$, Comm. ACM 10 (1967), pp. 430-432.

16. D.G. Moursund and G.D. Taylor, Optimal Starting Values for the Newton-Raphson Calculation of Inverses of Certain Functions, Siam J. Numer. Anal. 5 (1968), pp. 138-150.

17. ________, Uniform Rational Approximation Using a Generalized Weight Function, Siam J. Numer. Anal. 5 (1968), pp. 882-889.

18. I. Ninomiya, Best Rational Starting Approximations and Improved Newton Iteration for Square Root, Math. Comp. 24 (1970), pp. 391-404.

19. A. Perrie, Uniform Rational Approximation with Osculatory Interpolation, Doctoral Thesis, University of Oregon, 1969.

20. J.R. Rice, The Approximation of Functions, Vol. I, Addison-Wesley Co., Reading, Mass., 1964.

21. T.J. Rivlin, An Introduction to the Approximation of Functions, Blaisdell Publishing Co., Waltham, Mass., 1969.

22. H.E. Salzer, Note on Osculatory Rational Interpolation, Math. Comp. 16 (1962), pp. 486-491.

23. L. Schumaker and G.D. Taylor, On Approximation by Polynomials Having Restricted Ranges, MRC Tech. Summary Report 835, Madison, Wisconsin, January 1968.

24. P.H. Sterbenz and C.T. Fike, Optimal Starting Approximations for Newton's Method, Math. Comp. 23 (1969), pp. 313-318.

25. G.D. Taylor, Approximation by Functions Having Restricted Ranges III, J. Math. Anal. Appl. 27 (1969), pp. 241-248.

26. ________, On Approximation by Polynomials Having Restricted Ranges, Siam J. Numer. Anal. 5 (1968), pp. 258-268.

27. ________, An Improved Newton Iteration for Calculating Roots Which is Optimal (to appear).

28. ________, Optimal Starting Approximations for Newton's Method, J. Approx. Theory 3 (1970), pp. 156-163.

29. K. Taylor, Contributions to the Theory of Restricted Polynomial and Rational Approximation, Doctoral Thesis, Michigan State University, 1970.

30. J.F. Traub, Iterative Methods for the Solution of Equations, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1964.

31. ________, On a Class of Iteration Formulas and Some Historical Notes, Comm. ACM 4 (1961), pp. 276-278.