TECHNICAL REPORT

GENERAL RESEARCH IN FLIGHT SCIENCES

JANUARY 1959 - JANUARY 1960

VOLUME IV
MATHEMATICS AND STATISTICS

LMSD - 288139 JANUARY 1960

WORK CARRIED OUT UNDER
THE LOCKHEED GENERAL RESEARCH PROGRAM

Lockheed

MISSILES and SPACE DIVISION
LOCKHEED AIRCRAFT CORPORATION • SUNNYVALE, CALIF.

TECHNICAL REPORT

GENERAL RESEARCH IN FLIGHT SCIENCES

JANUARY 1959 - JANUARY 1960

VOLUME IV
MATHEMATICS AND STATISTICS

LMSD - 288139 JANUARY 1960

WORK CARRIED OUT UNDER
THE LOCKHEED GENERAL RESEARCH PROGRAM

Lockheed

MISSILES and SPACE DIVISION

LOCKHEED AIRCRAFT CORPORATION • SUNNYVALE, CALIF.

FOREWORD

The Lockheed Missiles and Space Division sponsors a comprehensive program of general research in connection with its defense contracts. A portion of the total research effort at LMSD is carried out in Flight Sciences in the fields of fluid mechanics, mechanics of deformable bodies, flight dynamics, space mechanics, and mathematics and statistics.

The results obtained from theoretical and experimental studies in Flight Sciences are felt to be of general interest and have been assembled into a collection of research reports. These reports comprise a series of papers which deal with various topics in the fields listed above.

This volume is concerned with mathematics and statistics.

CONTENTS

INTERVAL ANALYSIS 1

1

R.E. Moore
C.T. Yang

FOREWORD

This article is a reprint of the Technical Document LMSD-285875, bearing the same title and dated September 1959. It discusses work carried out under the Lockheed General Research Program.

CONTENTS

INTERVAL ANALYSIS I

INTRODUCTION

Digital computations by computers consist of finite sequences of psuedo-arithmetic operations. On the other hand, the exact numerical solution of a mathematical problem, if computable at all, often requires an infinite sequence of exact arithmetic operations.

The study of approximation by digital computations is the underlying motivation for the present study. A digital computation and the analysis of its error as an approximation are usually carried out separately. However, in the present study an interval arithmetic is devised which forms a basis for a concomitant analysis of error in a digital computation. In this system, computations are performed with intervals and intervals are so produced to contain, by construction, the exact numerical solutions sought Hence an approximation and its possible error will be obtained at the same time, choosing say the midpoint of an interval as the approximation.

This report is the first part of our study, in which we first examine some properties of exact or ideal interval arithmetic. After a preliminary discussion of the space of intervals (§ 1) we study addition, multiplication, subtraction and division of intervals (§§ 2-5). Then we construct arithmetic functions as compositions of these elementary operations (§ 6). As one may expect, arithmetic functions play exactly the same role in interval analysis as rational functions in real analysis, so that there is a relation between arithmetic functions and rational functions (§ 7).

In the present report we apply interval analysis to the study of the following problems. Let $f(x)$ be a continuous real-valued function defined on an interval $[a,b]$. (1) What is the image interval $f([a,b])$? (2) What is the definite integral $\int_a^b f(x)\,dx$? When $f(x)$ is a rational function, the approximations are given in theorems 1 and 2 (§§8,9). In general, if $f(x)$ is an arbitrary continuous function, we may still have approximation theorems 3 and 4 (§ 10), although they are not as precise as the first two theorems.

In forthcoming reports we shall apply interval analysis to differential equations and report on results of machine computations using a digital version of interval arithmetic modified to enable the computations to be carried out with pseudo-arithmetic operations. (See also, LMSD-48421, "Automatic Error Analysis in Digital Computation," by R. E. Moore.)

1. PRELIMINARIES

Throughout the whole study R denotes the real line. Whenever a and b are real numbers with $a \leqslant b$, $[a,b]$ denotes the subset of R consisting of all the real numbers x with $a \leqslant x \leqslant b$. In symbols,

$$[a,b] = \left\{x \in R \mid a \leqslant x \leqslant b\right\}.$$

Let $\mathcal{A}$ be the set of all such sets $[a,b]$.

That means,

$$\mathcal{A} = \left\{[a,b] \mid a,b \in R \text{ and } a \leqslant b\right\}.$$

Then we have natural functions

$$p : R \longrightarrow \mathcal{I} ,$$
$$\alpha : \mathcal{I} \longrightarrow R ,$$
$$\beta : \mathcal{I} \longrightarrow R ,$$
$$\gamma : \mathcal{I} \longrightarrow R ,$$
$$\sigma : \mathcal{I} \longrightarrow R ,$$

defined by

$$p(x) = [x,x] ,$$
$$\alpha([a,b]) = a ,$$
$$\beta([a,b]) = b ,$$
$$\gamma([a,b]) = \max\{|a| , |b|\} ,$$
$$\sigma([a,b]) = b - a$$

respectively, where $x \in R$ and $[a,b] \in \mathcal{I}$.

Whenever $x,x' \in R$, we let

$$\rho(x,x') = |x - x'| .$$

Then ρ is a metric on R , that means, ρ has the following properties.

(1) Whenever $x,x' \in R$, $\rho(x,x') = 0$ if and only if $x = x'$.

(2) Whenever $x,x' \in R$, $\rho(x,x') = \rho(x',x)$.

(3) Whenever $x,x',x'' \in R$, $\rho(x,x') + \rho(x',x'') \geq \rho(x,x'')$.

Whenever $A,A' \in \mathcal{A}$ we let

$$P(A,A') = \max \left\{ \rho\Big(\alpha(A)\ ,\ \alpha(A')\Big)\ ,\ \rho\Big(\beta(A)\ ,\ \beta(A')\Big) \right\} .$$

Then P has the same properties as those ρ has on R so that P is a metric on $\mathcal{A}$.

As direct consequences of the definitions of ρ, P, α, β, γ, σ we have

(1-1) The function $p : R \to \mathcal{A}$ is isometric; that means, for any $x,x' \in R$,

$$P\Big(p(x)\ ,\ p(x')\Big) = \rho(x,x') .$$

Hence p maps R homeomorphically onto $p(R)$.

(1-2) Whenever $A,A' \in \mathcal{A}$,

$$\rho\Big(\alpha(A)\ ,\ \alpha(A')\Big) < P(A,A') .$$

Hence the function $\alpha : \mathcal{A} \to R$ is uniformly continuous, that means, for any $\varepsilon > 0$ there is a $\delta > 0$ such that whenever $A,A \in \mathcal{A}$ with $P(A,A) < \delta$, we have $\rho\Big(\alpha(A)\ ,\ \alpha(A)\Big) < \varepsilon$.

Since uniformly continuous functions are continuous, it follows that:

(1-3) The function $\alpha : \mathcal{A} \to R$ is continuous, that means, for any $A \in \mathcal{A}$ and any $\varepsilon > 0$ there is a $\delta > 0$ such that whenever $A' \in \mathcal{A}$ with $P(A,A') < \delta$ we have $\rho\Big(\alpha(A)\ ,\ \alpha(A')\Big) < \varepsilon$.

Just as (1-2) and (1-3), we have

(1-4) Whenever $A,A' \in \mathcal{A}$,

$$\rho\Big(\beta(A)\ ,\ \beta(A')\Big) \leqslant P(A,A') .$$

Hence the function $\beta:\mathcal{A} \longrightarrow R$ is uniformly continuous and consequently it is continuous.

Using (1-2), (1-4) and well-known properties of real numbers, we have

(1-5) The functions $\gamma : \mathcal{A} \longrightarrow R$ and $\sigma : \mathcal{A} \longrightarrow R$ are uniformly continuous and hence they are continuous.

The following will be needed later

(1-6) Let $A,A' \epsilon \mathcal{A}$ and let $\alpha > 0$. Then $P(A,A') < \alpha$ if and only if the following two conditions hold.

(i) For every $x \epsilon A$ there is some $x' \epsilon A'$ with $\rho(x,x') < \alpha$.

(ii) For every $y' \epsilon A'$ there is some $y \epsilon A$ with $\rho(y',y) < \alpha$.

Whenever $I \epsilon \mathcal{A}$ we let

$$\mathcal{A}_I = \{A \epsilon \mathcal{A} \mid A \subset I\} \quad .$$

(1-7) Whenever $I \epsilon \mathcal{A}$, $\mathcal{A}_I$ is compact; that means, every sequence in $\mathcal{A}_I$ contains a convergent subsequence.

Let $\mathcal{J}$ be the subset of $\mathcal{A}$ consisting of all the elements of $\mathcal{A}$ <u>not</u> containing 0 .

(1-8) The set $\mathcal{J}$ is open in $\mathcal{A}$; that means, for every $A \epsilon \mathcal{J}$ there is a positive number γ_A such that every $A' \epsilon \mathcal{A}$ with $P(A,A') < \gamma_A$ belongs to $\mathcal{J}$. In fact, we may choose $\gamma_A = \min \{|\alpha(A)|, |\beta(A)|\}$.

2. ADDITION

There is a function

$$\oplus : \mathcal{L} \times \mathcal{L} \longrightarrow \mathcal{L}$$

defined by

$$\oplus (A,B) = \{x + y \mid x \in A \text{ and } y \in B\}$$

$$= \left[\alpha(A) + \alpha(B) , \beta(A) + \beta(B)\right] , \quad A,B \in \mathcal{L} .$$

The set $\oplus(A,B)$ is also written $A \oplus B$. The function $\oplus$ is called the addition on $\mathcal{L}$.

(2-1) The function $p : R \to \mathcal{L}$ preserves the addition; that means, for any $x,y \in R$,

$$p(x) \oplus p(y) = p(x + y) .$$

Because of (2-1), the addition $\oplus$ on $\mathcal{L}$ may be regarded as an extension of the addition $+$ on R. This is the reason for calling $\oplus$ the "addition" on $\mathcal{L}$.

(2-2) The addition $\oplus$ is commutative; that means, for any $A,B \in \mathcal{L}$,

$$A \oplus B = B \oplus A .$$

(2-3) The addition $\oplus$ is associative; that means, for any $A,B,C \in \mathcal{L}$,

$$A \oplus (B \oplus C) = (A \oplus B) \oplus C .$$

(2-4) Whenever $(A,B) , (A',B') \in \mathcal{L} \times \mathcal{L}$,

$$P(A \oplus B , A' \oplus B') \leqslant P(A,A') + P(B,B') .$$

Hence the addition $\oplus$ is uniformly continuous; that means, for any $\varepsilon > 0$ there is a $\delta > 0$ such that whenever (A,B) , $(A',B') \in \mathcal{I}x\mathcal{I}$ with $P(A,A') < \delta$ and $P(B,B') < \delta$, we have $P(A \oplus B , A' \oplus B') < \varepsilon$.

Proof. Let

$$A = [a,b] , A' = [a',b'] , B = [c,d] , B' = [c',d'] .$$

Then

$$P(A \oplus B , A' \oplus B') = P\Big([a + c , b + d] , [a' + c' , b' + d']\Big)$$

$$= \max \{ \rho(a + c , a' + c') , \rho(b + d , b' + d') \}$$

$$\leqslant \max \{ \rho(a,a') + \rho(c,c') , \rho(b,b') + \rho(d,d') \}$$

$$\leqslant \max \{ \rho(a,a') , \rho(b,b') \} + \max \{ \rho(c,c') , \rho(d,d') \}$$

$$= P(A,A') + P(B,B') .$$

To prove the uniform continuity we have only to pick $\delta = \varepsilon/2$. q.e.d.

Since uniformly continuous functions are continuous, it follows that

(2-5) The addition $\oplus$ is continuous; that means, for any $(A,B) \in \mathcal{I}x\mathcal{I}$ and any $\varepsilon > 0$ there is a $\delta > 0$ such that whenever $(A',B') \in \mathcal{I}x\mathcal{I}$ with $P(A,A') < \delta$ and $P(B,B') < \delta$, we have $P(A \oplus B , A' \oplus B') < \varepsilon$.

3. MULTIPLICATION

Whenever $A,B \in \mathcal{I}$ we let

$$\otimes(A,B) = \{xy \mid x \in A \text{ and } y \in B\} .$$

We claim that $\otimes(A,B) \in \mathcal{I}$.

Let $A = [a,b]$ and $B = [c,d]$. Then we have the following cases

(1) If $a \geq 0$ and $c \geq 0$, then

$$\otimes(A,B) = [ac,bd] .$$

(2) If $a \geq 0$ and $c < 0 < d$, then

$$\otimes(A,B) = [bc,bd] .$$

(3) If $a > 0$ and $d \leq 0$, then

$$\otimes(A,B) = [bc,ad] .$$

(4) If $a < 0 < b$ and $c \geq 0$, then

$$\otimes(A,B) = [ad,bd] .$$

(5) If $a < 0 < b$ and $c < 0 < d$, then

$$\otimes(A,B) = [\min\{bc,ad\} , \max\{ac,bd\}] .$$

(6) If $a < 0 < b$ and $d \leq 0$, then

$$\otimes(A,B) = [bc,ac] .$$

(7) If $b \leq 0$ and $c \geq 0$, then

$$\otimes(A,B) = [ad,bc] .$$

(8) If $b \leq 0$ and $c < 0 < d$, then

$$\otimes(A,B) = [ad,ac] .$$

(9) If $b \leq 0$ and $d \leq 0$, then

$$\otimes(A,B) = [bd,ac] .$$

Because of this result, we have a function

$$\otimes : \mathcal{A} \times \mathcal{A} \to \mathcal{A}$$

called the multiplication on $\mathcal{A}$. Whenever $A,B \in \mathcal{A}$, the set $\otimes(A,B)$ is also written $A \otimes B$.

(3-1) The function $p : R \longrightarrow \mathcal{A}$ preserves the multiplication.

Because of (3-1) the multiplication $\otimes$ on $\mathcal{A}$ may be regarded as an extension of the multiplication on R.

(3-2) The multiplication $\otimes$ is commutative; that means, for any $A,B \in \mathcal{A}$,

$$A \otimes B = B \otimes A .$$

(3-3) The multiplication $\otimes$ is associative; that means, for any $A,B,C \in \mathcal{A}$,

$$A \otimes (B \otimes C) = (A \otimes B) \otimes C$$

(3-4) For any $A,B,C \in \mathcal{A}$,

$$A \otimes (B \oplus C) \subset (A \otimes B) \oplus (A \otimes C)$$

but both sides may not be equal. Hence the distributive law does not hold.

(3-5) Whenever (A,B), $(A',B') \in \mathcal{A} \times \mathcal{A}$

$$P(A \otimes B , A' \otimes B') \leqslant \gamma(B)\ P(A,A') + \gamma(A')\ P(B,B') .$$

Hence the multiplication is continuous.

Proof. Let $xy \in A \otimes B$, where $x \in A$ and $y \in B$. By (1-6), there is some $x' \in A'$ and some $y' \in B'$ such that $\rho(x,x') \leqslant P(A,A')$ and $\rho(y,y') \leqslant P(B,B')$.

Therefore $x'y' \in A' \otimes B'$ and

$$\rho(xy, x'y') = |xy - x'y'|$$

$$= |y(x - x') + x'(y - y')|$$

$$\leq |y| \cdot |x - x'| + |x'| \cdot |y - y'|$$

$$\leq \gamma(B)\, P(A,A') + \gamma(A')\, P(B,B') .$$

Let $x'y' \in A' \otimes B'$, where $x' \in A'$ and $y' \in B'$. Similarly there is some $xy \in A \otimes B$ such that $x \in A$, $y \in B$ and

$$\rho(x'y', xy) < \gamma(B)\, P(A,A) + \gamma(A')\, P(B,B') .$$

Making use of (1-6) again, these results imply the first part of (3-5).

To prove the continuity of $\otimes$, we let $(A,B) \in \mathcal{A} \times \mathcal{A}$ and let $\varepsilon > 0$. Take

$$\delta = \min\left\{ \varepsilon / \left(\gamma(A) + \gamma(B) + 1 \right), 1 \right\} .$$

Then for any $(A',B') \in \mathcal{A} \times \mathcal{A}$ with $P(A,A') < \delta$ and $P(B,B') < \delta$ we have

$$P(A \otimes B, A' \otimes B') < \gamma(B)\delta + \gamma(A')\delta \leq \gamma(B)\delta + \left(\gamma(A) + \delta\right)\delta$$

$$\leq \left(\gamma(A) + \gamma(B) + 1\right)\delta < \varepsilon . \qquad \text{q.e.d.}$$

It is not hard to see that the multiplication is <u>not</u> uniformly continuous. However, since the restriction of a continuous function on a compact set is uniformly continuous, it follows from (1-7) and (3-5) that

(3-6) Let I and J be fixed elements of $\mathcal{A}$. Then the multiplication

$$\otimes : \mathcal{A}_I \times \mathcal{A}_J \longrightarrow \mathcal{A}$$

is uniformly continuous. In fact, for any $A,A' \in I$ and $B,B' \in J$ we have

$$P(A \otimes B , A' \otimes B') < \gamma(J)\, P(A,A') + \gamma(I)\, P(B,B') .$$

4. SUBTRACTION

By (3-5) we have

(4-1) Let E be a fixed element of $\mathcal{L}$. Then the function $\otimes_E$ of $\mathcal{L}$ into $\mathcal{L}$, defined by

$$\otimes_E(A) = E \otimes A , \qquad A \in \mathcal{L} ,$$

is uniformly continuous.

In particular, if $E = [-1, -1]$, we have

(4-2) The function $\otimes_{[-1,\,-1]} : \mathcal{L} \to \mathcal{L}$, defined by

$$\otimes_{[-1,\,-1]}(A) = [-1, -1] \otimes A , \qquad A \in \mathcal{L} ,$$

is uniformly continuous.

Whenever $A \in \mathcal{L}$, we shall abbreviate $\otimes_{[-1,\,-1]}(A)$ by $-A$. Since $-(-A) = A$, $\otimes_{[-1,\,-1]}$ is a homeomorphism.

Combining the addition $\oplus$ and the function $\otimes_{[-1,\,-1]}$ we define the <u>subtraction</u>

$$\ominus : \mathcal{L} \times \mathcal{L} \to \mathcal{L}$$

by

$$\ominus(A,B) = A \oplus (-B) = \{x - y \mid x \in A \text{ and } y \in B\} , \quad A,B \in \mathcal{L} .$$

The set $\ominus(A,B)$ is also written $A \ominus B$.

From (2-1) and (3-1), it follows

(4-3) The function $p : R \to \mathcal{A}$ preserves the subtraction.

(4-4) Whenever $A,B \in \mathcal{A}$,

$$(-A) \otimes B = A \otimes (-B) = -(A \otimes B) ,$$

$$(-A) \otimes (-B) = A \otimes B .$$

From (2-4) and (4-2) it follows

(4-5) The subtraction $\ominus$ is uniformly continuous and hence is continuous. In fact, for any (A,B) , $(A',B') \in \mathcal{A} \times \mathcal{A}$,

$$P(A \ominus B , A' \ominus B') \leqslant P(A,A') + P(B,B') .$$

5. DIVISION

For every $A = [a,b] \in \mathcal{J}$ (see § 1),

$$A^{-1} = \left\{ x^{-1} \mid x \in A \right\} = \left[b^{-1}, a^{-1} \right]$$

is in $\mathcal{J}$. Hence we have a function $\tau : \mathcal{J} \to \mathcal{J}$ defined by

$$\tau(A) = A^{-1} .$$

(5-1) Whenever $J \in \mathcal{J}$, $\mathcal{A}_J \subset \mathcal{J}$ and for any $A,A' \in \mathcal{A}_J$,

$$P\left(A^{-1} , A^{-1}\right) \leqslant \gamma\left(J^{-1}\right)^2 P(A,A') .$$

Hence τ is uniformly continuous on $\mathcal{A}_J$, $J \in \mathcal{J}$, and consequently it is continuous on $\mathcal{J}$.

Proof. Let $A,A' \in \mathcal{A}_J$. For any $x \in A$ there is, by (1-6), some $x' \in A'$ such that $\rho(x,x') \leqslant P(A,A')$ so that

$$\rho\left(x^{-1},x'^{-1}\right) = |\, x^{-1} - x'^{-1} \,| = |\, x^{-1} \,|\, |\, x'^{-1} \,|\, x - x' \,|$$

$$\leqslant \gamma(J)^2 P(A,A') \ .$$

Similarly, for any $y' \in A'$ there is some $y \in A$ such that $\rho\left(y'^{-1},y^{-1}\right) \leqslant \gamma(J)^2 P(A,A')$. Hence the first part is proved.

To prove the uniform continuity of τ on $\mathcal{A}_J$, $J \in \mathcal{J}$, we let $\varepsilon > 0$ and take

$$\delta = \varepsilon / \gamma\left(J^{-1}\right)^2 \ .$$

It is clear that for any $A,A' \in \mathcal{A}_J$ with $P(A,A') \leqslant \delta$, we have $P\left(A^{-1},A'^{-1}\right) < \gamma\left(J^{-1}\right)^2 \delta = \varepsilon$.

To prove the continuity of τ we let $A \in \mathcal{J}$ and let $\varepsilon > 0$. Take a $J \in \mathcal{J}$ such that

$$\alpha(J) < \alpha(A) \leqslant \beta(A) < \beta(J) \ .$$

Then for any $A' \in \mathcal{A}$ with $P(A,A') < \min\{\alpha(A) - \alpha(J) \ , \ \beta(J) - \beta(A)\}$, $A' \in \mathcal{J}$. Hence the continuity of τ on $\mathcal{A}_J$ implies the continuity of τ at A . q.e.d.

Since $\left(A^{-1}\right)^{-1} = A, A \in \mathcal{J}$, τ is a homeomorphism.

Combining the multiplication $\otimes$ and the function τ we define the <u>division</u>

$$\oplus \ : \ \mathcal{A} \times \mathcal{J} \to \mathcal{A}$$

by

$$\oplus\ (A,B) = A \otimes B^{-1} = \left\{ xy^{-1} \mid x \epsilon A \text{ and } y \epsilon B \right\},\ A \epsilon \mathcal{A} \text{ and } B \epsilon \mathcal{B} .$$

The set $\oplus(A,B)$ is also written A/B .

Since for any real number $x \neq 0$,

$$p(x)^{-1} = p\left(x^{-1}\right) ,$$

it follows from (3-1) that

(5-2) The function p preserves the division $\oplus$.

From (3-6) and (5-1) it follows

(5-3) Whenever $I \epsilon \mathcal{A}$ and $J \epsilon \mathcal{B}$,

$$\oplus : \mathcal{A}_I \times \mathcal{A}_J \longrightarrow \mathcal{A}$$

is uniformly continuous. In fact, for any (A,B) , $(A',B') \epsilon \mathcal{A}_I \times \mathcal{A}_J$ we have

$$P(A/B, A'/B') \leqslant \gamma\left(J^{-1}\right) P(A,A') + \gamma\left(J^{-1}\right)^2 \gamma(I)\, P(B,B') .$$

Hence $\oplus : \mathcal{A} \times \mathcal{B} \longrightarrow \mathcal{A}$ is continuous.

6. ARITHMETIC FUNCTIONS

Before giving the definition of an arithmetic function we remark that every arithmetic function has a <u>domain</u> contained in $\mathcal{A}$, an <u>order</u> which is a non-negative integer, and a finite number of <u>parameters</u> which are elements of $\mathcal{A}$. An arithmetic function of order n with parameters $A_1, A_2, \ldots A_m$ is written $F^{(n)}_{A_1 A_2 \ldots A_m}$ or simply F . It is a function from its domain $\mathfrak{D}(F)$ to $\mathcal{A}$.

An arithmetic function of order n is defined by induction on n. When $n = 0$, the number m of parameters is either 0 or 1. In the former case it is the identity function $F^{(o)} : \mathcal{L} \to \mathcal{L}$ given by

$$F^{(o)}(X) = X \quad , \qquad X \in \mathcal{L} \ .$$

In the latter case it is the constant function $F_A^{(o)} : \mathcal{L} \to \mathcal{L}$ given by

$$F_A^{(o)}(X) = A \quad , \qquad X \in \mathcal{L} \ ,$$

where A is an arbitrary element of $\mathcal{L}$. Notice that every arithmetic function of order 0 has $\mathcal{L}$ as its domain.

Let n be a positive integer and suppose that arithmetic functions of order $< n$ have been defined. Then every arithmetic function $F^{(n)}_{A_1A_2 \cdots A_m}$ of order n is defined as follows. There is an arithmetic function $F^{(s)}_{A_1A_2 \cdots A_\ell}$ of order s, $0 \leqslant s \leqslant n-1$, with parameters $A_1, A_2, \ldots, A_\ell$, $0 \leqslant \ell \leqslant m$, and an arithmetic function $F^{(n-1-s)}_{A_{\ell+1}A_{\ell+2} \cdots A_m}$ of order $n-1-s$ with parameters $A_{\ell+1}, A_{\ell+2}, \ldots, A_m$ such that $F^{(n)}_{A_1A_2 \cdots A_m} = F^{(s)}_{A_1A_2 \cdots A_\ell} \circ F^{(n-1-s)}_{A_{\ell+1}A_{\ell+2} \cdots A_m}$ is given by

$$F^{(n)}_{A_1A_2 \cdots A_m}(X) = F^{(s)}_{A_1A_2 \cdots A_\ell}(X) \circ F^{(n-1-s)}_{A_{\ell+1}A_{\ell+2} \cdots A_m}(X) \ ,$$

where $\circ$ is one of $\oplus$, $\ominus$, $\otimes$, $\circledast$ and X is such an element of $\mathcal{L}$ that the right side is well-defined. It is clear that if $\mathcal{D}_1$ and $\mathcal{D}_2$ are the respective domains of $F^{(s)}_{A_1A_2 \cdots A_\ell}$ and $F^{(n-1-s)}_{A_{\ell+1}A_{\ell+2} \cdots A_m}$, then the domain $\mathcal{D}$ of $F^{(n)}_{A_1A_2 \cdots A_m}$ is given by

$$\mathcal{D} = \begin{cases} \mathcal{D}_1 \cap \mathcal{D}_2 \cap \left(F^{(n-1-s)}_{A_{\ell+1}A_{\ell+2} \cdots A_m}\right)^{-1}(\mathcal{I}) & \text{if } \circ \text{ is } \circledast \ , \\ \mathcal{D}_1 \cap \mathcal{D}_2 & \text{if otherwise.} \end{cases}$$

Let F be an arithmetic function of parameters $A_1, A_2, \ldots, A_m$ and domain $\mathfrak{D}(F)$. We may write

$$F(A_1, A_2, \ldots, A_m ; X)$$

instead of $F(X)$ and consider F as a function of $\mathfrak{I}^m \times \mathfrak{D}(F)$ into $\mathfrak{I}$. Notice that $\mathfrak{D}(F)$ depends on $A_1, A_2, \ldots, A_m$.

(6-1) If $F(B_1, B_2, \ldots, B_m ; Y)$ is defined, then for any $C_1 \epsilon \mathfrak{I}_{B_1}$, $C_2 \epsilon \mathfrak{I}_{B_2}$, .., $C_m \epsilon \mathfrak{I}_{B_m}$, $Z \epsilon \mathfrak{I}_Y$, $F(C_1, C_2, \ldots, C_m ; Z)$ is defined and is contained in $F(B_1, B_2, .., B_m ; Y)$.

Proof. If F is of order 0, our assertion is trivial. Hence our assertion holds for arithmetic functions of order 0.

Now we proceed by induction. Let n be a positive integer and assume our assertion for all arithmetic functions of order $< n$. By definition every arithmetic function F of order n with parameters $A_1, A_2, \ldots, A_m$ is given by

$$F(A_1, A_2, \ldots, A_m ; X) = F_1(A_1, \ldots, A_\ell ; X) \circ F_2(A_{\ell+1}, \ldots, A_m ; X),$$

where F_1 and F_2 are arithmetic functions of order $< n$ and $\circ$ is one of $\oplus, \ominus, \otimes, \oplus$.

If $F(B_1, B_2, \ldots, B_m ; Y)$ is defined, then $F_1(B_1, \ldots, B_\ell ; Y)$ and $F_2(B_{\ell+1}, \ldots, B_m ; Y)$ is in $\mathfrak{J}$ when $\circ$ is $\oplus$. By the induction hypothesis, for any $C_1 \epsilon \mathfrak{I}_{B_1}$, $C_2 \epsilon \mathfrak{I}_{B_2}$, $\ldots$, $C_m \epsilon \mathfrak{I}_{B_m}$, $Z \epsilon \mathfrak{I}_Y$, $F_1(C_1, \ldots, C_\ell ; Z)$ and $F_2(C_{\ell+1}, \ldots, C_m ; Z)$ are defined and

$$F_1(C_1, \ldots, C_\ell ; Z) \subset F_1(B_1, \ldots, B_\ell ; Y),$$

$$F_2(C_{\ell+1}, \ldots, C_m ; Z) \subset F_2(B_{\ell+1}, \ldots, B_m ; Y).$$

So $F_2(C_{\ell+1}, \ldots, C_m ; Z) \epsilon \mathcal{J}$ when o is $\oplus$. Hence

$$F(C_1, C_2, \ldots, C_m ; Z) = F_1(C_1, \ldots, C_\ell ; Z) \circ F_2(C_{\ell+1}, \ldots, C_m ; Z)$$

is defined and is contained in $F(B_1, B_2, \ldots, B_m ; Y)$. q.e.d.

(6-2) If $F(B_1, \ldots, B_m ; Y)$ is defined, then there is a number $k > 0$ such that whenever $A_1, A_1' \epsilon \mathcal{A}_{B_1}, \ldots, A_m, A_m' \epsilon \mathcal{A}_{B_m}, X, X' \epsilon \mathcal{A}_Y$, we have

$$P\Big(F(A_1, \ldots, A_m ; X) , F(A_1', \ldots, A_m' ; X')\Big) \leqslant k \Big(P(A_1, A_1') + \ldots + P(A_m, A_m') + P(X, X')\Big) .$$

Hence F is uniformly continuous on $\mathcal{A}_{B_1} \times \ldots \times \mathcal{A}_{B_m} \times \mathcal{A}_Y$.

Proof. When F is of order o it is evident that the inequality holds with $k = 1$. Therefore we may proceed by induction and assume the inequality for arithmetic functions of order $< n$.

Every arithmetic function F of order n is given by

$$F(A_1, \ldots, A_m ; X) = F_1(A_1, \ldots, A_\ell ; X) \circ F_2(A_{\ell+1}, \ldots, A_m ; X) ,$$

where F_1 and F_2 are arithmetic functions of order $< n$ and o is one of $\oplus$, $\ominus$, $\otimes$, $\oplus$. By the induction hypothesis, there exist positive numbers k_1 and k_2 such that whenever $A_1, A_1' \epsilon \mathcal{A}_{B_1}, \ldots, A_m, A_m' \epsilon \mathcal{A}_{B_m}$, $x, x' \epsilon \mathcal{A}_Y$, we have

$$P\Big(F_1(A_1, \ldots, A_\ell ; X) , F_1(A_1', \ldots, A_\ell' ; X')\Big) \leqslant k_1 \Big(P(A_1, A_1') + \ldots + P(A_\ell, A_\ell') + P(X, X')\Big)$$

$$P\Big(F_2(A_{\ell+1}, \ldots, A_m ; X) , F_2(A_{\ell+1}', \ldots, A_m' ; X')\Big) \leqslant k_2 \Big(P(A_{\ell+1}, A_{\ell+1}) + \ldots + P(A_m, A_m') + P(X, X')\Big) .$$

By (2-4), (3-6), (4-5), (5-3), there exists a positive number k for which our desired inequality holds. In fact, we can let

$$k = \begin{cases} 2(k_1 + k_2)\left(\gamma(J^{-1}) + \gamma(J^{-1})^2\,\gamma(I)\right) & \text{if } o \text{ is } \oplus\,, \\ 2(k_1 + k_2)\left(\gamma(I) + \gamma(J) + 1\right) & \text{if otherwise,} \end{cases}$$

where

$$I = F_1(B_1, \ldots, B_\ell\,; Y)\,, \quad J = F_2(B_{\ell+1}, \ldots, B_m\,; Y)\,.$$

q.e.d.

(6-3) If $F(B_1, \ldots, B_m; Y)$ is defined, then there is a $\delta > 0$ such that whenever $A_1, \ldots, A_m, X \in \mathfrak{A}$ with

$$P(A_1,B_1) < \delta\,, \ \ldots, \ P(A_m,B_m) < \delta\,, \ P(X,Y) < \delta\,,$$

$F(A_1, \ldots, A_m; X)$ is defined and is continuous.

Proof. If F is of order 0, then our assertion is trivial. Therefore, we may proceed by induction and assume our assertion for arithmetic functions of order $< n$, $n > 0$.

Every arithmetic function F of order n is given by

$$F(A_1, \ldots, A_m; X) = F_1(A_1, \ldots, A_\ell\,; X)\ o\ F_2(A_{\ell+1}, \ldots, A_m; X)\,,$$

where F_1 and F_2 are arithmetic functions of order $< n$ and o is one of $\oplus$, $\ominus$, $\otimes$, $\oplus$.

By the induction hypothesis there is a positive number δ such that whenever $A_1, \ldots, A_m, X \in \mathfrak{A}$ with $P(A_1,B_1) < \delta$, ..., $P(A_m,B_m) < \delta$, $P(X,Y) < \delta$, both $F_1(A_1, \ldots, A_\ell\,; X)$ and $F_2(A_{\ell+1}, \ldots, A_m; X)$ are defined and continuous. If $F_2(B_{\ell+1}, \ldots, B_m; Y) \in \mathfrak{J}$, we may choose δ so small that $F_2(A_{\ell+1}, \ldots, A_m; X) \in \mathfrak{J}$. Hence $F(A_1, \ldots, A_m; X)$ is defined and is continuous.

q.e.d.

It follows from (6-1), (6-2) and (6-3) that

(6-4) Let F be an arithmetic function with m parameters and let

$$U = \{(A_1, \ldots, A_m; X) \in \mathfrak{A}^{m+1} \mid F(A_1, \ldots, A_m; X) \text{ defined}\} .$$

Then U is open in $\mathfrak{A}^{m+1}$ and F is a continuous function of U into $\mathfrak{A}$. Moreover, for every $(B_1, \ldots, B_m; Y) \in U$, $\mathfrak{A}_{B_1} x \ldots x \mathfrak{A}_{B_m} x \mathfrak{A}_Y \subset U$ and F is uniformly continuous on $\mathfrak{A}_{B_1} x \ldots x \mathfrak{A}_{B_m} x \mathfrak{A}_Y$. Furthermore, $(A_1, \ldots, A_m; X) \in \mathfrak{A}_{B_1} x \ldots x \mathfrak{A}_{B_m} x \mathfrak{A}_Y$ implies

$$F(A_1, \ldots, A_m; X) \subset F(B_1, \ldots, B_m; Y) .$$

When we write $F(X)$ in place of $F(A_1, \ldots, A_m; X)$ it is understood that parameters $A_1, \ldots, A_m$ are fixed. Since

$$\mathfrak{D}(F) = \{X \in \mathfrak{A} \mid F(X) \text{ defined}\} ,$$

it follows from (6-4) and (6-2) that

(6-5) For every arithmetic function F, $\mathfrak{D}(F)$ is open in $\mathfrak{A}$ and $F: \mathfrak{D}(F) \to \mathfrak{A}$ is continuous. Let $I \in \mathfrak{D}(F)$. Then $\mathfrak{A}_I \subset \mathfrak{D}(F)$ and $X \in \mathfrak{A}_I$ implies $F(X) \subset F(I)$. Moreover, there is a positive number k such that whenever $X, X' \in \mathfrak{A}_I$,

$$P\left(F(X), F(X')\right) \leqslant k P(X, X') .$$

7. RELATION BETWEEN ARITHMETIC FUNCTIONS AND RATIONAL FUNCTIONS.

In the construction of arithmetic functions, if we replace $\mathfrak{A}$ by R and replace $\oplus$, $\ominus$, $\otimes$, $\oplus$ by corresponding operations on R, then we obtain <u>rational functions</u> in place of arithmetic functions. Therefore, we can establish a relation between arithmetic functions and rational function.

An arithmetic function is called special if all of its parameters belong to $p(R)$.

(7-1) If F is a special arithmetic function of domain $\mathfrak{D}(F)$, then $p^{-1}\left(\mathfrak{D}(F)\right)$ is open in R and for every $x \in p^{-1}\left(\mathfrak{D}(F)\right)$, $Fp(x) \in p(R)$. Moreover, there is a unique rational function f whose domain contains $p^{-1}\left(\mathfrak{D}(F)\right)$ such that $pf = Fp$ or $f = p^{-1}Fp$.

$$\begin{array}{ccc} R & \xrightarrow{\ p\ } & \mathfrak{L} \\ \uparrow f & & \uparrow F \\ p^{-1}\left(\mathfrak{D}(F)\right) & \xrightarrow{\ p\ } & \mathfrak{D}(F) \end{array}$$

As the converse of (7-1), we have

(7-2) Given any rational function f of domain $D(f)$ there is a special arithmetic function F of domain $\mathfrak{D}(F) \supset p\left(D(f)\right)$ such that $f = p^{-1}Fp$.

Remark. It is possible to have two distinct special arithmetic functions F_1 and F_2 such that $p^{-1}F_1p = p^{-1}F_2p$. For example,

$$F_1(X) = (X \otimes X) \oplus X,$$
$$F_2(X) = X \otimes \left(X \oplus p(1)\right), \qquad X \in \mathfrak{L},$$

give two arithmetic functions F_1 and F_2 of domain $\mathfrak{L}$. It it clear that

$$(p^{-1}F_1p)(x) = x^2 + x = x(x+1) = (p^{-1}F_2p)(x).$$

Since

$$F_1\left([-1, 0]\right) = [-1, 1], \quad F_2\left([-1, 0]\right) = [-1, 0],$$

F_1 and F_2 are distinct.

Let F be an arithmetic function with parameters $A_1, \ldots, A_m$. Let G be the arithmetic function with $B_1, \ldots, B_m$ in place of $A_1, \ldots, A_m$ respectively; that means, it is given by

$$G(X) = F(B_1, \ldots, B_m; X) .$$

If $B_1 \subset A_1, \ldots, B_m \subset A_m$, then, by (6-4), $\mathfrak{D}(F) \subset \mathfrak{D}(G)$ and for every $X \epsilon \mathfrak{D}(F)$, $G(X) \subset F(X)$. Hence the relation that $B_1 \subset A_1, \ldots, B_m \subset A_m$ will be written

$$G \subset F .$$

Let f be a rational function and let F be an arithmetic function. If there is a special arithmetic function $G \subset F$ with $f = p^{-1}Gp$, we say that f is an <u>associated rational function</u> of F or that F is an <u>associated arithmetic function</u> of f. In particular, if $G = F$ and then $f = p^{-1}Fp$, we say that F is an associated <u>special</u> arithmetic function of f and f is <u>the</u> associated rational function of F.

8. FIRST APPROXIMATION THEOREM

Let $X \epsilon \mathfrak{L}$. By a <u>subdivision</u> of X we mean

$$\xi = \left\{\xi_1, \xi_2, \ldots, \xi_r\right\}$$

such that

$$\alpha(X) = \alpha(\xi_1) < \beta(\xi_1) = \alpha(\xi_2) < \beta(\xi_2) = \alpha(\xi_3) <$$

$$\ldots < \beta(\xi_{r-1}) = \alpha(\xi_r) < \beta(\xi_r) = \beta(X) .$$

For every subdivision $\xi = \left\{\xi_1, \xi_2, \ldots, \xi_r\right\}$ we let

$$\sigma(\xi) = \max\left\{\sigma(\xi_1), \sigma(\xi_2), \ldots, \sigma(\xi_r)\right\} .$$

Let f be a rational function of domain $D(f)$; let G be an associated special arithmetic function of f and let F be an associated arithmetic function of f with $F \supset G$ (see § 7).

Let I be an element of $\mathfrak{A}$ contained in $D(f)$. Then for every $x \in \mathfrak{A}$, $Gp(x) = pf(x)$ is well-defined. Since, by (6-5), the domain $\mathfrak{D}(G)$ of G is open, there is, for every $x \in I$, a positive number r_x such that whenever $Y \in \mathfrak{A}$ with $P(p(x), Y) < r_x$, $G(Y)$ is well-defined. Let

$$I_x = [x - r_x/2, x + r_x/2] .$$

Then $G(I_x)$ is well-defined.

Since I is compact, there exist a finite number of points of I, say $x_1, x_2, \ldots, x_t$, such that I is contained in the union of the interior

$$Q_i = \left(x_i - r_{x_i}/2, x_i + r_{x_i}/2\right)$$

of I_{x_i}, $i = 1, \ldots, r^t$. We abbreviate I_{x_i} by I_i.

Let $B_1, \ldots, B_m$ be the parameters of G and let $A_1, \ldots, A_m$ be the parameters of F. By definition,

$$B_1, \ldots, B_m \in p(R) ;$$

$$B_1 \subset A_1, \ldots, B_m \subset A_m ;$$

$$G(Y) = F(B_1, \ldots, B_m; Y) , \qquad Y \in \mathfrak{D}(G) ;$$

$$F(Y) = F(A_1, \ldots, A_m; Y) , \qquad Y \in \mathfrak{D}(F) .$$

By (6-3) there is, for every $i = 1, \ldots, t$, a $\delta_i > 0$ such that whenever $\sigma(A_1) < \delta_1, \ldots, \sigma(A_m) < \delta_i$, $F(I_i)$ is defined. Let

$$3\delta = \min\left\{\delta_1, \ldots, \delta_t\right\} .$$

Then whenever $\sigma(A_1) < 3\delta$, ..., $\sigma(A_m) < 3\delta$, $F(I_i)$ is defined for all i .

By a well-known theorem on compact metric spaces there exists a $\delta' > 0$ such that whenever $Y \in \mathcal{J}_I$ with $\sigma(Y) < \delta'$, Y is contained in one of I_1, ..., I_t so that $F(Y)$ is defined by (6-5).

Let $X \in \mathcal{J}_I$ and let $\xi = \{\xi_1, \ldots, \xi_r\}$ be a subdivision of X with $\sigma(\xi) < \delta'$. Then

$$F(X,\xi) = F(\xi_1) \cup \ldots \cup F(\xi_r) ,$$

$$\Sigma(F,X,\xi) = \left(F(\xi_1) \otimes p\sigma(\xi_1)\right) \oplus \ldots \oplus \left(F(\xi_r) \otimes p\sigma(\xi_r)\right)$$

are well-defined when the parameters A_1, ..., A_m satisfy

$$\sigma(A_1) < 3\delta , \ldots, \sigma(A_m) < 3\delta .$$

Clearly $\Sigma(F,X,\xi) \in \mathcal{J}$. Since for every $j = 2, \ldots, r$,

$$F(\xi_{j-1}) \cap F(\xi_j) \subset F\left(p\alpha(\xi_j)\right) \neq \phi ,$$

it follows that $F(X,\xi) \in \mathcal{J}$.

By (6-2), there is, for every $i = 1, \ldots, t$, a positive number k_i such that for any $Y, Y' \in \mathcal{J}_{I_i}$,

$$P\left(G(Y), G(Y')\right) \leqslant k_i P(Y,Y') ,$$

$$P\left(F(Y), G(Y)\right) \leqslant k_i\left(\sigma(A_1) + \ldots + \sigma(A_m)\right)$$

hold where the parameters satisfy

$$\sigma(A_1) < \delta, \ldots, \sigma(A_m) < \delta ,$$

and k_i is <u>independent of</u> the parameters.

Let $k = \max\left\{k_1, \ldots, k_t\right\}$.

Then for any $Y, Y' \in \mathcal{I}_I$ with $\sigma(Y) < \delta$ and $\sigma(Y') < \delta'$

$$P\Big(F(Y), G(Y')\Big) \leqslant k\Big(P(Y,Y') + \sigma(A_1) + \ldots + \sigma(A_m)\Big) \quad .$$

In fact, there is a finite sequence

$$Y_1 = Y, Y_2, \ldots, Y_s = Y'$$

in $\mathcal{I}_I$ such that

$$\sigma(Y_2) = \ldots = \sigma(Y_s) = \sigma(Y') \ ,$$

$$\alpha(Y_2) < \ldots < \alpha(Y_s) = \alpha(Y')$$

and Y_{j-1} and Y_j are contained in the same I_i for some i . Hence

$$\begin{aligned} kP(Y, Y') &= k\Big(P(Y_1,Y_2) + \ldots + P(Y_{s-1},Y_s)\Big) \\ &\geqslant P\Big(G(Y_1), G(Y_2)\Big) + \ldots + P\Big(G(Y_{s-1}), G(Y_s)\Big) \\ &\geqslant P\Big(G(Y), G(Y')\Big) \ . \end{aligned}$$

Moreover

$$k\Big(\sigma(A_{\ }) + \ldots + \sigma(A_m)\Big) \geqslant P\Big(F(Y), G(Y)\Big) \ .$$

Our assertion thus follows.

Now we are ready to prove

Theorem 1. Let f be a rational function of domain $D(f)$, let G be an associated special arithmetic function of f and let I be an element of $\mathcal{I}$ contained in $D(f)$. Then there are positive numbers δ, δ', k such that whenever F is an associated arithmetic function of f with

parameters $A_1, \ldots, A_m$ such that $F \supset G$ and $\sigma(A_1) < \delta, \ldots, \sigma(A_m) < \delta$, X is an element of $\mathcal{J}_I$ and $\xi = \{\xi_1, \ldots, \xi_r\}$ is a subdivision of X with $\sigma(\xi) < \delta'$,

$$F(X,\xi) = F(\xi_1) \cup \ldots \cup F(\xi_r)$$

is defined and satisfies

$$f(X) \subset F(X,\xi) \subset f(X) \oplus [-\kappa, \kappa] ,$$

where

$$f(X) = \{f(x) \mid x \in X\} ,$$

$$\kappa = k\left(\sigma(\xi) + \sigma(A_1) + \ldots + \sigma(A_m)\right) .$$

Proof. Let δ, δ', k be chosen as above. Let F be an associated arithmetic function of f with parameters $A_1, \ldots, A_m$ such that $F \subset G$ and $\sigma(A_1) < \delta, \ldots, \sigma(A_m) < \delta$. Let $X \in \mathcal{J}_I$ and let ξ be a subdivision of X with $\sigma(\xi) < \delta'$. We have shown that $F(X,\xi)$ is defined.

For every $y \in f(X)$ there is an $x \in X$ with $f(x) = y$. Let $x \in \xi_j$. Then, by (6-5),

$$y = f(x) \in F\,p(x) \subset F(\xi_j) \subset F(X,\xi) .$$

Hence $f(X) \subset F(X,\xi)$.

For every $y \in F(X,\xi)$, there is a ξ_j with $y \in F(\xi_j)$. Let $x \in \xi_j$. Then $P\left(\xi_j, p(x)\right) \leqslant \sigma(\xi_j) < \sigma(\xi)$. It follows that

$$P\left(F(\xi_j), Gp(x)\right) \leqslant k\left(\sigma(\xi) + \sigma(A_1) + \ldots + \sigma(A_m)\right)$$

or

$$P\left(F(\xi_j), pf(x)\right) \leqslant \kappa .$$

Therefore

$$y \epsilon F(\xi_j) \subset pf(x) \oplus [-\kappa, \kappa]$$

$$\subset f(X) \oplus [-\kappa, \kappa]$$

Hence $F(X,\xi) \subset f(X) \oplus [-\kappa, \kappa]$. q.e.d.

Corollary 1. Let f be a rational function of domain $D(f)$, let G be an associated special arithmetic function of f and let X be an element of $\mathcal{I}$ contained in $D(f)$. Let F be an associated arithmetic function of f with parameters $A_1, \ldots, A_m$ such that $F \supset G$. Then whenever ξ is a subdivision of X with small $\sigma(\xi)$ and $\sigma(A_1), \ldots, \sigma(A_m)$ are small, $F(X,\xi)$ is defined. Moreover, as $\sigma(\xi) + \sigma(A_1) + \ldots + \sigma(A_m) \longrightarrow 0$,

$$\lim F(X,\xi) = f(X) ,$$

that means,

$$\lim P\Big(F(X,\xi), f(X)\Big) = 0 .$$

9. SECOND APPROXIMATION THEOREM

Theorem 2. Let f be a rational function of domain $D(f)$, let G be an associated special arithmetic function of f and let I be an element of $\mathcal{I}$ contained in $D(f)$. Then there are positive numbers δ, δ', k such that whenever F is an associated arithmetic function of F with parameters $A_1, \ldots, A_m$ such that $F \supset G$ and $\sigma(A_1) < \delta, \ldots, \sigma(A_m) < \delta$, $X = [a,b]$ is an element of $\mathcal{I}_I$ and ξ is a subdivision of X,

$$\Sigma(F,X,\xi) = \Big(F(\xi_1) \otimes p\sigma(\xi_1)\Big) \oplus \ldots \oplus \Big(F(\xi_r) \otimes p\sigma(\xi_r)\Big)$$

is defined and satisfies

$$p\left(\int_a^b f(x)dx\right) \subset \Sigma(F,X,\xi) \subset p\left(\int_a^b f(x)dx\right) + \Big[-\kappa\sigma(X), \kappa\sigma(X)\Big] ,$$

where

$$\kappa = k\Big(\sigma(\xi) + \sigma(A_1) + \ldots + \sigma(A_m)\Big) \ .$$

Proof. Let δ, δ', k be as before. Let F be an associated arithmetic function of F with parameters $A_1, \ldots, A_m$ such that $F \supset G$ and $\sigma(A_1) < \delta, \ldots, \sigma(A_m) < \delta$. Let $X = [a,b]$ be an element of $\mathcal{I}_I$ and let $\xi = \{\xi_1, \ldots, \xi_r\}$ be a subdivision of X with $\sigma(\xi) < \delta'$. We have shown that $\Sigma(F,X,\xi)$ is defined.

Let

$$m_j = \inf_{x \in \xi_j} f(x) \ , \qquad M_j = \sup_{x \in \xi_j} f(x) \ .$$

Then

$$\Sigma_{j=1}^{r} m_j \sigma(\xi_j) \leqslant \int_a^b f(x)dx \leqslant \Sigma_{j=1}^{r} M_j \sigma(\xi_j) \ .$$

Since ξ_j is compact, there is a point x_j of ξ_j with $f(x_j) = m_j$. Then

$$m_j = f(x_j) \in F\,p(x_j) \subset F(\xi_j)$$

so that

$$m_j \sigma(\xi_j) \in F(\xi_j) \otimes p\sigma(\xi_j) \ .$$

Hence

$$\Sigma_{j=1}^{r} m_j \sigma(\xi_j) \in \Sigma(F,X,\xi) \ .$$

Similarly we can show that

$$\Sigma_{j=1}^{r} M_j \sigma(\xi_j) \in \Sigma(F,X,\xi) \ .$$

Since $\Sigma(F,X,\xi) \in \mathcal{A}$, it follows that

$$p\left(\int_a^b f(x)dx\right) \subset \Sigma(F,X,\xi) \quad .$$

Let x_j be as above. Since

$$P\left(\xi_j, p(x_j)\right) \leqslant \sigma(\xi_j) \leqslant \sigma(\xi) \quad ,$$

it follows that

$$P\left(F(\xi_j), Gp(x_j)\right) \leqslant k\left(\sigma(\xi) + \sigma(A_1) + \ldots + \sigma(A_m)\right) = \kappa$$

so that

$$F(\xi_j) \subset p(m_j) \oplus [-\kappa, \kappa]$$

Therefore

$$F(\xi_j) \otimes p\sigma(\xi_j) \subset p\left(m_j\ \sigma(\xi_j)\right) \oplus \left[-\kappa\ \sigma(\xi_j), \kappa\ \sigma(\xi_j)\right] \quad .$$

Hence

$$\Sigma(F,X,\xi) \subset p\left(\Sigma_{j=1}^{r}\ m_j\ \sigma(\xi_j)\right) \oplus \left[-\kappa\ \sigma(X), \kappa\ \sigma(X)\right] \quad .$$

Similarly we can prove that

$$\Sigma(F,X,\xi) \subset p\left(\Sigma_{j=1}^{r}\ m_j\ \sigma(\xi_j)\right) \oplus \left[-\kappa\ \sigma(X), \kappa\ \sigma(X)\right] \quad .$$

Since $\Sigma(F,X,\xi) \in \mathcal{A}$, it follows that

$$\Sigma(F,X,\xi) \subset p\left(\int_a^b f(x)dx\right) \oplus \left[-\kappa\ \sigma(X), \kappa\ \sigma(X)\right] \quad .$$

q.e.d.

Corollary. Let f be a rational function of domain $D(f)$, let G be an associated special arithmetic function of f and let $X = [a,b]$ be an

element of $\mathcal{A}$ contained in $D(f)$. Let F be an associated arithmetic function of f with parameters $A_1, \ldots, A_m$ such that $F \supset G$. Then whenever ξ is a subdivision of X with small $\sigma(\xi)$ and $\sigma(A_1), \ldots, \sigma(A_m)$ are small, $\Sigma(F,X,\xi)$ are defined. Moreover, as $\sigma(\xi) + \sigma(A_1) + \ldots + \sigma(A_m) \longrightarrow 0$,

$$\lim \Sigma(F,X,\xi) = \int_a^b f(x)dx \quad ,$$

that means

$$\lim P\left(\Sigma(F,X,\xi),\ p\left(\int_a^b f(x)dx\right)\right) = 0 \quad .$$

10. APPROXIMATION OF A CONTINUOUS FUNCTION BY ARITHMETIC FUNCTIONS

Let I be an element of $\mathcal{A}$ and let $f = I \longrightarrow R$ be a continuous function. Let

$$\{F_n\} = \{F_1, F_2, \ldots\}$$

be a sequence of arithmetic functions such that $p(I)$ is contained in the domain $\mathcal{D}(F_n)$ for all n . If for every $x \epsilon I$,

$$F_1 p(x) \supset F_2 p(x) \supset \ldots$$

and

$$\bigcap_{n=1}^{\infty} F_n p(x) = pf(x) \quad ,$$

we say that $\{F_n\}$ converges to f . In symbols,

$$\lim_{n \to \infty} F_n = f \quad .$$

(10-1) Let f be a rational function of domain $D(f)$ and let I be an element of $\mathcal{A}$ contained in $D(f)$. Let F be an associated arithmetic

function of f with parameters $A_1, \ldots, A_m$. Then for every $\varepsilon > 0$ there is a $\delta > 0$ such that if $\sigma(A_1) < \delta, \ldots, \sigma(A_m) < \delta$, then for every $x \epsilon I$, $Fp(x)$ is defined and $\sigma\big(Fp(x)\big) < \varepsilon$.

Proof. Let G be the associated special arithmetic function of f with $F \supset G$ and let $B_1, \ldots, B_m$ be the parameters of G. By (6-3) there is, for every $x \epsilon I$, a $\delta_x > 0$ such that if $P(A_1,B_1) < \delta_x, \ldots, P(A_m,B_m) < \delta_x$ and X is an element of $\mathcal{I}_I$ with $P\big(X,p(x)\big) < \delta_x$, then $F(X)$ is defined and $P\big(F(X), Gp(x)\big) < \varepsilon/2$. Since I is compact there exist a finite number of points of I, say $x_1, \ldots, x_t$, such that the union of

$$\left(x_j - \delta_{x_j}, x_j + \delta_{x_j}\right), \qquad j = 1, \ldots, t,$$

contains I. Let

$$\delta = \min\left\{\delta_{x_1}, \ldots, \delta_{x_t}\right\}.$$

It follows that if $P(A_1,B_1) < \delta \ldots, P(A_m,B_m) < \delta$, then for every $x \epsilon I$,

$$\sigma\big(Fp(x)\big) < \varepsilon.$$

In fact, there is an x_j with $P\big(p(x), p(x_j)\big) < \delta_j$ so that $P\big(Fp(x), pf(x)\big) < \varepsilon/2$. Hence

$$\sigma\big(Fp(x)\big) < \varepsilon. \qquad \text{q.e.d.}$$

Let f be a rational function of domain $D(f)$ and let I be an element of $\mathcal{I}$ contained in $D(f)$. By (10-1) we can easily construct a sequence

of arithmetic functions

$$F_1 \supset F_2 \supset \dots$$

such that for every $x \epsilon I$,

$$\sigma\left(F_n p(x)\right) < 1/n \quad , \qquad n = 1, 2, \dots \quad .$$

Hence

$$\lim_{n \to \infty} F_n = f \quad .$$

This result can be extended as follows.

(10-2) For every continuous function $f : I \to R$, $I \epsilon \mathcal{I}$, there is a sequence of arithmetic functions

$$F_n \quad , \qquad n = 1, 2, \dots$$

with

$$\lim_{n \to \infty} F_n = f \quad .$$

Proof. It is well-known that every continuous function can be approximated by polynomials. For every integer $n > 1$ we let f_n be a polynomial such that for every $x \epsilon I$, $\rho\left(f_n(x), f(x)\right) < 1/(6 \cdot 3^n)$. By (10-1), there is an arithmetic function G_n such that for every $x \epsilon I$, $G_n p(x)$ is defined and

$$P\left(G_n p(x), pf_n(x)\right) \leqslant 1/(6 \cdot 3^n)$$

Let F_n be the arithmetic function such that

$$F_n(X) = G_n(X) \oplus \left[-4/(6 \cdot 3^n), 4/(6 \cdot 3^n)\right] \quad , \qquad X \epsilon \mathcal{D}(G_n) \quad .$$

Then for every $x \epsilon I$, $F_n p(x)$ is defined. Since

$$P\Big(G_n p(x),\ pf(x)\Big) \leqslant P\Big(G_n p(x),\ pf_n(x)\Big) + P\Big(pf_n(x),\ pf(x)\Big)$$
$$\leqslant 1/(6\cdot 3^n) + 1/(6\cdot 3^n) = 2/(6\cdot 3^n) \ ,$$

it follows that

$$pf(x) \subset G_n p(x) \oplus \Big[-2/(6\cdot 3^n),\ 2/(6\cdot 3^n)\Big] \ ,$$
$$G_n p(x) \subset pf(x) \oplus \Big[-2/(6\cdot 3^n),\ 2/(6\cdot 3^n)\Big] \ .$$

Hence

$$pf(x) \oplus \Big[-2/(6\cdot 3^n),\ 2/(6\cdot 3^n)\Big] \subset F_n p(x)$$
$$\subset pf(x) \oplus \Big[-1/3^n,\ 1/3^n\Big]$$

Consequently

$$F_{n+1} p(x) \subset pf(x) \oplus \Big[-1/3^{n+1},\ 1/3^{n+1}\Big] \subset F_n p(x) \ .$$

Since $\sigma\Big(F_n p(x)\Big) < 2/3^n$ and $\lim_{n\to\infty} 2/3^n = 0$, it follows that $\lim F_n = f$.

q.e.d.

Let I be an element of $\mathcal{I}$ and let

$$\xi = \Big\{\xi_1,\ \ldots,\ \xi_r\Big\}, \qquad \eta = \Big\{\eta_1,\ \ldots,\ \eta_s\Big\}$$

be subdivisions of I . If there exist integers $1 \leqslant j(1) < j(2) < \ldots < j(r) = s$ such that for every $i = 1,\ \ldots,\ r$, $\Big\{\eta_{j(i-1)+1},\ \ldots,\ \eta_{j(i)}\Big\}$ is a subdivision of ξ_i , we write $\xi \prec \eta$ and call η a <u>refinement</u> of ξ .

Let $I \epsilon \mathcal{I}$ and let F be an arithmetic function with $\mathcal{D}(F) \supset p(I)$. As before, there is a positive number δ such that whenever $X \epsilon \mathcal{I}_I$ with

$\sigma(X) < \delta$, $F(X)$ is defined. Hence if ξ is a subdivision of I with $\sigma(\xi) < \delta$, both $F(I,\xi)$ and $\Sigma(F,I.\xi)$ are defined.

(10-3) Let F be an arithmetic function and let $\xi = \left\{\xi_1, \ldots, \xi_r\right\}$ be a subdivision of $I \in \mathcal{I}$ such that $F(\xi_1), \ldots, F(\xi_r)$ are defined. Then for every refinement $\eta = \left\{\eta_1, \ldots, \eta_s\right\}$ of ξ , $F(\eta_1), \ldots, F(\eta_s)$ are defined so that $F(I,\eta)$ and $\Sigma(F,I,\eta)$ are defined. Moreover,

$$F(I,\eta) \subset F(I,\xi) , \qquad \Sigma(F,I,\eta) \subset \Sigma(F,I,\xi) .$$

(10-4) Let $I \in \mathcal{I}$ and let F be an arithmetic function. Let $\xi = \left\{\xi_1, \ldots, \xi_r\right\}$ be a subdivision of I such that $F(\xi_1), \ldots, F(\xi_r)$ are defined, and let

$$\xi = \xi^{(1)} \prec \xi^{(2)} \prec \ldots$$

be a sequence of subdivisions of I . Then $F\left(I,\xi^{(n)}\right)$ and $\Sigma\left(F,I,\xi^{(n)}\right)$ are defined for all n and

$$F\left(I,\xi^{(1)}\right) \supset F\left(I,\xi^{(2)}\right) \supset \ldots ,$$

$$\Sigma\left(F,I,\xi^{(1)}\right) \supset \Sigma\left(F,I,\xi^{(2)}\right) \supset \ldots .$$

If $\lim_{n\to\infty} \sigma\left(\xi^{(n)}\right) = 0$, then $\bigcap_{n=1}^{\infty} F\left(I,\xi^{(n)}\right)$ and $\bigcap_{n=1}^{\infty} \Sigma\left(F,I,\xi^{(n)}\right)$ are independent of the choice of ξ and $\left\{\xi^{(n)}\right\}$.

Proof. Let $\eta = \left\{\eta_1, \ldots, \eta_s\right\}$ be a subdivision of I such that $F(\eta_1), \ldots, F(\eta_s)$ are defined and let $\zeta = \left\{\zeta_1, \ldots, \zeta_t\right\}$ be a refinement of η . By definition, there are integers $1 \leqslant j(1) < \ldots < j(s) = t$

such that for every $i = 1, \ldots, s$, $\left\{\zeta_{j(i-1)+1}, \ldots, \zeta_{j(i)}\right\}$ is a subdivision of η_i. It follows from (6-5) that

$$F\left(\zeta_{j(i-1)+1}\right) \subset F(\eta_i), \ldots, F\left(\zeta_{j(i)}\right) \subset F(\eta_i) \quad .$$

Hence

$$F(I,\zeta) = F(\zeta_1) \cup \ldots \cup F(\zeta_t)$$

$$\subset F(\eta_1) \cup \ldots \cup F(\eta_s) = F(I,\eta)$$

As a consequence of this result we have

$$F\left(I,\xi^{(1)}\right) \supset F\left(I,\xi^{(2)}\right) \supset \ldots$$

Since

$$\alpha\left(\left(F\left(\zeta_{j(i-1)+1}\right) \otimes p\sigma\left(\zeta_{j(i-1)+1}\right)\right) \oplus \cdots \oplus \left(F\left(\zeta_{j(i)}\right) \otimes p\sigma\left(\zeta_{j(i)}\right)\right)\right)$$

$$= \alpha\left(F\left(\zeta_{j(i-1)+1}\right)\right) \sigma\left(\zeta_{j(i-1)+1}\right) + \cdots + \alpha\left(F\left(\zeta_{j(i)}\right)\right) \sigma\left(\zeta_{j(i)}\right)$$

$$\geq \alpha\left(F(\eta_i)\right) \sigma\left(\zeta_{j(i-1)+1}\right) + \cdots + \alpha\left(F(\eta_i)\right) \sigma\left(\zeta_{j(i)}\right)$$

$$= \alpha\left(F(\eta_i)\right) \sigma(\eta_i) = \alpha\left(F(\eta_i) \otimes p\sigma(\eta_i)\right)$$

and similarly

$$\beta\left(\left(F\left(\zeta_{j(i-1)+1}\right) \otimes p\sigma\left(\zeta_{j(i-1)+1}\right)\right) \oplus \cdots \oplus \left(F\left(\zeta_{j(i)}\right) \otimes p\sigma\left(\zeta_{j(i)}\right)\right)\right)$$

$$\leq \beta\left(F(\eta_i) \otimes p\sigma(\eta_i)\right) \quad ,$$

it follows that

$$\left(F\left(\zeta_{j(i-1)+1}\right) \otimes p\sigma\left(\zeta_{j(i-1)+1}\right)\right) \oplus \cdots \oplus \left(F\left(\zeta_{j(1)}\right) \otimes p\sigma\left(\zeta_{j(1)}\right)\right)$$

$$\subset F(\eta_i) \otimes p\sigma(\eta_i) \quad .$$

Hence

$$\Sigma(F,I,\zeta) = \left(F(\zeta_1) \otimes p\sigma(\zeta_1)\right) \oplus \cdots \oplus \left(F(\zeta_t) \otimes p\sigma(\zeta_t)\right)$$

$$\subset \left(F(\eta_1) \otimes p\sigma(\eta_1)\right) \oplus \cdots \oplus \left(F(\eta_s) \otimes p\sigma(\eta_s)\right) = \Sigma(F,I,\eta) \quad .$$

From this result it follows that

$$\Sigma\left(F,I,\xi^{(1)}\right) \supset \Sigma\left(F,I,\xi^{(2)}\right) \supset \cdots$$

Now we assume $\lim_{n\to\infty} \sigma\left(\xi^{(n)}\right) = 0$. Let $\eta = \left\{\eta_1, \ldots, \eta_s\right\}$ be a subdivision of I such that $F(\eta_1), \ldots, F(\eta_s)$ are defined, and let

$$\eta = \eta^{(1)} \prec \eta^{(2)} \prec \cdots$$

be a sequence of subdivisions with $\lim_{n\to\infty} \sigma\left(\eta^{(n)}\right) = 0$.

Let $\varepsilon > 0$. For any $y \in \bigcap_{n=1}^{\infty} F\left(I,\xi^{(n)}\right)$ there is a sequence

$$\xi^{(1)}_{i(1)} \supset \xi^{(2)}_{i(2)} \supset \cdots .$$

in $\mathcal{I}$ such that for every $n = 1, 2, \ldots, \xi^{(n)}_{i(n)} \in \xi^{(n)}$ and $F\left(\xi^{(n)}_{i(n)}\right) \ni y$.

Since $\lim_{n\to\infty} \sigma\left(\xi^{(n)}\right) = 0$, $\bigcap_{n=1}^{\infty}\left(\xi^{(n)}_{i(n)}\right)$ contains a single point x . By (6-5) there is a $\delta > 0$ such that whenever $X \in \mathcal{I}_I$ with $P\left(X,p(x)\right) < \delta$,

$F(X)$ is defined and $P\left(F(X), Fp(x)\right) < \varepsilon$. Since $\lim_{n\to\infty} \sigma\left(\eta^{(n)}\right) = 0$, $\sigma\left(\eta^{(n)}\right) < \delta$ holds for all large n. Let $\eta_i^{(n)}$ be the element of $\eta^{(n)}$ containing x. Then $P\left(\eta_i^{(n)}, p(x)\right) < \delta$ so that $P\left(F\left(\eta_i^{(n)}\right) Fp(x)\right) < \varepsilon$. Hence for all large n

$$y \in Fp(x) \subset F\left(\eta_i^{(n)}\right) \oplus [-\varepsilon, \varepsilon] \subset F\left(I, \eta^{(n)}\right) \oplus [-\varepsilon, \varepsilon]$$

This proves that $y \in \bigcap_{n=1}^{\infty} F\left(I, \eta^{(n)}\right) \oplus [-\varepsilon, \varepsilon]$. Since y is an arbitrary point of $\bigcap_{n=1}^{\infty} F\left(I, \xi^{(n)}\right)$, it follows that

$$\bigcap_{n=1}^{\infty} F\left(I, \xi^{(n)}\right) \subset \bigcap_{n=1}^{\infty} F\left(I, \eta^{(n)}\right) \oplus [-\varepsilon, \varepsilon] .$$

Similarly,

$$\bigcap_{n=1}^{\infty} F\left(I, \eta^{(n)}\right) \subset \bigcap_{n=1}^{\infty} F\left(I, \xi^{(n)}\right) \oplus [-\varepsilon, \varepsilon] .$$

Applying (1-6), we have

$$\bigcap_{n=1}^{\infty} F\left(I, \xi^{(n)}\right) = \bigcap_{n=1}^{\infty} F\left(I, \eta^{(n)}\right) .$$

In order to prove that $\bigcap_{n=1}^{\infty} \Sigma\left(F, I, \xi^{(n)}\right) = \bigcap_{n=1}^{\infty} \Sigma\left(F, I, \eta^{(n)}\right)$, we may assume $\sigma(I) > 0$. It is sufficient to prove that for every integer $m \geqslant 1$ and every $\varepsilon > 0$,

$$\Sigma\left(F, I, \xi^{(m)}\right) \oplus [-\varepsilon, \varepsilon] \supset \Sigma\left(F, I, \eta^{(n)}\right)$$

holds for large n. In fact, if this is proved, then

$$\Sigma\left(F, I, \xi^{(m)}\right) \oplus [-\varepsilon, \varepsilon] \supset \bigcap_{n=1}^{\infty} \Sigma\left(F, I, \eta^{(n)}\right) .$$

Since ε is arbitrary, it follows that $\Sigma\left(F, I, \xi^{(m)}\right) \supset \bigcap_{n=1}^{\infty} \Sigma\left(F, I, \eta^{(n)}\right)$.

Since m is arbitrary, it follows that

$$\bigcap_{m=1}^{\infty} \Sigma\left(F, I, \xi^{(m)}\right) \supset \bigcap_{n=1}^{\infty} \Sigma\left(F, I, \eta^{(n)}\right) \quad .$$

Similarly,

$$\bigcap_{n=1}^{\infty} \Sigma\left(F, I, \eta^{(n)}\right) \supset \bigcap_{m=1}^{\infty} \Sigma\left(F, I, \xi^{(m)}\right) \quad .$$

Hence our assertion follows.

Now we let m be an arbitrary integer $\geqslant 1$ and let $\xi^{(m)} = \left\{A_1, \ldots, A_u\right\}$. Let ε be an arbitrary positive number < 1. By (6-3), there is, for every $x \epsilon I$, a positive number r_x such that whenever $X \epsilon \mathscr{L}$ with $P\left(X, p(x)\right) < r_x$. $F(X)$ is defined and $P\left(F(X), Fp(x)\right) < \varepsilon/u\left(\gamma\left(F\left(I, \xi^{(m)}\right)\right)+1\right)$. Since I is compact, there is a $\delta > 0$ such that whenever $X \epsilon \mathscr{L}_I$ with $\sigma(X) < \delta$, $P\left(X, p(x)\right) < r_x$ for some $x \epsilon I$.

Since $\lim_{n \to \infty} \sigma\left(\eta^{(n)}\right) = 0$, there is an integer n_o such that

$$\sigma\left(\eta^{(n)}\right) < \min\left(\delta, \varepsilon/u\left(\gamma\left(F\left(I, \xi^{(m)}\right)\right) + 1\right)\right)$$

for all integers $> n_o$. Let $n > n_o$ and let

$$\eta^{(n)} = \left\{B_1, \ldots, B_v\right\} \quad .$$

For every $i = 1, \ldots, u$, there is a largest integer $j(i)$ with $\beta(A_i) \epsilon B_{j(i)}$. Clearly

$$1 \leqslant j(1) \leqslant j(2) \leqslant \ldots \leqslant j(u) = v \quad .$$

Since $P\left(B_{j(i)}, p\beta(A_i)\right) < \delta$, it follows that

$$P\left(F\left(B_{j(i)}\right), Fp\,\beta(A_i)\right) < \varepsilon/u\left(\gamma\left(F\left(I, \xi^{(m)}\right)\right) + 1\right) < 1 \quad .$$

Therefore

$$F\left(B_{j(i)}\right) \subset F\, p\, \beta(A_i) \oplus [-1, 1] \subset F(A_i) \oplus [-1, 1]$$

and then

$$\gamma\left(F\left(B_{j(i)}\right) \otimes p\,\sigma\left(B_{j(i)}\right)\right) \leqslant \left(\gamma\left(F(A_i)\right) + 1\right) \sigma\left(B_{j(i)}\right)$$

$$\leqslant \left(\gamma\left(F\left(I, \xi^{(m)}\right)\right) + 1\right) \sigma\left(B_{j(i)}\right) \ .$$

Hence

$$\Sigma_{i=1}^{u}\, \gamma\left(F\left(B_{j(i)}\right) \otimes p\sigma\left(B_{j(i)}\right)\right) \leqslant \left(\gamma\left(F\left(I, \xi^{(m)}\right)\right) + 1\right) \Sigma_{i=1}^{m}\, \sigma\left(B_{j(i)}\right)$$

$$\leqslant \left(\gamma\left(F\left(I, \xi^{(m)}\right)\right) + 1\right) u\, \sigma\left(\eta^{(n)}\right) < \varepsilon$$

and consequently

$$\left(F\left(B_{j(1)}\right) \otimes p\sigma\left(B_{j(1)}\right)\right) \oplus \cdots \oplus \left(F\left(B_{j(u)}\right) \otimes p\sigma\left(B_{j(u)}\right)\right) \subset [-\varepsilon, \varepsilon]$$

Since, for every $k = j(i-1) + 1, \ldots, j(i) - 1$, $B_k \subset A_i$, it follows that

$$\left(F\left(B_{j(i-1)+1}\right) \otimes p\sigma\left(B_{j(i-1)+1}\right)\right) \oplus \cdots \oplus \left(F\left(B_{j(i)-1}\right) \otimes p\,\sigma\left(B_{j(i)-1}\right)\right)$$

$$\subset F(A_i) \otimes p\sigma(A_i) \ , \qquad i = 1, \ldots, u \ .$$

Hence, by adding these equations, we have

$$\Sigma\left(F, I, \eta^{(n)}\right) \subset \Sigma\left(F, I, \xi^{(m)}\right) \oplus [-\varepsilon, \varepsilon] \ .$$

This completes our proof. q.e.d.

Let F be an arithmetic function with domain $\mathfrak{D}(F)$. Let $\mathcal{I}^F$ be the subset of $\mathcal{I}$ consisting of all the elements X of $\mathcal{I}$ with $p(X) \subset \mathfrak{D}(F)$. Then, by (10-4), we may define a function

$$\overline{F} : \mathcal{I}^F \longrightarrow \mathcal{I}$$

by

$$\overline{F}(X) = \bigcap_{n=1}^{\infty} F\left(X, \xi^{(n)}\right),$$

where $X \in \mathcal{I}^F$ and $\xi^{(1)} \prec \xi^{(2)} \prec \ldots$ is a sequence of subdivisions of X with $\lim_{n\to\infty} \sigma\left(\xi^{(n)}\right) = 0$.

(10-5) Let X and Y be elements of $\mathcal{I}^F$ with $\beta(X) = \alpha(Y)$. Then $X \cup Y \in \mathcal{I}^F$ and

$$\overline{F}(X \cup Y) = \overline{F}(X) \cup \overline{F}(Y) .$$

Proof. Since $\beta(X) = \alpha(Y)$, $X \cup Y \in \mathcal{I}$. Since $X \cup Y \in \mathcal{I}^F$, $p(X) \subset \mathfrak{D}(F)$ and $p(Y) \subset \mathfrak{D}(F)$. Hence $p(X \cup Y) \subset \mathfrak{D}(F)$ and consequently $X \cup Y \in \mathcal{I}^F$.

If $Y \in p(R)$, then $\overline{F}(Y) = F(Y) \subset \overline{F}(X)$ so that our assertion is obvious. If $Y \notin p(R)$, then we may have a sequence of subdivisions

$$\xi^{(n)} = \left\{\xi_1^{(n)}, \ldots, \xi_{r(n)}^{(n)}\right\}$$

of $X \cup Y$ such that $\xi^{(1)} \prec \xi^{(2)} \prec \ldots$, $\lim_{n\to\infty} \sigma\left(\xi^{(n)}\right) = 0$ and for every integer n there is an integer $s(n)$ with $\beta\left(\xi_{s(n)}^{(n)}\right) = \beta(X)$. Therefore

$$\eta^{(n)} = \left\{\xi_1^{(n)}, \ldots, \xi_{s(n)}^{(n)}\right\}$$

is a sequence of subdivisions of X such that $\eta^{(1)} \prec \eta^{(2)} \prec \ldots$ and $\lim_{n\to\infty} \sigma\left(\eta^{(n)}\right) = 0$; and $\zeta^{(n)} = \left\{\xi_{s(n)+1}^{(n)}, \ldots, \xi_{r(n)}^{(n)}\right\}$ is a sequence of

subdivisions of Y such that $\zeta^{(1)} \prec \zeta^{(2)} \prec \cdots$ and $\lim_{n\to\infty} \sigma\left(\zeta^{(n)}\right) = 0$. Since

$$F\left(X \cup Y, \xi^{(n)}\right) = F\left(X, \eta^{(n)}\right) \cup F\left(Y, \zeta^{(n)}\right),$$

it follows that

$$\overline{F}(X \cup Y) = \overline{F}(X) \cup \overline{F}(Y) \quad . \qquad \text{q.e.d.}$$

(10-6) $\overline{F} : \mathcal{L}^F \longrightarrow \mathcal{L}$ is continuous.

Proof. Let $Y \in \mathcal{L}^F$ and let $\varepsilon > 0$. Since $F\left(\alpha(Y)\right)$ is defined, there is a $\delta > 0$ such that whenever $Z \in \mathcal{L}$ with $\sigma(Z) < \delta$, $P\left(Z, p\alpha(Y)\right) < \delta$, $F(Z)$ is defined and $P\left(F(Z), Fp\alpha(Y)\right) < \varepsilon/4$. Since

$$F(Z) \supset \overline{F}(Z) \supset Fp\alpha(Y) = \overline{F}p\alpha(Y) \quad ,$$

we have

$$P\left(\overline{F}(Z), \overline{F}p\alpha(Y)\right) < \varepsilon/4 \quad .$$

Similarly there is a $\delta' > 0$ such that whenever $Z' \in \mathcal{L}$ with $P\left(Z', p\beta(Z)\right) < \delta'$, $F(Z')$ is defined and

$$P\left(\overline{F}(Z'), \overline{F}p\beta(Y)\right) < \varepsilon/4 \quad .$$

For every $X \in \mathcal{L}_F$ with $P(X, Y) < \min(\delta, \delta')$ we have $Z, Z' \in \mathcal{L}$ such that

$$P\left(Z, p\alpha(Y)\right) < \delta \quad , \qquad P\left(Z', p\beta(Y)\right) < \delta'$$

and one of the following holds:

(1) $\alpha(X) = \alpha(Z)$, $\beta(Z) = \alpha(Y)$, $\beta(Y) = \alpha(Z')$, $\beta(Z') = \beta(X)$;

(2) $\alpha(X) = \alpha(Z)$, $\beta(X) = \alpha(Z')$, $\beta(Z) = \alpha(Y)$, $\beta(Z') = \beta(Y)$;

(3) $\alpha(Z) = \alpha(Y)$, $\beta(Z) = \alpha(X)$, $\beta(Y) = \alpha(Z')$, $\beta(X) = \beta(Z')$;

(4) $\alpha(Z) = \alpha(Y)$, $\beta(Z) = \alpha(X)$, $\beta(X) = \alpha(Z')$, $\beta(Z') = \beta(Y)$.

In case (1) we have, by (10-5),

$$\overline{F}(X) = \overline{F}(Z) \cup \overline{F}(Y) \cup \overline{F}(Z')$$

Then $\overline{F}(X) \supset \overline{F}(Y)$. Since

$$\overline{F}(Z) \subset \overline{F} \, p \, \alpha(Y) \oplus \left[-\varepsilon/2, \varepsilon/2\right] \subset \overline{F}(Y) \oplus \left[-\varepsilon/2, \varepsilon/2\right] ,$$

and similarly,

$$\overline{F}(Z') \subset \overline{F}(Y) \oplus \left[-\varepsilon/2, \varepsilon/2\right] ,$$

it follows that

$$\overline{F}(X) \subset \overline{F}(Y) \oplus \left[-\varepsilon, \varepsilon\right] .$$

Hence $P\left(\overline{F}(X), \overline{F}(Y)\right) < \varepsilon$.

In case (2), we have, by (10-5),

$$\overline{F}(X) \cup \overline{F}(Z') = \overline{F}(Z) \cup \overline{F}(Y) .$$

Since $\overline{F}(Z) \subset \overline{F}(Y) \oplus \left[-\varepsilon/2, \varepsilon/2\right]$ and $\overline{F}(Z') \subset \overline{F}(Y) \oplus \left[-\varepsilon/2, \varepsilon/2\right]$, it follows that $\overline{F}(X) \subset \overline{F}(Y) \oplus \left[-\varepsilon, \varepsilon\right]$. On the other hand,

$$P\left(\overline{F}(Z'), \overline{F} \, p \, \beta(X)\right) \leq P\left(\overline{F}(Z'), \overline{F} \, p \, \beta(Y)\right) \oplus P\left(\overline{F} \, p \, \beta(X), \overline{F} \, p \, \beta(Y)\right)$$

$$< \varepsilon/4 + \varepsilon/4 = \varepsilon/2$$

so that

$$\overline{F}(Z') \subset \overline{F}(X) \oplus \left[-\varepsilon/2, \varepsilon/2\right] .$$

Hence $\overline{F}(Y) \subset \overline{F}(X) + \left[-\varepsilon, \varepsilon\right]$. This again proves that $P\left(\overline{F}(X), \overline{F}(Y)\right) < \varepsilon$.

Similar argument shows that our assertion also holds for the other two cases. q.e.d.

(10-7) Let I be an element of $\mathcal{I}$ and let $f : I \to R$ be a continuous function. Let $\{F_n\}$ be a sequence of arithmetic functions with $\lim_{n \to \infty} F_n = f$. Then

$$\overline{F}_1(I) \supset \overline{F}_2(I) \supset \dots$$

and $\bigcap_{n=1}^{\infty} \overline{F}_n(I) = f(I)$.

Proof. Let n be an integer ≥ 1. Let $\varepsilon > 0$. For every $x \in \mathcal{I}$ there is an $r_x > 0$ such that whenever $X \in \mathcal{I}$ with $P\big(X, p(x)\big) < r_x$, $F_n(X)$ and $F_{n+1}(X)$ are defined and $P\big(F_{n+1}(X), F_{n+1}p(x)\big) < \varepsilon$. Let δ be a positive number such that whenever $X \in \mathcal{I}_I$ with $\sigma(X) < \delta$, we have $P\big(X, p(x)\big) < r_x$ for some $x \in X$.

Let $\xi = \{\xi_1, \dots, \xi_r\}$ be any subdivision of I with $\sigma(\xi) < \delta$. For every $i = 1, \dots, r$, there is an $x_i \in I$ such that $P\big(\xi_i, p(x_i)\big) < r_{x_i}$. Then

$$F_{n+1}(\xi_i) \subset F_{n+1}\, p(x_i) \oplus \big[-\varepsilon, \varepsilon\big] \subset F_n\, p(x_i) \oplus \big[-\varepsilon, \varepsilon\big]$$

$$F_n(I, \xi) \oplus \big[-\varepsilon, \varepsilon\big]$$

so that

$$F_{n+1}(I, \xi) \subset F_n(I, \xi) \oplus \big[-\varepsilon, \varepsilon\big] .$$

Hence

$$\overline{F}_{n+1}(I) \subset \overline{F}_n(I) \oplus \big[-\varepsilon, \varepsilon\big] .$$

Since ε is arbitrary, it follows that

$$\overline{F}_{n+1}(I) \subset \overline{F}_n(I) \quad .$$

This proves that

$$\overline{F}_1(I) \supset \overline{F}_2(I) \supset \dots \quad .$$

Clearly $\overline{F}_n(I) \supset f(I)$ for all n. Let $\varepsilon > 0$.

For every $n = 1, 2, \dots$ we let

$$I_n = \left\{ x \in I \mid P\left(F_n p(x), pf(x)\right) \geqslant \varepsilon \right\} .$$

It follows from the continuity of F_n that I_n is closed in I. Since $I_1 \supset I_2 \supset \dots$ and $\bigcap_{n=1}^{\infty} I_n = \emptyset$, we infer that there is an integer n_o such that whenever $n > n_o$, $P\left(F_n p(x), pf(x)\right) < \varepsilon$ holds for all $x \in I$. Let $n > n_o$. Then for any subdivision $\xi = \left\{\xi_1, \dots, \xi_r\right\}$ of I we have

$$F_n(\xi_i) \subset f(\xi_i) \oplus \left[-\varepsilon, \varepsilon\right] , \qquad i = 1, \dots, r .$$

Hence

$$\overline{F}_n(I) \subset F_n(I, \xi) \subset f(I) \oplus \left[-\varepsilon, \varepsilon\right] .$$

Since ε is arbitrary, we infer that

$$\bigcap_{n=1}^{\infty} \overline{F}_n(I) = f(I) \qquad \text{q.e.d.}$$

Theorem 3. Let $I \in \mathcal{A}$ and let $f : I \to R$ be a continuous function. Let $\left\{F_n\right\}$ be a sequence of arithmetic functions with $\lim_{n \to \infty} F_n = f$.

Then there is a sequence $\xi^{(1)} \prec \xi^{(2)} \prec \dots$ of subdivisions of I such that $\lim_{n\to\infty} \sigma\left(\xi^{(n)}\right) = 0$ and

$$\bigcap_{n=1}^{\infty} F_n\left(I, \xi^{(n)}\right) = f(I) \quad .$$

If, moreover, for every $x \in I$ and every integer $n \geq 1$, $F_{n+1}\, p(x)$ is contained in the interior of $F_n\, p(x)$, then there is a sequence $\xi^{(1)} \prec \xi^{(2)} \prec \dots$ of subdivisions of I such that $\lim_{n\to\infty} \sigma\left(\xi^{(n)}\right) = 0$, $F_1\left(I, \xi^{(1)}\right) \supset F_2\left(I, \xi^{(2)}\right) \supset \dots$ and $\bigcap_{n=1}^{\infty} F_n\left(I, \xi^{(n)}\right) = f(I)$.

Proof. By (10-7), we have $\overline{F}_1(I) \supset \overline{F}_2(I) \supset \dots$ and $\bigcap_{n=1}^{\infty} \overline{F}_n(I) = f(I)$. Let $\xi^{(1)}$ be a subdivision of I such that $\sigma\left(\xi^{(1)}\right) < 1$ and $P\left(F\left(I, \xi^{(1)}\right), \overline{F}(I)\right) < 1$. Suppose that we have subdivisions $\xi^{(1)} \prec \xi^{(2)} \prec \dots \prec \xi^{(k)}$ of I such that $\sigma\left(\xi^{(n)}\right) < 1/n$ and $P\left(F\left(I, \xi^{(n)}\right), \overline{F}(I)\right) < 1/n$, $n = 1, \dots, k$. We let $\xi^{(k+1)}$ be a refinement of $\xi^{(n)}$ with $P\left(F_{k+1}\left(I, \xi^{(k+1)}\right), \overline{F}_{k+1}(I)\right) < 1/(k+1)$. By induction, we have a sequence $\xi^{(1)} \prec \xi^{(2)} \prec \dots$ of subdivisions of I with $P\left(F_n\left(I, \xi^{(n)}\right), \overline{F}_n(I)\right) < 1/n$. Hence

$$\bigcap_{n=1}^{\infty} F_n\left(I, \xi^{(n)}\right) = f(I) \quad .$$

If for every $x \in I$ and every integer $n \geq 1$, $F_{n+1}\, p(x)$ is contained in the interior of $F_n\, p(x)$, as in the proof of (10-7), there is a $\delta_n > 0$ such that whenever ξ is a subdivision of I with $\sigma(\xi) < \delta_n$, $F_n(I, \xi) \supset F_{n+1}(I, \xi)$, $n = 1, 2, \dots$. Now we construct a sequence

$\xi^{(1)} \prec \xi^{(2)} \prec \ldots$ of subdivisions of I, just as above, satisfying the additional condition that

$$\sigma\left(\xi^{(n)}\right) < \delta_n , \qquad n = 1, 2, \ldots$$

Then our conclusion follows. q.e.d.

Let F be an arithmetic function with domain $\mathfrak{D}(F)$ and $\mathcal{I}^F$ be the subset of $\mathcal{I}$ consisting of all the elements I of $\mathcal{I}$ with $p(I) \subset \mathfrak{D}(F)$. Then, by (10-4), we may define a function

$$\Sigma_F : \mathcal{I}^F \longrightarrow \mathcal{I}$$

by

$$\Sigma_F(I) = \bigcap_{n=1}^{\infty} \Sigma\left(F, I, \xi^{(n)}\right) ,$$

where $I \in \mathcal{I}^F$ and $\xi^{(1)} \prec \xi^{(2)} \prec \ldots$ is a sequence of subdivisions of I with $\lim_{n\to\infty} \sigma\left(\xi^{(n)}\right) = 0$.

As (10-5), (10-6) and (10-7), we have

(10-8) Let I and J be elements of $\mathcal{I}^F$ with $\beta(I) = \alpha(J)$. Then $I \cup J \in \mathcal{I}^F$ and

$$\Sigma_F(I \cup J) = \Sigma_F(I) \oplus \Sigma_F(J) .$$

(10-9) $\Sigma_F : \mathcal{I}^F \longrightarrow \mathcal{I}$ is continuous.

(10-10) Let $I = [a,b] \in \mathcal{I}$ and let $f : I \longrightarrow R$ be a continuous function. Let $\{F_n\}$ be a sequence of arithmetic functions with $\lim_{n\to\infty} F_n = f$.

Then

$$\Sigma_{F_1}(I) \supset \Sigma_{F_2}(I) \supset \dots$$

and $\bigcap_{n=1}^{\infty} \Sigma_{F_n}(I) = \int_a^b f(x)\,dx$

Making use of (10-10) and the definition of Σ_F we can prove

Theorem 4. Let $I = [a,b] \in \mathcal{A}$ and let $f : I \to R$ be a continuous function. Let $\{F_n\}$ be a sequence of arithmetic functions with $\lim_{n\to\infty} F_n = f$. Then there is a sequence $\xi^{(1)} \prec \xi^{(2)} \prec \dots$ of subdivisions of I such that $\lim_{n\to\infty} \sigma\left(\xi^{(n)}\right) = 0$ and

$$\bigcap_{n=1}^{\infty} \Sigma\left(F_n, I, \xi^{(n)}\right) = \int_a^b f(x)\,dx \ .$$

If, moreover, for every $x \in I$ and every integer $n \geq 1$ $F_{n+1}p(x)$ is contained in the interior of $F_n p(x)$, then there is a sequence $\xi^{(1)} \prec \xi^{(2)} \prec \dots$ of subdivisions of I such that $\lim_{n\to\infty} \sigma\left(\xi^{(n)}\right) = 0$, $\Sigma\left(F_1, I, \xi^{(1)}\right) \supset \Sigma\left(F_2, I, \xi^{(2)}\right) \supset \dots$ and $\bigcap_{n=1}^{\infty} \Sigma\left(F_n, I, \xi^{(n)}\right) = \int_a^b f(x)\,dx$.

CONVERGENCE OF APPROXIMATE EIGENVECTORS IN JACOBI METHODS

2

R.L. Causey
P. Henrici

FOREWORD

The work covered in this paper was begun by the authors at the Space Technology Laboratories, Los Angeles, under the sponsorship of the U.S. Air Force. It was concluded at Lockheed Missiles and Space Division, Sunnyvale, California, under the Lockheed General Research Program.

The paper is being published in the Numerische Mathematik, Vol. 2, Heidelberg, Springer-Verlag, 1960.

The authors wish to express their appreciation to Prof. George E. Forsythe for calling their attention to some of the problems discussed in this paper.

ABSTRACT

The paper constitutes a study of the convergence of infinite products of unitary matrices connected with Jacobi methods for computing eigenvalues of Hermitian matrices. Bounds are obtained for the error in approximate eigenvectors resulting from a finite number of steps of three Jacobi processes, namely, the classical Jacobi method, quasi-cyclic restricted Jacobi methods and threshold cyclic Jacobi methods. The results apply only to Hermitian matrices which do not have repeated eigenvalues. Also, certain questions are answered concerning the representation of an arbitrary unitary matrix as an infinite product.

CONTENTS

CONVERGENCE OF APPROXIMATE EIGENVECTORS IN JACOBI METHODS

1. INTRODUCTION. In this paper we are interested in studying the convergence of infinite products of unitary matrices connected with Jacobi methods [4] for computing eigenvalues of Hermitian matrices. We obtain bounds for the error in approximate eigenvectors resulting from a finite number of steps of a Jacobi process. Also, we are able to answer certain questions concerning the representation of an arbitrary unitary matrix as an infinite product. Use is made of the classical results of Jacobi [5] and of the recent results of Henrici [4] and Pope and Tompkins [7]. Basically, we adhere to the notation and terminology used in [4].

Consider the following computational algorithm. Let $A = A_0 = (a_{rc})$ be a Hermitian matrix of order n with eigenvalues $\lambda_r (r = 1,2, \ldots, n)$. One calculates a sequence of matrices $A_1, A_2, \ldots, A_k = (a_{rc}^{(k)}), \ldots$ which are unitarily similar to A by the recurrence relation

$$A_{k+1} = U_k^* A_k U_k \qquad (k = 0, 1, 2, \ldots). \tag{1.1}$$

The $U_k = (u_{rc}^{(k)})$ are special unitary matrices of order n. For every value of k there is specified a pair $\pi_k = (i_k, j_k) = (i, j)$ of indices (we omit the subscript k in the sequel for notational simplicity) satisfying $1 \leq i < j \leq n$, such that the 2×2 matrix

* Bracketed numbers indicate References, page 21.

$$V_k = \begin{pmatrix} u_{ii}^{(k)} & u_{ij}^{(k)} \\ u_{ji}^{(k)} & u_{jj}^{(k)} \end{pmatrix} , \tag{1.2}$$

which is a principal submatrix of U_k, is unitary. All other elements of U_k satisfy

$$u_{rc}^{(k)} = \delta_{rc} = \begin{cases} 1 , & r = c \\ 0 , & r \neq c \end{cases} . \tag{1.3}$$

The matrices U_k are completely determined by the pairs π_k and the 2×2 unitary matrices V_k.

Any set of rules for choosing the U_k is called a Jacobi method. A Jacobi method is said to be convergent if

$$A_k \to \Lambda \qquad (k \to \infty) , \tag{1.4}$$

for all A under the adopted set of rules, where Λ is a diagonal matrix whose diagonal elements are the λ_r in some order. Either of the quantities

$$S_k = \sum_{r \neq c} \left| a_{rc}^{(k)} \right|^2 \tag{1.5}$$

or

$$\mu_k = \operatorname*{Max}_{r \neq c} \left| a_{rc}^{(k)} \right| \tag{1.6}$$

may be used as a measure of the closeness of A_k to Λ. For the Jacobi methods to be considered in this paper, a necessary and sufficient condition for convergence is that $S_k \to 0$ $(k \to \infty)$ or,

equivalently, $\mu_k \to 0$ $(k \to \infty)$. See either [2] or [4] for the proof. If the infinite product of matrices

$$U = \prod_{k=0}^{\infty} U_k = U_0 U_1 U_2 \cdots \tag{1.7}$$

converges, then

$$U^* A U = \Lambda \ , \tag{1.8}$$

and the columns of U are a complete set of normalized eigenvectors of A . We shall investigate the convergence of (1.7) for three convergent Jacobi methods: (a) the classical Jacobi method, (b) the quasicyclic restricted Jacobi methods [4] and (c) the threshold cyclic Jacobi method [7]. Throughout the paper we shall restrict V_k to be of the form

$$V_k = \begin{pmatrix} \cos\theta_k & -e^{i\phi_k}\sin\theta_k \\ e^{-i\phi_k}\sin\theta_k & \cos\theta_k \end{pmatrix} \tag{1.9}$$

By the classical Jacobi method we shall mean a Jacobi method with the following set of rules for determining the U_k of (1.1). Choose π_k such that

$$\left|a_{ij}^{(k)}\right| = \mu_k \ , \tag{1.10}$$

and choose the V_k of (1.9) such that

$$a_{ij}^{(k+1)} = 0 \ . \tag{1.11}$$

The equation (1.11) will always be satisfied if the relations

$$\phi_k = \arg a_{ij}^{(k)} \ , \tag{1.12}$$

$$\tan 2\theta_k^* = \frac{2\left|a_{ij}^{(k)}\right|}{a_{ii}^{(k)} - a_{jj}^{(k)}} \quad (\theta_k = \theta_k^*) \tag{1.13}$$

hold simultaneously. Since Jacobi dealt only with real symmetric matrices A and orthogonal matrices U_k, the set of rules (1.10) - (1.13) constitute a generalization of his original method.

A Jacobi method is called cyclic if in every segment of $N = n(n-1)/2$ consecutive elements of the sequence $\{\pi_k\}$ every pair (p, q) $(1 \leq p < q \leq n)$ occurs exactly once. This is a special case of a quasicyclic Jacobi method with period K where the following condition is imposed. In every segment of $K \geq N$ consecutive elements of the sequence $\{\pi_k\}$ every pair (p, q) $(1 \leq p < q \leq n)$ occurs at least once. We next define a restricted Jacobi method. Let θ_k^* be an angle satisfying (1.13) and belonging to the closed interval $[-\pi/4, \pi/4]$, and let $\psi > 0$ be a constant angle not depending on k. Then any method in which ϕ_k is chosen to satisfy (1.12) and θ_k is chosen such that

$$\text{sign } \theta_K = \text{sign } \theta_k^* \ , \ |\theta_k| = \text{Min } (|\theta_k^*| \ , \psi) \tag{1.14}$$

is called a restricted Jacobi method with bound ψ. In [4] it was proved that any quasicyclic, restricted Jacobi method with suitable bound converges. Furthermore it is shown in [4] that, if A has n distinct eigenvalues, any quasicyclic Jacobi method in which θ_k, ϕ_k satisfy (1.12), (1.13) and $\theta_k^* \epsilon [-\pi/4, \pi/4]$ converges quadratically provided the offdiagonal elements are already sufficiently small. More precisely, if $4n\mu_k \leq d$, then*

---

* Mr. E. R. Hansen has shown (oral communication) that the exponent in the exponential function may be replaced by $\sqrt{2}nKd^{-1}\mu_k$.

$$\mu_{k+K} \leq 2^{\frac{1}{2}} n(K-1) e^{1.334nK^{\frac{3}{2}} d^{-1}\mu_k} d^{-1}\mu_k^2 \quad , \tag{1.15}$$

where

$$d = \underset{r \neq c}{\text{Min}} |\lambda_r - \lambda_c| \quad .$$

The threshold cyclic Jacobi method is a modification of the cyclic Jacobi method due to Pope and Tompkins [7]. It depends on the choice of a sequence of positive threshold values $t_\nu (\nu = 0, 1, 2, \ldots)$ such that $t_{\nu+1} < t_\nu$ holds for all ν. A fixed cyclic ordering is adopted for the π_k; Φ_k always satisfies (1.12), and for each ν, one chooses

$$\theta_k = \begin{cases} \theta_k^* \ , & \text{for } |a_{ij}^{(k)}| \geq t_\nu \\ 0 \ , & \text{for } |a_{ij}^{(k)}| < t_\nu \end{cases} \quad . \tag{1.16}$$

When all $|a_{rc}^{(k)}| < t_\nu$ $(r \neq c)$, t_ν is replaced by $t_{\nu+1}$. Pope and Tompkins prove that the threshold method converges if $\lim_{\nu \to \infty} t_\nu = 0$. In the present paper we shall take the sequence $\{t_\nu\}$ to be a geometric progression; i.e., we shall assume

$$t_\nu = \alpha q^\nu \qquad (0 < q < 1) \ . \tag{1.17}$$

2. NORMS OF MATRICES. In this section we summarize for convenience of reference certain known results concerning norms of matrices. Let $H = (h_{rc})$ and G denote square matrices of order n. The norm (cf. [1]) of a square matrix H is defined to be a non-negative number $N(H)$ satisfying the conditions

$$N(H) > 0 \quad \text{for} \quad H \neq 0 \ , \tag{2.1}$$

$$N(cH) = |c| \ . \ N(H) \quad \text{for all scalars } c \ , \tag{2.2}$$

$$N(H + G) \leq N(H) + N(G) , \tag{2.3}$$

$$N(HG) \leq N(H) \cdot N(G) . \tag{2.4}$$

This definition is equivalent to that given by Ostrowski [6] for his "multiplicative norm." By (2.4)

$$N(H) = N(HI) \leq N(H) \cdot N(I) ,$$

hence

$$N(I) \geq 1 \tag{2.5}$$

for any norm. There exist norms for which $N(I) = 1$, an example being the familiar spectral norm

$$N_1(H) = \underset{x \neq 0}{\text{Max}} \left(\frac{x^*H^*Hx}{x^*x} \right)^{\frac{1}{2}} . \tag{2.6}$$

Note that for any unitary matrix U, $N_1(U) = 1$. The Euclidean norm is defined by the relation

$$N_2(H) = \left(\sum_{r,c=1}^{n} |h_{rc}|^2 \right)^{\frac{1}{2}} . \tag{2.7}$$

It is well known that, for all H,

$$n^{-\frac{1}{2}} \leq \frac{N_1(H)}{N_2(H)} \leq 1 . \tag{2.8}$$

Inequalities such as (2.8) hold quite generally. In fact, it was shown in [6] that if $N(H)$ and $N'(H)$ are any two matrix norms, then

$$a(N,N') < \frac{N(H)}{N'(H)} < b(N,N') \tag{2.9}$$

where $a(N,N')$ and $b(N,N')$ are positive constants which do not depend on H. It is easy to show that

$$\text{Max } |h_{rc}| \leq N_1(H) , \tag{2.10}$$

and by (2.9) there exists a positive constant $b(N_1,N)$, independent of H , so that

$$|h_{rc}| \leq b(N_1,N) \cdot N(H) \tag{2.11}$$

holds for all r and c and for any norm.

3. PRELIMINARY THEOREM ON INFINITE PRODUCTS. A sequence of $n \times n$ matrices $B_1, B_2, \ldots, B_m, \ldots$ with $B_k = (b_{rc}^{(k)})$ is said to be convergent if a limit exists for each element of B_m . That is, if $\lim_{m\to\infty} b_{rc}^{(m)} = b_{rc}$, then the matrix $B = (b_{rc})$ is said to be the limit of the sequence $\{B_k\}$. If the sequence $\{B_k\}$ is not convergent, it is said to be divergent. An infinite product

$$(I + E_1)(I + E_2)(I + E_3) \ldots \tag{3.1}$$

of $n \times n$ matrices is here* called convergent if the partial product

$$P_m = \prod_{k=1}^{m} (I + E_k) \tag{3.2}$$

converges to a limit. We now prove the following

THEOREM 1. Let $N(H)$ denote some norm of the matrix H . Let the factors of the infinite product (3.1) be $n \times n$ unitary matrices, and let the series $\Sigma N(E_k)$ converge. Then (3.1) converges to a unitary matrix U . Furthermore, for all m

$$N(U - P_m) \leq c(N) \sum_{k=m+1}^{\infty} N(E_k) , \tag{3.3}$$

where $c(N) > 0$ depends only on N and n .

---

* In defining convergence of infinite products of matrices in general, it is necessary to exclude nonsingular factors in the product and also to distinguish between singular and nonsingular limits. However, since we are dealing only with unitary matrices in the present paper, it is unnecessary to complicate the discussion with such distinctions.

Proof: We first establish that if the sequence $\{P_m\}$ has a limit U, then U must be a unitary matrix. Let $P_m = U + Q_m$. We have that $Q_m \to 0$ $(m \to \infty)$. Since each factor $I + E_k$ is unitary, P_m is unitary for all m. Thus

$$P_m^* P_m = I = (U + Q_m)^*(U + Q_m) ,$$

or

$$I - U^*U = U^*Q_m + Q_m^*U + Q_m^*Q_m .$$

The right hand side of the last expression tends to the null matrix as $m \to \infty$, and this shows that U is unitary. Let $r > m$ and note that

$$P_r - P_m = P_m[(I + E_{m+1}) \ \ldots \ (I + E_r) - I] . \tag{3.4}$$

We shall prove that P_m has a limit with the help of the following

LEMMA. If $H_1, H_2, \ldots, H_p$ and $G_1, G_2, \ldots, G_p$ are any two sets of $n \times n$ matrices, then the following identity holds

$$H_1H_2 \ldots H_p - G_1G_2 \ldots G_p = (H_1 - G_1)G_2 \ldots G_p \tag{3.5}$$
$$+ H_1(H_2 - G_2)G_3 \ldots G_p + \ldots + H_1 \ldots H_{p-1}(H_p - G_p) .$$

The truth of (3.5) is easily established by induction on p. The details are left to the reader. Letting $H_i = I + E_{m+i}$, $G_i = I$ and combining (3.4) and (3.5) we have

$$P_r - P_m = P_m [E_{m+1} + (I + E_{m+1})E_{m+2} \tag{3.6}$$
$$+ (I + E_{m+1})(I + E_{m+2})E_{m+3} + \ldots + (I + E_{m+1}) \ldots (I+E_{r-1})E_r] .$$

Products of the matrices $(I + E_k)$ are unitary, hence by (2.4), (3.6)

$$N_1(P_r - P_m) \leq N_1(E_{m+1}) + \ldots + N_1(E_r) , \tag{3.7}$$

since the spectral norm of a unitary matrix is one. Now for an arbitrary norm N, there exist* $a(N) > 0$ and $b(N) > 0$ such that

$$a(N) \cdot N(H) \leq N_1(H) \leq b(N) \cdot N(H)$$

holds for all H. We have therefore that

$$a(N) \cdot N(P_r - P_m) \leq N_1(P_r - P_m) \tag{3.8}$$

and

$$\sum_{k=m+1}^{r} N_1(E_k) \leq b(N) \sum_{k=m+1}^{r} N(E_k) . \tag{3.9}$$

Letting $c(N) = b(N)/a(N)$ and combining (3.7) - (3.9), we get

$$N(P_r - P_m) \leq c(N) \sum_{k=m+1}^{r} N(E_k) . \tag{3.10}$$

By hypothesis the series $\Sigma N(E_k)$ is convergent. Therefore, given any $\epsilon > 0$, there exists an integer $\mu > 0$ such that

$$\sum_{\mu+1}^{\infty} N(E_k) < \frac{\epsilon}{b(N) \cdot c(N)} . \tag{3.11}$$

We see from (3.10) and (3.11) that

$$N(P_{m+\rho} - P_m) < \frac{\epsilon}{b(N)} \tag{3.12}$$

holds for all $m \geq \mu$ and for $\rho = 1, 2, \ldots$; but (3.12) is precisely the condition that

$$\left|(P_{m+\rho} - P_m)_{rc}\right| < \epsilon \tag{3.13}$$

holds for all r and c by (2.11). Hence P_m tends to a limit U as $m \to \infty$. Finally letting $r \to \infty$ in (3.10) we obtain (3.3) and this proves Theorem 1.

---

* The spectral norm N_1 is fixed; hence, the $a(N_1, N)$ and $b(N_1, N)$ of (2.9) do not depend on it. Therefore, for simplicity, we have here written $a(N)$ for $a(N_1, N)$ and $b(N)$ for $b(N_1, N)$.

4. CONVERGENCE OF EIGENVECTORS AND ERROR ESTIMATES. We first give a rather general result about convergence and estimation of the error in all the eigenvectors. Following this we give estimates for the three Jacobi methods considered in this paper.

THEOREM 2. Let the $n \times n$ Hermitian matrix A have n distinct eigenvalues, and let the sequence of matrices (1.1) be generated by a Jacobi method such that the angles θ_k satisfy (1.13), every $\theta_k \in [-\pi/4, \pi/4]$, and such that the series

$$\sum_{k=0}^{\infty} S_k^{1/2}$$

converges. Then the infinite product (1.7) converges. Furthermore, for all $m > m_0$ we have

$$N_2(U - P_m) \leq \frac{2n^{1/2}}{d} \sum_{k=m}^{\infty} S_k^{1/2} \quad , \tag{4.1}$$

where

$$d = \min_{r \neq c} |\lambda_r - \lambda_c| \quad , \tag{4.2}$$

and where m_0 is the smallest integer such that for all $m > m_0$

$$S_m^{1/2} < \frac{d}{4} \quad . \tag{4.3}$$

Proof: We shall make use of the fact that, for some ordering of the eigenvalues λ_r of A,

$$|\lambda_r - a_{rr}^{(k)}| \leq S_k^{1/2} \quad (r = 1, 2, \ldots, n) \tag{4.4}$$

hold for all k. The inequalities (4.4) follow from Theorem 1 of [4]. Note that, while (4.4) is valid for any value of k, the ordering of the

eigenvalues λ_r is generally not the same for all k. However, our use of (4.4) does not depend on the ordering of the λ_r.

Let the U_k of (1.1) be expressed in the form

$$U_k = I + E_k \,, \tag{4.5}$$

and let $E_k = (e_{rc}^{(k)})$. By (1.2), (1.3) and (1.9) we see that the 2×2 principal submatrix of E_k corresponding to V_k is given by

$$\begin{pmatrix} e_{ii}^{(k)} & e_{ij}^{(k)} \\ e_{ji}^{(k)} & e_{jj}^{(k)} \end{pmatrix} = \begin{pmatrix} \cos\theta_k - 1 & -e^{i\phi_k}\sin\theta_k \\ e^{-i\phi_k}\sin\theta_k & \cos\theta_k - 1 \end{pmatrix}, \tag{4.6}$$

and all other elements of E_k are zero. From the definition (2.7) we get

$$N_2(E_k) = [2(\cos\theta_k - 1)^2 + 2\sin^2\theta_k]^{1/2} \,,$$

and with the help of some standard trigonometric identities, we obtain

$$N_2(E_k) = 8^{1/2}\left|\sin\frac{\theta_k}{2}\right| \,. \tag{4.7}$$

By the definitions (1.5) and (1.6)

$$\mu_k \le \frac{1}{\sqrt{2}} S_k^{1/2} \tag{4.8}$$

Hence, by (4.3) we have

$$\mu_m < \frac{1}{\sqrt{2}}\frac{d}{4} < \frac{d}{4} \qquad (m > m_o) \,. \tag{4.9}$$

Using (1.6), (1.13) and (4.2)—(4.4) it follows that

$$\left|\tan 2\theta_k\right| = \frac{2\left|a_{ij}^{(k)}\right|}{\left|a_{ii}^{(k)} - a_{jj}^{(k)}\right|} < \frac{2\mu_k}{\frac{d}{2}} = \frac{\mu_k}{\frac{d}{4}} \tag{4.10}$$

holds for $k > m_0$. Hence, by (4.9), $|\tan 2\theta_k| < 1$ and $|2\theta_k| < \pi/4$ $(k > m_0)$, since $\theta_k \in [-\pi/4, \pi/4]$. For these small angles it is easily verified that $4|\tan(\theta/2)| \leq |\tan 2\theta|$. Hence for $k > m_0$, we have from (4.7), (4.8) and (4.10)

$$N_2(E_k) = 8^{1/2}\left|\sin\frac{\theta_k}{2}\right| \leq \frac{8^{1/2}}{4}\,|\tan 2\theta_k| < \frac{8^{1/2}}{d}\,\mu_k \leq \frac{2}{d}\,S_k^{1/2} . \tag{4.11}$$

By hypothesis the series $\Sigma S_k^{1/2}$ converges. Therefore $\Sigma N_2(E_k)$ converges and by Theorem 1 the infinite product (1.7) converges. Now applying (3.3) to the case where $N = N_2$, we see from (2.8) and (2.9) that

$$c(N_2) = \frac{b(N_1, N_2)}{a(N_1, N_2)} = n^{\frac{1}{2}} . \tag{4.12}$$

Combining (3.3), (4.11) and (4.12) we obtain (4.1) and this completes the proof of Theorem 2 .

As before we let N denote the number of elements on one side of the diagonal of an $n \times n$ matrix, viz.

$$N = \frac{1}{2}\,n(n-1) , \tag{4.13}$$

and for convenience we define

$$h = (1 - N^{-1})^{\frac{1}{2}} . \tag{4.14}$$

We let $[x]$ denote the integral part of x , i.e., the largest integer not exceeding x . Finally we let $\sigma = [N/q^2]$ and we state our main result as follows.

THEOREM 3. Let the $n \times n$ Hermitian matrix A have n distinct eigenvalues, and let the sequence of matrices (1.1) be generated by

either Method (a): the classical Jacobi method, Method (b): a quasicyclic restricted Jacobi method, or Method (c): a threshold cyclic Jacobi method in which the thresholds t_ν satisfy (1.17). Then the infinite product of matrices (1.7) converges. Furthermore, for Method (a)

$$N_2(U - P_m) \leq \frac{2}{d}(nS_o)^{\frac{1}{2}} \frac{h^{m-1}}{1-h} \qquad (m > m_o) \quad , \tag{4.15}$$

for Method (b)

$$N_2(U - P_m) \leq \frac{4Kn^{\frac{s}{2}}}{d} \mu_m \qquad (m > \rho) \quad , \tag{4.16}$$

where ρ is given by (4.23) below, and for Method (c)

$$N_2(U - P_m) \leq \frac{2\sigma}{d} q^{[m/\sigma]} (2nN\mu_o)^{\frac{1}{2}} \frac{1}{1-q} \qquad (m > m_o) \quad , \tag{4.17}$$

where in each case the integer m_o has the same meaning as it had in Theorem 2.

Proof: It is easy to show that, for any step in a Jacobi method in which the parameters θ_k, Φ_k satisfy (1.12) and (1.13), the following equation holds

$$S_{k+1} = S_k - 2|a_{ij}^{(k)}|^2 \quad . \tag{4.18}$$

In the classical Jacobi method we have

$$|a_{ij}^{(k)}|^2 \geq \tfrac{1}{2}N^{-1} S_k \tag{4.19}$$

always satisfied. Hence

$$S_{k+1} \leq (1 - N^{-1})S_k = h^2 S_k \qquad (k = 0, 1, 2 \ldots) \tag{4.20}$$

holds for Method (a). It follows therefore that

$$\sum_{k=0}^{\infty} S_k^{1/2} = S_0^{1/2} (1 + h + h^2 + \ldots) = S_0^{1/2} \frac{1}{1-h} \quad , \tag{4.21}$$

hence by Theorem 2 the product (1.7) converges. By (4.20)

$$\sum_{k=m}^{\infty} S_k^{1/2} = S_0^{1/2} \frac{h^{m-1}}{1-h}$$

holds for all m. For $m > m_0$ we may combine the last expression with (4.1) and obtain the inequality (4.15).

Now consider Method (b). We note that $\mu_k \to 0$ as $k \to \infty$, since a quasicyclic restricted Jacobi method is convergent. Since A has n distinct eigenvalues, there exists a smallest integer p_1 such that

$$\text{Min } (|\theta_k^*|, \psi) = |\theta_k^*| \qquad (k \geq p_1) \ ,$$

and by (1.14) the restricted method behaves like an ordinary quasicyclic method from the step p_1 on. There also exists a smallest integer $p_2 \geq p_1$ such that $4n\mu_k \leq d$ $(k \geq p_2)$ and hence (1.15) holds for $k \geq p_2$. Since the sequence $\{S_k^{1/2}\}$ is monotone decreasing and $\mu_k \to 0$, there is a smallest integer p_3 such that (4.3) is satisfied, viz.

$$S_k^{1/2} \leq \frac{d}{4} \qquad (k \geq p_3) \ ,$$

and therefore (4.11) holds for $k \geq p_3$. Finally, a smallest integer p_4 exists such that

$$\mu_k < \tfrac{1}{2}g^{-1} \qquad (k \geq p_4) \ , \tag{4.22}$$

where g is the coefficient of μ_k^2 in (1.15). Now let

$$\rho = \text{Max } (p_2, p_3, p_4) \ . \tag{4.23}$$

For $k \geq \rho$ we have that (1.15) and (4.11) hold, hence

$$\sum_{k=\rho}^{\infty} N_2(E_k) < \frac{8^{1/2}}{d} \sum_{k=\rho}^{\infty} \mu_k \ ,$$

$$= \frac{8^{1/2}}{d} \sum_{i=0}^{K-1} \sum_{j=0}^{\infty} \mu_{\rho+i+jK} \ ,$$

$$\leq \frac{8^{1/2}}{d} \{(\mu_\rho + g\mu_\rho^2 + g^3\mu_\rho^4 + \ldots)$$

$$+ (\mu_{\rho+1} + g\mu_{\rho+1}^2 + g^3\mu_{\rho+1}^4 + \ldots) + \ldots$$

$$+ (\mu_{\rho+K-1} + g\mu_{\rho+K-1}^2 + g^3\mu_{\rho+K-1}^4 + \ldots)\} \ ,$$

$$\leq \frac{8^{1/2}}{dg} \left(\frac{g\mu_\rho}{1-g\mu_\rho} + \frac{g\mu_{\rho+1}}{1-g\mu_{\rho+1}} + \ldots + \frac{g\mu_{\rho+K-1}}{1-g\mu_{\rho+K-1}} \right),$$

and using (4.22) we get

$$\sum_{k=\rho}^{\infty} N_2(E_k) \leq \frac{4\sqrt{2}}{d} (\mu_\rho + \mu_{\rho+1} + \ldots + \mu_{\rho+K-1}) \ . \tag{4.24}$$

Since S_k is monotone decreasing,

$$\mu_{\rho+\nu} \leq \frac{1}{\sqrt{2}} S_{\rho+\nu}^{1/2} \leq \frac{1}{\sqrt{2}} S_\rho^{1/2} \leq \sqrt{N}\mu_\rho \leq \frac{n}{\sqrt{2}} \mu_\rho \ (\nu = 0, 1, 2, \ldots),$$

we combine this with (4.24) to obtain

$$\sum_{k=\rho}^{\infty} N_2(E_k) \leq \frac{4nK}{d} \mu_\rho \ . \tag{4.25}$$

Now combining (3.3), (4.12) and (4.25) we get (4.16), and since $\mu_m \to 0$ $(m \to \infty)$ this proves the convergence of (1.7) for Method (b).

Turning our attention to Method (c), we see by (1.16) that $U_k = I$, $E_k = 0$ whenever $|a_{ij}^{(k)}| < t_\nu$ so that a step in the process (1.1) counts only when a non-trivial transformation is made, i.e., whenever

$|a_{ij}^{(k)}| \geq t_\nu$. We let the scalar α in (1.17) be equal to μ_o so that

$$|a_{rc}| \leq \mu_o = \mu_o q^o = t_o \quad (r \neq c) .$$

There may be no non-trivial transformations made at any given threshold level, but if there are, we can easily bound their number. Assume that after the r-th transformation $\mu_r < t_{\nu-1}$. Then

$$S_r < 2Nt_{\nu-1}^2 , \tag{4.26}$$

and by (4.18)

$$S_{k+1} \leq S_k - 2t_\nu^2 \tag{4.27}$$

holds for each transformation made at the ν-th threshold level, since $|a_{ij}^{(k)}| \geq t_\nu$. Combining (4.26) and (4.27) we see that

$$S_{r+p} \leq 2Nt_{\nu-1}^2 - 2pt_\nu^2$$

holds after p transformations at the ν-th level. Letting $[x]$ denote the integral part of x , we have that there are at most

$$\sigma_\nu = \left[\frac{Nt_{\nu-1}^2}{t_\nu^2}\right]$$

steps at the ν-th level, for

$$S_{r+\sigma_\nu} < 2Nt_{\nu-1}^2 - 2\left[\frac{Nt_{\nu-1}^2}{t_\nu^2}\right]t_\nu^2 ,$$

$$< 2Nt_{\nu-1}^2 - 2\left(\frac{Nt_{\nu-1}^2}{t_\nu^2} - 1\right)t_\nu^2 ,$$

$$< 2t_\nu^2 ,$$

so that $\mu_{r+\sigma_\nu} < t_\nu$, i.e., the next threshold level has been reached.

From (1.17)

$$\sigma_\nu = \left[\frac{N}{q^2}\right] = \sigma \qquad (\nu = 1, 2, \ldots) .$$

We have proved that

$$S_{\nu\sigma} \leq 2Nt_\nu^2 \qquad (\nu = 1, 2, \ldots) ,$$

so that, using (1.17)

$$S_{\nu\sigma}^{1/2} \leq (2N\mu_o)^{1/2} q^\nu . \tag{4.28}$$

Now, since $s[t/s] \leq t$ for all positive integers s and t, we may write

$$\sum_{k=m}^{\infty} S_k^{1/2} \leq \sum_{k=m}^{\infty} S_{\sigma\left[\frac{k}{\sigma}\right]}^{1/2} , \tag{4.29}$$

$$\leq \sum_{k=\sigma\left[\frac{m}{\sigma}\right]}^{\infty} S_{\sigma\left[\frac{k}{\sigma}\right]}^{1/2} .$$

Let i be one of the integers $0, 1, \ldots, \sigma-1$ and let

$$k = \sigma\left[\frac{m}{\sigma}\right] + i + j\sigma .$$

We have

$$\left[\frac{k}{\sigma}\right] = \left[j + \left[\frac{m}{\sigma}\right] + \frac{i}{\sigma}\right] ,$$

and since $i < \sigma$, we have

$$\left[\frac{k}{\sigma}\right] = j + \left[\frac{m}{\sigma}\right] \qquad (i = 0, 1, \ldots, \sigma-1) .$$

Thus we may write (4.29) as follows

$$\sum_{k=m}^{\infty} S_k^{1/2} \leq \sum_{i=0}^{\sigma-1} \sum_{j=0}^{\infty} S_{\sigma\left(j + \left[\frac{m}{\sigma}\right]\right)}^{1/2} ,$$

$$= \sigma \sum_{j=0}^{\infty} S_{\sigma\left(j + \left[\frac{m}{\sigma}\right]\right)}^{1/2} ,$$

and using (4.28) we have

$$\sum_{k=m}^{\infty} S_k^{\frac{1}{2}} \leq \sigma(2N\mu_o)^{\frac{1}{2}} \sum_{j=0}^{\infty} q^{j+\left[\frac{m}{\sigma}\right]} , \qquad (4.30)$$

$$= \sigma q^{\left[\frac{m}{\sigma}\right]} (2N\mu_o)^{\frac{1}{2}} \frac{1}{1-q} .$$

Finally, combining (4.1) with (4.30) we get immediately the expression (4.17). Obviously,

$$q^{\left[\frac{m}{\sigma}\right]} \to 0 \qquad (m \to \infty) ,$$

and thus the product (1.7) converges also for Method (c). This completes the proof of Theorem 3.

5. REMARKS. If we refer to unitary matrices of the general type used in the process (1.1) as <u>elementary unitary matrices</u>, then it is a rather easy matter to show that an arbitrary unitary matrix can be decomposed into a product of a finite number of elementary unitary matrices. In [3] (footnote (5), page 8) Givens gave a proof that every orthogonal matrix with positive determinant can be expressed as a product of a finite number of elementary orthogonal matrices. We shall not give the details here, but a proof of the corresponding result on the decomposition of an arbitrary unitary matrix can be constructed by a method very much the same as that used in [3].

In spite of the result just mentioned, it is perhaps not without some interest that we can use Theorem 3 to prove that an arbitrary unitary matrix can be expressed as an <u>infinite product</u> of elementary

unitary matrices. The decomposition is, of course, not unique. We sketch a proof as follows.

Suppose W is any $n \times n$ unitary matrix. Let D be a fixed $n \times n$ diagonal matrix with n distinct and real diagonal elements. Let the Hermitian matrix

$$A = WDW^* \tag{5.1}$$

be diagonalized by (say) Method (a) of Theorem 3, so that

$$U^*AU = D' ,$$

where U is the limit of the infinite product (1.7). The matrix D' has the same diagonal elements as D but possibly with a different arrangement. By a finite number of elementary unitary transformations with V_k of the form

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} ,$$

we can transform D' into D, so that

$$V^*AV = D , \tag{5.2}$$

where V is an infinite product of elementary unitary matrices. From (5.1) and (5.2) we see that

$$VDV^* = WDW^* . \tag{5.3}$$

It is easy to show from (5.3) that W can be expressed as VY where Y is a <u>diagonal</u> unitary matrix, i.e.

$$y_{rr} = e^{i\alpha_r} \quad (r = 1, 2, \ldots, n) . \tag{5.4}$$

Since any diagonal matrix Y satisfying (5.4) can be decomposed as a

product of at most $n-1$ elementary unitary matrices with V_k of the form

$$\begin{pmatrix} e^{i\beta} & 0 \cdot \\ 0 & e^{i\beta} \end{pmatrix},$$

the desired result is established.

Section 6
REFERENCES

1. V. N. Faddeeva, Vychislitel'nye Metody Lineĭnoĭ Algebry, Gostekhizdat, Moscow-Leningrad, 1950. (Translated by Curtis D. Benster under title Computational Methods of Linear Algebra, Dover, New York, 1959

2. G. E. Forsythe and P. Henrici, "The Cyclic Jacobi Method for Computing the Principal Values of a Complex Matrix," Trans. Amer. Math. Soc., (to be published)

3. W. Givens, Numerical Computation of the Characteristic Values of a Real Symmetric Matrix, Oak Ridge National Laboratory, Rep. ORNL 1574, 1954. (Purchasable from Office of Technical Services, U. S. Dept. of Commerce, Washington 25, D. C.)

4. P. Henrici, "On the Speed of Convergence of Cyclic and Quasicyclic Jacobi Methods for Computing Eigenvalues of Hermitian Matrices," J. Soc. Indust. Appl. Math., Vol. 6, 1958, pp. 144–162

5. C. G. J. Jacobi, "Über ein leichtes Verfahren, die in der Theorie der Säkularstörungen vorkommenden Gleichungen numerisch aufzulosen," J. Reine Angew. Math., Vol. 30, 1846, pp. 51–95

6. A. Ostrowski, "Über Normen von Matrizen," Math. Z., Vol. 63, 1955, pp. 2–18

7. D. A. Pope and C. B. Tompkins, "Maximizing Functions of Rotations--Experiments Concerning Speed of Diagonalization of Symmetric Matrices Using Jacobi's Method," J. Assoc. Comput. Mach., Vol. 4, 1957, pp. 459–466

WORD-CORRELATION STUDY

3

R.P. Mitchell

FOREWORD

This document, originally published as LMSD 311696, September 1959, was the final report for Word Correlation Study, Contract No. AF 30(602)-1889, Rome Air Development Center.

The author would like to express his appreciation to members of Applied Mathematics for their suggestions and encouragement in this research.

ABSTRACT

A method of syntactic analysis of declarative English sentences, proposed by Bar-Hillel and Lambek, was operationally extended in such a manner that the product of the analysis is a machine retrieval language. Syntactic analysis thus becomes a recursive translation from the natural language to the proposed machine language. This method is called "microsyntactic analysis," and the smallest meaningful linguistic units, or morphemes, of the retrieval language are termed "microsyntactic indices."

A three-value storage logic was devised for storage and retrieval operations. Various relations between syntactic strings and algebraic strings were defined. A characterization of semantic correspondences in terms of algebraic relations was proposed and is currently under study.

CONTENTS

Section 1

INTRODUCTION

The first quarter of research on this contract resulted in a proposal for studying word correlation in a sublanguage of English and presented a method by which correlations could be computed for the given sublanguage The method of syntactic analysis conceived first by Y. Bar-Hillel [5] and studied by J. Lambek [6] was arbitrarily extended to a larger part of the language than intended by these authors. This extension, termed by us "macrosyntax" because it provides a basis for classification of lexical units into function-sets, may not be justified on logical grounds; it was taken only as a working basis for further research.

The second and third quarters of research resulted in a refinement of syntactic analysis called "microsyntax." The chief advance consisted in the abstraction of syntactic forms from semantic and/or functional interpretations given them in the macrosyntax, and opening the way to mathematical characterizations of syntactic forms and their construction. Microsyntactic indexing was also introduced, providing a means for very fine semantic destinctions between words whenever such interpretations are desired by the analyst. Concurrently with these studies, we made serious efforts to find number-theoretic analogues of syntactic structure in the hope that syntactic analysis could be accomplished on standard digital computers. In this we were only partially successful, and these efforts had been abandoned at the close of the third quarter.

A list of basic English words was abstracted from West [7], their syntactic forms listed, and frequencies of occurrence per million

computed from data given by West [7]. This list was computer-programmed for future work; the program is described in the appendix of this report.

The fourth quarter of research has resulted in the conception and outline of a storage algebra of microsyntactic indices. The most important result of this work is the characterization of semantic correspondences in terms of relations defined in this algebra. We feel that this characterization gives us a firm basis for the derivation of a correlation function.

Section 2

A GENERALIZED SYNTAX FOR L

In [3] we presented a method for indexing strings of words which are analyzable in terms of a microsyntax of English. This method has been generalized and is presented here in a more systematic form. In the reference cited we defined heuristically the order of a syntactic form as the highest number of strokes separating elements of the form. It will be observed that forms of order greater than zero are constructed by repeated application of left or right stroke functions. Thus, we need only two mappings and the concept of order of syntactic forms to characterize all possible constructions. In [2], when we defined the microsyntax of L, we defined two sets of elements, $A = \{n_1, s_1, e_1\}$ and $B = \{n_2, s_2, e_2\}$, and constructed all elements of the microsyntax from these sets by means of appropriate mappings. The indicial feature of this method was extended in [3], and the method of microsyntactic indexing was introduced. In the process of systematizing the method of syntactic analysis originally proposed by Bar-Hillel in [4], we have abstracted the constructive properties of the method and discarded the interpretations of strings of these elements as "noun phrases," "verb phrases," "sentences," etc. These interpretations may serve a convenient classification purpose in descriptive semantics, but they are meaningless syntactically. In microsyntactic indexing, n and s serve an indicial function only, and may be replaced by any convenient symbol. Let Z be the set of zero-order indices. Indices of higher order may, in general, be either entirely associative or non-associative between certain elements. We denote by X_n the set constructed by n consecutive applications of the left stroke function on elements of Z; thus, X_1 is the set of elements of the form $a\backslash b$, $X_2 = \{a\backslash b\backslash c\}$, etc. Similarly, we denote by Y_m the

set of elements formed by m successive applications of the right stroke function. All members of X_n and Y_m are first-order and entirely associative. Forms containing elements which are non-associative relative to a left or a right stroke are constructed from the sets X_n and Y_m. These forms are the second-order forms, and are in turn used in the construction of third and higher order forms. We denote the many possible sets of second-order forms by $X^{m,n,p,\ldots}_{i,j,k,\ldots}$ and $Y^{m,n,p,\ldots}_{i,j,k,\ldots}$ where X or Y indicates that the left or right stroke, respectively, is the stroke function relative to which the elements are non-associative; and the elements of the superscript sequence indicate the number of right or left strokes, respectively, in the forms following the forms indicated in the paired subscript sequence. For example, $X^2_{1,0} = \{a\backslash b\backslash\backslash c/d/e\backslash\backslash f\}$, $Y^{1,1} = \{a\backslash b/\ /c\backslash d\}$, $Y^{1,1}_1 = \{a/b//c\backslash d//e\backslash f\}$, $X^{1,0}_{,1} = \{a/b\backslash\backslash c\backslash d\backslash\backslash e\}$, etc. Most of the possible second-order forms do not occur in practice; among these are the forms which are non-associative relative to both stroke functions. Where such forms should be necessary, and for forms of higher order, it is notationally simpler to write the required sets in terms of the left and right strokes, as was done in the above examples. A slightly more cumbersome notation is to define the stroke functions as functions of ordered pairs of sets (as we have implicitly done above), then write the actual function as required. For example, an element belonging to $Y^{1,1}$ would be written $\rho[\lambda(a,b), \lambda(c,d)]$, where λ and ρ are the left and right stroke functions respectively, and a, b, c, d are all elements of Z. Some convention must be used to characterize association between elements; capital Greek letters, Λ and P, may be used for this purpose.

We require an identity element $e \in Z$, such that

$$\lambda(e,a) = \rho(a,e) = a$$

where a is an element of any set. Sequences of indices are said to "reduce" to an index when any of the following sequences occur:

$$\lambda(a,b)\ \lambda\ (b,c) = \lambda\ (a,c)$$
$$\Lambda(a,b)\ \Lambda\ (b,c) = \Lambda\ (a,c)$$
$$\rho(a,b)\ \rho\ (b,c) = \rho\ (a,c)$$
$$P(a,b)\ P\ (b,c) = P\ (a,c)$$

where a, b, c are elements of any set. We define a "syntactic string" as any order sequence of indices. A string is said to reduce to an index when sequential pairs of indices reduce according to the transitive property of the stroke functions given above. Conversely, an expansion of any index is any string obtained by the converse application of the transitive property, providing the indices used in the expansion are members of sets required for analysis of the language. In particular, if q is any index, we denote by $\dashv q$ the set of allowable expansions of q .

Section 3

A STORAGE-DEPENDENT ALGEBRA OF INDICES

The central idea in developing a storage-dependent algebra of indices is the assignment of values to indices and algebraic strings of indices according to the results of a storage-verification procedure. We call the values thus assigned "s-values" for the indices, and denote them by "0," "1," and "δ." It is assumed that the storage of indices is a continually changing storage, but that storage-verification occurs for any fixed interval while the storage is static. The s-values 0, 1 are interpreted to mean negative and positive results of the verification procedures, while δ is interpreted as the (intentional or unintentional) non-application of the verification-procedure.

The stipulation of s-values for algebraic strings of indices is to some extent arbitrary, depending upon the purpose of the algebraic analysis. Once this purpose has been clarified, then the algebraic relations among indices can be defined in terms of assignments of s-values for the primitive relations. For the heuristic reason that, for a sufficiently large collection of sentences in storage, the negations of a large number of words and phrases are likely also to be in storage, we define, for each index p in L, an index $-p$ in L such that the following conditions of storage verification hold; if $p = 1$, $-p = 1$; if $p = 0$, $-p = \delta$; and if $p = \delta$, $-p = 0$. The element $-p$ is the antonym of p in a descriptive semantics.

We define a binary relation D between indices of L in terms of the following stipulation of s-values.

p	q	pDq
0	0	0
0	1	1
0	δ	δ
1	0	1
1	1	1
1	δ	1
δ	0	δ
δ	1	1
δ	δ	δ

We also define a binary relation C among indices of L such that C satisfies the identity pCq = - (-pD-q). From the definitions it follows that pCq satisfies the following table of s-values.

p	q	pCq
0	0	0
0	1	1
0	δ	0
1	0	1
1	1	1
1	δ	1
δ	0	0
δ	1	1
δ	δ	δ

Examination of the s-table for pDq shows that

$$pDq = qDp \ . \tag{1}$$

Further, from the definition for -p , we have

$$-(-p) = p \ . \tag{2}$$

Using the definition for C and (1), we have

$$pCq = qCp \ . \tag{3}$$

Through the s-tables, it can be shown that

$$pD(qDr) = (pDq)Dr \ , \tag{4}$$

and it is easily proved that

$$pC(qCr) = (pCq)Cr \tag{5}$$

Again through the s-tables, the following property may be verified:

$$pC(qDr) = (pCq)D(pCr) \tag{6}$$

We have shown that, for the given stipulation of s-values the relations C and D are commutative, associative, and distributive with respect to C .

Algebraic strings are linearly related to the stroke functions λ and ρ . Let x, y, z be any indices such that, if α denotes either λ or ρ , $\alpha(x,y)$, $\alpha(x,z)$, $\alpha(y,z)$ are all indices in L . Then we define:

$$\alpha(xCy,z) = \alpha(x,z)C\alpha(y,z) \tag{7}$$
$$\alpha(x,yCz) = \alpha(x,y)C\alpha(x,z) \tag{8}$$
$$\alpha(xDy,z) = \alpha(x,z)D\alpha(y,z) \tag{9}$$
$$\alpha(x,yDz) = \alpha(x,y)D\alpha(x,z) \tag{10}$$

Since C and D are commutative, we have

$$\alpha(xCy,z) = \alpha(yCx,z) \tag{11}$$
$$\alpha(x,yCz) = \alpha(x,zCy) \tag{12}$$
$$\alpha(xDy,z) = \alpha(yDx,z) \tag{13}$$
$$\alpha(x,yDz) = \alpha(x,zDy) \tag{14}$$

We define

$$\alpha(-x,z) = \alpha(x,-z) = -\alpha(x,z) \tag{15}$$

and show that this definition is consistent with (7) - (10).

$$\alpha(xCy,z) = \alpha[-(-xD-y),z]$$
$$= -\alpha(-xD-y,z) = -[\alpha(-x,z)D\alpha(-y,z)]$$
$$= -[-\alpha(x,z)D-\alpha(y,z)] = \alpha(x,z)C\alpha(y,z),$$

and similarly for $\alpha(x,yCz)$. From (15), we have the identies:

$$\alpha(xCy,z) = \alpha(-xD-y,-z) \tag{16}$$
$$\alpha(xDy,z) = \alpha(-xC-y,-z) \tag{17}$$
$$\alpha(x,yCz) = \alpha(-x,-yD-z) \tag{18}$$
$$\alpha(x,yDz) = \alpha(-x,-yC-z) \tag{19}$$

Further, from (7) - (10) and associativity, we have

$$\alpha(uCv,xCy) =$$
$$\alpha(u,x)C\alpha(u,y)C\alpha(v,x)C\alpha(v,y) \tag{20}$$
$$\alpha(uDv,xDy) = \alpha(u,x)D\alpha(u,y)D\alpha(v,x)D\alpha(v,y) \tag{21}$$

Indices containing both C and D are allowable if and only if all indices in the resulting decomposition have the same s-value. To prove this take, for example, the index $\alpha(uCv,xDy)$. We have

$$\alpha(uCv,xDy) = \alpha[-(uCv), -xC-y] = -\alpha(uCv, -xC-y) =$$
$$-[-\alpha(u,x)C-\alpha(u,y)C-\alpha(v,x)C-\alpha(v,y)]$$

On the other hand, we also have

$$\alpha(uCv, xDy) = \alpha[-uD-v, -(xDy)] = -\alpha(-uD-v, xDy) =$$
$$-[-\alpha(u,x)D-A(u,y)D-\alpha(v,x)D-\alpha(v,y)]$$

The two decompositions can be equal only if all indices in the strings are either 0, δ, or 1.

From the transitive property of the stroke functions, we have

$$\alpha(xCy,z)\alpha(z,u) = \alpha(xCy,u) \tag{22}$$

Since $\alpha(xCy,z) = \alpha(x,z)C\alpha(y,z)$, the sequence of indices must be distributive with respect to C and D ; thus

$$\alpha(xCy,z)\alpha(z,u) = [\alpha(x,z)C\alpha(y,z)]\alpha(z,u) =$$
$$\alpha(x,z)\alpha(z,u)C\alpha(y,z)\alpha(z,u) = \alpha(x,u)C\alpha(y,u) =$$
$$\alpha(xCy,u)$$

However, $\alpha(u,xCy)\alpha(xCy,v) = \alpha(u,v)$ if and only if $x = y$, and similarly if C is replaced by D in one or both indices.

Using (2) and (15), we have

$$\alpha(-x,y)\ \alpha\ (y,-z) = \alpha(x,-y)\ \alpha\ (-y,z) =$$
$$[-\alpha(x,y)]\ [-\alpha(y,z)] = \alpha(x,z) \qquad (23)$$

So far, we have not used any property of δ in our results, and hence the definitions and consequent properties of both algebraic and syntactic strings hold for a Boolean algebra as well. An interesting application of these results would be the analysis of the predicate calculus in terms of this theory; however, we have only fragmentary results on this application, and shall not include them here.

Section 4
SYNTACTIC THEORY OF DEFINITIONS

It was proposed, in the first report on this study [1], to compile a list of words to be taken as undefined and define all other words in the general lexicon in terms of the primitive words. The novel feature of our proposal was to define words in such a manner that the string of the definition is an allowable expansion of the syntactic form of the word defined. This feature would permit free substitution of definitions in the context of the words defined, and hence would permit correlation of sentences. The theory of definition is the heart of descriptive semantics, and it is in this area of research that any method or proposal for word correlation must be conceived. As was indicated in [4], we have encountered considerable difficulty in writing satisfactory definitions within the proposed syntactic restrictions. A "satisfactory" definition is simply a matter of the conscience and experience of the person writing the definition, and part of the difficulty lies herein--we have no criteria for an acceptable definition. The other part of the difficulty is that the theory of definition at this level is not properly a scientific problem, even for those who choose to recognize it as a problem. It is a philosophical problem in the present state of our knowledge of human behaviour and communication, and subject to endless debate.

The purpose of a definition in a natural language is to establish a correspondence between a word and whatever the word names or whatever function it serves in the language. The manner in which this purpose is accomplished is by stating a correspondence between the word and a collection of other words in the language, placing reliance on the human interpreter to perform the actual process of naming objects of experience,

and whatever other functions are required. Since the construction of a dictionary or thesaurus involves these complex human activities to a sophisticated degree, and since in our present state of knowledge these activities are so little understood, the proper objects of our research are simply those correspondences between words and collections of words which are already established in conventional dictionaries, rather than the construction of a new dictionary. We have limitations imposed by the method of syntactic analysis, of course, and these restrictions will induce changes in the precise wording of definitions. But we can study the problems in abstracto, and the results obtained ought to tell us whether any useful concept of word correlation is feasible.

Generally, a definition is a correspondence between a syntactic string and a sequence of syntactic strings. We consider here the correspondence between a single word and a single string, and write $p \rightarrow \sigma$ for the correspondence, where p is the word defined and σ the string constituting the definition. p denotes a set of indices each of which stands in the given relation to the string σ. Since we have no criterion for substitutability of σ in any string containing an index belonging to p, the relation $\rightarrow$ is not an equivalence relation. We define the equivalence relation $\equiv$ between indices as follows:

$$p \equiv \sigma \text{ if and only if } p \rightarrow \sigma \text{ and } \sigma \rightarrow p ,$$

the relation holding for all indices belonging to p. We define the relation $\rightarrow$ in terms of the storage-dependent algebra:

$$p \rightarrow \sigma = - pD\sigma \qquad (24)$$

The relation $\rightarrow$ may hold for any two indices in L. In the thesaurus-type of definition, where a word is "defined" in terms of a collection of words, the relation holds between the word defined and each word in the collection. From the definitions, we have the following table of s-values for $\rightarrow$:

p	q	$p \rightarrow q$
0	0	δ
0	1	1
0	δ	δ
1	0	1
1	1	1
1	δ	1
δ	0	0
δ	1	1
δ	δ	δ

The first set of values reflects the fact that the semantic correspondence between any two indices does not depend upon the storage of the indices. Further, if $q = \sigma$, a string, then the correspondence may implicitly be in storage through the words in δ, though σ itself need not be in storage.

The semantic correspondence relation has the following elementary properties:

$$p \rightarrow q = -(pC\text{-}q) \tag{25}$$

$$\alpha(x, y \rightarrow z) = \alpha(x,y) \rightarrow \alpha(x,z) \tag{26}$$

$$\alpha(x \rightarrow y, z) = \alpha(x,z) \rightarrow \alpha(y,z) \tag{27}$$

The form $\alpha(x \rightarrow y, u \rightarrow v)$, if it should occur, implies $x = y$ and $u = v$; this is easily seen by application of (26) and (27).

The purpose of the three-valued algebra we have outlined is to provide a basis for semantic correlation relative to a given set of correspondences between words and strings of words. It is our intention to define the s-value δ in such a manner that it provides a numerical measure of correlation relative to the given correspondences. This function will also provide a basis for semantic inference in L relative

to that sub-language of L which is in storage. Our speculations regarding the form of this numerical function center around statistical properties of indices subject to the syntactic restraints of acceptable strings and to the imposed structure of the storage algebra. Results obtained at this time are too indefinite for inclusion in this report.

Section 5

TECHNICAL SUMMARY AND CONCLUSIONS

We present in this section a summary of the more important technical results obtained for this contract.

The subject of analysis is the English language which occurs in meaningful discourse; in particular, the analysis of meaningful declarative sentences. The syntax of this language was analyzed by the method proposed by Bar-Hillel [5] and Lambek [6]. The result was a sublanguage of the original language consisting of sentences for which the method applies and the words which occur in those sentences. This method of analysis was extended operationally to include a larger set of sentences for experimental purposes. This extension is termed "macrosyntax" in our reports.

One of the weaknesses of the macrosyntax is the ambiguity of association in the more complex forms. For example, the form $n\backslash s/n$ can be taken as either of the forms $(n\backslash s)/n$ or $n\backslash(s/n)$. A resolution of this ambiguity was found by defining two semigroups $(A,/)$ and $(B,\backslash)$, with $A = \{n_1, s_1, e_1\}$ and $B = \{n_2, s_2, e_2\}$, then defining mappings which carry cross-products of the two sets into sets whose members are the non-associative forms. This syntax was termed "microsyntax."

It was then recognized that the indicial feature of the microsyntax could be used to index words in context. Thus, n_1, n_2, n_1/n_1, $n_1\backslash s_2$, $n_2\backslash s_2$, s_1, s_2, etc. are all sets whose elements are words or strings of words in the natural language used in particular sentences with meanings intended in those sentences. This method of indexing was called "microsyntactic

indexing," and represents a recursive translation of the natural language into a machine language suitable for storage and retrieval purposes. The analysis of the syntax of the natural language, with this recursive translation into the microsyntactic indices, was felt to be an essential part of the groundwork for word correlation. This is because the meanings of words depend upon their usage and functions in context, and a method for preserving context while simultaneously permitting algebraic operations with individual words was essential. We feel that microsyntactic indexing satisfies this technical requirement.

An algebra suitable for operations among indices in storage is outlined in this report. A three-value algebra was chosen so that one of the values could be used as a correlation function between any pair of indices. Relations between the stroke functions of the microsyntax and the algebraic relations were defined and simple consequences of the definitions were derived. A characterization of the correspondence between a word and a string which defines the word was proposed and defined in terms of the algebra. This is felt to be the most important technical result of our research, as it is through the analysis of these correspondences that the theory of word correlation must be constructed.

To each word in the natural language there corresponds a set of microsyntactic indices, at most one index for each occurrence of the word in the material analyzed. In addition, each sentence which is analyzable in the microsyntax is indexed. The storage can be manipulated at will using a Boolean algebra; for example, various hypotheses could be constructed and consequences of the hypotheses derived automatically by machine. However, the technique of using a Boolean algebra presupposes a precise knowledge of the contents of the storage, and when storage is large, the Boolean algebra becomes impractical. An automatic program for word correlation is essential as a guide for the analyst. If a correlation function were

available its application would tell the analyst in which direction to proceed in storage manipulation without reading out the entire storage. In other words, the correlation function serves as a tool for plausible inference; from the value of the correlation function in each particular instance the analyst is able to infer the contents of storage and construct the Boolean search function accordingly. These are the reasons we have chosen a three-value storage algebra.

It is unclear at present what form the correlation function must have to be useful in the intended way. It must be independent of the cardinality of index sets corresponding to words, yet it seems reasonable that the function be related to the probability that a given word is in storage. On the other hand, since we are seeking a semantic correlation function, we have no a priori reason to assume that probability enter the picture at all. A scheme for computing word correlation similar to the simple ideas originally proposed in [1] may serve the intended purpose, with suitable criteria for substitution.

Section 6

REFERENCES

1. Lockheed Missiles and Space Division, Word Correlation Study, I, by R. P. Mitchell and E. Greer, Contract No. AF 30(602)-1889, Sunnyvale, Calif.

2. Lockheed Missiles and Space Division, Word Correlation Study, II, by R. P. Mitchell, Contract No. AF 30(602)-1889, Sunnyvale, Calif.

3. Lockheed Missiles and Space Division, Word Correlation Study, III, by R. P. Mitchell, Contract No. AF 30(602)-1889, Sunnyvale, Calif.

4. Lockheed Missiles and Space Division, Proposal for Continuation of Word Correlation Study, Contract No. AF 30(602)-1889, LMSD-49901, Sunnyvale, Calif., 27 Jul 1959.

5. Bar-Hillel, Y. A., "Quasi-Arithmetical Notation for Syntactic Description," Language, pp. 47-58, 1953.

6. Lambek, J., "The Mathematics of Sentence Structure," American Mathematical Monthly, pp. 154-170, Mar 1958.

7. West, M. P., A General Service List of English Words, New York, Longmans, Green and Co., 1957.

Appendix A

WORD CORRELATION STUDY PROGRAM

General Description:

Given a list of words with their frequencies of occurrence in five million words and the percentages of occurrence as different syntactic types, this program will produce a readable list consisting of the words, their frequencies of occurrence in one million words, and the frequencies of occurrence as the different syntactic types. Where the percentage is unknown, the program will provide the number — 1111 1111 - 11 as an indicative symbol that the word is known to occur as a certain syntactic type but the programmer must put in the symbol in the desired location in the output buffer. See Note 2.

The program requires all of first core and up to location 10240 of second core. The remainder of second core may be used to enlarge the program or make comparisons between the words and/or numbers associated with them.

Output Format:

Each 'page' is headed by the page number and five words. The syntactic types are represented by numbers from 30 to 60, and appear to the left. Each word heads a column of numbers, the first of which is the frequency per million words, and the others, opposite the syntactic types, represent the occurrence of the word as the various types.

Input:

Input to the program is by tape. For every page, the five words to be printed on that page must be punched on one card in the following format:

column 1	begin first word
column 19	second word
column 37	third word
column 55	fourth word
column 73	fifth word

(Note: If the fifth word has more than eight letters, start in the number of column to the left of column 73 that equals the number of letters greater than eight.) e.g., a nine-letter word must start in column 72; a twelve-letter word should start in column 69. If possible, words with more than twelve letters should not be the fifth word on a page.

The numbers associated with every set of five words are punched on cards separate from the word card. Using the standard VARAB or VARCAR format (the same), punch the five frequencies, followed by the five percentages for the first syntactic type, next the five percentages for the second syntactic type, and so on. Zeros must be punched when a word doesn't occur as a particular type. There should be five cards of numbers of output desired, for each one card of words for each page. This is for 15 syntactic types. If more types are given, the argument word in location 00131)B must be changed from 03 00005 00010 to 03 000nn 00010, n = number if octal, of cards to be read for each set of five words.

Operation:

Card-to-tape program deck I; card-to-tape 'data deck II' (the word cards); card-to-tape 'data deck III' (the number cards). Place II on uniservo No. 4, III on uniservo No. 3; load octal program tape number. Output is on uniservo No. 5. The program starts at cell 00250)B and should stop at 05043)B. The original program provides for 100 pages of output, or 500 words. For fewer pages an MJ to 05043B) should be provided at the end of

the number of pages required. This may be accomplished by an octal corrector with the program deck, in which case the program must be card-to-taped; or by a four-field octal loader corrector, which must be loaded after the program tape (before program is started).

Notes:

1. The output buffer consists of an 84 word BUFF and 324 word HØLD. The syntactic type symbols go to HØLD + 60(30), HØLD + 66(31), HØLD + 72, HØLD + 78, HØLD + 84; HØLD + 96, HØLD + 102, HØLD + 108, HØLD + 114, HØLD + 120, HØLD + 132, etc. The numbers associated with these types are in locations called HØLD + 61 through HØLD + 65; HØLD + 67 through HØLD + 71; and so on.

2. If the indicative number - 1111 1111 - 11 is desired, for example, opposite type 31, under the second word on a page, then TP STAR HØLD + 68 must be inserted by programmer in 20 word buffer provided for that purpose for each page before PREDIT is called to output the page.

EFFICIENT ESTIMATION OF REGRESSION PARAMETERS FOR CERTAIN SECOND-ORDER STATIONARY PROCESSES

4

C.T. Striebel

FOREWORD

This reprint was published originally as a technical report, LMSD-288079, under the same title, November 1959.

ABSTRACT

The problem considered is that of estimating the single regression parameter k in the presence of a second-order stationary disturbing process $X(t)$ with rational spectral density. If the process

$$y(t) = k\phi(t) + X(t)$$

is observed in the continuous interval $o \leq t \leq T$, a linear unbiased estimate $\bar{k}^T$ of k is said to be efficient if its variance is asymptotically minimum among all linear unbiased estimates.

For known mean value functions of the form

$$\phi(t) = t^2 e^{at} \sum_{\alpha=1}^{n} \phi_\alpha e^{i\lambda_\alpha t}$$

efficient estimates are given for the two cases $a = o$ and $a > o$. They are shown to be economical of information concerning the spectral density in that no estimate exists which is efficient for a wider class of spectra.

CONTENTS

Section 1
THE PROBLEM AND ITS BACKGROUND

Let $x(t)$ be a second order stationary stochastic process with mean value function zero and covariance

$$E[x(t)\,\overline{x(s)}] = R(t-s)\,, \tag{1.1}$$

and suppose that the process

$$y(t) = k\varphi(t) + x(t) \tag{1.2}$$

is observed for the continuous time parameter t in the interval $0 \leq t \leq T$. The function $\varphi(t)$ is known, and the unknown parameter k is to be estimated. This process can be thought to consist of a systematic component $k\varphi(t)$, which is completely predictable except for the unknown scale factor k, plus a random disturbing component $x(t)$. The problem is to estimate this scale factor from the observed process.

Only linear unbiased estimates will be considered, and they will be represented as linear functionals

$$\overline{k}^T = \overline{k}^T\,[y(t),\ 0 \leq t \leq T] \tag{1.3}$$

which are limits in quadratic mean of finite linear combinations of the process $y(t_i)$, $i = 1, \ldots, N$ where the arguments t_i are in the interval $0 \leq t_i \leq T$. Such an estimate $\overline{k}^T$ is said to be

asymptotically efficient, or simply efficient, for the problem $[\varphi(t), R(t)]$ if it satisfies the condition

$$\frac{\text{variance } \bar{k}^T}{\text{variance } \hat{k}^T} \to 1 \qquad \text{as } T \to \infty , \tag{1.4}$$

where $\hat{k}^T$ is the minimum variance linear unbiased estimate of k computed for the problem determined by the mean value function $\varphi(t)$ and the covariance $R(t)$.

Interest in efficient estimates arises from the fact that the "best" estimate $\hat{k}^T$ is usually very inconvenient. If the estimate $\hat{k}^T$ is represented by the linear functional $\hat{k}^T[y(t), 0 \leq t \leq T]$ then it must satisfy the linear equation

$$\hat{k}^T[R(t-s), 0 \leq t \leq T] = \hat{M}^T\overline{\varphi(s)} \qquad 0 \leq s \leq T , \tag{1.5}$$

where $\hat{M}^T$ is a constant to be determined by the condition of unbiasedness,

$$\hat{k}^T[\varphi(t), 0 \leq t \leq T] = 1 . \tag{1.6}$$

For most combinations of functions $\varphi(t)$ and $R(t)$ it is difficult to exhibit this solution explicitly; and provided it can be exhibited at all, the solution usually depends on complete knowledge of $R(t)$ and is cumbersome to compute. Efficient estimates are provided by linear functional $\bar{k}^T$ which are in a sense asymptotic solutions to the linear equation (1.5).

The principal alternative estimate, which has been proposed, is the least square estimate,

$$\overline{k}_L^T = \int_0^T \overline{\varphi(t)} y(t) dt \Big/ \int_0^T |\varphi(t)|^2 dt \ . \tag{1.7}$$

This estimate has the advantages that it is easy to compute and requires no knowledge whatever of the covariance $R(t)$. Previous work on the problem of efficient estimates has been primarily devoted to determining those combinations of functions $\varphi(t)$ and $R(t)$ for which the least square estimate is efficient.

For the Ornstein Uhlenbeck process, that is,

$$R(t) = e^{-\rho|t|} \ , \tag{1.8}$$

and for mean value functions

$$\varphi(t) = t^r \text{ or } e^{-i\lambda_o t} \ , \tag{1.9}$$

where r is a non-negative integer and λ_o is a real frequency, Mann and Moranda in reference [10] proved that the least square estimate is efficient. The author in reference [13] extended this result to include mean value functions of the form

$$\varphi(t) = t^r e^{-i\lambda_o t} \tag{1.10}$$

and showed further that for the more general function

$$\varphi(t) = \sum_{\alpha=1}^{n} \varphi_\alpha t^r e^{-i\lambda_\alpha t} \tag{1.11}$$

where the φ_α are non-zero constants and $n > 1$, the least square estimate is not efficient.

For a much broader class of covariance functions $R(t)$ and essentially the same mean value functions $\varphi(t)$, this problem was first discussed by Grenander in reference [3]. Further work was carried out by Grenander and Rosenblatt in references [4] and [5]. Rosenblatt considered some of the same problems in the case of vector valued time series in reference [11] and extended his results in [12]. Most of these results together with some examples appear in Chapter 7 of reference [6]. In this work only the discrete parameter case is considered, and the mean value function $\varphi(t)$ is assumed to be "slowly increasing." This assumption requires that the function $\varphi(t)$ be essentially of form (1.11). All restrictions on the class of covariances are imposed on the equivalent class of spectral densities $f(\lambda)$, which by assumption exist and satisfy the relation

$$R(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i\lambda t} f(\lambda) d\lambda \tag{1.12}$$

for a discrete parameter process and

$$R(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\lambda t} f(\lambda) d\lambda \tag{1.13}$$

for a continuous parameter process. Thus a problem $[\varphi, R]$ can be referred to by the equivalent $[\varphi, f]$. For positive continuous spectral density and "slowly increasing" mean value function, a necessary and

sufficient condition is given in reference [6] for the least square estimate to be efficient. The condition, which is the same as that found in reference [13] for the continuous parameter Ornstein Uhlenbeck process and mean value function of the form (1.11), is that $f(-\lambda_\alpha)$ be constant for $\alpha = 1, \ldots, n$.

In Chapter 1.3 of reference [7] Grenander reproduces a few of these results using the methods of Toeplitz forms. In Chapter 1.4, under certain regularity conditions on $f(\lambda)$, he extends his results to the continuous parameter case for the single example

$$\varphi(t) = 1 \ . \tag{1.14}$$

With the exception of those in reference [7], all the above mentioned results are derived for the more general problem

$$E[y(t)] = \sum_{i=1}^{p} k_i \varphi_i(t) \tag{1.15}$$

where the k_i are unknown parameters and the $\varphi_i(t)$ are known functions. For $p > 1$, the definition of efficiency used by Mann, Moranda, and Striebel is different from that used by Rosenblatt and Grenander. For the case $p = 1$, both agree with definition (1.4) made above. In the present paper only the case $p = 1$ will be considered though it is believed that the results obtained could be generalized to larger p using the methods and definitions of efficiency given by Rosenblatt and Grenander.

In the present paper $f(\lambda)$ is assumed to be a rational function with no real zeros. Again it is believed that this requirement could be slightly relaxed so as to include the case considered in Chapter 1.4 of reference [7] though this has not been attempted.

Under these two limitations, for the continuous parameter process the results in references [10], [13], and [7], mentioned above, are special cases of the theory to be presented here. The continuous parameter equivalents of the important examples discussed in reference [6] are also covered.

Section 2

SUMMARY

In the results summarized above it has been shown that in certain circumstances the least square estimate is efficient for a large class of spectral densities. It shall be the purpose here to discuss efficiency classes of spectral densities defined as follows: For a fixed mean value function $\varphi(t)$, a class of spectral densities $\mathcal{F}$ and a member $f(\lambda)$ of $\mathcal{F}$, the <u>efficiency class</u> $\mathcal{F}(f,\varphi)$ of the density $f(\lambda)$ in the class $\mathcal{F}$ with respect to the mean value function $\varphi(t)$ is defined to be all members $g(\lambda)$ of $\mathcal{F}$ for which there exists an estimate (not necessarily the least square) that is efficient for the two problems $[\varphi,f]$ and $[\varphi,g]$. Thus for $\varphi(t) = 1$ and a certain class of densities $\mathcal{F}$, Grenander in Chapter 1.3 of reference [7] shows that $\mathcal{F}(f,\varphi) = \mathcal{F}$ for all $f \in \mathcal{F}$, since in this case he establishes that the least square estimate is efficient for all members of his class $\mathcal{F}$.

Section 3 is devoted to establishing the efficiency of certain proposed estimates (2.10), (2.11), and (2.13). It is also shown that for questions of efficiency the problem based on the process $y(t)$ for t is the finite interval $0 \leq t \leq T$ can be extended to an equivalent problem with t in the semi-infinite interval $-\infty < t \leq T$. This property is used in

section 4 to apply known results for the latter problem in obtaining a necessary condition that a density $g(\lambda)$ belong to an efficiency class $\mathfrak{F}(f,\varphi)$.

The mean value functions $\varphi(t)$ which will be considered are of the form

$$\varphi(t) = \sum_{\gamma=1}^{m} \sum_{j=0}^{r_\gamma} \varphi_{\gamma j}\, t^j e^{-i\lambda_\gamma t} \tag{2.1}$$

where r_γ is a non-negative integer, φ_γ and λ_γ are complex,

$$\max_{1\le\gamma\le m} \Re(\lambda_\gamma) = a \ge 0 , \tag{2.2}$$

and $\varphi_{\gamma r_\gamma} \neq 0$. The spectral densities considered are assumed to be rational functions with no real zeros. For spectral densities of this type, the solution of the linear equation (1.5) for the finite interval $0 \le t \le T$ is outlined by Laning and Battin in Chapter 8.4 of reference [8]. However, this estimate $\hat{k}^T[\varphi(t),\ 0 \le t \le T]$ will not be used here, since it will be shown that efficient estimates can instead be compared with the much more convenient estimate $\hat{k}^T[y(t),\ -\infty < t \le T]$.

In order to exhibit the estimates, whose efficiency is considered in section 3, some notation must first be developed. The one-sided Laplace transform,

$$\Phi(\lambda) = \int_0^\infty e^{-i\lambda t}\varphi(t)dt , \tag{2.3}$$

of $\varphi(t)$ given by (2.1) can be written

$$\Phi(\lambda) = \sum_{\gamma=1}^{m} \sum_{j=1}^{r_\gamma+1} \frac{\Phi_{rj}}{(\lambda+\lambda_\alpha)^j} \quad , \tag{2.4}$$

where $\Phi_{\gamma j} = (-i)^{j+1} j! \varphi_{\gamma j+1}$. Let r be the maximum of r_γ among subscripts γ for which $\Re\lambda_\gamma = a$, and indicate by $\alpha = 1, \ldots, n$ those subscripts for which $\Re(\lambda_\alpha) = a$ and $r_\alpha = r$. It will be seen that for questions of efficiency only those terms of the mean value function whose absolute value increases as $t^r e^{at}$ will be pertinent. Thus the mean value function (2.1) can in effect be replaced by

$$\varphi(t) = t^r e^{at} \sum_{\alpha=1}^{n} \varphi_{\alpha r} e^{-it\Re(\lambda_\alpha)} \quad . \tag{2.5}$$

Using the method described by Doob on page 542 of reference [2], the spectrum can be factored and written in the form

$$f(\lambda) = |F(\lambda)|^2 \quad , \tag{2.6}$$

where $F(\lambda)$ is a rational function with zeros and poles in the upper half-plane, $(\Re\lambda > 0)$. By the division algorithm the quotient of polynomials $1/F(\lambda)$ can be expanded as

$$\frac{1}{F(\lambda)} = E(\lambda) + M(\lambda) \tag{2.7}$$

where $E(\lambda)$ is a polynomial of degree e and $M(\lambda)$ is a proper rational function with poles in the upper half-plane. By <u>proper</u> it is meant that the degree of the numerator is less than the degree of the denominator. Similarly, the following expansions can be made:

$$\frac{\overline{\Phi}_{\alpha r+1}}{\overline{F(-\lambda_\alpha)}(\lambda+\overline{\lambda}_\alpha)F(\lambda)} = E_\alpha(\lambda) + M_\alpha(\lambda) \qquad \alpha = 1, \cdots, n \tag{2.8}$$

where $E_\alpha(\lambda)$ is a polynomial of degree $e-1$; and, provided $a > 0$, $M_\alpha(\lambda)$ is a proper rational function with poles in the upper half-plane.

Let

$$E(\lambda) = \sum_{j=0}^{e} e_j \lambda^j$$

$$E_\alpha(\lambda) = \sum_{j=0}^{e-1} e_{\alpha j} \lambda^j \qquad \alpha = 1, \cdots, n \tag{2.9}$$

$$m(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\lambda t} M(\lambda) d\lambda$$

$$m_\alpha(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\lambda t} M_\alpha(\lambda) d\lambda \qquad \alpha = 1, \cdots, n .$$

In section 3 the following estimates are proved efficient:

$$\tilde{M}^{T}\tilde{k}_1^{T}[y(t),\ 0 \leq t \leq T;\ a > 0] \tag{2.10}$$

$$= \sum_{\alpha=1}^{n} e^{i\overline{\lambda}_\alpha T}\Big[\sum_{j=0}^{e-1} (-i)^j e_{\alpha j} y^{(j)}(T) + \int_0^T m_\alpha(T - t) y(t) dt\Big] ;$$

$$\tilde{M}^{T}\tilde{k}_2^{T}[y(t),\ 0 \leq t \leq T;\ a > 0] \tag{2.11}$$

$$= \sum_{j=0}^{e} (-i)^j e_j Y_T^{(j)}(T) + \int_0^T m(T - t) Y_T(t) dt ,$$

where

$$Y_T(t) = \sum_{\alpha=1}^{n} \frac{\overline{\Phi}_{\alpha r+1}}{F(-\lambda_\alpha)} e^{i\overline{\lambda}_\alpha(T-t)} \int_0^t e^{i\overline{\lambda}_\alpha u} y(u)du \ ; \qquad (2.12)$$

and

$$\tilde{M}^T \tilde{k}^T[y(t),\ 0 \leq t \leq T;\ a = 0] \qquad (2.13)$$

$$= \sum_{\alpha=1}^{n} \frac{\overline{\Phi}_{\alpha r+1}}{f(-\lambda_\alpha)} \int_0^T t^r e^{i\lambda_\alpha t} y(t)dt \ .$$

In each case $\tilde{M}^T$ denotes a constant to be determined so that the estimate is unbiased. It is evaluated by replacing $y(t)$ with $\varphi(t)$ in the right side of the expressions (2.10), (2.11), and (2.13). As the notation indicates, estimates (2.10) and (2.11) are efficient in the case $a > 0$, and estimate (2.13) is efficient when all the frequencies λ_α are real, that is, $a = 0$. In order to compute the estimates (2.10) and (2.11), $\tilde{k}_1^T[a > 0]$ and $\tilde{k}_2^T[a > 0]$, it is necessary that the entire spectral density $f(\lambda)$ be known. Estimate $\tilde{k}^T[a = 0]$, (2.13), requires knowledge of the spectrum only at the points $-\lambda_\alpha$, that is, $f(-\lambda_\alpha)$, $\alpha = 1, \cdots, n$.

Results concerning the least square estimate can be obtained by letting $F(\lambda) = 1$, for which expressions (2.10), (2.11) and (2.13) reduce to the least square estimate for the equivalent mean value function (2.5).

Section 3
EFFICIENT ESTIMATES

Since $f(\lambda)$ is a positive rational function it is clear that the condition

$$\int_{-\infty}^{\infty} \frac{\log f(\lambda)}{1+\lambda^2}\, d\lambda > -\infty \tag{3.1}$$

is satisfied, and hence by Theorem 10.12b of reference [7] the process $x(t)$ can be represented as a moving average

$$x(t) = \int_{-\infty}^{t} \gamma(t-s)\, d\xi(s) \qquad -\infty < t < \infty \tag{3.2}$$

where $\xi(s)$ is a fundamental random process with zero mean value and $\gamma(t)$ is the Fourier inverse of $F(\lambda)$,

$$\gamma(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\lambda t} F(\lambda)\, d\lambda \ . \tag{3.3}$$

Suppose there exists a function $\psi(t)$ which satisfies the conditions

$$\int_{-\infty}^{t} \gamma(t-s)\dot{\psi}(s)\, ds = \varphi(t) \qquad 0 \leq t < \infty \tag{3.4}$$

and

$$\int_{-\infty}^{T} |\dot{\psi}(s)|^2\, ds < \infty \qquad \text{all } T < \infty \ , \tag{3.5}$$

then the process $y(t)$ can be represented as the moving average

$$y(t) = \int_{-\infty}^{t} \gamma(t-s)d\eta(s) \qquad 0 \leq t < \infty \tag{3.6}$$

where $\eta(s)$ is the process defined by

$$\eta(s) = k\psi(t) + \xi(t) \qquad -\infty < t < \infty \ . \tag{3.7}$$

Thus there are three processes which will be considered

I	$y(t) = k\Phi(t) + x(t)$	$0 \leq t < \infty$
II	$y(t) = k\zeta(t) + x(t)$	$-\infty < t < \infty$
III	$\eta(t) = k\psi(t) + \xi(t)$	$-\infty < t < \infty$.

The $x(t)$ processes are second order stationary with zero mean value and spectral density $f(\lambda)$. The $\xi(t)$ process is a fundamental random process. The mean value functions satisfy the conditions

$$\int_{-\infty}^{t} \gamma(t-s)\dot{\psi}(s)ds = \zeta(t) \qquad -\infty < t < \infty \tag{3.8}$$

$$\zeta(t) = \Phi(t) \qquad 0 \leq t < \infty$$

and condition (3.5). Their transforms $\Phi(\lambda)$, given by (2.3), (3.9)

$$Z(\lambda) = \int_{-\infty}^{\infty} e^{-i\lambda t}\zeta(t)dt \ , \tag{3.10}$$

and

$$\Psi(\lambda) = \int_{-\infty}^{\infty} e^{-i\lambda t}\dot{\psi}(t)dt \ , \tag{3.11}$$

are assumed to converge to proper rational functions in some non-degenerate strip. They satisfy conditions (2.4) and from (3.8)

$$Z(\lambda) = \Psi(\lambda)F(\lambda) \ . \tag{3.12}$$

The integral equation (3.4) can be solved using Laplace transforms and the Wiener-Hopf technique. (See, for example, page 313 of [14].) Taking two-sided Laplace transforms of equation (3.4) yields the equation

$$\Phi(\lambda) + H(\lambda) = \Psi(\lambda)F(\lambda) \tag{3.13}$$

where

$$H(\lambda) = \int_{-\infty}^{0} e^{-i\lambda t} \int_{-\infty}^{t} \gamma(t-s)\dot{\psi}(s)dsdt \ . \tag{3.14}$$

Solving (3.13), the transform $\Psi(\lambda)$ must satisfy

$$\Psi(\lambda) = \frac{\Phi(\lambda) + H(\lambda)}{F(\lambda)} \tag{3.15}$$

The integral $\Phi(\lambda)$, (2.3), converges in the region $\mathcal{I}(\lambda) < -a$. In order that the Laplace inverse of $\Psi(\lambda)$ in (3.15) exist and satisfy conditions (3.4) and (3.5), it is sufficient that $H(\lambda)$ be a proper rational function with poles in the region $\mathcal{I}(\lambda) < -b$ for some $-b < -a$ and that $\Psi(\lambda)$ also be a proper rational function. Such a function $H(\lambda)$ can be found, for example, as follows: Let $D_H(\lambda)$ be any polynomial with zeros whose imaginary parts are less than $-b$. By the division algorithm the rational function $\Phi(\lambda)D_H(\lambda)$ can be written

$$\Phi(\lambda)D_H(\lambda) = Q(\lambda) + R(\lambda) \ , \tag{3.16}$$

where $R(\lambda)$ is a proper rational function and $Q(\lambda)$ is a polynomial of degree equal to that of $\Phi(\lambda)D_H(\lambda)$ —that is, the degree of the numerator of $\Phi(\lambda)D_H(\lambda)$ minus the degree of its denominator. Since $\Phi(\lambda)$

is proper, the degree of $Q(\lambda)$ is less than that of $D_H(\lambda)$. Now define $H(\lambda)$ by

$$H(\lambda) = -\frac{Q(\lambda)}{D_H(\lambda)} \quad , \tag{3.17}$$

so that

$$H(\lambda) + \Phi(\lambda) = \frac{R(\lambda)}{D_H(\lambda)} \quad . \tag{3.18}$$

By taking the degree of $D_H(\lambda)$ larger than e, the degree of $1/F(\lambda)$, a proper function $\Psi(\lambda)$ is obtained. From (3.12) $Z(\lambda)$ can be defined by

$$Z(\lambda) = \Phi(\lambda) + H(\lambda) \quad . \tag{3.19}$$

Thus, given a process I, the existence of processes II and III has been established.

In reference [9], Mann has solved the problem of estimating the parameter k for the process III. The minimum variance linear unbiased estimate is given by

$$\hat{M}^T\hat{k}^T[\eta(t),\ -\infty < t \leq T] = \int_{-\infty}^{T} \overline{\dot{\psi}(t)}d\eta(t) \quad , \tag{3.20}$$

and

$$\text{variance } \hat{k}^T[\eta(t),\ -\infty < t \leq T] = 1\Big/\int_{-\infty}^{T} |\dot{\psi}(t)|^2 dt \quad . \tag{3.21}$$

The remainder of the section will be devoted to obtaining approximate expressions for the estimate (3.20), which are more convenient but are still efficient for the problem. The efficiency of proposed estimates $\tilde{k}_1^T[a > 0]$, $\tilde{k}_2^T[a > 0]$, and $\tilde{k}^T[a = 0]$ given by (2.10), (2.11), and (2.13) is proved in Theorem 1. However, before this is done, a non-rigorous derivation of the estimates will be given by way of motivation.

First equation (3.6) will be solved for $d\eta(s)$ by transform methods. Since the poles of $F(\lambda)$ are in the upper half-plane, its inverse transform $\gamma(t)$ vanishes for $t < 0$. Thus the process $y(t)$ can be written

$$y(t) = \int_{-\infty}^{\infty} \gamma(t-s)d\eta(s) , \tag{3.22}$$

and

$$\begin{aligned} \int_{-\infty}^{T} e^{-i\lambda t} y(t)dt &= \int_{-\infty}^{T} e^{-i\lambda t} \int_{-\infty}^{\infty} \gamma(t-s)d\eta(s)dt \\ &= \int_{-\infty}^{T} \left[\int_{-\infty}^{\infty} e^{-i\lambda(\tau+s)}\gamma(\tau)d\tau\right]d\eta(s) \\ &= F(\lambda) \int_{-\infty}^{T} e^{-i\lambda s} d\eta(s) . \end{aligned} \tag{3.23}$$

Substituting the Laplace inversion formula for both $\dot{\psi}(t)$ and $d\eta(t)$ in formula (3.20) gives the alternative form for the estimate

$$\hat{M}^T\hat{k}^T[\eta(t),\ -\infty < t \le T] = \int_{-\infty}^{T} \overline{\Psi(t)}\,d\eta(t)$$

$$= \frac{1}{(2\pi)^2}\int_{-\infty}^{T}\left\{\left[\int_{-\infty-ic}^{+\infty-ic} e^{i\lambda t}\,\frac{\overline{\Phi}(-\lambda) + \overline{H}(-\lambda)}{\overline{F}(-\lambda)}\,d\lambda\right]\right. \tag{3.24}$$

$$\left.\left[\int_{-\infty+id}^{+\infty+id} \frac{e^{i\omega t}}{\overline{F}(\omega)}\left(\int_{-\infty}^{T} e^{-i\omega u}y(u)du\right)d\omega\right]\right\}dt \quad ,$$

where c is in the region of convergence of the integral $\Psi(\lambda)$ in (3.11), $b > c > a$; d is taken in the interval $0 < d < c$ where the mean and variance of the process $\int_{-\infty}^{T} e^{-i\omega t}y(t)dt$ and the integral $\int_{-\infty}^{T} e^{(i\omega+i\lambda)t}dt$ are all finite. The notation $\overline{\Phi}(\lambda)$ indicates that complex conjugates of the coefficients in the rational function $\Phi(\lambda)$ are taken.

The most important change that will be made in (3.24) is truncation of $y(t)$ for $t < 0$. Thus (3.24) becomes

$$\tilde{M}^T\tilde{k}^T[y(t),\ 0 \le t \le T] = \frac{1}{2\pi}\int_{-\infty+id}^{+\infty+id} \frac{\overline{\Psi^T(\omega)}}{\overline{F}(\omega)}\left[\int_{0}^{T} e^{-i\omega t}y(t)dt\right]d\omega \tag{3.25}$$

where

$$\overline{\Psi^T(\omega)} = \frac{1}{2\pi}\int_{-\infty-ic}^{+\infty-ic} \frac{e^{T(i\omega+i\lambda)}}{i(\omega+\lambda)}\left[\frac{\overline{\Phi}(-\lambda) + \overline{H}(-\lambda)}{\overline{F}(-\lambda)}\right]d\lambda$$

$$= \overline{\int_{-\infty}^{T} e^{-i\omega t}\dot{\psi}(t)dt} \quad . \tag{3.26}$$

Additional changes in the estimate (3.25) are made by expanding $\overline{\Psi^T(\omega)}$ as a sum of residues at the poles of $\Phi(\lambda)$, $H(\lambda)$ and $1/F(\lambda)$ and then picking out those terms which dominate at $T \to \infty$. When terms of $\overline{\Psi^T(\omega)}$ are omitted, the complex integral (3.25) will in general cease to exist in the common sense. It must be replaced by a Ceasàro limit of order $e + 1$. This limit is defined and its important properties are stated in Theorem 2. The final estimates $\tilde{k}_1^T[a > 0]$, $\tilde{k}_2^T[a > 0]$ and $\tilde{k}^T[a = 0]$ given by (2.10), (2.11), and (2.13) are then obtained by interpreting the resulting Ceasàro limit as the Laplace inverse of appropriate functions of $y(t)$ for $0 \le t \le T$.

The simplification of $\overline{\Psi^T(\omega)}$ will be different for the two cases $a > 0$ and $a = 0$. If $a > 0$, the dominant terms of $\overline{\Psi^T(\omega)}$ have magnitude $T^r e^{(a+\omega)}$. They are contributed by the poles of $\overline{\Phi}(-\lambda)$, $\lambda = \lambda_\alpha$, $\alpha = 1, \cdots, n$ and are of the form

$$\overline{\Psi^T(\omega)} \sim \frac{T^r}{r!} \sum_{\alpha=1}^{n} \frac{\overline{\Phi}_{\alpha r+1}}{\overline{F(-\lambda_\alpha)}} \frac{e^{iT(\omega+\bar{\lambda}_\alpha)}}{(\omega+\bar{\lambda}_\alpha)} \quad . \tag{3.27}$$

Making this substitution in (3.25) gives the estimate

$$\begin{aligned} &\tilde{M}^T\tilde{k}^T[y(t),\ 0 \le t \le T;\ a > 0] \\ &= \frac{1}{2\pi} \int_{(C,e+1)}^{+\infty+id} \left[\sum_{\alpha=1}^{n} \frac{\overline{\Phi}_{\alpha r+1}}{\overline{F(-\lambda_\alpha)}} \frac{e^{i(\bar{\lambda}_\alpha+\omega)T}}{(\omega+\bar{\lambda}_\alpha)F(\omega)} \right] \left[\int_0^T e^{-i\omega t} y(t)\,dt \right] d\omega \quad . \end{aligned} \tag{3.28}$$

Estimates (2.10) and (2.11) are two slightly different interpretations of this Ceasàro limit as a Laplace inverse. The existence in quadratic mean of the derivatives of $y(t)$ involved in estimates (2.10) and (2.11) can easily be established. In order to show that (3.28) itself provides a bona fide estimate, it would first have to be shown that the Ceasàro limit involved exists in quadratic mean.

In the case $a = 0$, the magnitude of the dominant terms in the variance of $\tilde{k}^T$ is T^{2r+1}. They are contributed by terms in $\overline{\Psi^T(\omega)}$ of the form

$$\overline{\Psi^T(\omega)} \sim \sum_{\alpha=1}^{n} \frac{\overline{\Phi}_{\alpha r+1} e^{iT(\lambda_\alpha+\omega)}}{\overline{F(-\lambda_\alpha)}(\omega+\lambda_\alpha)^{r+1}} \cdot \tag{3.29}$$

In this case the dominant terms in (3.25) are the residues at the poles $\omega = -\lambda_\alpha$, which are evaluated simply in the estimate $\tilde{k}^T[a = 0]$ given by (2.13).

Theorem 1. Let $y(t)$, $0 \leq t \leq T$, be a process of type I where $x(t)$ is a second order stationary process whose spectral density $f(\lambda)$ is a rational function with no real zeros, and the mean value function $\varphi(t)$ is of the form (2.1). Then for $a > 0$, the estimates $\tilde{k}_1^T[y(t), 0 \leq t \leq T; a > 0]$ and $\tilde{k}_2^T[y(t), 0 \leq t \leq T; a > 0]$ given by (2.10) and (2.11) are efficient for estimating the parameter k. If $a = 0$, then estimate $\tilde{k}^T[y(t), 0 \leq t \leq T; a = 0]$ given by (2.13) is efficient.

Proof: Since any linear unbiased estimate based on process I $(y(t),\ 0 \leq t \leq T)$ is also a linear unbiased estimate for process III $(\eta(t),\ -\infty < t \leq T)$, it follows from (3.21) that the inequality

$$\text{variance } \hat{k}^T[y(t),\ 0 \leq t \leq T] \geq 1 \Big/ \int_{-\infty}^{T} |\dot{\psi}(t)|^2 dt \tag{3.30}$$

must hold. Thus it is sufficient to show that the limit

$$[\text{variance } \tilde{k}^T][\int_{-\infty}^{T} |\dot{\psi}(t)|^2 dt] \to 1 \quad \text{as} \quad T \to \infty \tag{3.31}$$

holds in order to establish efficiency of an estimate $\tilde{k}^T$. The remainder of the proof is devoted to obtaining asymptotic expressions for the quantities $\int_{-\infty}^{T} |\dot{\psi}(t)|^2 dt$, $\tilde{M}^T$, and variance $\tilde{M}^T\tilde{k}^T$ computed for the three estimates considered. The notation

$$A^T \sim B^T \tag{3.32}$$

will be used to mean

$$\frac{A^T}{B^T} \to 1 \quad \text{as} \quad T \to \infty . \tag{3.33}$$

In addition to the standard Laplace transform inversion formula, the following result due to Amerio [1] will be used:

Theorem 2. If the function $\varphi(t)$ and its first n derivatives are absolutely continuous in the interval $0 \leq t \leq b$, then the derivative $\varphi^{(n)}(t)$ satisfies the inversion formula

$$\varphi^{(n)}(t) = \lim_{R\to\infty} \frac{1}{2\pi} \int_{-R+ic}^{R+ic} (i\lambda)^n e^{i\lambda t} \Phi(\lambda) \left(1 - \frac{\lambda^2}{R^2}\right)^{n+1} d\lambda \quad a \le t \le b \tag{3.34}$$

where $\Phi(\lambda)$ is the transform of $\varphi(t)$

$$\Phi(\lambda) = \int_{-\infty}^{\infty} e^{-i\lambda t} \varphi(t) dt \tag{3.35}$$

and the line $\quad = ic$ is in the strip of convergence of this integral for $\Phi(\lambda)$.

In addition, if $\Phi(\lambda)$ is a proper rational function, the Ceasàro limit (3.34) can be evaluated by the contour integral on a path C around the poles of $\Phi(\lambda)$ above ic if $t > 0$ and below ic if $t < 0$,

$$\lim_{R\to\infty} \frac{1}{2\pi} \int_{-R+ic}^{+R+ic} (i\lambda)^n e^{i\lambda t} \Phi(\lambda) \left(1 - \frac{\lambda^2}{R^2}\right)^{n+1} d\lambda = \frac{1}{2\pi} \oint (i\lambda)^n e^{i\lambda t} \Phi(\lambda) d\lambda \ . \tag{3.36}$$

From the Laplace inversion formula

$$\begin{aligned} \int_{-\infty}^{T} |\dot{\psi}(t)|^2 \, dt &= \frac{1}{(2\pi)^2} \int_{-\infty}^{T} \int_{-\infty-ic}^{+\infty-ic} \int_{-\infty-ic}^{+\infty-ic} e^{(i\lambda+i\omega)t} \Psi(\lambda) \overline{\Psi}(-\omega) d\lambda \, d\omega \, dt \\ &= \frac{1}{(2\pi)^2} \int_{-\infty-ic}^{+\infty-ic} \int_{-\infty-ic}^{+\infty-ic} \frac{e^{T(i\lambda+i\omega)}}{i(\lambda+\omega)} \left(\frac{\Phi(\lambda)+H(\lambda)}{F(\lambda)}\right)\left(\frac{\overline{\Phi}(-\omega)+\overline{H}(-\omega)}{\overline{F}(-\omega)}\right) d\lambda \, d\omega \ . \end{aligned} \tag{3.37}$$

The complex integrals are taken in the region of analyticity which separates the poles of $\Phi(\lambda)/F(\lambda)$ from those of $H(\lambda)$. These integrals can be taken as contour integrals closings them upward around the poles of $\Phi(\lambda)/F(\lambda)$. Examination of the residues at the poles of $\Phi(\lambda)$ and

of $1/F(\lambda)$ shows that for large T, the dominant terms are contributed by the poles with minimum imaginary part and among these, by the poles with maximum order—that is, by $\lambda = -\lambda_\alpha$, $\alpha = 1, \cdots, n$. The residue of $H(\lambda)/F(\lambda)$ at poles of $\Phi(\lambda)$ is zero, so the asymptotic expression

$$\int_{-\infty}^{T} |\dot{\psi}(t)|^2 dt \sim \frac{1}{(2\pi)^2} \oint \oint \frac{e^{T(i\lambda+i\omega)}}{i(\lambda+\omega)} \frac{\Phi(\lambda)}{F(\lambda)} \frac{\overline{\Phi}(-\omega)}{\overline{F}(-\omega)} d\lambda d\omega \quad (3.38)$$

is obtained, where the contours include only the poles $\lambda = -\lambda_\alpha$ and $\omega = \overline{\lambda}_\alpha$, $\alpha = 1, \cdots, n$. Evaluating this integral for $a > 0$ and picking out the terms of magnitude $T^{2r}e^{2a}$, gives the asymptotic expression

$$\int_{-\infty}^{T} |\dot{\psi}(t)|^2 dt \sim \frac{T^{2r}e^{2a}}{(r!)^2} \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} \frac{\Phi_{\alpha r+1}\overline{\Phi}_{\beta r+1} e^{iT\rho(\lambda_\beta - \lambda_\alpha)}}{i(\overline{\lambda}_\beta - \lambda_\alpha)F(-\lambda_\alpha)\overline{F(-\lambda_\beta)}} \quad (3.39)$$

This can be summarized by the form

$$\int_{-\infty}^{T} |\dot{\psi}(t)|^2 dt \sim T^{2r}e^{2a}C(T) \quad (3.40)$$

where $C(T)$ is bounded from above and away from zero. This is clear because $C(T)$ is a positive definite form in the vector

$$\frac{\Phi_{\alpha r+1} e^{-iT\rho\lambda_\alpha}}{F(-\lambda_\alpha)} \qquad \alpha = 1, \cdot \ , n \quad (3.41)$$

whose terms have finite positive absolute values which do not depend on T. For $a = 0$, terms of order T^{2r+1} dominate and asymptotically the integral (3.38) becomes

$$\int_{-\infty}^{T} |\dot{\psi}(t)|^2 dt \sim \frac{T^{2r+1}}{r!r!(2r+1)} \sum_{\alpha=1}^{n} \frac{|\Phi_{\alpha r+1}|^2}{f(-\lambda_\alpha)} \quad . \tag{3.42}$$

For $a > 0$, $M_1^T[a > 0]$ can be evaluated from (2.10),

$$\tilde{M}_1^T[a > 0] = \sum_{\alpha=1}^{n} e^{i\bar{\lambda}_\alpha T} \left[\sum_{j=0}^{e-1} (-i)^j e_{\alpha j} \varphi^{(j)}(T) + \int_0^T m_\alpha(T-t)\varphi(t)dt \right] . \tag{3.43}$$

From Theorem 2 this can be written

$$\begin{aligned} \tilde{M}_1^T[a > 0] &= \frac{1}{2\pi} \sum_{\alpha=1}^{n} e^{i\bar{\lambda}_\alpha T} \int_{\substack{-\infty-ic \\ (C,e+1)}}^{+\infty+ic} e^{i\lambda T} \left[\sum_{j=0}^{e-1} (-i)^j e_{\alpha j} (i\lambda)^j \phi(\lambda) \right. \\ &\qquad \left. + M_\alpha(\lambda)\phi(\lambda) \right] d\lambda \\ &= \frac{1}{2\pi} \int_{\substack{-\infty-ic \\ (C,e+1)}}^{+\infty+ic} \Phi(\lambda) \sum_{\alpha=1}^{n} e^{iT(\bar{\lambda}_\alpha+\lambda)} \left(E_\alpha(\lambda) + M_\alpha(\lambda) \right) d\lambda \ . \end{aligned} \tag{3.44}$$

From (2.8) this becomes

$$\tilde{M}_1^T[a > 0] = \frac{1}{2\pi} \int_{\substack{-\infty-ic \\ (C,e+1)}}^{+\infty+ic} \Phi(\lambda) \sum_{\alpha=1}^{n} \frac{\bar{\Phi}_{\alpha r+1}}{\overline{F(-\lambda_\alpha)}} \frac{e^{iT(\bar{\lambda}_\alpha+\lambda)}}{(\lambda+\bar{\lambda}_\alpha)F(\lambda)} \, d\lambda \quad . \tag{3.45}$$

This can be evaluated by a contour integral. Only terms of maximum order $T^r e^{2aT}$ are retained,

$$\tilde{M}_1^T[a > 0] \sim \frac{(i)^{r+1} T^r e^{2aT}}{r!} \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} \frac{\Phi_{\alpha r+1} \bar{\Phi}_{\beta r+1}}{(\bar{\lambda}_\beta - \lambda_\alpha)F(-\lambda_\alpha)} \frac{e^{iT\ell(\lambda_\beta-\lambda_\alpha)}}{\overline{F(-\lambda_\beta)}} \ . \tag{3.46}$$

Applying Theorem 2 to (2.11), the same expression is obtained for $\tilde{M}_2^T[a > 0]$.

Next, variance $\tilde{M}^{T}\tilde{k}_1^{T}[a > 0]$ will be computed for estimate (2.10).

variance $\tilde{M}^{T}\tilde{k}_1^{T}[a > 0]$

$$= \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} e^{iT(\bar{\lambda}_\alpha - \lambda_\beta)}\, E\left[\sum_{j=0}^{e-1}(-i)^j e_{\alpha j} x^{(j)}(T) + \int_0^T m_\alpha(T-u)x(u)du\right] \tag{3.47}$$

$$\left[\sum_{k=0}^{e-1}(i)^k \overline{e_{\beta k}}\,\overline{x^{(k)}(T)} + \int_0^T \overline{m_\beta(T-v)}\;\overline{x(v)}dv\right]$$

$$= \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} e^{iT(\bar{\lambda}_\alpha - \lambda_\beta)}\left\{\sum_{j=0}^{e-1}\sum_{k=0}^{e-1}(-i)^j(i)^k e_{\alpha j}\,\overline{e}_{\beta k}\,\frac{\partial^j}{\partial u^j}\frac{\partial^k}{\partial v^k}R(u-v)\Big|_{u=v=T}\right.$$

$$+ \sum_{j=0}^{e-1}(-i)^j e_{\alpha j}\int_0^T \overline{m_\beta(T-v)}\,\frac{\partial^j}{\partial u^j}R(u-v)\Big|_{u=T}dv$$

$$+ \sum_{k=0}^{e-1}(i)^k \overline{e}_{\beta k}\int_0^T m_\alpha(T-u)\,\frac{\partial^k}{\partial v^k}R(u-v)\Big|_{v=T}du$$

$$\left.+ \int_0^T\int_0^T m_\alpha(T-u)\overline{m_\beta(T-v)}\,R(u-v)dudv\right\}$$

$$= \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} e^{iT(\bar{\lambda}_\alpha - \lambda_\beta)}\,\frac{1}{2\pi}\int_{-\infty}^{\infty}_{(C,e+i)} f(\lambda)$$

$$\left\{\sum_{j=0}^{e-1}\sum_{k=0}^{e-1}(-i)^j(i)^k e_{\alpha j}\,\overline{e}_{\beta k}\frac{\partial^j}{\partial u^j}\frac{\partial^k}{\partial v^k}e^{i\lambda(u-v)}\Big|_{u=v=T}\right.$$

$$+ \sum_{j=0}^{e-1}(-i)^j e_{\alpha j}\int_0^T \overline{m_\beta(T-v)}\frac{\partial^j}{\partial u^j}e^{i\lambda(u-v)}\Big|_{u=T}dv$$

$$\left.+ \sum_{k=0}^{e-1}(i)^k \overline{e}_{\beta k}\int_0^T m_\alpha(T-u)\frac{\partial^k}{\partial v^k}e^{i\lambda(u-v)}\Big|_{v=T}du + \int_0^T\int_0^T m_\alpha(T-u)m_\beta(T-v)e^{i\lambda(u-v)}dudv\right\}d\lambda$$

$$= \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} e^{iT(\bar{\lambda}_\alpha - \lambda_\beta)} \frac{1}{2\pi} \int_{-\infty}^{\infty} f(\lambda)[E_\alpha(\lambda)e^{i\lambda T}$$

$$+ \int_0^T m_\alpha(T-u)e^{i\lambda u}du]\,\overline{[E_\beta(\lambda)e^{i\lambda T} + \int_0^T m_\beta(T-v)e^{i\lambda v}dv]}d\lambda \ .$$

The factor in the square bracket will be examined in detail. Using the substitutions

$$e^{i\lambda T}E_\alpha(\lambda) = \frac{1}{2\pi} \int_{\substack{-\infty-id \\ (C,e+1)}}^{+\infty-id} \frac{E_\alpha(\lambda)e^{ipT}}{(ip-i\lambda)}dp \ ,$$

$$\int_0^T m_\alpha(T-u)e^{i\lambda u}du = \frac{1}{2\pi} \int_{-\infty-id}^{+\infty-id} \frac{e^{ipT}M_\alpha(p)}{(ip-i\lambda)}\ dp \ , \tag{3.48}$$

where $d > 0$, and (2.8); the bracket can be written

$$E_\alpha(\lambda)e^{i\lambda T} + \int_0^T m_\alpha(T-u)e^{i\lambda u}du = \frac{1}{2\pi} \int_{\substack{-\infty-id \\ (C,e+1)}}^{+\infty-id} \frac{e^{ipT}(E_\alpha(p)+M_\alpha(p))}{(ip-i\lambda)}\ dp$$

$$= \frac{1}{2\pi} \int_{\substack{-\infty-id \\ (C,e+1)}}^{+\infty-id} \frac{\bar{\Phi}_{\alpha r+1}}{\overline{F(-\lambda_\alpha)}} \frac{e^{ipT}}{(p+\bar{\lambda}_\alpha)F(p)(ip-i\lambda)}\ dp \ . \tag{3.49}$$

Let $\beta_1, \cdots, \beta_k$ be the zeros of $F(p)$. By assumption $\Re\beta_\gamma > 0$. The contour in (3.49) will be closed upward around the poles $p = \lambda, -\bar{\lambda}_\alpha$, $\beta_1, \cdots, \beta_k$ and evaluated by residues,

$$E_\alpha(\lambda)e^{i\lambda T} + \int_0^T m_\alpha(T-u)e^{i\lambda u}du$$

$$= \frac{\bar{\Phi}_{\alpha r+1}}{\overline{F(-\lambda_\alpha)}} \left[\frac{e^{i\lambda T}}{(\lambda+\bar{\lambda}_\alpha)F(\lambda)} - \frac{e^{-i\bar{\lambda}_\alpha T}}{F(-\bar{\lambda}_\alpha)(\bar{\lambda}_\alpha+\lambda)} + \frac{C_{\gamma\delta\nu\alpha}T^\delta e^{i\beta_\gamma T}}{(\beta_\gamma-\lambda)^\nu} \right] \ . \tag{3.50}$$

Summation on γ, δ, and ν is understood. The constants $C_{\gamma\delta\nu\alpha}$ will not be explicitly evaluated. Making the substitutions (2.6) and (3.50), the integral (3.47) can now be evaluated,

$$\text{variance } \tilde{M}^{T}\tilde{k}_1^{T}[a > 0]$$

$$= \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} \frac{\bar{\Phi}_{\alpha r+1}\Phi_{\beta r+1}e^{iT(\bar{\lambda}_\alpha-\lambda_\beta)}}{\overline{F(-\lambda_\alpha)}F(-\lambda_\beta)} \frac{1}{2\pi}\int_{(C,e+1)}^{\infty} F(\lambda)\overline{F(\lambda)}\left[\frac{1}{(\lambda+\bar{\lambda}_\alpha)(\lambda+\lambda_\beta)F(\lambda)\overline{F(\lambda)}}\right.$$

$$+\frac{e^{iT(\lambda_\beta-\bar{\lambda}_\alpha)}}{F(-\bar{\lambda}_\alpha)\overline{F(-\lambda_\beta)}(\bar{\lambda}_\alpha+\lambda)(\lambda_\beta+\lambda)} \quad \frac{C_{\gamma\delta\nu\alpha}C_{\gamma'\delta'\nu'\beta}T^{\delta+\delta'}e^{iT(\beta_\gamma-\bar{\beta}_{\gamma'})}}{(\beta_\gamma-\lambda)^{\nu}(\bar{\beta}_{\gamma'}-\lambda)^{\nu'}} \tag{3.51}$$

$$-\frac{e^{iT(\lambda+\lambda_\alpha)}}{(\lambda+\bar{\lambda}_\alpha)F(\lambda)\overline{F(-\lambda_\beta)}(\lambda_\beta+\lambda)} - \frac{e^{-iT(\bar{\lambda}_\alpha+\lambda)}}{\overline{F(\lambda)}(\lambda+\lambda_\beta)F(-\bar{\lambda}_\alpha)(\bar{\lambda}_\alpha+\lambda)}$$

$$+\frac{C_{\gamma'\delta'\nu'\beta}T^{\delta'}e^{iT(\lambda-\bar{\beta}_{\gamma'})}}{(\lambda+\bar{\lambda}_\alpha)F(\lambda)(\bar{\beta}_{\gamma'}-\lambda)^{\nu'}} + \frac{C_{\gamma\delta\nu\alpha}T^{\delta}e^{iT(\beta_\gamma-\lambda)}}{\overline{F(\lambda)}(\lambda+\lambda_\beta)(\beta_\gamma-\lambda)^{\nu}}$$

$$\left.-\frac{C_{\gamma'\delta'\nu'\beta}T^{\delta'}e^{-iT(\bar{\lambda}_\alpha+\bar{\beta}_{\gamma'})}}{F(-\bar{\lambda}_\alpha)(\bar{\lambda}_\alpha+\lambda)(\bar{\beta}_{\gamma'}-\lambda)^{\nu'}} - \frac{C_{\gamma\delta\nu\alpha}T^{\delta}e^{iT(\lambda_\beta+\beta_\gamma)}}{\overline{F(-\lambda_\beta)}(\lambda_\beta+\lambda)(\beta_\gamma-\lambda)^{\nu}}\right] d\lambda \ .$$

Examination of the residues of each term shows that, with the exception of the first, all terms contain one of the factors $e^{iT(\lambda_\beta-\bar{\lambda}_\alpha)}$, $e^{iT(\beta_\gamma-\bar{\beta}_\gamma)}$, $e^{-iT(\bar{\lambda}_\alpha+\bar{\beta}_\gamma)}$, or $e^{iT(\lambda_\beta+\beta_\gamma)}$. Thus they go to zero as $T \to \infty$. Hence, the asymptotic expression,

$$\text{variance } \tilde{M}^{T}\tilde{k}_1^{T}[a>0] \sim e^{2aT}\sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} \frac{\bar{\Phi}_{\alpha r+1}\Phi_{\beta r+1}e^{iT\ell(\lambda_\alpha-\lambda_\beta)}}{\overline{F(-\lambda_\alpha)}F(-\lambda_\beta)i(\bar{\lambda}_\alpha-\lambda_\beta)} \tag{3.52}$$

is provided by the first term of (3.51). Computation of variance $\tilde{M}^T\tilde{k}_2^T[a > 0]$ for estimate (2.11) is entirely analogous and yields the same result.

Asymptotic expressions (3.39), (3.46), and (3.52) can now be combined to compute the efficiency of estimates (2.10) and (2.11),

$$\text{variance } \tilde{k}^T[a > 0] \int_{-\infty}^{T} |\dot{\psi}(t)|^2 dt \tag{3.53}$$

$$= \{\text{variance } \tilde{M}^T\tilde{k}^T[a > 0]\}\{\int_{-\infty}^{T} |\psi(t)|^2 dt\} \Big/ |\tilde{M}^T[a > 0]|^2 \sim 1 .$$

Next, $\tilde{M}^T[a = 0]$ will be computed from (2.13),

$$\tilde{M}^T[a = 0] = \sum_{\alpha=1}^{n} \frac{\bar{\Phi}_{\alpha r+1}}{f(-\lambda_\alpha)} \int_0^T t^r e^{i\bar{\lambda}_\alpha t} \varphi(t) dt \tag{3.54}$$

$$= \sum_{\alpha=1}^{n} \frac{\bar{\Phi}_{\alpha r+1}}{f(-\lambda_\alpha)} \sum_{\gamma=1}^{m} \sum_{j=1}^{r_\gamma+1} \frac{\Phi_{\beta j}}{2\pi} \int_{-\infty-ic}^{+\infty-ic} \frac{\left[\int_0^T t^r e^{(i\bar{\lambda}_\alpha + i\lambda)t} dt\right]}{(\lambda+\lambda_\beta)^j} d\lambda$$

$$= \sum_{\alpha=1}^{n} \sum_{\gamma=1}^{m} \sum_{j=1}^{r_\gamma+1} \frac{\bar{\Phi}_{\alpha r+1}}{f(-\lambda_\alpha)} \Phi_{\beta j} \frac{i}{(j-1)!} \int_0^T t^r (it)^{j-1} e^{i(\lambda_\alpha - \lambda_\beta)t} dt .$$

The dominant terms are for $\gamma = \alpha$ and $j = r + 1$, so that

$$\tilde{M}^T[a = 0] \sim \frac{i^{r+1} T^{2r+1}}{r!(2r+1)} \sum_{\alpha} \frac{|\Phi_{\alpha r+1}|^2}{f(-\lambda_\alpha)} \tag{3.55}$$

For estimate (2.13) the variance can be computed as follows:

$$\text{variance } \tilde{M}^T\tilde{k}^T[a=0] \tag{3.56}$$

$$= \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} \frac{\bar{\Phi}_{\alpha r+1}\Phi_{\beta r+1}}{f(-\lambda_\alpha)f(-\lambda_\beta)} \int_0^T\int_0^T u^r v^r e^{i\lambda_\alpha u - i\lambda_\beta v} R(u-v)\,du\,dv$$

$$= \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} \frac{\bar{\Phi}_{\alpha r+1}\Phi_{\beta r+1}}{f(-\lambda_\alpha)f(-\lambda_\beta)} \frac{1}{2\pi} \int_{-\infty}^{\infty} f(\lambda)$$

$$\left[\int_0^T\int_0^T u^r v^r e^{i\lambda_\alpha u - i\lambda_\beta v + iu\lambda - iv\lambda}\,du\,dv\right] d\lambda$$

$$= \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} \bar{\Phi}_{\alpha r+1}\Phi_{\beta r+1} \frac{1}{2\pi} \int_{-\infty}^{\infty} f(\lambda)$$

$$\left[\sum_{j=0}^{r} \frac{e^{iT(\lambda+\lambda_\alpha)}T^{r-j}r!(-1)^j}{(i\lambda_\alpha+i\lambda)^{j+1}(r-j)!} + \frac{r!(-1)^{r+1}}{(i\lambda_\alpha+i\lambda)^{r+1}}\right]$$

$$\left[\sum_{k=0}^{r} \frac{e^{-iT(\lambda+\lambda_\beta)}T^{r-j}r!}{(i\lambda_\beta+i\lambda)^{j+1}(r-k)!} + \frac{r!}{(i\lambda_\beta+i\lambda)^{r+1}}\right] d\lambda \quad .$$

Evaluating each term by residues, the only terms of order T^{2r+1} are the cross product terms for $\alpha = \beta$. The contour must be deformed around the poles and then closed upward for one cross product term and downward for the other. Thus only one of them contains the poles of interest. As before, the poles of $F(\lambda)$ and $\bar{F}(\lambda)$ will contribute terms containing the factors $e^{iT(\eta_\gamma+\lambda_\alpha)}$ where η_γ are the poles of $F(\lambda)$ and hence $\Im(\eta_\gamma) > 0$, so that terms of this type are asymptotically negligible. Thus the asymptotic expression becomes

$$\text{variance } \tilde{M}^T\tilde{k}^T[a = 0]$$

$$\sim \sum_{\alpha=1}^{n} \frac{|\Phi_{\alpha r+1}|^2}{[f(-\lambda_\alpha)]^2} \, if(-\lambda_\alpha) \sum_{j=0}^{n} \frac{T^{r-j}(iT)^{j+r-1}(r')^2(-1)^j}{(j+r+1)!\,i^{r+j+2}(r-j)!} \tag{3.57}$$

$$= \frac{T^{2r+1}}{2r+1} \sum_{\alpha=1}^{n} \frac{|\Phi_{\alpha r+1}|^2}{f(-\lambda_\alpha)}$$

Combining expressions (3.42), (3.55), and (3.57), it is seen that estimate (2.13) is also efficient This concludes the proof of Theorem 1.

Corollary 1. Under the assumptions of Theorem 1, it follows that

$$\frac{\text{variance } \hat{k}^T[y(t),\ 0 \leq t \leq T]}{\text{variance } \hat{k}^T[y(t),\ -\infty < t \leq T]} \to 1 \quad \text{as} \quad T \to \infty \tag{3.58}$$

Proof: In the proof of Theorem 1 it was shown that

$$\{\text{variance } \tilde{k}^T[y(t),\ 0 \leq t \leq T]\} \left\{\int_{-\infty}^{T} |\dot{\psi}(t)|^2 dt\right\} \to 1 \tag{3.59}$$

and hence

$$1 \leq \frac{\text{variance } \hat{k}^T[y(t),\ 0 \leq t \leq T]}{\text{variance } \hat{k}^T[y(t),\ -\infty < t \leq T]} \leq \frac{\text{variance } \tilde{k}^T[y(t),\ 0 \leq t \leq T]}{\text{variance } \hat{k}^T[\eta(t),\ -\infty < t \leq T]}$$

$$= \{\text{variance } \tilde{k}^T[y(t),\ 0 \leq t \leq T]\}\left\{\int_{-\infty}^{T} |\dot{\psi}(t)|^2 dt\right\} \to 1 \,. \tag{3.60}$$

Corollary 2. If a linear unbiased estimate is efficient for process I, then it is efficient for processes II and III.

Proof: This follows immediately from Corollary 1.

Section 4
EFFICIENCY CLASSES

In this section a necessary condition that a density g be in the efficiency class $\mathcal{F}(f,\varphi)$ will be established.

Lemma 1. If there exists a linear unbiased estimate $\overline{k}^T$ that is efficient for either problem I or II, then it follows that

$$\int_{-\infty}^{\infty} \left| \frac{\overline{\Psi^T(\lambda)}}{b^T} - \frac{F(\lambda)K^T(\lambda)}{a^T} \right|^2 d\lambda \to 0 \quad \text{as } T \to \infty \tag{4.1}$$

where

$$K^T(\lambda) = \overline{k}^T[e^{i\lambda t},\ -\infty < t \leq T] , \tag{4.2}$$

$$\Psi^T(\lambda) = \int_{-\infty}^{T} e^{-i\lambda t}\dot{\psi}(t)dt , \tag{4.3}$$

$$b^T = \sqrt{\int_{-\infty}^{\infty} |\Psi^T(\lambda)|^2 d\lambda} , \tag{4.4}$$

and

$$a^T = \sqrt{\int_{-\infty}^{\infty} |F(\lambda)K^T(\lambda)|^2 d\lambda} . \tag{4.5}$$

Proof: From reference [2] page 534 it is clear that the linear estimate $\overline{k}^T$ can be represented as

$$\bar{k}^T[y(t),\ -\infty < t \leq T] = \bar{k}^T\left[\frac{1}{2\pi}k\int_{-\infty}^{\infty} e^{i\lambda t}Z^T(\lambda)d\lambda\right.$$

$$\left. + \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\lambda t}dz(\lambda),\ -\infty < t \leq T\right] \qquad (4.6)$$

$$= k\,\frac{1}{2\pi}\int_{-\infty}^{\infty} K^T(\lambda)Z^T(\lambda)d\lambda + \frac{1}{2\pi}\int_{-\infty}^{\infty} K^T(\lambda)dz(\lambda)$$

where $dz(\lambda)$ is the spectral process of $x(t)$ with $E[|dz(\lambda)|^2] = f(\lambda)d\lambda$, $K^T(\lambda)$ is given by (4.2), and $Z^T(\lambda)$ by

$$Z^T(\lambda) = \int_{-\infty}^{T} e^{-i\lambda t}{}_{,}(t)dt = \Phi^T(\lambda) + H(\lambda)\ . \qquad (4.7)$$

Unbiasedness of the estimate implies that

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} K^T(\lambda)Z^T(\lambda)d\lambda = 1\ . \qquad (4.8)$$

The variances of the estimates $\bar{k}^T$ and $\hat{k}^T[\eta(t)]$ are given by

$$\text{variance } \bar{k}^T = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty} |K^T(\lambda)|^2 f(\lambda)d\lambda \qquad (4.9)$$

$$= \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty} |K^T(\lambda)F(\lambda)|\ d\lambda$$

and

$$1/\text{variance } \hat{k}^T[\eta(t)] = \int_{-\infty}^{T} |\dot{\psi}(t)|^2 dt = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty} |\Psi^T(\lambda)|^2 d\lambda\ . \qquad (4.10)$$

Applying the argument used to derive equation (3.23) to equation (3.8), it is seen that

$$Z^T(\lambda) = F(\lambda)\Psi^T(\lambda) \ . \tag{4.11}$$

From Corollary 2 the estimate $\overline{k}^T$ is also efficient for process III, so that

$$a^T b^T = \sqrt{\frac{\text{variance } \overline{k}^T[y(t),\ -\infty < t \leq T]}{\text{variance } \hat{k}^T[n(t),\ -\infty < t \leq T]}} \to 1 \quad \text{as } T \to \infty \ . \tag{4.12}$$

The result can now be derived as follows:

$$\int_{-\infty}^{\infty} \left| \frac{\overline{\Psi^T(\lambda)}}{b^T} - \frac{F(\lambda)K^T(\lambda)}{a^T} \right|^2 d\lambda = \int_{-\infty}^{\infty} \left[\frac{\overline{\Psi^T(\lambda)}}{b^T} - \frac{F(\lambda)K^T(\lambda)}{a^T} \right] \left[\frac{\Psi^T(\lambda)}{b^T} - \frac{\overline{F(\lambda)K^T(\lambda)}}{a^T} \right] d\lambda . \tag{4.13}$$

$$= \left(\frac{b^T}{b^T}\right)^2 + \left(\frac{a^T}{a^T}\right)^2 - \frac{2}{a^T b^T} \mathcal{R} \int_{-\infty}^{\infty} F(\lambda)K^T(\lambda)\Psi^T(\lambda)d\lambda$$

$$= 2 - \frac{2}{a^T b^T} \to 1 \ .$$

Lemma 2. If there exists a linear unbiased estimate that is efficient for either process I or II and for the two problems $[\varphi(t), f(\lambda)]$ and $[\varphi(t), g(\lambda)]$; that is, if $g \in \mathcal{F}(\varphi, f)$, then

$$\int_{-\infty}^{\infty} \left| \frac{\overline{\Psi_F^T(\lambda)}G(\lambda)}{b_F^T} - \frac{a_G^T}{a_F^T} \frac{\overline{\Psi_G^T(\lambda)}}{b_G^T} F(\lambda) \right|^2 d\lambda \to 0 \quad \text{as } T \to \infty \ . \tag{4.14}$$

Proof: From an elementary inequality, it is clear that

$$\int_{-\infty}^{\infty} \left| \frac{\overline{\Psi_F^T(\lambda)} G(\lambda)}{b_F^T} - \frac{a_G^T}{a_F^T} \frac{\overline{\Psi_G^T(\lambda)} F(\lambda)}{b_F^T} \right|^2 d\lambda$$

$$\leq 2 \int_{-\infty}^{\infty} \left\{ \left| G(\lambda) \left[\frac{\overline{\Psi_F^T(\lambda)}}{b_F^T} - \frac{K^T(\lambda) F(\lambda)}{a_F^T} \right] \right|^2 + \left| \frac{a_G^T F(\lambda)}{a_F^T} \left[\frac{\overline{\Psi_G^T(\lambda)}}{b_G^T} - \frac{K^T(\lambda) G(\lambda)}{a_G^T} \right] \right|^2 \right\} d\lambda \quad . \tag{4.15}$$

Since $G(\lambda)$ and $F(\lambda)$ are rational functions with no real poles, they are bounded

$$\begin{aligned} |G(\lambda)| &\leq M \\ |F(\lambda)| &\leq M \end{aligned} \tag{4.16}$$

for λ real and hence

$$\int_{-\infty}^{\infty} \left| \frac{\Psi_F^T(\lambda) G(\lambda)}{b_F^T} - \frac{a_G^T}{a_F^T b_F^T} \overline{\Psi_G^T(\lambda)} F(\lambda) \right|^2 d\lambda$$

$$\leq 2 \left\{ M^2 \int_{-\infty}^{\infty} \left| \frac{\overline{\Psi_F(\lambda)}}{b_F^T} - \frac{K^T(\lambda) F(\lambda)}{a_F^T} \right|^2 d\lambda + \left(\frac{a_G^T}{a_F^T} \right)^2 M^2 \int_{-\infty}^{\infty} \left| \frac{\overline{\Psi_G^T(\lambda)}}{b_G^T} - \frac{K^T(\lambda) G(\lambda)}{a_G^T} \right|^2 d\lambda \right\} \tag{4.17}$$

From (3.40) and (3.42)

$$\left(b_F^T \right)^2 \sim T^{2r} e^{2a} C_F(T) \quad \text{or} \quad T^{2r+1} C \tag{4.18}$$

where $C(T)$ is bounded from above and away from zero. From efficiency $a_F^T b_F^T \to 1$ so that

$$a_F^T \sim \frac{T^{-2r} e^{-2a}}{C_F(T)} \quad \text{or} \quad \frac{T^{-(2r+1)}}{C} \quad . \tag{4.19}$$

Thus the expression

$$\left(\frac{a_G^T}{a_F^T}\right)^2 \sim \frac{C_F(T)}{C_G(T)} \quad \text{or} \quad \frac{C_F}{C_G} \tag{4.20}$$

is bounded and the result follows from Lemma 1.

Theorem 3. If there exists a linear unbiased estimate that is efficient for either process I or II and for the two problems $[\varphi(\lambda), f(\lambda)]$ and $[\varphi(\lambda), g(\lambda)]$; that is, if $g \in \mathcal{F}[\varphi, f]$, then (i) if $a > 0$, it follows that

$$\left| \frac{C_1}{G(\lambda)} \sum_{\alpha=1}^{n} \frac{\overline{\Phi}_{\alpha r+1}}{\overline{G(-\lambda_\alpha)}} \frac{e^{iT\rho\lambda_\alpha}}{(\overline{\lambda}_\alpha+\lambda)} - \frac{1}{F(\lambda)} \sum_{\alpha=1}^{n} \frac{\overline{\Phi}_{\alpha r+1}}{\overline{F(-\lambda_\alpha)}} \frac{e^{iT\rho\lambda_\alpha}}{(\overline{\lambda}_\alpha+\lambda)} \right| \to 0 \tag{4.21}$$

as $T \to \infty$,

and (ii) if $a = 0$,

$$F(-\lambda_\alpha) = C_2 G(-\lambda_\alpha) \qquad \alpha = 1, \cdots, n , \tag{4.22}$$

where C_1 and C_2 are constants.

Proof: (i) For any sequence T_n there exists a subsequence T_j, and complex numbers ℓ_α such that

$$e^{iT_j\rho\lambda_\alpha} \to \ell_\alpha \quad \text{as } j \to \infty , \ \alpha = 1, \cdots, n . \tag{4.23}$$

For this subsequence, it follows from (3.27), (4.18) and (4.19) that

$$\left| \frac{\overline{\Psi_F^{T_j}(\lambda)}G(\lambda)}{b_F^{T_j}} - \frac{a_G^{T_j}\overline{\Psi_G^{T_j}(\lambda)}F(\lambda)}{a_F^{T_j}b_G^{T_j}} \right|$$

$$\sim \left| \frac{\frac{T_j^r e^{iT_j\lambda} e^{aT_j}}{r!}}{T_j^r e^{aT_j}} \left[\frac{G(\lambda)}{\sqrt{C_F(T_j)}} \sum_{\alpha=1}^{n} \frac{\overline{\Phi}_{\alpha r+1} e^{iT_j\overline{\lambda}_\alpha}}{(\lambda+\overline{\lambda}_\alpha)\overline{F(-\lambda_\alpha)}} \right.\right. \tag{4.24}$$

$$\left.\left. - \frac{\sqrt{C_F(T_j)}}{C_G(T_j)} F(\lambda) \sum_{\alpha=1}^{n} \frac{\overline{\Phi}_{\alpha r+1}}{(\lambda+\overline{\lambda}_\alpha)} \frac{e^{iT_j\overline{\lambda}_\alpha}}{\overline{G(-\lambda_\alpha)}} \right] \right|$$

$$\sim \left| \frac{G(\lambda)}{\sqrt{C_F}} \sum_{\alpha=1}^{n} \frac{\overline{\Phi}_{\alpha r+1}}{(\lambda+\overline{\lambda}_\alpha)} \frac{\ell_\alpha}{\overline{F(-\lambda_\alpha)}} - \frac{\sqrt{C_F}}{C_G} F(\lambda) \sum_{\alpha=1}^{n} \frac{\overline{\Phi}_{\alpha r+1}\ell_\alpha}{\overline{G(-\lambda_\alpha)}(\overline{\lambda}_\alpha+\lambda)} \right|$$

From Lemma 2 this limit must be identically zero.

(ii) For $a = 0$, the integral (4.14) will be evaluated. Making the substitution $\Psi^T(\lambda) = Z^T(\lambda) / F(\lambda)$ for $\Psi_F^T(\lambda)$ and $\Psi_G^T(\lambda)$, this becomes

$$\int_{-\infty}^{\infty} \left| \frac{\overline{\Psi_F^T(\lambda)}G(\lambda)}{b_F^T} - \frac{a_G^T}{a_F^T} \frac{\overline{\Psi_G^T(\lambda)}F(\lambda)}{b_G^T} \right|^2 d\lambda$$

$$= \int_{-\infty}^{\infty} \left| \frac{G(\lambda)}{F(\lambda)} - \frac{a_G^T b_F^T}{a_F^T b_G^T} \frac{F(\lambda)}{G(\lambda)} \right|^2 \left| \frac{Z^T(\lambda)}{b_F^T} \right|^2 d\lambda \ . \tag{4.25}$$

It is clear from the discussion in section 3 that the extension of $\varphi(t)$ to $\mathcal{J}(t)$ depends only on the degree of $F(\lambda)$ and hence the same extension $Z^T(\lambda)$ will do for both $F(\lambda)$ and $G(\lambda)$. The function $Z(\lambda)$ will be evaluated by closing the contour in the integral

$$Z^T(\lambda) = \frac{1}{2\pi} \int_{-\infty - ic}^{+\infty - ic} \frac{e^{(i\omega - i\lambda)T}}{(i\omega - i\lambda)} [\Phi(\omega) + H(\omega)] d\omega \ , \tag{4.26}$$

upward around the poles of $\Phi(\omega)$ and $\omega = \lambda$ and computing the residues

$$Z^T(\lambda) = \Phi(\lambda) + H(\lambda) + \sum_{\gamma=1}^{m} \sum_{j=1}^{r_\gamma+1} \frac{\Phi_{\gamma j} e^{-i\lambda T}}{(j-1)!} \frac{\partial^{j-1}}{\partial \omega^{j-1}} \frac{e^{i\omega T}}{(\omega - \lambda)} \Bigg|_{\omega = -\lambda_\gamma} \tag{4.27}$$

$$= H(\lambda) + \sum_{\gamma=1}^{m} \sum_{j=1}^{r_\gamma+1} \Phi_{\gamma j} \left[\frac{1}{(\lambda+\lambda_\gamma)^j} \right.$$

$$\left. + \sum_{k=0}^{j-1} \frac{e^{-i\lambda T}}{(j-1)!} \binom{j-1}{k} \frac{(iT)^k e^{-i\lambda_\gamma T}}{(\lambda+\lambda_\gamma)^{j-k}} (j-1+k)!(-1) \right] .$$

Since $b_F^T \sim T^{2r+1} C_F$, terms of magnitude T^{2r+1} must be found in (4.25). Terms containing the factor $H(\lambda)$ will not be of this magnitude. Thus making the substitution (4.27) in (4.25)

$$\int_{-\infty}^{\infty} \left| \frac{\overline{\Psi_F^T(\lambda) G(\lambda)}}{b_F^T} - \frac{a_G^T}{a_F^T b_G^T} \overline{\Psi_G^T(\lambda)} F(\lambda) \right|^2 d\lambda$$

$$\sim \frac{1}{(b_F^T)^2} \sum_{\gamma=1}^{m} \sum_{\delta=1}^{m} \sum_{i=1}^{r_\gamma+1} \sum_{j=1}^{r_\delta+1} \Phi_{\gamma i} \overline{\Phi}_{\delta i} \int_{-\infty}^{\infty} \left| \frac{G(\lambda)}{F(\lambda)} \right. \tag{4.28}$$

$$\left. - \frac{a_G^T b_F^T F(\lambda)}{a_F^T b_G^T G(\lambda)} \right|^2 \left[\frac{1}{(\lambda+\lambda_\gamma)^i} + e^{-i\lambda T} \sum_{k=0}^{i-1} \frac{(iT)^k e^{-i\lambda_\gamma T}}{(\lambda+\lambda_\gamma)^{i-k}} \frac{(-1)}{k!} \right]$$

$$\left[\frac{1}{(\lambda+\lambda_\delta)^{j+1}} + e^{i\lambda T} \sum_{\ell=0}^{j-1} \frac{(-iT)^\ell e^{i\lambda_\delta T}}{(\lambda+\lambda_\delta)^{j-\ell}} \frac{(-1)}{\ell!} \right] d\lambda$$

Terms of maximum order are obtained when $\gamma = \delta = \alpha$ and $i = j = r + 1$ for the cross product terms. The contour may be deformed downward around the poles at $\lambda = -\lambda_\alpha$ $\alpha = 1, \cdots, n$. Again the poles of $\left|\frac{G(\lambda)}{F(\lambda)} - \frac{a_G^T b_F^T}{a_F^T b_G^T}\frac{F(\lambda)}{G(\lambda)}\right|^2$ will not contribute terms of maximum order.

Evaluating by residues,

$$\int_{-\infty}^{\infty}\left|\frac{\overline{\Psi_F^T(\lambda)}G(\lambda)}{b_F^T} - \frac{a_G^T}{a_F^T b_G^T}\overline{\Psi_G^T(\lambda)}F(\lambda)\right|^2 d\lambda \tag{4.29}$$

$$\sim \frac{2\pi i}{C_F T^{2r+1}} \sum_{\alpha=1}^{n} |\Phi_{\alpha r+1}|^2 \left|\frac{G(-\lambda_\alpha)}{F(-\lambda_\alpha)} - \frac{C_F}{C_G}\frac{F(-\lambda_\alpha)}{G(-\lambda_\alpha)}\right|^2 \sum_{\ell=0}^{r} \frac{(-iT)^{\ell}(-1)(iT)^{2r+1-\ell}}{\ell!(2r+1-\ell)!}$$

$$= \frac{2\pi}{(2r+1)(r!)^2} \sum_{\alpha=1}^{n} |\Phi_{\alpha r+1}|^2 \left|\frac{G(-\lambda_\alpha)}{F(-\lambda_\alpha)} - \frac{C_F}{C_G}\frac{F(-\lambda_\alpha)}{G(-\lambda_\alpha)}\right|^2 .$$

From Lemma 2 this limit must be zero, and since $\Phi_{\alpha r+1} \neq 0$, it follows that (4.22) is satisfied.

In the case $a = 0$, it is clear that the estimate $\tilde{k}^T[y(t), 0 \leq t \leq T; a = 0]$ given by (2.13) is efficient for all densities $g(\lambda)$ which satisfy condition (ii) of the theorem. Thus condition (ii) is a necessary and sufficient condition that $g(\lambda)$ belong to $\mathcal{F}(f,\varphi)$, and the estimate (2.13) is economical of information concerning the spectrum $f(\lambda)$ in the sense that there exists no linear unbiased estimate which is efficient for a wider class of spectra.

In the case $a > 0$, the theorem states that if there exists an estimate that is efficient for $[\varphi(t), f(\lambda)]$ and $[\varphi(t), g(\lambda)]$ on a sequence T_j for which condition (4.23) holds, then the spectra $f(\lambda)$ and $g(\lambda)$ are related by

$$\frac{C}{G(\lambda)} \sum_{\alpha=1}^{n} \frac{\overline{\Phi}_{\alpha r+1} \ell_\alpha}{G(-\lambda_\alpha)(\lambda+\bar{\lambda}_\alpha)} = \frac{1}{F(\lambda)} \sum_{\alpha=1}^{n} \frac{\overline{\Phi}_{\alpha r+1} \ell_\alpha}{\overline{F(-\lambda_\alpha)}(\lambda+\bar{\lambda}_\alpha)} \quad . \tag{4.30}$$

For $n = 1$ or 2 it can be shown that this implies that $g(\lambda) = Cf(\lambda)$ and hence that estimates $\tilde{k}_1^T[y(t), 0 \leq t \leq T; a > 0]$ and $\tilde{k}_2^T[y(t), 0 \leq t \leq T; a > 0]$ given by (2.10) and (2.11) are economical in the same sense stated above for $\tilde{k}^T[y(t), 0 \leq t \leq T, a = 0]$. However, the economy of estimates (2.10) and (2.11) is of little practical interest since it is very unlikely that the function

$$\frac{1}{F(\lambda)} \sum_{\alpha=1}^{n} \frac{\overline{\Phi}_{\alpha r+1} \ell_\alpha}{\overline{F(-\lambda_\alpha)}(\lambda+\bar{\lambda}_\alpha)} \tag{4.31}$$

would in practice be known if the function $F(\lambda)$ were unknown. This is not the case for the estimate $\tilde{k}^T[y(t), 0 \leq t \leq T; a = 0]$, since here a real reduction is made in the information required concerning $f(\lambda)$. For instance, if $n = 1$, $\tilde{k}^T[a = 0]$ becomes essentially the least square estimate and no knowledge of $f(\lambda)$ is required. For larger n, if the spectrum is to be estimated from an independent experiment, then estimates need only be made at the frequencies of interest $-\lambda_\alpha$, $\alpha = 1, \cdots, n$, considerably reducing the work.

For the least square estimate $\bar{k}_L^T$ given by (1.4), the function $K^T(\lambda)$ becomes

$$K^T(\lambda) = \frac{\int_0^T \overline{\varphi(t)} e^{i\lambda t} dt}{\int_0^T |\varphi(t)|^2 dt} = \frac{\overline{\Phi^T(\lambda)}}{\int_0^T |\varphi(t)|^2 dt} \tag{4.32}$$

For the pseudo-spectral density $g(\lambda) = 1$,

$$\Psi_G^T(\lambda) = Z^T(\lambda) = \Phi^T(\lambda) + H(\lambda) , \tag{4.33}$$

and

$$K^T(\lambda) = \frac{\overline{\Psi_G^T(\lambda)} - \overline{H(\lambda)}}{\int_0^T |\varphi(t)|^2 dt} . \tag{4.34}$$

If this expression is substituted in (4.1) and terms involving $H(\lambda)$ are neglected, the expression (4.25) is obtained and the following results concerning the least square estimate can be deduced: If the least square estimate is efficient for $f(\lambda)$; then if $a = 0$

$$f(-\lambda_\alpha) = C_2 \qquad \alpha = 1, \cdots, n , \tag{4.35}$$

and if $a > 0$

$$\left| \frac{1}{F(\lambda)} \sum_{\alpha=1}^{n} \frac{\overline{\Phi}_{\alpha r+1} e^{iT\varphi(\lambda_\alpha)}}{\overline{F(-\lambda_\alpha)}(\lambda+\overline{\lambda}_\alpha)} - C_1 \sum_{\alpha=1}^{n} \frac{\overline{\Phi}_{\alpha r+1} e^{iT\varphi(\lambda_\alpha)}}{(\lambda+\overline{\lambda}_\alpha)} \right| \to 0 \tag{4.36}$$

as $T \to \infty$.

Section 5

REFERENCES

[1] Luigi Amerio, "Sull' inversione della transformata di Laplace," Rendiconti delle Acc. della Scienze Fisiche e Matematiche, Napoli, Series IV, Vol. X (1940), p. 232

[2] J. L. Doob, Stochastic Processes, John Wiley and Sons, Inc., New York, 1953.

[3] U. Grenander, "On the estimation of regression coefficients in the case of an autocorrelated disturbance," Annals of Math. Stat. Vol. 25 (1954), p. 252.

[4] U. Grenander and M. Rosenblatt, "An extension of a theorem of G. Szegö and its application to the study of stochastic processes," Trans. Amer. Math. Soc. Vol. 76 (1954), p. 112.

[5] U. Grenander and M. Rosenblatt, "Regression analysis of time series with stationary residuals," Proc. Nat. Acad. Sci. Vol 40 (1954), p. 812.

[6] U. Grenander and M. Rosenblatt, Statistical Analysis of Stationary Time Series, John Wiley and Sons, Inc., New York, 1957 .

[7] U. Grenander and Gabor Szegö, Toeplitz Forms and Their Application, University of California Press, Berkeley and Los Angeles, 1958

[8] J. H. Laning, Jr. and R. H. Battin, Random Processes in Automatic Control, McGraw-Hill Book Company, New York, 1956

[9] H. B. Mann, "A theory of estimation for the fundamental random process and the Ornstein Uhlenbeck process," Sankhyā, Vol. 13 (1954), p. 325.

[10] H. B. Mann and P. B. Moranda, "On the efficiency of the least square estimates of parameters in the Ornstein Uhlenbeck process," Sankhyā, Vol. 13 (1954), p. 351.

[11] M. Rosenblatt, "On the estimation of regression coefficients of a vector-valued time series with stationary residuals," Annals of Math. Stat., Vol. 27 (1956), p. 99.

[12] M. Rosenblatt, "Some regression problems in time series analysis," Third Berkeley Symposium, Vol. 1, p. 165, University of California Press, Berkeley and Los Angeles, 1956.

[13] C. T. Striebel, "On the efficiency of estimates of trend in the Ornstein Uhlenbeck process," Annals of Math. Stat., Vol. 29 (1958), p. 192.

[14] B. van der Pol and H. Bremmer, Operational Calculus, Cambridge University Press, 1955.

A MODEL OF TURBULENCE DISPLAYING AUTONOMOUS OSCILLATIONS

5

R.J. Dickson

FOREWORD

This paper was published originally under the same title, as LMSD-48446, 25 March 1959.

The principal result of Section 3 was presented at the annual meeting of the American Mathematical Society, Philadelphia, Penn., 20–22 January 1959. A technical abstract (#553-84) appears in Notices of the A. M. S., Vol. 5, No. 7, Dec 1958, pp. 822–823.

ABSTRACT

This report presents the results of the author's attempt to find and treat simplified examples of problems governed by nonlinear space-time systems. The work was carried out under the General Research Program of Lockheed Missiles and Space Division.

The introduction sketches the general setting of a "Model of Turbulence" in the sense of J. M. Burgers, and directs attention to the character of the particular problem considered here.

Section 2 introduces a special class of nonlinear ordinary differential equations and explains its relationship to a partial differential system studied by Burgers.

Section 3 takes the first step in a detailed study of this class of equations. Specifically, it is proved that a certain fourth-order autonomous system of ordinary equations possesses an unstable equilibrium solution and a stable periodic solution; and that with increasing time all other solutions tend asymptotically to one of these two possibilities.

CONTENTS

Section 1
INTRODUCTION

Hydrodynamic turbulence is a variety of parasitic oscillation. Its qualitative features are easily observed in nature and in the laboratory, yet their mathematical description has remained for many years at a primitive level. Almost no progress has been made toward proving – or disproving – that the dynamical (Navier-Stokes) equations of fluid mechanics possess solutions consistent with observed turbulent flows.

To gain insight into the mathematical structure of turbulence, J. M. Burgers initiated in 1939 the study of "models of turbulence" (Refs. 1 and 2). Typically, one of Burgers' models is a mathematical problem which preserves selected features of the hydrodynamical problem yet remains simple enough to admit analytical study; it does not correspond to any real physical problem, but is studied formally as a mathematical problem for its suggestiveness and for possible insight into the nature of its more complex parent.

The relation of Burgers' work to the existing literature on the subject of "isotropic turbulence" may be clarified by the following quotation (Ref. 1, p. 47):

> It should be observed that the theories developed by Taylor, von Kármán and others are concerned mainly with a different problem, viz., the character of the turbulence found, e.g., in an air current in the wake of a grid, and its gradual decay. These theories do not consider the side of the problem to which attention has been given here, viz., the development of a dissipative secondary phenomenon, which grows by detracting energy from a given primary phenomenon and in this process gains such an intensity, that finally a balance is obtained between the energy detracted and the energy dissipated. This latter subject is the one which has been traced here through stages of successive complexity . . . and it will be clear that it has a

> generality much wider than the field of hydrodynamics. The problem can be compared to that of the development of relaxational oscillations, investigated by van der Pol. The classical example treated by van der Pol, however, refers to a system with a single variable, whereas the present case is characterized by the appearance of a dissipative secondary phenomenon embracing an infinite number of degrees of freedom . . .

Referring to phase flows generated by the Navier-Stokes equations, E. Hopf (Ref. 3) described the motivation for the study of models when he wrote:

> How do the solutions which represent the observed turbulent motions fit into the phase picture? The great mathematical difficulties of these important problems are well known and at present the way to a successful attack on them seems hopelessly barred. There is no doubt, however, that many characteristic features of the hydrodynamical phase flow occur in a much larger class of similar problems governed by nonlinear space-time systems. In order to gain insight into the nature of hydrodynamical phase flows we are, at present, forced to find and to treat simplified examples within that class. The study of such models has been originated by J. M. Burgers in a well known memoir . . .

This paper presents results of the author's attempt "to find and to treat simplified examples within that class." More specifically, the attempt was to find a model which compromised the shortcomings of two of Burgers' models. The first of these, which will be called here the "preliminary model," is the following differential boundary value problem for a (real valued) function $u(x,t)$ defined for positive "time," $t>0$, on the one-dimensional "space," $0 \leq x \leq 1$:

$$u_t = \nu u_{xx} + u - 2uu_x \qquad (1.1)$$

$$u(0,t) = u(1,t) = 0 \ , \ t>0 \qquad (1.2)$$

The subscripts here denote partial derivatives, while ν is a positive constant which Burgers calls the viscosity. In Burgers' "interpretation," the function u is thought of as the turbulent perturbation of an Eulerian velocity field; thus, the obvious solution $u(x,t) = 0$ is referred to as the "laminar" solution. If Eq. (1.1) is multiplied

by u and integrated over $0 \leq x \leq 1$, the contribution of the nonlinear term vanishes because of Eq. (1.2) The contribution of the viscosity term can be integrated by parts, yielding formally the relation

$$\frac{1}{2}\frac{d}{dt}\int_0^1 u^2\,dx = -\nu\int_0^1 u_x^2\,dx + \int_0^1 u^2\,dx \qquad (1.3)$$

which Burgers calls the "energy equation" because it reveals that the "energy of turbulence," $\int_0^1 u^2\,dx$, tends to be diminished (or dissipated) by the viscosity term and increased by the "driving" term u of Eq. (1.1), which plays a role analogous to the pressure gradient term of the Navier-Stokes equations. (The question of which of these influences predominates is discussed in Ref. 4, where it is shown, in particular, that Eq. (1.1) belongs to a class of equations for which a modified maximum principle holds.)

If ν is sufficiently large, it can be proved that the viscous dissipation is predominant; i.e., all solutions tend to the stationary, laminar solution as $t \to \infty$. As is decreased, however, a positive critical value is reached at which this total stability of the laminar solution is lost. On the basis of hydrodynamical evidence, it would be reasonable to conjecture that this loss of stability marks the entry into a "turbulent regime" in which a time-dependent periodic solution appears and that further decrease of ν will lead to successive branchings at which more and more complicated oscillatory solutions become possible. (For a detailed conjecture on the nature of the branching process and illustrative examples, see Ref. 3.) So far as is known, however, the model under consideration is not sufficiently sophisticated to display such behavior. A branching process does indeed occur, but the "turbulent" solutions which appear are stationary, i.e., time independent. (For details see Refs. 1 and 2.)

In constructing a more complicated turbulence model, Burgers ingeniously excluded the possibility of stationary solutions other than the laminar one. Specifically, he

proposed the system

$$v_t = \nu v_{xx} + v - w + (w^2 - v^2)_x$$
$$w_t = \nu w_{xx} + v + w + (2vw)_x \tag{1.4}$$

for two functions $v(x,t)$, $w(x,t)$ on $0 \le x \le 1$, $t \ge 0$ subject to

$$v(0,t) = v(1,t) = w(0,t) = w(1,t) = 0, \quad t \ge 0 \tag{1.5}$$

It is convenient to combine Eqs. (1.4) into one equation for the complex valued function

$$u(x,t) = v(x,t) + iw(x,t)$$

Burgers' model is then

$$u_t = \nu u_{xx} + (1+i)u - 2\bar{u}\bar{u}_x, \quad \bar{u} = v - iw \tag{1.6}$$

$$u(0,t) = u(1,t) = 0 \tag{1.7}$$

Multiplication of Eq. (1.6) by $\bar{u}$, integration over $0 \le x \le 1$, and integration by parts yield

$$\int_0^1 \bar{u}\, u_t \, dx = -\nu \int_0^1 |u_x|^2 \, dx + (1+i) \int_0^1 |u|^2 \, dx \tag{1.8}$$

The real and imaginary parts of Eq. (1.8) are

$$\frac{1}{2}\frac{d}{dt}\int_0^1 |u|^2 \, dx = -\nu \int_0^1 |u_x|^2 \, dx + \int_0^1 |u|^2 \tag{1.9}$$

and

$$\frac{1}{2i}\int_0^1 (\bar{u}\,u_t - u\,\bar{u}_t)dx = \int_0^1 |u|^2\,dx \qquad (1.10)$$

Equation (1.9) is the energy equation, exactly analogous to Eq. (1.3); Eq. (1.10) shows that $u = 0$ is the only solution independent of the time. If the model has any turbulent solutions at all, they are not stationary.

But are there any turbulent solutions? As with the preliminary model, it is found that the laminar solution loses stability below a positive critical value of viscosity. For subcritical ν, there is presumably a turbulent regime. Although Burgers presented heuristic arguments favoring the existence and boundedness of solutions for all time, essentially nothing is known with certainty about the behavior of solutions in the turbulent range of ν.

Whereas the preliminary model was too simple to display time-dependent turbulent solutions, this model is not simple enough to permit analysis. It was an attempt to formulate models of intermediate complexity which generated the results of this paper.

Section 2

A MODIFIED MODEL

INTRODUCTION

Numerical approximations to the solution of a partial differential boundary value problem are ordinarily obtained with digital or electrical analog computers. Such machines will accept the problem only in a modified or analogous form. For a digital computer, the problem must be replaced by one of its approximate "discretized" versions requiring only algebraic operations. For an analog machine, the discretization need not be complete; in a space-time system, for example, it is sufficient to "lump" space alone, leaving time as a continuous variable.

In either case, metamorphosis of the problem requires that functions of a continuous variable be replaced by functions defined on a finite set of points, and that derivatives be replaced by algebraic differences. There is freedom of choice in the number "n" of points, i.e., the "fineness" of the discretization, and in the particular difference scheme employed. No criteria for making these choices are dictated by the original differential problem. Mathematical problems of considerable depth arise from the questions of how to make these choices in an optimal way and how to assess the goodness of the resulting approximation.

The more or less arbitrary choice of a particular difference scheme is often justified by the observation that the distinction between formally admissible schemes vanishes as $n \to \infty$, i.e., in the limit of refinement of the discretization. It may occur, however, that for <u>small</u> values of n certain special schemes are most effective in preserving some kind of <u>analogy</u> between the exact and approximate problems. In discretizing the "space" $0 \leq x \leq 1$ involved in Burgers' model Eq. (1.6), the most obvious and straightforward choice is poor since it does not yield for each n an exact analog of Eq. (1.8).

Specifically, let $u_k(t)$ denote an approximation to $u(k/n,t)$, $k = 1, \ldots, n-1$ and set $u_0(t) = u(0,t) \equiv 0$, $u_n(t) = u(1,t) \equiv 0$. Suppose the u_k to be solutions of the following system of ordinary differential equations obtained from Eq. (1.6) by replacing spatial derivatives by central divided differences:

$$\frac{du_k}{dt} \equiv \dot{u}_k = \nu \frac{u_{k+1} - 2u_k + u_{k-1}}{1/n^2} + (1+i)\, u_k - 2\bar{u}_k \frac{\bar{u}_{k+1} - \bar{u}_{k-1}}{2/n}$$

$$k = 1, \ldots, n-1$$

Multiplying by $\bar{u}_k$ and summing on k, it is easily found that the exact analog of Eq. (1.8) does not hold because the contribution from the nonlinear terms does not vanish identically, i.e.,

$$\sum_{k=1}^{n-1} u_k^2 (u_{k+1} - u_{k-1}) \not\equiv 0 .$$

In the attempt to find an approximating system without this imperfection, it has been found sufficient to approximate the function $\bar{u}$ in the nonlinear term $\bar{u}\bar{u}_x$ by the average $(\bar{u}_{k+1} + \bar{u}_k + \bar{u}_{k-1})/3$. Properties of the resulting system will be summarized as Theorem 2.1.

THEOREM 2.1:

If

$$\nu > 0, \quad n \geq 2, \quad u_o = u_n = 0, \tag{2.1}$$

the system

$$\dot{u}_k = \nu n^2 (u_{k+1} - 2u_k + u_{k-1}) + (1+i)\, u_k - \frac{n}{3} (\bar{u}_{k-1} + \bar{u}_k + \bar{u}_{k+1})(\bar{u}_{k+1} - \bar{u}_{k-1}) \tag{2.2}$$

for the (complex valued) functions $u_k(t)$, $k = 1, 2, \ldots, n-1$ has the following properties:

- Every solution satisfies the relations

$$\mathrm{Rl} \sum_{k=1}^{n-1} \bar{u}_k \dot{u}_k \equiv \frac{1}{2} \frac{d}{dt} \sum_{k=1}^{n-1} |u_k|^2$$

$$= -\nu n^2 \sum_{k=1}^{n} |u_k - u_{k-1}|^2 + \sum_{k=1}^{n-1} |u_k|^2 \qquad (2.3)$$

$$\mathrm{Im} \sum_{k=1}^{n-1} \bar{u}_k \dot{u}_k \equiv \frac{1}{2i} \sum_{k=1}^{n-1} (\bar{u}_k \dot{u}_k - u_k \dot{\bar{u}}_k) \equiv \sum_{k=1}^{n-1} |u_k|^2 \dot{\theta}_k$$

$$= \sum_{k=1}^{n-1} |u_k|^2 \qquad (2.4)$$

where

$$u_k = |u_k| \exp(i\theta_k)$$

- The solution taking arbitrarily assigned initial values exists uniquely and is bounded over any finite time interval $0 \leqslant t \leqslant T < \infty$.

- $u_k(t) = 0$, $k = 1, 2, \ldots, n-1$ is the only time-independent, i.e., equilibrium, solution. Define $\nu_{mn} = \left[2n \sin (m\pi/2n) \right]^{-2}$, $m = 1, \ldots, n-1$. (As $n \to \infty$ $\nu_{1n} \to \pi^{-2}$, $\nu_{n-1n} \to 0$). If $\nu > \nu_{1n}$, every solution approaches the zero solution as $t \to \infty$. If $0 < \nu < \nu_{n-1n}$, every nonzero solution approaches infinity, i.e.,

$$\sum_{k=1}^{n-1} |u_k|^2 \to \infty \quad \text{as} \quad t \to \infty .$$

Proof.

1. After multiplying Eq. (2.1) by u_k and summing on k, Eqs. (2.3) and (2.4) follow from the

Lemma: If u_k, $k = 0, 1, \ldots, n$ are complex numbers with $u_o = u_n = 0$, then

$$\sum_{k=1}^{n-1} u_k (u_{k+1} - 2u_k + u_{k-1}) = - \sum_{k=1}^{n} |u_k - u_{k-1}|^2 \qquad (2.5)$$

and

$$\sum_{k=1}^{n-1} u_k (u_{k-1} + u_k + u_{k+1})(u_{k+1} - u_{k-1}) = 0 . \qquad (2.6)$$

Equation (2.5) follows at once from the Abel partial summation formula, quite analogously to the integration by parts used in obtaining Eq. (1.8). Equation (2.6) follows from the observation that the $k^{\underline{th}}$ summand can be written as $\alpha_k - \alpha_{k-1} + \beta_k - \beta_{k-1}$, where $\alpha_k = u_k^2 u_{k+1}$, $\beta_k = u_{k+1}^2 u_k$, so that the sum telescopes to $\alpha_{n-1} - \alpha_o + \beta_{n-1} - \beta_o = 0$.

2. It is apparent from classical theorems that a unique solution to the initial value problem exists for some interval $0 \leq t < T$. That T may be assumed arbitrarily large follows from the classical extension theorem once it is shown that

$$c(t) = \sum_{k=1}^{n-1} |u_k|^2$$

is bounded over any finite interval. To prove this, observe that the right member of Eq. (2.3) is a hermitian form (call it H) in the variables $u_1, \ldots, u_{n-1}$.

In the sequel, it will appear that the eigenvalues of this form are given explicitly, in decreasing order, by $\lambda_k = 1 - 4\nu n^2 \sin^2(k\pi/2n)$, $k=1,2, \ldots, n-1$. By a well-known characterization of the least and greatest eigenvalues,

$$\lambda_{n-1} c \leq H \leq \lambda_1 c$$

yielding, from Eq. (2.3), the differential inequality

$$\lambda_{n-1}\, c(t) \leq \frac{1}{2}\dot{c}(t) \leq \lambda_1\, c(t)$$

Integration gives

$$c(o)\, e^{2\lambda_{n-1}t} \leq c(t) \leq c(o)\, e^{2\lambda_1 t} \qquad (2.7)$$

which exhibits the desired bound.

3. The uniqueness of the equilibrium solution is obvious from Eq. (2.4). The remaining assertions are implied by Eq. (2.7) upon noting that $\nu > \nu_{1n}$ implies $\lambda_1 < 0$ and $\nu < \nu_{n-1\,n}$ implies $\lambda_{n-1} > 0$.

As in the models considered by Burgers, Eq. (2.2) exhibits a "laminar" regime, $\nu > \nu_{1n}$, in which all solutions tend to a totally stable unique laminar solution. For fixed n and small ν, however, the model encounters a deficiency inherent in finite discretization schemes. Of necessity, a function defined on a finite mesh cannot adequately approximate spatial fluctuations on a scale smaller than the mesh spacing; in a model of turbulence like Burgers', however, fluctuations on a finer and finer spatial scale are to be expected with diminishing viscosity. It is, therefore, not surprising that solutions of a model like Eq. (2.2) are found to "blow up" if ν is too small. As a consequence, it is apparent that Eq. (2.2) can display "interesting" behavior only

in some range $\nu' < \nu < \nu_{1n}$, and then Eq. (2.1) gives ν_{n-1n} as a lower bound on ν'.

In the case $n = 2$, the model is essentially trivial, since Eq. (2.2) reduces to a single linear equation and the interval of interest degenerates to a point, i.e., $\nu_{n-1n} = \nu_{1n}$. It is suggestive, however, that for $\nu = \nu_{1n}$, all solutions are periodic.

The simplest nontrivial case, $n = 3$, will be analyzed in Section 3. In the remainder of this section, the effect upon Eq. (2.2) of a particular change of dependent variables will be described.

One of the tools used by Burgers in his study of Eq. (1.8) and similar equations is the expansion of the solution function in Fourier series; the partial differential equation yields formally an infinite system of ordinary differential equations for the determination of the Fourier coefficients as functions of time. At a fixed time, the coefficients may be called a "spectrum" which describes the harmonic content of the spatial distribution of "turbulent" fluctuations; the significance of this spectrum and various questions concerned with its change with time are discussed in Burgers' papers (Refs. 1 and 2).

To obtain an analogous (but finite) spectrum for the discretized model Eq. (2.2), the solution $u_k(t)$ may be expanded (with respect to the spatial variable k) in a finite sine series; in other words, new complex coordinates ζ_k may be introduced in place of the u_k in the (complex, n - 1 dimensional) phase space of the system (2.2) by means of the transformation

$$u_k = \sum_{j=1}^{n-1} \zeta_j \sin j \frac{k\pi}{n} \quad , \quad k = 1, 2, \ldots , n - 1 \tag{2.8}$$

Abbreviating $\sin j\frac{k\pi}{n} = \sigma_{kj}$ and noting the orthogonality relation

$$\sum_{k=1}^{n-1} \sigma_{ki}\,\sigma_{kj} = \frac{n}{2}\,\delta_{ij} = \begin{cases} n/2 & i = j \\ 0 & i \neq j \end{cases}$$

the inverse transformation is easily found to be

$$\zeta_i = \frac{2}{n} \sum_{k=1}^{n-1} u_k\,\sigma_{ki} \tag{2.9}$$

Inserting Eq. (2.8) in Eq. (2.2) and performing algebraic reductions lead to the following systems of equations for the ζ_k

$$\ddot{\zeta}_k = (\lambda_k + i)\zeta_k - \frac{n}{3} \sum_{1 \le i < i+k \le n-1} \alpha_{i,i+k}\,\tilde{\zeta}_i\,\tilde{\zeta}_{i+k} - \frac{n}{6} \sum_{1 \le i,\,k-i \le n-1} \beta_{i,k-i}\,\tilde{\zeta}_i\,\tilde{\zeta}_{k-i} \tag{2.10}$$

$$k = 1, 2, \ldots, n-1$$

where

$$\alpha_{jk} = \sin j\frac{\pi}{n}\left(1 + 2\cos k\frac{\pi}{n}\right) - \sin k\frac{\pi}{n}\left(1 + 2\cos j\frac{\pi}{n}\right)$$

$$\beta_{jk} = \sin j\frac{\pi}{n}\left(1 + 2\cos k\frac{\pi}{n}\right) + \sin k\frac{\pi}{n}\left(1 + 2\cos j\frac{\pi}{n}0\right.$$

$$\lambda_k = 1 - 4\nu n^2 \sin^2\frac{k\pi}{2n}$$

i.e., the linear part of the system has been diagonalized by the change of coordinates.

Section 3
THE CASE n = 3

INTRODUCTION

If n = 3, Eqs. (2.2), (2.8), (2.9), and (2.10) have the following forms:

$$\begin{aligned} u_1 &= (1 - 18\nu + i)\, u_1 + 9\nu\, u_2 - (\bar{u}_1 + \bar{u}_2)\, \bar{u}_2 \\ \dot{u}_2 &= 9\nu\, u_1 + (1 - 18\nu + i)\, u_2 + (\bar{u}_1 + \bar{u}_2)\, \bar{u}_1 \end{aligned} \tag{3.1}$$

$$\begin{aligned} u_1 &= \frac{\sqrt{3}}{2}\,(\zeta_1 + \zeta_2) \\ u_2 &= \frac{\sqrt{3}}{2}\,(\zeta_1 - \zeta_2) \end{aligned} \tag{3.2}$$

$$\begin{aligned} \zeta_1 &= \frac{1}{\sqrt{3}}\,(u_1 + u_2) \\ \zeta_2 &= \frac{1}{\sqrt{3}}\,(u_1 - u_2) \end{aligned} \tag{3.3}$$

$$\begin{aligned} \dot{\zeta}_1 &= (1 - 9\nu + i)\, \zeta_1 + \sqrt{3}\, \bar{\zeta}_1 \bar{\zeta}_2 \\ \dot{\zeta}_2 &= (1 - 27\nu + i)\, \zeta_2 - \sqrt{3}\, \bar{\zeta}_1^{\,2} \end{aligned} \tag{3.4}$$

The most evident property of the system (3.4) will be stated as Theorem 3.1.

THEOREM 3.1

If $\zeta_1(t)$, $\zeta_2(t)$ is a solution of Eq.(3.4) in an interval containing a point t_o such that $\zeta_1(t_o) = 0$, then

$$\zeta_1(t) \equiv 0 \quad , \quad \zeta_2(t) = \zeta_2(t_o) \exp\left[(1 - 27\nu + i)\,(t - t_o)\right] \tag{3.5}$$

is the unique extension of the given solution to $-\infty < t < +\infty$.

This theorem implies that the phase point sets $\xi_1 = 0$ and $\xi_1 \neq 0$ are invariant manifolds; solution trajectories never pass from one to the other. Solutions of the form Eq. (3.5) will be subsequently referred to as trivial. For $\nu > 1/27$, it is apparent that all trivial solutions approach the zero solution.

Because of the complex notation, Eq.(3.5) is a convenient condensation for a system of four first-order differential equations for four real-valued functions. For discussion of the nontrivial solutions, real coordinates will be introduced which permit the reduction of the initial-value problem for Eq. (3.4) to the initial-value problem for a system of three real differential equations plus a quadrature.

THEOREM 3.2

The solution of Eq. (3.4) taking initial values $\zeta_1(0) \neq 0$, $\zeta_2(0)$ is given by

$$\zeta_1(t) = \frac{\zeta_1(0)}{\sqrt{3}\,|\zeta_1(0)|}\ \sqrt{z(t)}\ \exp\left\{i \int_o^t \left[1 - y(s)\right] ds\right\}$$

$$\zeta_2(t) = \frac{\overline{\zeta_1(0)}}{\sqrt{3}\,\zeta_1(0)}\ \left[x(t) + i\,y(t)\right]\ \exp\left\{-2i \int_o^t \left[1 - y(s)\right] ds\right\} \tag{3.6}$$

in which $x(t)$, $y(t)$, $z(t)$ is the solution of the system

$$\dot{x} = (1 - 27\nu)x - 3y - z + 2y^2 \tag{3.7a}$$

$$\dot{y} = 3x + (1 - 27\nu)y - 2xy \tag{3.7b}$$

$$\dot{z} = 2(1 - 9\nu)z + 2xz \tag{3.7c}$$

satisfying the initial conditions dictated by the special case $t = 0$ of the transformation relations

$$x(t) + i\,y(t) = \sqrt{3}\,\frac{\zeta_1(t)\,\zeta_2(t)}{\overline{\zeta_1(t)}}\,,\quad z(t) = 3\,|\,\zeta_1(t)\,|^2 \tag{3.8}$$

Proof. The proof consists of straightforward verification and will be omitted. In its place the formal processes leading to Eqs. (3.7) will be sketched. ("Formal" here refers specifically to the fact that possible discontinuities of the function $\varphi_2(t)$ defined below are ignored.)

Into Eqs. (3.4) introduce real polar coordinates by

$$\sqrt{3}\,\zeta_k = \rho_k\,\exp(i\,\varphi_k)\,,\quad k = 1,2 \tag{3.9}$$

The following differential equations are found for ρ_k, φ_k:

$$\dot{\rho}_1 = (1 - 9\nu)\,\rho_1 + \rho_1\,\rho_2 \cos(2\varphi_1 + \varphi_2)$$

$$\dot{\rho}_2 = (1 - 27\nu)\,\rho_2 - \rho_1^2 \cos(2\varphi_1 + \varphi_2) \tag{3.10}$$

$$\rho_1 \dot{\varphi}_1 = \rho_1 - \rho_1 \rho_2 \sin(2\varphi_1 + \varphi_2)$$

$$\rho_2 \dot{\varphi}_2 = \rho_2 + \rho_1^2 \sin(2\varphi_1 + \varphi_2) \qquad (3.11)$$

By defining a new variable $\psi = 2\varphi_1 + \varphi_2$, Eqs. (3.11) yield formally

$$\dot{\psi} = 3 - \left(2\rho_2 - \frac{\rho_1^2}{\rho_2}\right) \sin \psi \qquad (3.12)$$

Now Eqs. (3.10) and (3.12) constitute a third-order system for ρ_1, ρ_2, ψ as functions of t. Given a solution of this system the functions φ_1, φ_2 can be found by quadrature from Eqs. (3.11).

The system of Eqs. (3.7) now results from transformation of the system (3.10) and (3.12) according to

$$x = \rho_2 \cos \psi$$

$$y = \rho_2 \sin \psi$$

$$z = \rho_1^2$$

The combined transformation from Eqs. (3.4) to (3.7) is given by Eq. (3.8). The first of Eqs. (3.11) yields

$$\varphi_1(t) = \varphi_1(o) + \int_o^t [1 - y(s)]\, ds$$

or

$$\frac{\zeta_1(t)}{|\zeta_1(t)|} = \frac{\zeta_1(o)}{|\zeta_1(o)|} \exp \int_o^t [1 - y(s)]\, ds$$

The transformation employed can be interpreted so as to yield also a correspondence between the trivial solutions of Eq. (3.4) and solutions of Eqs. (3.7) with $z(t) = 0$; simplicity is lost in the process, however, since the restriction of Eq. (3.4) to $\zeta_1 = 0$ leaves a linear system while the restriction of Eqs. (3.7) to $z = 0$ is non-linear.

It is easily seen that in the (invariant) half-space $z > 0$ the system (3.7) can have at most one critical point. In fact, setting the right-hand members equal to zero, Eq. (3.7) implies

$$x = x_c = 9\nu - 1$$

Eq. (3.7b) then requires

$$y = y_c = \frac{1 - 9\nu}{1 - 15\nu}$$

and Eq. (3.7a) requires

$$z = z_c = (1 - 9\nu)\,(27\nu - 1)\,\left[\, 1 + (1 - 15\nu)^{-2} \,\right]$$

This point (x_c, y_c, z_c) belongs to the half-space $z > 0$ if ν lies in either of the intervals $1/27 < \nu < 1/15$ or $1/15 < \nu < 1/9$. That the first interval represents values of ν too small to yield "interesting" results in the case $n = 3$ is suggested by the following negative result.

THEOREM 3.3

If $1/27 < \nu < 1/15$, every solution of Eqs. (3.7) which intersects the region $z > 0$, $y < 3/2$ is unbounded as $t \to + \infty$.

Proof. Using Eqs. (3.7), it is readily verified that for any solution

$$\frac{d}{dt}\left[z\left(y - \frac{3}{2}\right) e^{-3(1 - 15\nu)t} \right] = \frac{3}{2}(1 - 27\nu)\, z\, e^{-3(1 - 15\nu)t}$$

For a nontrivial solution ($z(t) > 0$) and $\nu > 1/27$, this implies

$$\frac{d}{dt}\left[z\left(y - \frac{3}{2}\right) e^{-3(1 - 15\nu)t} \right] < 0$$

Hence for $t_1 < t_2$

$$z(t_2)\left(y(t_2) - \frac{3}{2}\right) < z(t_1)\left(y(t_1) - \frac{3}{2}\right) e^{-3(1 - 15\nu)(t_1 - t_2)}$$

If $y(t_1) - 3/2 < 0$, $\nu < 1/15$, and $z(t)$ is bounded, this inequality implies $y(t_2) \to -\infty$ as $t_2 \to \infty$. Under the hypothesis of the theorem, therefore, at least one of the functions $z(t)$, $y(t)$ is not bounded.

Against this negative result for $1/27 < \nu < 1/15$, the theorem following reveals interesting behavior for solutions of Eqs. (3.7) for at least a part of the interval $1/15 < \nu < 1/9$.

THEOREM 3.4

There is a number ν', $1/15 \leq \nu' < 0.0836 < 1/9$ such that for $\nu' < \nu < 1/9$ every nontrivial solution of the system (3.7) approaches the equilibrium solution (x_c, y_c, z_c) as $t \to \infty$: Eqs. (3.6) then imply that every nontrivial solution of Eqs. (3.4) approaches a periodic solution.

Proof. The proof depends on the following properties of the function

$$V(x, y, z) = (x - x_c)^2 + (y - y_c)^2 + (z - z_c) - z_c \log \frac{z}{z_c} \tag{3.13}$$

defined in the half-space $z > 0$, $-\infty < x, y < +\infty$.

- $V \geq 0$ $V = 0$ if and only if $x = x_c$, $y = y_c$, $z = z_c$.
- $V \to \infty$ if either $z \to 0$ or $x^2 + y^2 \to \infty$ in $z > 0$.
- If $0 \leq \alpha < \infty$, the set $E(\alpha) = \{ (x,y,z) \mid V(x,y,z) \leq \alpha \}$ is a closed and bounded subset of $z > 0$ and $E(\alpha) \subset E(\beta)$ if $\alpha < \beta$.
- Along a solution curve of the system (3.7).

$$-\frac{1}{2}\frac{dV}{dt} = (2 + 3x_c)(x - x_c)^2 - 2y_c (x - x_c)(y - y_c) + (2 + 5x_c)(y - y_c)^2 \tag{3.14}$$

The right-hand member of this relation is a positive definite quadratic form in the variables $(x - x_c)$, $(y - y_c)$ provided that $\nu' < \nu < 1/9$, where $\nu' = 0.0836$ is the l[illegible] real root of the discriminant (which is a rational function of ν). The [illegible]y implication of Eq.(3.14)needed in the proof is the inequality

$$-\frac{1}{2}\frac{dV}{dt} \leq B\left[(x - x_c)^2 + (y - y_c)^2\right] \tag{3.15}$$

where

$$B = 2(1 - 2x_c) - \sqrt{x_c^2 + y_c^2}$$

is the positive smallest eigenvalue of the form.

These properties are either obvious or can be proved by direct calculation.

If $x(t)$, $y(t)$, $z(t)$ is the solution of Eqs. (3.7)with initial values x_o , y_o , $z_o > 0$ and if $\alpha_o = V(x_o , y_o , z_o)$, the phase point $(x(t), y(t), z(t))$ belongs to $E(\alpha_o)$ for $t \geq 0$ since $V = V[x(t) , y(t) , z(t)]$ is a nonincreasing function of t by Eq. (3.15). Thus, the phase point is bounded; from Eqs. (3.7), it follows that the phase-point velocity $(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)^{1/2}$ is also uniformly bounded for $t \geq 0$.

It will be shown first that $x(t) \rightarrow x_c$ and $y(t) \rightarrow y_c$ as $t \rightarrow \infty$, or, equivalently, that no point of the limit set of a solution lies off the half-line $z > 0, (x - x_c)^2 + (y - y_c)^2 = 0$. Assuming the contrary, there is point P of the limit set, a neighborhood, $N(P)$, of P , and an $\epsilon > 0$ such that $N(P)$ does not intersect the cylinder $(x - x_c)^2 + (y - y_c)^2 \leq \epsilon$. During the time intervals that the phase point lies in $N(P)$, Eq. (3.15) implies $dV/dt \leq - 2B\epsilon$. Since V is nonnegative and nonincreasing, this inequality implies that the time which a phase point spends in $N(P)$ is bounded. The boundedness of the phase point velocity and the assumption that P belongs to the limit set imply, however, that the time spent in $N(P)$ is unbounded. The contradiction proves that the limit set lies on the half-line and hence that $x(t) \rightarrow x_c$ and $y(t) \rightarrow y_c$.

It remains to prove that $\lim z(t)$ exists and equals z_c. Along a trajectory, V decreases and thus has a limit. From Eq. (3.13) and the existence of limits for V, x, and y, there follows the existence of $\lim (z - z_o \log z)$ and, hence, that of $\lim z$. From Eq. (3.7a), the existence of $\lim \dot{x}$ is then immediate. But if $\lim x$ and $\lim \dot{x}$ exist, it follows at once that $\lim \dot{x} = 0$. Hence, the right-hand member of Eq. (3.7a) converges to zero as $t \rightarrow \infty$, which implies that $\lim z = z_c$ and concludes the proof.

Section 4
REFERENCES

1. J. M. Burgers, "Mathematical Examples Illustrating Relations Occurring in the Theory of Turbulent Fluid Motion," Verhandel. Kon. Nederl. Akad. Wetenschappen Amsterdam, Afdeel. Natuurkunde (1st Sect.), Deel 17, No. 2, 1939 pp. 1–53

2. J. M. Burgers, "A Mathematical Model Illustrating the Theory of Turbulence," Advances in Applied Mechanics, Vol. 1, New York, Academic Press, 1948, pp. 171–199

3. E. Hopf, "A Mathematical Example Displaying Features of Turbulence," Comm. on App. Math., Vol. 1, No. 4, 1948, pp. 303–322

4. R. J. Dickson, "Bounds for Solutions of Some Parabolic Problems," Unpublished Ph.D. Dissertation, California Institute of Technology, 1954