CS 105

# ε-CALCULUS

BY

PAUL L. RICHMAN

TECHNICAL REPORT NO. CS 105
AUGUST 16, 1968

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

# $\varepsilon$-CALCULUS*

By

Paul Lawrence Richman

ABSTRACT: We use recursive function theory to lay the basis for a partially constructive theory of calculus, which we call the $\varepsilon$-calculus. This theory differs from other theories that have grown out of recursive function theory in that

(1) it is directly related to the variable-precision computations used in scientific computation today, and

(2) it deals explicitly with intermediate results rather than ideal answers.

As $\varepsilon \to 0$, intermediate results in the $\varepsilon$-calculus approach their corresponding answers in the calculus. Thus we say "the $\varepsilon$-calculus approaches the calculus, as $\varepsilon \to 0$ ." It is hoped that investigations in the $\varepsilon$-calculus will lead to a better understanding of numerical analysis. Several new results in this direction are presented, concerning instability and also machine numbers. Discrete notions of limit, convergence, continuity, arithmetic, derivative and integral are also presented and analyzed.

*This research was supported by the National Science Foundation under contract NSF GP 5962, by the Office of Naval Research under contract N00014-67-A-0112-0029, and by the Air Force under contract AF1047-66.

Acknowledgment: Deepest thanks go to my advisor, George Forsythe, for suggesting the idea of an $\epsilon$-calculus and for his encouragement and direction during the initial stages of my research. I owe a special debt of gratitude to Dana Scott for his many excellent conceptual and notational contributions to the $\epsilon$-calculus.

Their fast and accurate typing has earned Gail D. Schwartz, Lynne E. Leontovich, and Phyllis A. Winkler special mention. Thanks also go to Dorothy T. McGrath for her artistic illustrations.

Above all, I thank my wife, Jackie, for her unfailing patience, encouragement and confidence. I dedicate this work to her.

This research was supported by the National Science Foundation under contract NSF GP 5962, by the Office of Naval Research under contract N00014-67-A-0112-0029, and Air Force AF1047-66.

## Table o:f Contents

Chapter Page

1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . 1

1.1 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . 1

1.2 A Motivating Example . . . . . . . . . . . . . . . . . . . . 4

1.3 An Outline . . . . . . . . . . . . . . . . . . . . . . . . . 8

1.4 Notation . . . . . . . . . . . . . . . . . . . . . . . . . . 11

1.5 Recursive Natural Functions and Operators . . . . . . . . . . . 14

Remarks . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

2 Basic Concepts . . . . . . . . . . . . . . . . . . . . . . . . . 16

2.1 Variable-Precision Computations . . . . . . . . . . . . . . . . 16

2.2 Machine Numbers and Their Comparison . . . . . . . . . . . . 17

2.3 Real Inputs . . . . . . . . . . . . . . . . . . . . . . . . . 23

2.4 Multiple-Precision Subroutines . . . . . . . . . . . . . . . . 25

2.5 ε-Function . . . . . . . . . . . . . . . . . . . . . . . . . 29

2.6 An Example: $e^x$ . . . . . . . . . . . . . . . . . . . . . . 33

2.7 Operators . . . . . . . . . . . . . . . . . . . . . . . . . . 35

2.8 Rounding Subroutines . . . . . . . . . . . . . . . . . . . . 38

Remarks , . . . . . . . . . . . . . . . . . . . . . . . . . . 42

2.A Appendix: Maximum Relative Error . . . . . . . . . . . . . . 43

3 Numerical Instability . . . . . . . . . . . . . . . . . . . . . 47

3.1 A Definition . . . . . . . . . . . . . . . . . . . . . . . . 47

Chapter Page

3.2 A Geometric Characterization . . . . . . . . . . . . . . . . 51

3.3 A Stopping Criterion That Works . . . . . . . . . . . . . . 52

3.4 An Algorithm for Overcoming Instability . . . . . . . . . 54

3.5 Applications . . . . . . . . . . . . . . . . . . . . . . . . 57

Remarks . . . . . . . . . . . . . . . . . . . . . . . . . . . . 65

4 ε-Limit, ε-Comparison, ε-Convergence and ε-Continuity . . . 66

4.1 ε-Limit and Truncation-Error Bounds . . . . . . . . . . . 66

4.2 ε-Comparison Relations, $<_\varepsilon$ and $=_\varepsilon$ . . . . . . . . . . . 72

4.3 ε-Convergence and ε-Continuity: Pointwise . . . . . . . 75

4.4 Discontinuities . . . . . . . . . . . . . . . . . . . . . . 80

4.5 ε-Convergence and ε-Continuity: Over Intervals . . . . . 84

Remarks . . . . . . . . . . . . . . . . . . . . . . . . . . . . 89

5 More ε-Operators . . . . . . . . . . . . . . . . . . . . . . . 92

5.1 Identity ε-Functions . . . . . . . . . . . . . . . . . . . 95

5.2 Constant ε-Functions . . . . . . . . . . . . . . . . . . . 96

5.3 ε-Arithmetic . . . . . . . . . . . . . . . . . . . . . . . 97

5.4 ε-Limit . . . . . . . . . . . . . . . . . . . . . . . . . . 102

5.5 ε-Composition . . . . . . . . . . . . . . . . . . . . . . . 104

5.6 ε-Recursion . . . . . . . . . . . . . . . . . . . . . . . . 106

5.7 An ε-Function Corresponding to $e^x$ Over $\widetilde{R}$ . . . . . . 114

Remarks . . . . . . . . . . . . . . . . . . . . . . . . . . . .118

Chapter Page

6 **ε-Derivative and ε-Integral** . . . . . . . . . . . . . . . . . . . . . 119

6.1 **ε-Differentiability and ε-Derivative** . . . . . . . . . . . . 119

6.2 **ε-Integrability and ε-Integral** . . . . . . . . . . . . . . . . 121

6.3 **The Fundamental Theorem of the ε-Calculus** . . . . . . . . 127

7 Computable Real Functions and Completeness . . . . . . . . . . . . 128

7.1 Computable Real Functions . . . . . . . . . . . . . . . . . . . 128

7.2 Completeness . . . . . . . . . . . . . . . . . . . . . . . . . 133

Summary and Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . 135

References . . . . . . . . . . . . . . . . . . . . . . . . . . . . 137

## Table of Illustrations

| Figure Number | Page |
|---|---|
| 1.2-1 | 5 |
| 1.2-2 | 6 |
| 2.5-1 | 30 |
| 3.2-1 | 51 |
| 3.5-1 | 59 |
| 4.2-1 | 72 |

# Chapter 1: Introduction

## 1.1 Summary

By a "notion" we mean a property of or an operation defined on a function or functions. The calculus can be thought of as a collection of elementary notions such as limit, convergence, continuity, derivative and integral, together with certain proved relations between notions, such as the reciprocal relationship between integration and differentiation. Fundamental to all this are the concepts of a real number and a real function. In the usual textbook developments, these basic concepts are not presented constructively, the notions are not necessarily effective or computable in any sense and relations between notions are often proved unconstructively. This is in direct contrast to E. Bishop's Foundations of Constructive Analysis [B1] and in partial contrast to recursive or computable analysis (Turing, Mazur, Grzegorczyk, Goodstein, Specker, Klaua, Aberth and Kreisel, to mention a few researchers in the area). Bishop defines constructive concepts of real number and real function, develops constructive notions and proves relations between notions constructively. (He then goes on to constructive theories of sets, metric spaces, complex analysis, measure, integration, normed linear spaces, locally compact abelian groups and commutative algebras.) His work is based on Brouwer's intuitionistic mathematics. In their work, "constructive" is an undefined or primitive term. Recursive analysis also has constructive concepts of real number and real function (see [G2, pp. 61-2]) and deals with constructive notions, but it allows unconstructive proofs (see Kreisel [Kl, p. 101]) It is based on recursive function theory, initiated by Church. In recursive analysis, "constructive" is defined in terms of recursive functions.

Both of these constructive theories are presented in a way which makes them foreign to numerical computation as it is done on today's computers. Here, we use recursive function theory to develop a theory of not only constructive, but even finitely computable real functions and defined notions, which we call $\varepsilon$-functions and $\varepsilon$-notions; these represent the intermediate results which arise from numerical computation. We call the resulting theory $\varepsilon$-calculus. This theory is directly related to modern day numerical computation. $\varepsilon$-Functions are essentially defined over a finite set, $\mathbf{R}(\varepsilon)$ , of $\varepsilon$-precision machine numbers. $\mathbf{R}(\varepsilon)$ approximates the real numbers and each $\varepsilon$-function and $\varepsilon$-notion approximates respectively a function and a notion from calculus. And, as $\varepsilon \to 0$ , $\mathbf{R}(\varepsilon)$ approaches (i.e. becomes dense in) the reals and each $\varepsilon$-function or $\varepsilon$-notion approaches (in a sense to be defined) its corresponding function or notion. Thus we say the $\varepsilon$-calculus is a discretization of the calculus such that, as $\varepsilon \to 0$ , the $\varepsilon$-calculus approaches the calculus.

The value of the $\varepsilon$-calculus to numerical analysis is that it presents a model of variable-precision computations. The study of $\mathbf{R}(\varepsilon)$ , $\varepsilon$-functions and $\varepsilon$-notions within the context of this model should lead to a better understanding of numerical computation. Our principal results in this direction are

(1) a new and simple definition of numerical instability (the kind caused by propagation of roundoff-error) together with

a suggestive geometric characterization (ch. 3), and

(2) an algorithm for overcoming such instabilities (ch. 3 and ch. 4).

Other new results presented here include

(1) a characterization of the concept of <u>variable-precision machine numbers</u> (sec. 2.2), and

(2) two new definitions of computable real functions, one allowing functions with discontinuities (ch. 7).

Before we present the ε-calculus, we give a motivating example to point out some of the basic problems involved in forming such a theory (i.e., involved in going from ideal mathematics to actual numerical computation), and to develop some of our basic notation.

## 1.2 A Motivating Example

Let us use "precision of computation" in a general way to mean the accuracy of a given mathematical approximation together with the precision of the arithmetic used to evaluate this approximation. It is often said that "numerical analysis is not very interesting because all you have to do to get more accuracy in a numerical result is increase the precision of computation." As a broad and optimistic point of view, the above statement is quite reasonable. But, when applied to particular cases, it can be quite false. Increasing the precision of computation can drastically decrease the accuracy of the result! For example, consider an algorithm which uses

$$f(x,y) = (g(y) - g(x))/(y-x)$$

to approximate $\ell \equiv \frac{d}{dt} g(t)|_{t=x}$ . Fix $x$ . For simplicity, suppose $f(x,y) \to \ell$ monotonically as $|y-x| \to 0$ , and that $f(x,y_1)$ is computed in a certain form of single-precision arithmetic to give a single-precision approximation, $F(\varepsilon_1;\ x,\ y_1)$ , to $\ell$ (here, "$\varepsilon_1$" denotes "single-precision"). This would be the value of the $\varepsilon_1$-limit corresponding to $\lim_{y \to x} f(x,y)$ . We can increase the precision of computation by

(1) replacing $y_1$ by $y_2$ with $0 < |y_2-x| < |y_1-x|$ , yielding a more accurate mathematical approximation, $f(x,y_2)$ (more accurate because of the monotonicity assumption), and

(2) evaluating $f(x,y_2)$ in a certain form of double-precision arithmetic, yielding a double-precision approximation,

$F(\varepsilon_2; x, y_2)$ , to $\ell$ (here, "$\varepsilon_2$" denotes "double-precision". This would be the value of the $\varepsilon_2$-limit. But $F(\varepsilon_2; x, y_2)$ is not necessarily closer to $\ell$ than $F(\varepsilon_1; x, y_1)$ ; in fact, if $y_2$ is too close to $x$ , $F(\varepsilon_2: x, y_2)$ may be much worse than $F(\varepsilon_1; x, y_1)$ (e.g., see example 3.1-2, where $g(x)$ is taken to be $x + 1$ ) This is illustrated in figure 1 2-1 as three graphs (with $x$ fixed)

(a) $f(x,y)$ versus $1/(y-x)$ ,

(b) $F(\varepsilon_1; x, y)$ versus $1/(y-x)$ , and

(c) $F(\varepsilon_2; x, y)$ versus $1/(y-x)$ ,

where $y$ varies in the interval $(x, x+1]$ .

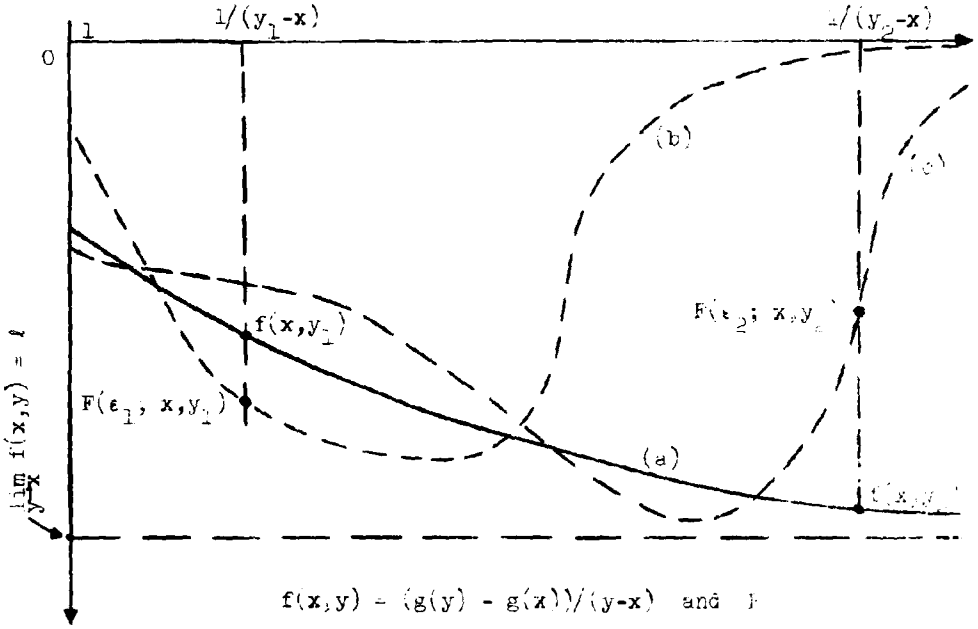


FIGURE 1.2-1

Notice that graph (b) stays close to $f$ for awhile, but then falls off sharply to zero. Graph (c) stays close to $f$ for awhile longer, but then it too falls off to zero. In general, $F$ which exhibit such behavior are called unstable (this is discussed in detail in ch. 3). See Riesel [R1] for a similar example.

The tools normally used to deal with such instabilities are roundoff-error bounds, RF, and truncation-error bounds, TF. RF bounds the error incurred by using $F$ in place of $f$ ; TF bounds the error incurred by using $f(x,Y)$ in place of $\lim_{y \to x} f(x,y)$ . And RF + TF bounds the error incurred by using $F$ in place of $\lim_{y \to x} f(x,y)$ . RF and TF are shown in figure 1.2-2, which is a redrawing of figure 1.2-1.

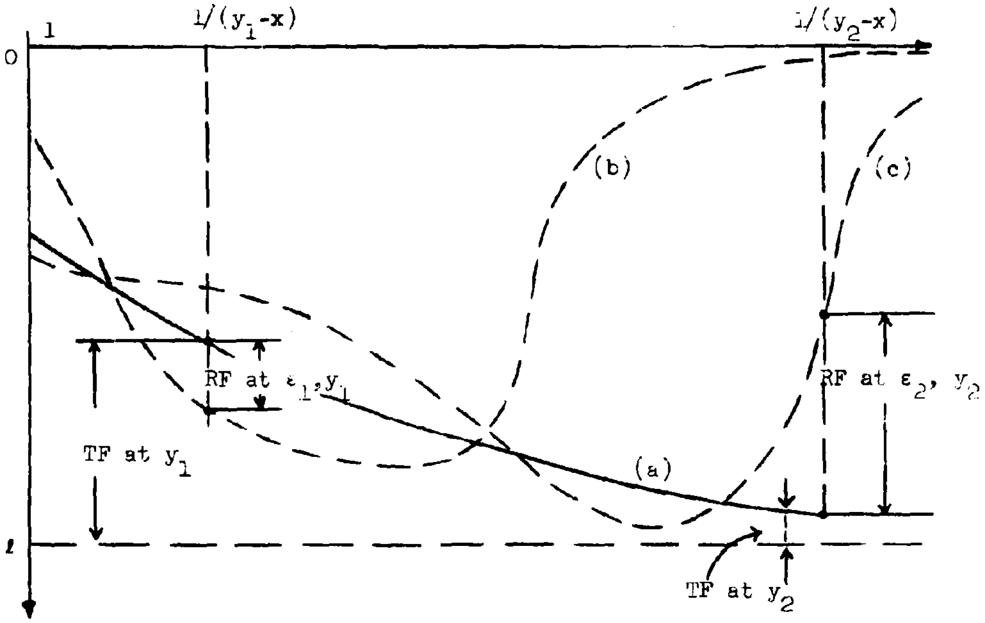


FIGURE 1.2-2
RF and TF

Of course, roundoff error is just a particular kind of truncation error; namely, it is the truncation error caused by using $F(\varepsilon; x, y)$ in place of $\lim_{\varepsilon \to 0} F(\varepsilon; x, y)$ (which is $= f(x,y)$). Both errors are caused by replacing infinite processes by finite processes. However, it is useful to distinguish roundoff error from truncation error so that they can be dealt with separately. The motivation for introducing these bounds comes from R. E. Moore's theory of interval analysis [M3, M4] . The key idea is that such bounds can be used to give precise information about a numerical result; i.e., an interval which contains the result.

## 1.3 An Outline

In the next two sections we give our basic notation and we discuss recursive natural functions and recursive natural operators. In chapter 2, we present the basic concept of a variable-precision computation, including the concepts of machine numbers, real inputs, subroutines, $\varepsilon$-functions and $\varepsilon$-operators. We give three examples of machine-numbers: floating-point, logarithmic and rational. And we give our main reason for introducing truncation-error bounds.

In chapter 3, we define and discuss numerical instability. A geometric characterization of instability is given which leads to the concept of "an $\varepsilon$-wave", to a proof that there is some desirable behavior even in the presence of instability (thm. 3.3-1), and finally to a (very inefficient) algorithm for overcoming instability (def. 3.3-4). This motivates our definition of $\varepsilon$-limit, given in the first section of chapter 4 (defs. 4.1-1, 4.1-2). We prove that, under certain conditions, the $\varepsilon$-limit of an $\varepsilon$-function approaches the limit of its corresponding ideal function as $\varepsilon \to 0$ (thm. 4.1-1). This $\varepsilon$-limit is shown to be a potentially efficient algorithm for overcoming instabilities of an approximation function by using a stably convergent truncation-error bound.

In section 4.2, we define $\varepsilon$-comparison relations, $<_\varepsilon$ and $=_\varepsilon$, and we prove that the truth-value of the $\varepsilon$-comparison of two real inputs approaches the truth-value of their comparison as $\varepsilon \to C$. These considerations are basic to what follows, and must be understood. In section 4.3, we use these $\varepsilon$-comparison relations to define $\varepsilon$-convergence and $\varepsilon$-continuity (pointwise). Again we prove that these $\varepsilon$-notions approach (in a certain sense) their corresponding notions as $\varepsilon \to 0$ . In section 4.5, we do the same for $\varepsilon$-convergence and $\varepsilon$-continuity over intervals. In preparation for this, we prove in section 4.4 some theorems about the kinds of discontinuities an ideal function can have while there exists an $\varepsilon$-function corresponding to it. These latter results are also made use of when we define $\varepsilon$-integrability in section 6.2.

In chapter 5, we define $\varepsilon$-operators for $\varepsilon$-arithmetic, $\varepsilon$-limit, $\varepsilon$-composition and $\varepsilon$-recursion. We also define two initial $\varepsilon$-functions, the identity and the constant $\varepsilon$-functions. The choice of these $\varepsilon$-operators and initial $\varepsilon$-functions was motivated by the operations and initial functions used in Mendelson [M1, pp. 120-1] to define the recursive natural functions. We illustrate the use of these $\varepsilon$-operators and $\varepsilon$-functions by using them to define an $\varepsilon$-function corresponding to $e^x$ .

In chapter 6, we use the $\varepsilon$-operators and initial $\varepsilon$-functions of chapter 5 together with the $\varepsilon$-convergence of section 4.3 to define $\varepsilon$-differentiability and then $\varepsilon$-derivative, $\varepsilon$-integrability and $\varepsilon$-integral. In section 6.3, we prove the $\varepsilon$-calculus analog to the fundamental theorem of calculus.

In chapter 7, we define two notions of computable real function (based on $\varepsilon$-functions), and we prove that one of them is equivalent to one of the standard definitions from recursive analysis. We also prove that the operators and initial functions of chapter 5 are complete, in a certain sense.

The discussions of $\varepsilon$-convergence and $\varepsilon$-continuity in sections

4.3 and 4.5 and of ε-derivative and ε-integral in chapter 6 are only of definitional interest. Chapter 5 and the rest of chapter 4 are of more general interest; developments presented there should be useful in extending our theory.

## 1.4 Notation

Next our basic notation is presented. We begin with a list, using S and T to denote sets and m , an integer $\geq 0$ :

| Symbol or Expression | Meaning |
|---|---|
| $=$ | equal in numeric value |
| $\equiv$ | equivalent |
| $\Rightarrow$ | implies |
| $\Leftrightarrow$ | if and only if |
| $\in$ | set membership |
| $\cup$ | union |
| $\cap$ | intersection |
| $\subset$ | inclusion: $S \subset T \Leftrightarrow (x \in S \Rightarrow x \in T)$ |
| S-T | $S \cap$ (the complement of T) |
| "{" , "}" | used only in defining sets |
| { } | the null set |
| $\bar{x}_m$ | if $m \geq 1$, the list $x_1, x_2, \ldots, x_m$ ; $\bar{x}_0$ is the empty list |
| $S^{(m)}$ | if $m \geq 1$, $S\times\ldots\times S$ (m-fold); $S^{(0)}$ is $\{\bar{x}_0\}$ |
| $f:S^{(m)} \rightarrow T$ | f is a function from $S^{(m)}$ to T |
| $-$ (bar) | generally indicates repetition on a subscript, as in $\bar{x}_m$ |
| $[x]$ | greatest integer in x |
| $n$ | $\{0, 1, 2, \ldots\}$ |
| $R(\epsilon)$ | $\epsilon$-precision machine numbers (sec. [illegible]) |

| | |
|---|---|
| $\mathcal{M}$ | the set of all machine numbers (sec. 2 2) |
| $\hat{I}(\cdot)$, $\check{I}(\cdot)$ | conversion functions (sec. 2.8) |
| $\hat{+}$, $\underset{\smile}{+}$, $\hat{-}$, ... | roundup and rounddown machine arithmetic (sec. 2.8) |
| $N_\varepsilon(a)$ | machine neighborhood of a (sec. 2.8) |
| $=_\varepsilon$, $<_\varepsilon$ | $\varepsilon$-comparison relations (sec. 4.2) |
| $\mathcal{F} \approx f(P)$ | $\mathcal{F}$ approximates f over P (sec. 2.5) |
| bool [statement] | = 1 if statement is true,<br>= 0 otherwise (sec. 4.2) |

Note that $S^{(0)} \neq \{\ \}$, since $\bar{x}_0 \in S^{(0)}$. Define R and $\tilde{R}$ by

$$R \equiv \{x:\ x \text{ is a finite real number}\},$$
$$\tilde{R} \equiv R \cup \{-\infty, \infty, \omega\};$$

$\omega$ stands for "undefined". Thus "$x = \omega$" means "x is undefined in terms of the members of $R \cup \{-\infty, \infty\}$". We will treat $\omega$ like any other point in $\tilde{R}$, except that $\underline{\omega > x\ \text{ for all }\ x \in \tilde{R} - \{\omega\}}$, and $\omega$ is isolated from the rest of $\tilde{R}$ (the null set is the only neighborhood of $\omega$). For example, $\infty - \infty = \omega$, $0/0 = \omega$, $\omega + 3 = \omega$, $(-1) \times \omega = \omega$, $\lim_{i \to \infty} (-1)^i = \omega$, etc. All our constants, variables and functions will take values in $\tilde{R}$.

We will use the usual neighborhood definition of limit for the doubly extended real line, with the addition that a limit which does not exist in the usual sense has the value $\omega$.

To simplify inequalities, we let

$$|a-b| = 0 \qquad \text{if } a = b \ ,$$

even when $a = b \in \{-\infty, \infty, \omega\}$ . We do this because we use $|a-b|$ to measure the distance between $a$ and $b$ . It is easy to show that this distance function satisfies the triangle inequality,

$$|a-b| \leq |a-c| + |c-b| \ ,$$

for $a, b, c \in \tilde{R}$ . (In showing this, it is best to refer to the special rules for arithmetic involving $\pm\infty$ and $\omega$ given in sec. 2.8).

We will use notation of the form

$$B = \begin{cases} A_1 & \text{if } R_1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ A_{n-1} & \text{if } R_{n-1} \\ A_n & \text{otherwise,} \end{cases}$$

to indicate that $B = R_j$ $(1 \leq j \leq n)$ when $j$ is the smallest integer such that $R_j$ is true, and $R_n$ is defined to be true always.

## 1.5 Recursive Functions and Operators

Inductive schemes for defining recursive natural functions can be found in Mendelson [M1, pp. 120-1] and elsewhere. Let $\mathfrak{n}$ be the set of nonnegative integers, 0, 1, .... The recursive functions from $\mathfrak{n}^{(m)} \to \mathfrak{n}$ $(m \geq 0)$ are those mappings from $\mathfrak{n}^{(m)} \to \mathfrak{n}$ which obtain the image point via constructive operations on the domain point. Recursive functions essentially characterize the input/output of Turing machines.

By recursive operator we mean a standard recursive functional with its integer arguments left unspecified. Thus a recursive operator $\varphi$ of $n$ function arguments constructively maps $n$ functions, $\alpha_i$: $\mathfrak{n}^{(m_i)} \to \mathfrak{n}$ $(i = 1, \ldots, n)$ , into a function, $\varphi(\alpha_1, \ldots, \alpha_n) : \mathfrak{n}^{(m)} \to \mathfrak{n}$ for some $m \geq 0$ . Inductive schemes for defining recursive functionals may be found in Schoenfield [S2], Grzegorczyk [G1], Klaua [K1] and elsewhere.

The reader does not need to know any more about recursive functions and recursive operators than what we have just stated. We will not use their inductive definitions. As usual, we say a process is "effective" or "constructive" precisely when that process can be carried out by purely mechanical means (i.e., by a Turing machine).

<u>**REMARKS**</u>: In section 1.2 we saw that increasing the precision of computation may decrease the accuracy of a numerical result. In section 5.3 we show that this does not apply to purely arithmetic processes, i.e., rational function evaluation. There, increasing precision ultimately leads to convergence. The trouble arises when limits are involved: e.g., $\lim\limits_{y\to x} f(x,y)$ in section 1.2, and $\lim\limits_{n\to\infty} \sum\limits_{i=1}^{n} a_i$ in Riesel's example [R2]. The root of this trouble is an interchange of limits which may not work. This merits further explanation. (Keep the example of section 1.2 in mind for the following.) In general, we will have

$$\lim_{\varepsilon\to 0} F(\varepsilon;\ x,y) = f(x,y)$$

for $y \neq x$ . This implies

$$\lim_{y\to x}\ \lim_{\varepsilon\to 0} F(\varepsilon;\ x,y) = \lim_{y\to x} f(x,y) \quad .$$

But in order to compute successive approximations to this limit, we must define an $\varepsilon$-limit $\varepsilon$-operator, $\underset{y\to x}{\mathrm{LIM}}$, such that $\underset{y\to x}{\mathrm{LIM}}\ F(\varepsilon;\ x,y)$ is the finitely computable $\varepsilon$-approximation to $\lim\limits_{y\to x} f(x,y)$ and such that

$$\lim_{\varepsilon\to 0}\ \underset{y\to x}{\mathrm{LIM}}\ F(\varepsilon;\ x,y) = \lim_{y\to x} f(x,y) \quad .$$

This interchange of limits is investigated in chapter 3.

BLANK PAGE

# Chapter 2: Basic Concepts

## 2.1 Variable-Precision Computations

Before launching into a description of our model, we first loosely describe the kinds of finite computations which we are interested in modeling. These are characterized by having the following facilities:

(1) variable-precision *machine numbers*;

(2) the ability to make decisions based on a *comparison* of the values of machine numbers;

(3) variahle-precision *input routines* for inputting the members of $\tilde{R}$ into machine numbers (at a specified precision), and for giving roundoff-error bounds on these inputted values; and

(4) variable-precision arithmetic (*$\epsilon$-arithmetic*) for machine numbers.

We formalize (1) - (3) in this chapter; (4) is formalized in section 5.3.

## 2.2 Machine Numbers and Their Comparison

Let $S_i (i = 1, 2, \ldots)$ be a finite subset of $\tilde{R}$ with $\{-\infty, \infty, \omega\} \subset S_i$. In order to formalize (1) and (2) above, we define the concept "$S_i$ is effectively generable from $i$." Essentially we mean by this that there is an algorithm, with input parameter $i$, which produces as output all the elements of $S_i$. But this is not precise because we as yet have not said what "produces as output a member $x$ of $\tilde{R}$" means: e.g., $x$ may have a nonrepeating decimal expansion, so our algorithm cannot in general produce $x$ by producing a decimal expansion for $x$. We define this concept precisely as follows. Let $\mathfrak{n}$ be as in section 1.5. Define the function, $\theta\colon \mathfrak{n}^{(2)} \rightarrow \{-\infty, \ldots, -1, 0, 1, \ldots, \infty, \omega\}$, by

$$(2.2\text{-}1) \qquad \theta(i, j) = \begin{cases} -\infty & \text{if } i = j = 1 \\ \infty & \text{if } i = j = 2 \\ \omega & \text{if } i = j = 3 \\ i-j & \text{otherwise.} \end{cases}$$

Let $\alpha_1$ and $\alpha_2$ be functions from $\mathfrak{n}$ to $\mathfrak{n}$ and define $\alpha(\cdot) \equiv \theta(\alpha_1(\cdot), \alpha_2(\cdot))$. We say $\alpha_1, \alpha_2$ compute $a \in \tilde{R}$ (or $\alpha$ computes $a$) precisely when

$$|a - \alpha(n)/n| \leq 1/n \qquad \text{for } n = 1, 2, \ldots,$$

and we use $a$ and $<\alpha>$ interchangeably. See Grzegorczyk [G2, p. 61] for a similar definition. When $\alpha_1$ and $\alpha_2$ are recursive and $\alpha_1, \alpha_2$ compute $a$, we say a (or $<\alpha>$ is a computable number. For given $\alpha_1, \alpha_2$: $\mathfrak{n}^{(m+1)} \rightarrow \mathfrak{n}$ $(m \geq 1)$ and a fixed $\bar{x}_m \in \mathfrak{n}^{(m)}$, we treat $\alpha(\bar{x}_m, \cdot) \equiv \theta(\alpha_1(\bar{x}_m, \cdot), \alpha_2(\bar{x}_m, \cdot))$ like a function of one variable, so that when $\alpha(\bar{x}_m, \cdot)$ computes $a \in \tilde{R}$ we use $< \alpha(\bar{x}_m, \cdot) >$ interchangeably with $a$ .

Let $S_1, S_2, \ldots$ be as described above. Let $\ell(i)$ be the number of elements in $S_i$ . We say $S_i$ is effectively generable from $i$ precisely when $\ell(\cdot)$ is recursive and there are recursive functions, $\alpha_1, \alpha_2$: $\mathfrak{n}^{(3)} \rightarrow \mathfrak{n}$ , such that, with $\alpha(\cdot) \equiv \theta(\alpha_1(\cdot), \alpha_2(\cdot))$ , we have

$$S_i = \{ < \alpha(i, 1, \cdot) > , < \alpha(i, 2, \cdot) > , \ldots , < \alpha(i, \ell(i), \cdot) > \} ,$$
$$< \alpha(i, 1, \cdot) > = -\infty, < \alpha(i, 2, \cdot) > = \infty, < \alpha(i, 3, \cdot) > = \omega .$$

We call the pair $(\alpha, \ell)$ a generator of $S_i$ .

The concept of variable-precision machine numbers is formalized as follows. We use the positive real constants, $\varepsilon_1, \varepsilon_2, \ldots$, to denote the possible levels of precision: $\varepsilon_1$ denotes single-precision, etc. We use $\varepsilon$ to denote a variable which takes values in the set

$$\mathcal{E} \equiv \{\varepsilon_1, \varepsilon_2, \ldots \} . \tag{2.2-2}$$

DEFINITION 2.2-1: A machine number system, $(\mathcal{R}, \mathcal{E})$ , is a set $\mathcal{E}$ of constants $\varepsilon_1, \varepsilon_2, \ldots$ together with a mapping, $\mathcal{R}$: $\mathcal{E} \rightarrow$ (set of subsets of $\tilde{R}$), where

(I) $\varepsilon_i \rightarrow 0$ strictly monotonically as $i \rightarrow \infty$ ,

(II) $\bigcup_{i \geq 1} \mathcal{R}(\varepsilon_i)$ is dense in $R$ ,

(III) $\mathcal{R}(\varepsilon_1) \subset \mathcal{R}(\varepsilon_2) \subset \ldots$,

(IV) $\mathfrak{R}(\varepsilon)$ is finite, for each $\varepsilon$ in $\mathcal{E}$ ,

(V) $\{0, 1, -\infty, \infty, \omega\} \subset \mathfrak{R}(\varepsilon_1)$ , and

(VI) $\mathfrak{R}(\varepsilon_i)$ is effectively generable from $i$ .

$\mathfrak{R}(\varepsilon)^{(n)}$ represents a discretization of Euclidean n-space. Condition I reflects the statement "decreasing $\varepsilon$ increases the precision"; no other use of the values of the $\varepsilon_i$ will be made in this paper. Conditions II and VI are the only really essential restrictions; if $\mathfrak{R}$ and $\mathcal{E}$ satisfy them, minor modifications will produce an $\mathfrak{R}'$ and $\mathcal{E}'$ which satisfy I-VI. (If I is violated, replace $\mathcal{E}$ by any $\mathcal{E}'$ satisfying it; if III is violated, let $\mathfrak{R}'(\varepsilon_i) \equiv \bigcup_{j=1}^{i} \mathfrak{R}(\varepsilon_j)$, etc.) Condition II allows us to get at any number in R through the exclusive use of machine numbers. The nesting condition, III, says that we may reuse, at precision $\varepsilon_{i+1}$ , any machine number that we used at precision $\varepsilon_i$ ; this will be used in dealing with instability and in the proof that our $\varepsilon$-limit approaches limit as $\varepsilon \to 0$ . Condition IV will simplify our treatment of instability. Condition VI means that $\mathfrak{R}$ could really be used as the basis of a variable-precision number system on a computer; it insures that the switching of precision can be done automatically. (We investigate other implications of this condition below.) That 0 and 1 are in $\mathfrak{R}(\varepsilon)$ will prove convenient in many situations, but never will this be a necessity. However, having $-\infty, \infty, \omega \in \mathfrak{R}(\varepsilon)$ greatly simplifies our model. We give three examples to clarify these ideas and to show the variety of machine number systems which satisfy I-VI.

Example 2.2-1: Let $\beta$ be an integer $\geq 2$. Let $0.a_1 a_2 \ldots a_i$

denote a base $\beta$ fraction Define $\varepsilon_i^{\odot} = 1/(2\beta^{i-1}+1)$.

A <u>base $\beta$ normalized floating-point number system</u> is given by

$$\mathcal{E}^{\odot} \equiv \{\varepsilon_1^{\odot}, \varepsilon_2^{\odot}, \dots\},$$

$$\mathcal{R}^{\odot}(\varepsilon_i^{\odot}) \equiv \{0, -\infty, \infty, \omega\} \cup \{x: \ |x| = 0.a_1 a_2 \dots a_i \times \beta^e,$$

$$a_1 > 0, \text{ and } e \text{ is an integer, } |e| < \beta^i\}.$$

We chose these $\varepsilon_i^{\odot}$ because each real number $x$ such that

$\beta^{-\beta^i} \le |x| \le (\beta^i-1)\,\beta^{\beta^i-i-1}$ can (in principle) be inputted into $\mathcal{R}^{\odot}(\varepsilon_i^{\odot})$ with a relative error $\le$ this $\varepsilon_i^{\odot}$.

<u>**EXAMPLE 2.2-2:**</u> Let $\beta_1$ be some finite number $> 1$. Define $\beta_{i+1} = \beta_i^{1/10}$ and $\varepsilon_i^{*} = 1-2/(\beta_i+1)$. A base $\beta_1$ <u>logarithmic number system is given by</u>

$$\mathcal{E}^{*} \equiv \{\varepsilon_1^{*}, \varepsilon_2^{*}, \dots\}$$

$\mathcal{R}^{*}(\varepsilon_i^{*}) \equiv \{0, -\infty, \infty, \omega\} \cup \{x: \ |x| = \beta_i^e, \ e \text{ an integer, } |e| < 10^{2i}\}$

We chose these $\varepsilon_i^{*}$ because each real number $x$, with $\beta_1^{-10^{i+1}+10^{1-i}} \le |x| \le \beta^{10^{i+1}-10^{1-i}}$, can (in principle) be inputted into $\mathcal{R}^{*}(\varepsilon_i^{*})$ with a relative error $\le$ this $\varepsilon_i^{*}$. We had to use different bases, $\beta_i$, approaching 1 as $i \to \infty$, so that condition II is met. The fact that some of the $\beta_i$ will be irrational causes no difficulties.

<u>**EXAMPLE 2.2-3:**</u> Define $\varepsilon_i^{\#} = 1/(10^i+1)$ and

$$\mathcal{E}^{\#} \equiv \{\varepsilon_1^{\#}, \varepsilon_2^{\#}, \dots\},$$

$$\mathcal{R}^{\#}(\varepsilon_i^{\#}) \equiv \{0, -\infty, \infty, \omega\} \cup \{x: \ x = p/q \text{ for integers } p \text{ and } q$$

$$\text{with } |p|, |q| < 10^i\}.$$

We chose these $\varepsilon_i^{\#}$ because each real number $x$, with $1/(10^i-1) \le |x| \le 10^i-1$ can (in principle) be inputted into $\mathfrak{R}^{\#}(\varepsilon_i^{\#})$ with a relative error $\le$ this $\varepsilon_i^{\#}$. (This last statement is more difficult to prove than the corresponding statements of the other examples, so a proof is included in the appendix to this chapter. The other statements are also proved there as simple corollaries.)

Another important property which $\mathfrak{R}$ must possess is that the members of $\mathfrak{R}(\varepsilon_i)$ must be representable in some simple form which varies with $i$ in a simple way. This is necessary so that the members of $\mathfrak{R}(\varepsilon)$ can be represented simply in the computer. For example, any $x \in \mathfrak{R}^{\odot}(\varepsilon_i^{\odot}) - \{0, -\infty, \infty, \omega\}$ can be represented by a pair of integers $(a, e)$ with $\beta^{i-1} \le |a| < \beta^i$ and $|e| < \beta^i - i$ since $x$ must equal $(a\,\beta^{-i})\,\beta^e$ for some such $a$ and $e$. And any $x$ in $\mathfrak{R}^*(\varepsilon_i^*) - \{0, -\infty, \infty, \omega\}$ can be represented by a pair of integers $(\pm\, i, e)$ with $|e| < 10^{2i}$, since $x$ must equal $\pm\,\beta_i^e$ for some such $e$. A similar statement can be made about $\mathfrak{R}^{\#}$. In fact, any $\mathfrak{R}$ which satisfies condition VI possesses this representability property. Suppose $(\alpha_{\mathfrak{R}}, \ell_{\mathfrak{R}})$ is a generator for $\mathfrak{R}(\varepsilon_i)$. Then any $x$ in $\mathfrak{R}(\varepsilon_i)$ can be represented by a pair of integers $(i, j)$, with $1 \le j \le \ell_{\mathfrak{R}}(i)$, since $x$ must equal some $< \alpha_{\mathfrak{R}}(i, j, \cdot) >$. So much for representability. Next we consider comparison of machine numbers.

The fact that each member of $\mathfrak{R}(\varepsilon_i)$ has a unique representation in terms of $\alpha_{\mathfrak{R}}$ means that, given i, j and k, we can effectively decide whether $< \alpha_{\mathfrak{R}}(i, j, \cdot) >$ is $>$, $<$ or $= < \alpha_{\mathfrak{R}}(i, k, \cdot) >$;

when $j \neq k$ we have $< \alpha_{\mathcal{R}}(i, j, \cdot) > \neq < \alpha_{\mathcal{R}}(i, k, \cdot) >$ and we can determine which is greater by computing $\alpha_{\mathcal{R}}(i, j, n)$ and $\alpha_{\mathcal{R}}(i, k, n)$ for some finite number of values of $n$. This gives us (2) of section 2.1. Further, the following two conditions imply the existence of a generator for $S_i$:

(1) there are recursive functions $\alpha_1', \alpha_2', \ell'$ sucn that, with $\alpha'(\cdot) \equiv \theta(\alpha_1'(\cdot), \alpha_2'(\cdot))$, we have

$$S_i \equiv \bigcup_{j=1}^{\ell'(i)} \{ < \alpha'(i, j, \cdot) > \} \qquad \text{for } i = 1, 2, \dots ,$$

(2) the relation $< \alpha'(i, j, \cdot) > = < \alpha'(i, k, \cdot) >$ is effectively decidable from $j$ and $k$.

Thus condition VI on $\mathcal{R}$ is not too restrictive in (implicitly) requiring $(\alpha_{\mathcal{R}}, \ell_{\mathcal{R}})$ to be nonredundant.

Throughout the rest of this paper, we assume that $\mathcal{R}$, $\mathcal{E}$ and a corresponding generator $(\alpha_{\mathcal{R}}, \ell_{\mathcal{R}})$ are given and fixed. All of the following definitions are implicitly relative to these $\mathcal{R}$, $\mathcal{E}$ and $(\alpha_{\mathcal{R}}, \ell_{\mathcal{R}})$.

For later use, we define the <u>machine number set</u>, $\mathcal{M}$, by

$$\mathcal{M} = \bigcup_{i \geq 1} \mathcal{R}(\varepsilon_i) . \tag{2.2-3}$$

If $\mathcal{R}$ and $\mathcal{E}$ are the $\mathcal{R}^{\#}$ and $\mathcal{E}^{\#}$ of example 2.2-3, then $\mathcal{M}$ is $\mathcal{M}^{\#}$, the set of rationals together with $-\infty$, $\infty$ and $\omega$.

## 2.3 Input Routines

We handle the inputting of members of $\tilde{\mathfrak{R}}$ essentially by assuming that members of $\tilde{\mathfrak{R}}$ are given by giving an input routine. To avoid confusing a number with its input routine, we introduce a new concept, that of *a real input*.

DEFINITION 2.3-1: *A real input $x \equiv (X, RX)$ is a pair of mappings, $X, RX: \mathcal{E} \to \mathfrak{M}$, such that*

(1) $X(\varepsilon)$, $RX(\varepsilon) \in \mathfrak{R}(\varepsilon)$ *and* $RX(\varepsilon) \geq 0$ *for all* $\varepsilon \in \mathcal{E}$,

(2) $|X(\varepsilon) - X(\eta)| \leq RX(\varepsilon) + RX(\eta)$ *for all* $\varepsilon, \eta \in \mathcal{E}$,

(3) $\lim_{\varepsilon \to 0} RX(\varepsilon) = 0$.

*If* $x \equiv (X, RX)$ *satisfies* (1) *and* (2), *we call* $x$ *a poor real input. We call* $(X, RX)$ *an input routine for $x$*.

It follows that the numeric value corresponding to the real input $x$ is just $\lim_{\varepsilon \to 0} X(\varepsilon)$. *When a real input $x$ is used in a context that calls for a numeric value, we let that value be $\lim_{\varepsilon \to 0} X(\varepsilon)$.* Thus for each $\varepsilon$ we have

$$RX(\varepsilon) \geq |X(\varepsilon) - x| , \tag{2.3-1}$$

and so $RX(\varepsilon)$ is just a roundcff-error bound, bounding the error caused by using $X(\varepsilon)$ in place of (the numeric value of) $x$.

A real input can be thought of as a variable ranging not only over $\tilde{\mathfrak{R}}$ but also over input routines. We will find it unnecessary to distinguish notationally between a real variable (ranging over $\tilde{\mathfrak{R}}$ ) and a real input, or between a real constant and a fixed real input. When a real input is named $x$, its input routine will be named (X, RX), etc. Note that for real inputs $x$ and $y$, $x = y$ ("=" means "equal in numeric value" throughout this paper, see sec 1.4) precisely when

$$(2.3\text{-}2) \qquad |X(\varepsilon) - Y(\varepsilon)| \leq RX(\varepsilon) + RY(\varepsilon) \quad \text{for all} \quad \varepsilon \in \mathcal{E} \ .$$

This relation will be useful in defining $\varepsilon$-comparison relations.

The following conventions will simplify notation later:

(1) when, in a given context, the value of the real input $x$ is known to be in $\mathbf{R}(\varepsilon)$ and $(X, RX)$ has not been explicitly specified, it will be assumed that $X(\delta) = x$ (in value) and $RX(\delta) = 0$ for all $\delta \leq \varepsilon$ ,

(2) when we state that $x \in \mathcal{M}$, for a real input $x$, we mean that $\{RX(\varepsilon) = 0$ or $RX(\varepsilon) < \infty = |X(\varepsilon)|$ for some $\varepsilon\}$ and $x \notin \mathcal{M}$ means $\{|x| \neq \infty$ and $RX(\varepsilon) \neq 0$ for all $\varepsilon\}$, and

(3) we will use the same notation for sets of numbers and sets of real inputs. If $P$ is (in a given context) a set of numbers and it is not known that $P \subset \mathcal{M}$, then the set, $P$, of real inputs contains all real inputs $x$ with value in the number set $P$ . If the number set $P$ is known to be a subset of $\mathcal{M}$, then the real input set $P$ contains all real inputs $x$ with value in the number set $P$ such that $x \in \mathcal{M}$ (under convention (2) above).

By (3), $\widetilde{R}$ may denote the set of all numbers or the set of all real inputs, depending on context. We use a rule analogous to (3) when [illegible] is a set of m-tuples of numbers.

## 2.4 Multiple-Precision Subroutines

Let $\overline{x}_0$ denote the empty list and for $m \geq 1$ let $\overline{x}_m$ denote $x_1, x_2, \ldots, x_m$. For the moment, let $\overline{x}_m$ be $m \geq 0$ variable poor real inputs and let $x_{m+1}, \ldots, x_{m+n}$ be $n \geq 0$ fixed real inputs. A multiple-precision subroutine of $m \geq 0$ variables and $n \geq 0$ constants is essentially a computer subroutine with input $\varepsilon$ and with access to any finite number of values of the input routines for $\overline{x_{m+n}}$, say

$$\text{(2.4-1)} \quad \left.\begin{array}{l} X_j(\varepsilon_1), X_j(\varepsilon_2), \ldots, X_j(\varepsilon_M) \\ \\ RX_j(\varepsilon_1), RX_j(\varepsilon_2), \ldots, RX_j(\varepsilon_M) \end{array}\right\} \quad \text{for } j = 1, \ldots, m+n \quad ,$$

and with output in $\mathfrak{R}(\varepsilon)$, with the requirement that if any $X_j(\varepsilon)$ or $RX_j(\varepsilon)$ is $\omega$ then the output is $\omega$. We call $\varepsilon$ and $\overline{x}_m$ the inputs. With inputs $\varepsilon$ and $\overline{x}_m$, we denote the output value of the subroutine $F$ by $F(\varepsilon; \overline{x}_m)$. If for some $j \leq m$ we have $\limsup_{\varepsilon \to 0} RX_j(\varepsilon) > 0$, then we allow $F$ to not halt when its inputs are $\overline{x}_m$ and (any) $\varepsilon$. When this happens we write $F(\varepsilon; \overline{x}_m) = \omega$ because $F(\varepsilon; \overline{x}_m)$ is undefined in terms of the members of $\widetilde{R} - \{\omega\}$ (see sec. 1.4).

This can be formalized as follows. For $n = 1, 2, \ldots$ let $p_n$ be the $n^{th}$ prime: $p_1 = 2$ etc. Let $x$ be a poor real input and suppose that for $i \geq 1$,

$$X(\varepsilon_i) = < \alpha_{\mathfrak{R}}(i, k_{2i-1}, \cdot) > \quad RX(\varepsilon_i) = < \alpha_{\mathfrak{R}}(i, k_{2i}, \cdot) > \quad .$$

For $j \geq 1$ define

$$gn_j(x) = p_1^{k_1} \times p_2^{k_2} \times \ldots \times p_{2j}^{k_{2j}} \quad .$$

Define the $j^{th}$ *Gödel number of $\bar{x}_m$* $(m \geq 0)$ by

$$(2.4\text{-}2) \qquad GN_j(\bar{x}_m) = p_1^{gn_j(x_1)} \times p_2^{gn_j(x_2)} \times \ldots \times p_m^{gn_j(x_m)} ,$$

where the empty product is 1 (i.e., $GN_j(\bar{x}_0) = 1$) . $GN_M(\bar{x}_m)$ contains complete information about the numbers shown in (2.4-1). We say that *$\gamma$ is m-determining* $(m \geq 1)$ precisely when $\gamma$ is recursive and, for $i = 1, 2, \ldots,$

(1) $0 \leq \gamma(i, j) \leq \ell_R(i)$ for any $j$ ,

(2) for any poor real inputs $\bar{x}_m$, $\gamma(i, GN_k(\bar{x}_m)) = 0$

for $1 \leq k < i$, and if any $X_j(\varepsilon_i)$ or $RX_j(\varepsilon_i)$ is $\omega$ then $\gamma(i, GN_i(\bar{x}_m)) = 3$, and

(3) for any real inputs $\bar{x}_m$ there is an $M \geq i$ with $\gamma(i, GN_M(\bar{x}_m)) \neq 0$ .

Thus $\gamma$ waits until sufficient information about $\bar{x}_m$ has been collected (in $GN_M(\bar{x}_m)$), and then $\gamma$ returns a nonzero value. When $\liminf_{\varepsilon \to 0} RX_j(\varepsilon) > 0$ for some $j$, $GN_k(\bar{x}_m)$ may never contain enough information about $x_j$ for $\gamma$ to return a nonzero value. (Of course, even when $\liminf_{\varepsilon \to 0} RX_j(\varepsilon) = 0 < \limsup_{\varepsilon \to 0} RX_j(\varepsilon)$ for some $j$ we may have $\gamma(i, GN_k(\bar{x}_m)) = 0$ for some $i$ and *all* $k$, but this will not be due to lack of information about $x_j$ .) We say *$\gamma$ is 0-determining* precisely when $\gamma$ is recursive and $1 \leq \gamma(i, 1) \leq \ell_R(i)$ . Let $\tilde{\mu}_n(\gamma(i, GN_n(\bar{x}_m)) \neq 0)$ be $\gamma(i, GN_M(\bar{x}_m))$ where $M$ is the least value of $n$ such that $\gamma(i, GN_n(\bar{x}_m)) \neq 0$, or let it be 3 when there is no such $n$ .

DEFINITION 2.4-1: A multiple-precision subroutine of $m \geq 0$ variables and $n \geq 0$ constants, $x_{m+1}, \ldots, x_{m+n}$, is a mapping, $F : \mathcal{E} \times \{\text{poor real inputs}\}^{(m)} \to \mathcal{M}$, such that there is an $m+n$-determining $\gamma$ which satisfies the following for any poor real inputs $\bar{x}_m$ and any $i \geq 1$ :

$$(2.4\text{-}3) \qquad F(\varepsilon_i; \bar{x}_m) = < \alpha_{\mathbf{R}}(i, \tilde{\mu}_n(\gamma(i, GN_n(\bar{x}_{m+n})) \neq 0), \cdot) > .$$

We say $\gamma$ determines $F$ relative to $(\alpha_{\mathbf{R}}, \ell_{\mathbf{R}})$ . We stress that so long as the $\bar{x}_m$ are real inputs (not just poor), the computation of $F(\varepsilon; \bar{x}_m)$ via $\gamma$ will always halt. Essentially the only subroutines whose computation may fail to halt are those which, with inputs $\varepsilon$ and $\bar{x}_m$, try to find an $\eta \leq \varepsilon$ such that, say $RX_1(\eta) \leq t$, for some tolerance level $0 < t < \infty$ ; for example, when $x_1$ is a poor real input with $RX_1(\varepsilon) = \infty$ for all $\varepsilon$, such a subroutine will fail to halt. We will use "subroutine" as an abbreviation for "multiple-precision subroutine." An immediate consequence of the above definitions is

$$F(\varepsilon; \bar{x}_m) = \omega \quad \text{if any } X_j(\varepsilon) \text{ or } RX_j(\varepsilon) \text{ is } \omega .$$

This convention is taken from Scott [S1]. Note that these definitions have all been relative to $(\alpha_{\mathbf{R}}, \ell_{\mathbf{R}})$ .

THEOREM 2.4-1: If $F$ is a subroutine relative to $(\alpha_{\mathbf{R}}, \ell_{\mathbf{R}})$ and $(\alpha'_{\mathbf{R}}, \ell_{\mathbf{R}})$ is some other generator of $\mathbf{R}(\varepsilon_i)$ then $F$ is a subroutine relative to $(\alpha'_{\mathbf{R}}, \ell_{\mathbf{R}})$ .

This means that the concept of subroutine is independent of which generator of $\mathfrak{R}(\varepsilon_i)$ one uses.

Proof: Let $GN_j(\overline{x}_m)$ and $GN'_j(\overline{x}_m)$ denote the $j^{th}$ Gödel numbers of $\overline{x}_m$ relative to $(\alpha_{\mathfrak{R}}, \ell_{\mathfrak{R}})$ and $(\alpha'_{\mathfrak{R}}, \ell_{\mathfrak{R}})$, respectively. The considerations of section 2.2 show that there is an effective procedure which, when given any generator for $\mathfrak{R}(\varepsilon_i)$ and any $j \geq 1$, can order the members of $\mathfrak{R}(\varepsilon_j)$ . This means that there are recursive functions $\varphi_1$ and $\varphi_2$ such that $\varphi_1(i, 0) = 0$ and

$$< \alpha'_{\mathfrak{R}}(i, \varphi_1(i, j), \cdot) > = < \alpha_{\mathfrak{R}}(i, j, \cdot) > \quad ,$$

$$\varphi_2(i, GN'_i(\overline{x}_m)) = GN_i(\overline{x}_m) \quad ,$$

for any $i \geq 1$ and $1 \leq j \leq \ell_{\mathfrak{R}}(i)$ . Define $\gamma'$ by

$$\gamma'(i, GN'_j(\overline{x}_{m+n})) = \varphi_1(i, \gamma(i, \varphi_2(j, GN'_j(\overline{x}_{m+n})))) \quad ,$$

for $i \geq 1$, $j \geq 1$ and any $\overline{x}_m$ . If $\gamma$ determines $F$ relative to $(\alpha_{\mathfrak{R}}, \ell_{\mathfrak{R}})$ then $\gamma'$ determines $F$ relative to $(\alpha'_{\mathfrak{R}}, \ell_{\mathfrak{R}})$ . This completes the proof.

## 2.5 $\epsilon$-Function

An ideal function of $m$ variables $(m \geq 0)$ is a mapping, $f: \widetilde{R}^{(m)} \rightarrow \widetilde{R}$, with the constraint that $f(\overline{x}_m) = \omega$ if any $x_i = \omega$.

DEFINITION 2.5-1: For $m \geq 0$, an $\epsilon$-function, $\mathfrak{F}$, of $m$ variables corresponding to an ideal function $f$ (of $m$ variables) over a set $P$ of $m$-tuples of real inputs is a triple $(F, RF, TF)$ of subroutines, where for any real inputs $\overline{x}_m$ we have

(1) for each $\epsilon$, $RF(\epsilon; \overline{x}_m) \geq |F(\epsilon; \overline{x}_m) - f(\overline{x}_m)|$ ,

(2) $[\overline{x}_m \in P$ and $f(\overline{x}_m) \neq \omega] \Rightarrow \lim_{\epsilon \to 0} RF(\epsilon; \overline{x}_m) = 0$, and

(3) if $m = 0, 1$ then $TF \equiv \omega$; otherwise, for all $\epsilon$,

$$TF(\epsilon; \overline{x}_m) \geq |f(\overline{x}_m) - \lim_{y \to x_{m-1}} f(\overline{x}_{m-1}, y)| \quad . \tag{2.5-1}$$

We call $P$ a domain set of $\mathfrak{F}$ and we write

$$\mathfrak{F} \approx f(P) \quad ,$$

to be read "$\mathfrak{F}$ corresponds to (or approximates) $f$ over (or mod) $P$." This definition is illustrated in figure 2.5-1 for the case $m = 2$ and $x = \infty$ ,

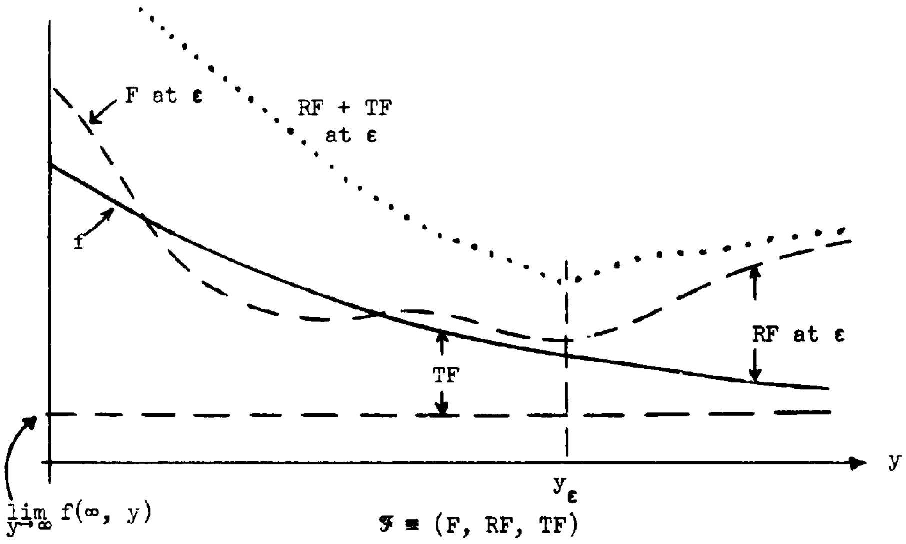


FIGURE 2.5-1

RF is a <u>roundoff-error bound</u>, bounding the error incurred by using F in place of f . TF is a <u>truncation-error bound</u>, bounding the error incurred by using $f(\bar{x}_m)$ in place of $\lim_{y\to\bar{x}_{m-1}} f(\bar{x}_{m-1}, y)$ . For example, if $f(x, n) = \sum_{i=1}^{n} g(i)$, then $TF(\varepsilon; \infty, n)$ bounds the truncation error $|\sum_{i=n+1}^{\infty} g(i)|$ . $RF(\varepsilon; \bar{x}_{m-1}, Y) + TF(\varepsilon; \bar{x}_{m-1}, Y)$ bounds the error incurred by using $F(\varepsilon; \bar{x}_{m-1}, Y)$ in place of $\lim_{y\to\bar{x}_{m-1}} f(\bar{x}_{m-1}, y)$ . For the above graph, this bound is smallest when $Y = y_\varepsilon$ .

Conditions (1) and (3) on $\mathcal{F}$ require that the bounds RF and TF work properly for <u>any</u> real inputs $\bar{x}_m$ and <u>any</u> ε . Condi^ton (2) requires the convergence of F to f and RF to 0 at $\bar{x}_m \in P$ for which $f(\overline{x}_m)$ is defined. (We must have $RF \to 0$ in order to <u>effectively</u> compute, via $F$, an approximation to $f$ that is correct to within some desired and arbitrarily small tolerance.) If instead of (2) we have

$$(2')[\overline{x}_m \in P \text{ and } f(\overline{x}_m) \neq \omega] \Rightarrow \lim_{\varepsilon \to 0} f(\varepsilon; \overline{x}_m) = f(\overline{x}_m) \quad ,$$

we say <u>$\mathfrak{F}$ weakly corresponds to $f$ over $P$</u> and we write

$$\mathfrak{F} \sim f(P) \quad .$$

We call $P$ <u>a weak domain set of $\mathfrak{F}$</u>. An immediate consequence of these definitions is

<u>THEOREM 2.5-1</u>: <u>If</u> $Q \subset P$ <u>and</u> $\mathfrak{F} \approx f(P)$ (<u>or</u> $\mathfrak{F} \sim f(P)$) <u>then</u> $\mathfrak{F} \approx f(Q)$ (<u>or</u> $\mathfrak{F} \sim f(Q)$) .

We will use $\mathfrak{F}(\varepsilon; \overline{x}_m)$ to denote the triple of values,

$$(F(\varepsilon; \overline{x}_m), RF(\varepsilon; \overline{x}_m), TF(\varepsilon; \overline{x}_m)) \quad .$$

For any triple $\overline{a}_3$ of numbers, we will use $(\overline{a}_3)_i$ to denote $a_i (1 \leq i \leq 3)$ . Thus we have

$$(\mathfrak{F}(\varepsilon; \overline{x}_m))_1 = F(\varepsilon; \overline{x}_m) \quad ,$$

etc. $\varepsilon$-Functions are finitely computable in the following sense: there is a Turing machine which, when given an object[1/] for computing the $X_j(\varepsilon)$ and $RX_j(\varepsilon)$ for any given $\varepsilon$, can output the triple of values, $\mathfrak{F}(\varepsilon; \overline{x}_m)$ for any given $\varepsilon$ .

Dealing with the instabilities alluded to in the introduction was not our main reason for introducing the truncation-error bounds (TF). We had to introduce them because any definition of "$\mathcal{F}$ $\varepsilon$-converges at $x$," which is based only on the values of $F(\varepsilon; x, y)$ and $RF(\varepsilon; x, y)$ ($\varepsilon$ and $x$ are fixed here), cannot have much to do with "f converges at $x$) (which is true when $\lim_{y \to x} f(x,y) \neq \omega$) ; remember that $F$ and $RF$ can only take on a finite number of different values for each fixed $\varepsilon$. The truncation-error bound, TF, gives us the needed local information about $f$.

Suppose $\gamma_1$, $\gamma_2$, and $\gamma_3$ determine the subroutines F, RF and TF respectively. Then we say *$(\gamma_1, \gamma_2, \gamma_3)$ determines $\mathcal{F} \equiv (F, RF, TF)$* and *$(\gamma_1, \gamma_2)$ partially determines $\mathcal{F}$*. When we say "given $\mathcal{F}$" we mean "given $(\gamma_1, \gamma_2, \gamma_3)$ determining $\mathcal{F}$."

[1]/We call the thing which computes $(X_i, RX_i)$ an "object" rather than a Turing machine because there may be no such Turing machine; there are only a countable number of Turing machines, but there are an uncountable number of values of $\bar{x}_m$. See Shoenfield [S1, p. 248] for similar considerations.

## 2.6 An Example: $e^x$

Following is an example of how one might go about defining an $\epsilon$-function corresponding to $e^x$ over $\widetilde{R}$. This example is formalized in section 5.7.

Let $[y]$ denote the greatest integer in $y$ and let $\operatorname{sgn}(x)$ denote the sign function at $x$ (which is $\omega$ if $x$ is $\omega$, and otherwise is $-1$ if $x < 0$, is $0$ if $x = 0$, and is $1$ if $x > 0$).

Define

$$f(x,k,y) = \begin{cases} \omega & \text{if } k=\omega, \text{ or } y \geq \infty \\ 0 & \text{if } y < 1 \\ \sum\limits_{n=1}^{[y]} (-|x|)^{n-1}/(n-1)! & \text{otherwise}, \end{cases}$$

$$tf(x,k,y) = \begin{cases} |x|^{[y]}/[y]! & \text{if } k = \infty \text{ and } |x| + 1 < y < \omega \\ \omega & \text{otherwise}, \end{cases}$$

$$f_{\exp}(x) = \begin{cases} \infty & \text{if } x = \infty \\ 0 & \text{if } x = -\infty \\ (\lim\limits_{y \to \infty} f(x,\infty,y))^{-\operatorname{sgn}(x)} & \text{otherwise}. \end{cases}$$

$tf(x,\infty,y)$ bounds the remainder term, $\left| \sum\limits_{n=[y]}^{\infty} (-|x|)^n/n! \right|$. The point here is that $tf$ can be computed using only arithmetic, $[\cdot]$, $|\cdot|$ and numerical comparisons. It should not surprise the reader that there is a subroutine, $F$, such that $F(\epsilon; \bar{x}_3) \to f(\bar{x}_3)$ as $\epsilon \to 0$ for most $\bar{x}_3$. We can use the methods of interval analysis, or an error analysis in the style of Wilkinson [W2], to obtain a subroutine RF.

We can use interval analysis to obtain a subroutine, TF, which satisfies

$$TF(\varepsilon;\, \bar{x}_3) \geq tf(\bar{x}_3) \quad .$$

The result is $\mathcal{F} \equiv (F, RF, TF)$, an $\varepsilon$-function corresponding to $f$ over most of $\tilde{R}^{(3)}$. For each $\varepsilon$, let $y_\varepsilon$ be some member of $\mathcal{R}(\varepsilon)$ with $y_\varepsilon \neq \infty$ and $\lim\limits_{\varepsilon \to 0} y_\varepsilon = \infty$. Let $RG(\varepsilon;\, x)$ be either of the two smallest numbers in $\mathcal{R}(\varepsilon)$ which are $\geq RF(\varepsilon;x,\infty,y_\varepsilon) + TF(\varepsilon;x,\infty,y_\varepsilon)$. We can define an $\varepsilon$-function corresponding to

$$g(x) = \begin{cases} \omega & \text{if } \; |x| \geq \infty \\ e^{x} & \text{if } \; x \leq 0 \\ e^{-x} & \text{if } \; x \geq 0 \quad , \end{cases}$$

by

$$\mathcal{G}(\varepsilon;x) \equiv (F(\varepsilon;x,\infty,y_\varepsilon),\, RG(\varepsilon;x),\, \omega) \quad .$$

Thus $\mathcal{G}(\varepsilon_{i+1};x)$ is computed at a higher precision of computation (see sec. 1.2) than is $\mathcal{G}(\varepsilon_i;x)$ and we will have $\mathcal{G} \approx g(P)$ for some set $P \subset \tilde{R}$. It is easy to get $\mathcal{F}_{\exp}$ from $\mathcal{G}$.

This method of approximating $e^x$ by an alternating series has the numerical disadvantage of involving cancellation, but it affords the use of the simple and rapidly convergent truncation-error bound, $|x|^n/n!$ (when $n \geq |x|$). A method based on $\Sigma |x|^n/n!$ would involve no cancellation (so lower precision arithmetic could be used) but we would have to use a more complicated and more slowly converging truncation-error bound of the form $|x|^n\, 2.74^{[|x|+1]}/n!$ (valid for any $n \geq 0$). We use the former method here because it simplifies the formalization in section 5.7.

## 2.7 Operators

Operators and ε-operators will be our principal vehicles for defining notions and ε-notions in chapters 4-6. Let $S_f$ be *the set of all ideal functions of 0, 1, 2,.. variables*. *An operator of $n \geq 0$ ideal functions over $S \subset S_f^{(n)}$* is a mapping, $\emptyset : S \rightarrow S_f$ . Let $\overline{f}_n$ denote the list of ideal functions, $f_1, \ldots, f_n$, and likewise for $\overline{\mathfrak{F}}_n$ and $\overline{P}_n$ . We say *$\overline{\mathfrak{F}}_n$ corresponds to $\overline{f}_n$ over $\overline{P}_n$* precisely when $\mathfrak{F}_i \approx f_i(P_i)$ $(1 \leq i \leq n)$, and we write

$$\overline{\mathfrak{F}}_n \approx \overline{f}_n(\overline{P}_n) \quad .$$

Let S be the set of all $\overline{\mathfrak{F}}_n$ such that there is a $\overline{f}_n \in S$ and a $\overline{P}_n$ with $\overline{\mathfrak{F}}_n \approx \overline{f}_n(\overline{P}_n)$ . Let $S_w$ be *the set of all weak ε-functions*. *A weak ε-operator, $(\Phi,Q)$, corresponding to $\emptyset$ over $S'$* $\subset S$ is a mapping, $\Phi : S' \rightarrow S_w$, together with a set function, Q, which depends on $\overline{\mathfrak{F}}_n, \overline{f}_n, \overline{P}_n$, such that

(1) if $\overline{\mathfrak{F}}_n \approx \overline{f}_n(\overline{P}_n)$, $\overline{\mathfrak{F}}_n \in S'$ and $\overline{f}_n \in S$, then

$\Phi(\overline{\mathfrak{F}}_n) \sim \emptyset(\overline{f}_n)(Q(\overline{\mathfrak{F}}_n, \overline{f}_n, \overline{P}_n))$, and

(2) there exist recursive operators $\Psi_1$ and $\Psi_2$ such that,

if $\overline{\mathfrak{F}}_n \approx \overline{f}_n(\overline{P}_n)$, $\overline{\mathfrak{F}}_n \in S'$, $\overline{f}_n \in S$ and $(Y_{3i-2}, Y_{3i-1}, Y_{3i})$ determines $\mathfrak{F}_i$ $(1 \leq i \leq n)$, then $(\Psi_1(\overline{Y_{3n}}), \Psi_2(\overline{Y_{3n}}))$ partially determines $\Phi(\overline{\mathfrak{F}}_n)$ .

When these conditons hold, we write

$$(\Phi, Q) \sim \emptyset(S') \quad .$$

Condition (1) requires that $\Phi$ gives $\varepsilon$-approximations to $\emptyset$ and that $\Phi \to \emptyset$ as $\varepsilon \to 0$ in the sense that we at least have

$$\lim_{\varepsilon \to 0} (\Phi(\bar{\mathfrak{F}}_n)(\varepsilon; \bar{x}_p))_1 = \emptyset(\bar{f}_n)(\bar{x}_p)$$

for $\bar{x}_p \in Q(\bar{\mathfrak{F}}_n, \bar{f}_n, \bar{F}_n)$ at which $\emptyset(\bar{f}_n)(\bar{x}_p) \neq \omega$. Conditon (2) requires that $\Phi$ be finitely computable from its arguments; it requires $\Phi$ to constructively map the determiners of $\bar{\mathfrak{F}}_n$ into a partial determiner of $\Phi(\bar{\mathfrak{F}}_n)$. We have left truncation-error bounds out of (2) because, for the $\varepsilon$-operators which we will present later, we do not believe there is an automatic way to define a good truncation-error bound for $\Phi(\bar{\mathfrak{F}}_n)$ from the determiners of the $\bar{\mathfrak{F}}_n$ (see def. 4.1-1). In general, such bounds depend on certain analytic properties of $\emptyset(\bar{f}_n)$, properties which cannot be effectively recovered from the numeric information given by the determiners of $\bar{\mathfrak{F}}_n$. We avoid this problem in most cases by assuming such bounds to be given. If $\emptyset(\bar{f}_n)$ is a function of 0 or 1 variables, then the TF part of $\mathfrak{F} \equiv \Phi(\bar{\mathfrak{F}}_n)$ is, by definition, identically $\omega$, and this problem does not arise. We will have more to say about this in chapter 5.

If condition (1) above holds with "$\Phi(\bar{\mathfrak{F}}_n) \approx \emptyset(\bar{f}_n)$", we say (Φ, Q) is an ε-operator corresponding to ∅ over S' (and not just a weak $\varepsilon$-operator), and we write

$$(\Phi, Q) \approx \emptyset(S') \quad .$$

The "goodness" of $(\Phi, Q)$ depends on how nontrivial the relation between $\bar{\mathfrak{F}}_n$, $\bar{f}_n$, $\bar{F}_n$ and $Q(\bar{\mathfrak{F}}_n, \bar{f}_n, \bar{F}_n)$ is, how large S' is, and especially on how efficient $\Phi$ is, in terms of the number of evaluations of the $\overline{\mathcal{F}}_n$ required to evaluate $\Phi(\overline{\mathcal{F}}_n)(\varepsilon; \overline{x}_m)$, and the accuracy achieved (i.e., the size of $(\Phi(\overline{\mathcal{F}}_n)(\varepsilon; \overline{x}_m))_2)$. For example, let $\Phi_{bad}(\overline{\mathcal{F}}_n) \equiv (\omega, \omega, \omega)$ and $Q_{bad}(\overline{\mathcal{F}}_n, \overline{f}_n, \overline{P}_n) \equiv \{\}$, the null set. Then $(\Phi_{bad}, Q_{bad}) \approx \emptyset(S)$ for any operator $\emptyset$ over $S$ and its corresponding $S$. Of course this is not a good $\varepsilon$-operator in any sense. The formalization of a measure of the goodness of $(\Phi, Q)$ is a worthwhile and as yet unsolved problem. When this is satisfactorily solved, the rules for $\varepsilon$-izing a notion will be complete.

For simplicity, when we present particular $\varepsilon$-operators, we will give a constructive analytic definition of $\Phi$, rather than giving $\Psi_1$ and $\Psi_2$. It is a simple but tedious task to construct particular $\Psi_1$ and $\Psi_2$ from such a definition.

## 2.8 Rounding Subroutines

**Roundup and rounddown ε-arithmetic subroutines**, $A_{1,*}$ and $A_{2,*}$, for $*$ being $+, -, \times, \div$, are multiple-precision subroutines which essentially give upper and lower bounds on the ideal arithmetic operations, $+, -, \times, \div$. We make this precise as follows. For any $a \in R$, define **an ε-neighborhood of $a$, $N_\varepsilon(a)$,** by

$$(2.8\text{-}1) \qquad N_\varepsilon(a) = \{a_1, a_2, a_3, a_4\} \quad ,$$

where $[a_1$ and $a_2$ are the largest members of $\Re(\varepsilon)$ satisfying $a_1 < a_2 \leq a$, or, if this is not possible, then $a_1 = a_2 \leq a]$, and $[a_3$ and $a_4$ are the smallest members of $\Re(\varepsilon) - \{\omega\}$ satisfying $a \leq a_3 < a_4$ or, if this is not possible, then $a \leq a_3 = a_4]$. The fact that $-\infty$ and $\infty$ are in $\Re(\varepsilon)$ means that $N_\varepsilon(a)$ is well defined and nonempty for all $a \in R$. For $a \in \{-\infty, \infty, \omega\}$, define

$$N_\varepsilon(a) = \{a\} \quad .$$

For each $\varepsilon$, we require the $A_{n,*}$ to satisfy

(1) $RX_i(\varepsilon) \neq \omega\,(i=1,2) \Rightarrow A_{n,*}(\varepsilon; \overline{X}_2) = A_{n,*}(\varepsilon; \overline{X_2(\varepsilon)})$, and

(2) for any $a, b \in \Re(\varepsilon)$ we have $c_{n,*} = A_{n,*}(\varepsilon; a,b)$ in $N_\varepsilon(a*b)$ and $c_{1,*} \leq a \leq c_{2,*}$ .

Condition (1) states that the $A_{n,*}$ do not use the imputed roundoff-error bounds. Condition (2) requires $A_{1,*}$ and $A_{2,*}$ to give such upper and lower approximations to $a * b$. It is easy to verify that such subroutines exist. For example, subroutines for $A_{1,+}$, [illegible] $A_{1,-}$

and $A_{2,-}$ can be defined as follows. (The $\hat{I}(\cdot)$ and $\check{I}(\cdot)$ defined here will be used later.)

EXAMPLE 2.8-1: Define $\varepsilon$-precision roundup and rounddown converted values, $\hat{I}(\varepsilon,\beta)$ and $\check{I}(\varepsilon,\beta)$, for the number $<\beta>$ as followc. Let $n_0$ be the least integer such that the interval $[(\beta(n_0)-1)/n_0, (\beta(n_0)+1)/n_0]$ overlaps at most one of the intervals $[(\alpha_{\mathbf{R}}(i, \ell, n_0)-1)/n_0, (\alpha_{\mathbf{R}}(i, \ell, n_0)+1)/n_0](1 \leq \ell \leq \ell_{\mathbf{R}}(i))$. If $\beta(n_0)$ is $-\infty$, $\infty$ or $\omega$ then let $\hat{\ell}$ and $\check{\ell}$ be 1, 2 or 3, respectively. Otherwise let the intervals about $\alpha_{\mathbf{R}}(i, \hat{\ell}, n)$ and $\alpha_{\mathbf{R}}(i, \check{\ell}, n_0)$ be the first ones lying completely to the right and left, respectively, of the one about $\beta(n_0)$. Define

$$(2.8\text{-}2) \quad \overset{*}{I}(\varepsilon_i, \beta) = <\alpha_{\mathbf{R}}(i, \overset{*}{\ell}, \cdot)> \quad \text{for } * \text{ being } \frown \text{ and } \smile \ .$$

Then $\overset{*}{I}(\varepsilon, \beta) \in N_{\varepsilon}(<\beta>)$ and $\check{I}(\varepsilon, \beta) \leq (<\beta>) \leq \hat{I}(\varepsilon, \beta)$ .
Define $\beta^{\pm}_{i,j,k}$ by

$$\beta^{\pm}_{i,j,k}(n) = \alpha_{\mathbf{R}}(i, j, 2n) \pm \alpha_{\mathbf{R}}(i, k, 2n) \quad \text{for } n \geq 1 \ .$$

Then $\beta^{\pm}_{i,j,k}$ computes $<\alpha_{\mathbf{R}}(i, j, \cdot)> \pm <\alpha_{\mathbf{R}}(i, k, \cdot)>$. (See Bishop [B1, pp. 16, 21] for similar definitions cf $+$, $-$, $\times$ and $\div$.) Let $a = <\alpha_{\mathbf{R}}(i, j, \cdot)>$ and $b = <\alpha_{\mathbf{R}}(i, k, \cdot)>$ . We can define the $A_{n,\pm}$ by

$$(2.8\text{-}3) \quad A_{1,\pm}(\varepsilon_i; a, b) = \hat{I}(\varepsilon_i, \beta^{\pm}_{i,j,k}), \quad A_{2,\pm}(\varepsilon_i; a, b) = \check{I}(\varepsilon_i, \beta^{\pm}_{i,j,k}) \ .$$

For the rest of the paper, we assume particular $A_{n,*}$ to be given.
For $\varepsilon$, $b \in \mathbf{R}(\varepsilon)$ we will use $a *_{\varepsilon} b$ and $a *_{\varepsilon} b$ to denote $A_{1,*}(\varepsilon; a,b)$

and $A_{2,*}(\varepsilon; a,b)$, respectively, omitting the subscript $\varepsilon$ whenever no confusion can arise. For subroutines F and G, we will abbreviate $F(\varepsilon; \bar{x}_m) \,\widehat{+}\, G(\varepsilon; \bar{x}_m)$ by $(F \,\widehat{+}\, G)(\varepsilon; \bar{x}_m)$, etc. In general, we will factor out arguments as much as possible, calling the resulting form *argument factored form*.

In order to prove that the $A_{n,*}$ converge to ideal arithmetic as $\varepsilon \to 0$, we must first state explicitly the special rules for arithmetic involving $\pm\infty$ and $\omega$. Let $x,y,z \in \widetilde{R}$ satisfy $-\infty < y < \infty$ and $0 < z \leq \infty$. The special rules are

*General*: $x * \omega = \omega * x = \omega$ .

*Addition*: $\infty + (-\infty) = \omega$ ; $\infty + \infty = \infty + y = \infty$ .

*Multiplication*: $\infty \times 0 = \omega$ ; $\infty \times z = \infty$ .

*Division*: $x \div 0 = \infty \div \infty = \omega$; $y \div \infty = 0$ .

These, combined with the usual definition of real arithmetic (see, for example, Bishop [Bi, pp. 16, 21]) and the usual associative, commutative and distributive laws, completely define the arithmetic of $\widetilde{\mathbf{R}}$ . For example, $\infty - \infty = \infty + (-\infty) = \omega$ and $\infty \times (-10) = (-1) \times (\infty \times 10) = -\infty$ .

THEOREM 2.8-1: *Suppose* $\lim_{\varepsilon \to 0} X_i(\varepsilon) = x_i (i = 1,2)$ *and* $x_1 * x_2 \neq \omega$ . *Then* $X_1(\varepsilon) \,\widehat{*}\, X_2(\varepsilon)$ *and* $X_1(\varepsilon) \underset{\smile}{*} X_2(\varepsilon)$ *approach* $x_1 * x_2$ *as* $\varepsilon \to 0$ .

*Proof*: If $x_1, x_2 \in R$ then $x_1 * x_2 \neq \omega$ implies that $x_1 * x_2$ is finite. We have that $X_1(\varepsilon) \,\widehat{*}\, X_2(\varepsilon)$ and $X_1(\varepsilon) \underset{\smile}{*} X_2(\varepsilon)$ are in $N_\varepsilon(X_1(\varepsilon) * X_2(\varepsilon))$ and property II of $\mathbf{R}$ together with the continuity of arithmetic give us convergence. If $x_1$ or $x_2$ is infinite then

$x_1 * x_2 \neq \omega$ and the above special rules for arithmetic involving $\pm\infty$ yield that $|x_1 * x_2|$ is one of

$$\infty + \infty,\ \infty + y,\ \infty \times z,\ y \div \infty \quad .$$

Convergence is clear in these cases. This completes the proof.

For $A, B \in \mathfrak{R}(\varepsilon)$, let

$$|A \overset{\frown}{-}_{\varepsilon} B| = \begin{cases} 0 & \text{if } A = B \\ A \overset{\frown}{-}_{\varepsilon} B & \text{if } A \geq B \\ B \overset{\frown}{-}_{\varepsilon} A & \text{otherwise,} \end{cases} \tag{2.8-4}$$

$$|A \underset{\smile}{-}_{\varepsilon} B| = \begin{cases} 0 & \text{if } A = B \\ \max(0, A \underset{\smile}{-}_{\varepsilon} B) & \text{if } A \geq B \\ \max(0, B \underset{\smile}{-}_{\varepsilon} A) & \text{otherwise} \quad . \end{cases} \tag{2.8-5}$$

This simplifies inequalities, because $|A \overset{\frown}{-}_{\varepsilon} B|$ and $|A \underset{\smile}{-}_{\varepsilon} B|$ are effective upper and lower bounds on the distance, $|A - B|$, between $A$ and $B$ (see sec. 1.4).

REMARKS: In our model, we have assumed that it is possible to use arbitrary levels of precision (arbitrarily small $\varepsilon$). But in practice, we almost always use single- or double-precision, and there is a finite upper limit on how high the precision can be (precisions higher than double-precision being provided via software). However, our model does not preclude an emphasis on single- and double-precision computations. We feel it is conceptually correct to keep arbitrary precision in mind in the design and analysis of algorithms; doing so helps keep algorithms machine independent and is kind to the occasional user who requires high accuracy.

In our definition of roundoff-error bounds and truncation-error bounds (sec. 2.3 and 2.5), we have taken the stand that numerical analysis should concern itself with rigorous approximation rather than just estimation. However, it should be possible to form an "$\varepsilon$-calculus of estimation" by defining these bounds to be statistical quantities. In fact, it should be possible to form an "$\varepsilon$-calculus of stable $\varepsilon$-functions" which involves no error bounds, as we indicate in the remarks after chapters 4 and 5. (Such an $\varepsilon$-calculus would not have very interesting $\varepsilon$-notions of $\varepsilon$-comparison, $\varepsilon$-convergence and $\varepsilon$-continuity.) The last two $\varepsilon$-calculi should be interesting to explore. The last one will probably resemble current scientific computation more closely than the $\varepsilon$-calculus developed in this thesis. In the "$\varepsilon$-calculus of stable $\varepsilon$-functions", a "poor real input" $x$ would be a mapping $X\colon \mathcal{E} \to \mathcal{M}$ such that $X(\varepsilon) \in \mathcal{R}(\varepsilon)$. A "real input" $x$ with value $c \in \tilde{R}$ would be a "poor real input" such that for every $\delta > 0$ there is an $\varepsilon$ with $|X(\varepsilon) - c| < \delta$. A "subroutine" would essentially remain as before. And "an $\mathcal{F} \approx f(P)$" would be a "subroutine" $F$ with $\lim_{\varepsilon \to 0} F(\varepsilon; \bar{x}_m) = f(\bar{x}_m)$ for all $\bar{x}_m \in P$ at which $f(\bar{x}_m) \neq \omega$.

## 2.A Appendix: Maximum Relative Error

Here we prove statements made about the examples of section 2.2

THEOREM 2.A-1: *Assume that, for each* $\varepsilon$, $\mathcal{R}(\varepsilon) - \{\omega\}$ *is symmetric about* $0$ . *Let* $\sigma(\varepsilon)$ *and* $\tau(\varepsilon)$ *be the smallest and largest finite positive numbers in* $\mathcal{R}(\varepsilon)$ *and define*

$$(2.A\text{-}1) \qquad E(\varepsilon) = \sup_{\sigma(\varepsilon) \le |x| \le \tau(\varepsilon)} \ \min_{y \in \mathcal{R}(\varepsilon)} \left|\frac{x-y}{x}\right| .$$

*Letting* $a < a'$ *range over positive finite neighbors in* $\mathcal{R}(\varepsilon)$ *yields*

$$(2.A\text{-}2) \qquad E(\varepsilon) = \max_a (a'-a)/(a'+a) .$$

Proof: By symmetry it suffices to consider only $x$ with $\sigma(\varepsilon) \le x \le \tau(\varepsilon)$ . For $a \in \mathcal{R}(\varepsilon)$ with $\sigma(\varepsilon) \le a < \tau(\varepsilon)$, let $a'$ denote the successor of $a$ in $\mathcal{R}(\varepsilon)$ . We have

$$E(\varepsilon) = \max_a \left( \sup_{a \le x \le a'} \ \min_{y \in \mathcal{R}(\varepsilon)} \left|\frac{x-y}{x}\right| \right)$$

$$= \max_a \ \sup_{a \le x \le a'} \ \min\left(\frac{x-a}{x}, \frac{a'-x}{x}\right) .$$

The facts that $(x-a)/x$ is monotone increasing and $(a'-x)/x$ is monotone decreasing for $x \neq 0$, and that these functions intersect at $(a + a')/2$ yield (2.A-2). This completes the proof.

This theorem gives immediate results for examples 2.2-1 and 2.2-2. For $a \in \mathcal{R}^{\odot}(\varepsilon_i^{\odot})$ with $a = (b\beta^{-i})\beta^e$ and $\beta^{i-1} \le b < \beta^i$ we have

$a'-a = \beta^{e-i}$ and $\beta^{e-i}/(a'+a)$ is smallest for such $a$ when $a = \beta^{-1}\beta^{e}$, yielding $a'+a = \beta^{-1}\beta^{e} + (\beta^{-1} + \beta^{-i})\beta^{e} = 2\beta^{e-1} + \beta^{e-i}$ and

$$E^{\odot}(\varepsilon_i^{\odot}) = \max_{|e| < \beta^{i}} \beta^{e-i}/(2\beta^{e-1} + \beta^{e-i}) = \varepsilon_i^{\odot} .$$

For $a \in R^{*}(\varepsilon_i^{*})$ with $a = \beta_i^{e}$ we have $a' = \beta_i^{e+1}$ and

$$E^{*}(\varepsilon_i^{*}) = \max_{|e| < 10^{2i}} (\beta_i^{e+1} - \beta_i^{e})/(\beta_i^{e+1} + \beta_i^{e}) = \varepsilon_i^{*} .$$

To prove this for $R^{\#}$ we need more machinery. First we note that it suffices to take the maximum in (2.A-2) over $\Theta(\varepsilon) \leq a < 1$ (rather than $\Theta(\varepsilon) \leq a < \tau(\varepsilon))$ for $R^{\#}$: suppose $a = p/q \geq 1$ and $a' = p'/q'$; then the successor of $b = q'/p'$ is $b' = q/p$ and we have

$$\frac{b'-b}{b'+b} = \frac{1/a - 1/a'}{1/a + 1/a'} = \frac{a'-a}{a'+a} .$$

For $n = 1, 2, \ldots$ define the Farey series of order $n$, $F_n$, to be the sequence of rationals, $p/q$, with $0 \leq p \leq q \leq n$ and G.C.D.$(p,q) = 1$, written in increasing order. We shall require the following two well known lemmas (see Niven and Zuckerman [NZ, pp. 128-130]).

LEMMA 2.A-1: *If $p/q$ and $p'/q'$ are consecutive fractions in $F_n$ then* $p'q - pq' = 1$ .

LEMMA 2.A-2: *If $p/q$ and $p'/q'$ are consecutive in $F_n$, then among all rationals with value between these two,* $(p+p')/(q+q')$ *is the unique one with smallest denominator.*

We will also need the following two new results.

LEMMA 2.A-3: For $n \geq 2$, if $p/q < p'/q'$ are consecutive in $F_n$ and $p > 0$, then $pq' \geq [\frac{1}{2}(n+1)]$ .

Proof: $F_2$ is $< 0, \frac{1}{2}, \frac{1}{1} >$, so the theorem is true for $F_2$ by inspection. Suppose it is true for $F_{n-1}$. Any consecutive fractions in $F_n$ will be either $\frac{p}{q}, \frac{p'}{q'}$ or $\frac{p}{q}, \frac{p+p'}{q+q'}$ or $\frac{p+p'}{q+q'}, \frac{p'}{q'}$ where $\frac{p}{q}, \frac{p'}{q'}$ are consecutive fractions in $F_{n-1}$ (by lemma 2.A-2). We have $pq' \geq [n/2]$ . In the last two cases this implies

$$p(q+q') = pq + pq' \geq 1 + [n/2] \geq [\frac{1}{2}(n+1)] \quad ,$$

$$(p+p')q' = pq' + p'q' \geq [n/2] + 1 \geq [\frac{1}{2}(n+1)] \quad ,$$

and the induction step follows. In the first case, if $n$ is even then $[n/2] = [\frac{1}{2}(n+1)]$ so $pq' \geq [\frac{1}{2}(n+1)]$ . Or, if $pq' > [n/2]$ then $pq' \geq [\frac{1}{2}(n+1)]$ . Suppose $n$ is odd and $pq' = [n/2]$ . Then we have

$$q' \leq pq' = (n-1)/2 \quad ,$$

$$q \leq p'q = pq' + 1 = (n-1)/2 + 1 \quad ,$$

$$q + q' \leq n \quad .$$

It follows that $(p+p')/(q+q')$ is in $F_n$ and so $p/q$, $p'/q'$ could not have been consecutive in $F_n$ . This completes the proof.

LEMMA 2.A-4: *Let* $n$ *be* $\geq 2$ *and let* $a < a'$ *run through all consecutive fractions in* $F_n$ *with* $a > 0$ . *Then we have*

$$\max_a \frac{a'-a}{a'+a} = \frac{1}{2[\frac{1}{2}(n+1)] + 1} .$$

*Proof*: Let $a = p/q$ and $a' = p'/q'$ . We have

$$\frac{p'/q' - p/q}{p'/q' + p/q} = \frac{p'q - pq'}{p'q + pq'} = \frac{1}{2pq' + 1} \leq \frac{1}{2[\frac{1}{2}(n+1)] + 1} .$$

Further, the fractions $1/([\frac{1}{2}(n+1)] + 1)$, $1/[\frac{1}{2}(n+1)]$ are consecutive in $F_n$ because $([\frac{1}{2}(n+1)] + 1) + [\frac{1}{2}(n+1)] > n$, and for these we have

$$\frac{1}{2pq' + 1} = \frac{1}{2[\frac{1}{2}(n+1)] + 1} .$$

This completes the proof.

Taking $n = 10^4 - 1$ yields

$$E^{\#}(\varepsilon_1^{\#}) = \max_{0 \neq a \in F_n} (a'-a)/(a'+a) = \varepsilon_1^{\#} ,$$

completing our task.

BLANK PAGE

# Chapter 3. Numerical Instability

## 3.1 A Definition

To simplify notation, we restrict this discussion to $\varepsilon$-functions and functions of two variables. Let $x$ be a real input and suppose that $\lim_{y \to x} f(x, y)$ exists. Let $F$ be a subroutine satisfying

$$(3.1\text{-}1) \qquad \lim_{\varepsilon \to 0} F(\varepsilon;\, x,\, y) = f(x,\, y) \ ,$$

for $y \in \mathfrak{M} - \{x\}$ at which $f(x, y) \neq \omega$. This means that, with $\mathfrak{F} \equiv (F, \omega, \omega)$, we have $\mathfrak{F} \sim f(\{x\} \times (\mathfrak{M} - \{x\}))$. We do not require (3.1-1) for $y = x$ because (1) we do not need to, and (2) this could cause problems when $f(x, y)$ is discontinuous at $y = x$ ($x$ fixed).

We are interested here in a computation of $\lim_{y \to x} f(x, y)$ which proceeds at precision $\varepsilon$ by selecting a $y_\varepsilon \in \mathfrak{R}(\varepsilon)$ and then using $F(\varepsilon;\, x,\, y_\varepsilon)$ to approximate $\lim_{y \to x} f(x, y)$. We call the rule used to select these $y_\varepsilon$'s (as a function of $F$, $x$ and possibly other things) <u>a stopping criterion</u> because it tells us where to stop at precision $\varepsilon$. We say that this stopping criterion <u>works at $x$</u> precisely when

$$(3.1\text{-}2) \qquad \lim_{\varepsilon \to 0} F(\varepsilon;\, x,\, y_\varepsilon) = \lim_{y \to x} f(x,\, y)$$

This framework is quite general. $f(\infty, n)$ might be the $n^{th}$ iterate of an iterative procedure for evaluating $\lim_{n \to \infty} f(\infty, n)$, as in Newton's or Bernoulli's method for finding zeros of a polynomial, or as in numerical integration methods for ordinary differential equations. The assumption that $\lim_{n \to \infty} f(\infty, n)$ exists means that the discrete method converges in exact arithmetic, with exact starting values. (This is weaker than "convergence" as defined in Ralston [R1, p. 171].

We are now ready to discuss numerical stability. As used in numerical analysis, stability deals with the way local rounding errors of some iterative procedure propogate and effect the total accumulated error. (See Henrici [H1, pp. 11, 302, 309] and Ralston [R1, p. 175 (under 1). Let us consider an example.

<u>EXAMPLE 3.1-1:</u> Let $q_1, q_2, \ldots$ be defined by

$$q_n = Q_n(q_0, q_1, \ldots, q_{n-1}) .$$

At precision $\varepsilon$, let $\tilde{q}_0$ approximate $q_0$, $\tilde{Q}_i$ approximate the $i^{th}$ recurrence relation, $Q_i$, and define $\tilde{q}_1, \tilde{q}_2, \ldots,$ by

$$\tilde{q}_n = \tilde{Q}_n(\tilde{q}_0, \tilde{q}_1, \ldots, \tilde{q}_{n-1}) .$$

The $n^{th}$ local rounding error of this iterative procedure is

$$\tilde{q}_n - Q_n(\tilde{q}_0, \tilde{q}_1, \ldots, \tilde{q}_{n-1}) ,$$

and the total accumulated error is $\tilde{q}_n - q_n$. Let $f(x, y) = q_n$ when $n = [y] \geq 0$ and $x \neq \omega$; otherwise let $f(x, y) = \omega$. Suppose we are interested in the finite limit, $\lim_{y \to \infty} f(\infty, y) = \lim_{n \to \infty} q_n$. Let $F(\varepsilon; x, y)$ be defined in terms of the $\tilde{q}_n$ so that (3.1-1) is satisfied for $x = \infty$. (This is easy, but tedious, to do; F will be effective so long as the $\tilde{Q}_n$ are.) Then $F(\varepsilon; \infty, y) - f(\infty, y)$ is the total accumulated error. If, as $y \to \infty$ through finite values in $R(\varepsilon)$, $|F(\varepsilon; \infty, y) - f(\infty, y)|$ becomes large, it would be said that "numerical instability has set in at precision $\varepsilon$."

If this happens for infinitely many values of $\varepsilon$, it would be said that $F$ is unstable at $\infty$. Suppose this happens, and let $y_\varepsilon$ be some value in $\mathcal{R}(\varepsilon)$ where the total error has become large. Then we would have $\lim_{\varepsilon \to 0} F(\varepsilon; \infty, y_\varepsilon) \neq \lim_{y \to \infty} f(\infty, y)$, even though we may have $\lim_{\varepsilon \to 0} y_\varepsilon = \infty$ and $y_\varepsilon \neq \infty$. This is the tragedy of numerical instability; *when $F$ is unstable at $\infty$, there will be seemingly quite reasonable stopping criteria that do not work at $\infty$*. On the other hand, if $F$ is stable at $\infty$ (in the sense usual to numerical analysis), then any such reasonable stopping criteria should work at $\infty$.

This idea of stability generalizes easily to any $F$, $x$ and $f$ satisfying the assumptions at the beginning of this section, whether or not they involve iterative methods and local rounding errors. In this generalization, it is important that "reasonable" stopping criteria choose $y_\varepsilon$'s that satisfy

$$|X(\varepsilon) - y_\varepsilon| > RX(\varepsilon)$$

so that $y_\varepsilon$ is effectively distinct from $x$ at precision $\varepsilon$. This is necessary so that $F$ is not unstable just because $f(x, y)$ is discontinuous at $y = x$, e.g., when $f(x, y)$ involves divisions by $(x-y)$. Define *the set, $\rho(\varepsilon; x)$, of members of $\mathcal{R}(\varepsilon)$ that are effectively distinct from $x$ at precision $\varepsilon$* by

$$(3.1\text{-}3) \qquad \rho(\varepsilon; x) \equiv \{Y : Y \in \mathcal{R}(\varepsilon) \text{ and } |X(\varepsilon) - Y| > RX(\varepsilon)\}$$

*DEFINITION 3.1-1. We say a stopping criterion is reasonable at $x$ precisely when its $y_\varepsilon$'s satisfy*

(3.1-4) $$y_\varepsilon \in \rho(\varepsilon; x) \cup \{\omega\} \ , \ \lim_{\varepsilon \to 0} y_\varepsilon = x \ .$$

If $RX(\varepsilon) = \omega$ then (3.1-4) forces the choice $y_\varepsilon = \omega$. This cannot happen when $\varepsilon$ is sufficiently small.

*DEFINITION 3.1-2*: *Suppose* $\lim_{y \to x} f(x, y)$ *is finite and* $\mathfrak{F} \sim f(\{x\} \times (\mathfrak{M} - \{x\}))$. *We say* ***F is stable at x*** *precisely when any stopping criterion which is reasonable at* $x$, *works at* $x$. *Otherwise, we say* ***F is unstable at x***.

Following is an example of an $F$ unstable at $0$.

*EXAMPLE 3.1-2:* Let $\Theta(\varepsilon)$ be the smallest positive number in $\mathcal{R}(\varepsilon)$. Suppose a certain form of $\varepsilon$-arithmetic is to be used and that in this $\varepsilon$-arithmetic we have $0 + 1 = 1$, $\Theta(\varepsilon) + 1 = 1$, $1 - 1 = 0$ and $0/(0 - \Theta(\varepsilon)) = 0$ (see section 5.3 for a detailed discussion of $\varepsilon$-arithmetic). Suppose a subroutine $F$, evaluated at $(\varepsilon; x, y)$, approximates $f(x, y) = (x + 1 - (y + 1))/(x - y)$ by replacing $x$ and $y$ by $X(\varepsilon)$ and $Y(\varepsilon)$ and by replacing arithmetic by this $\varepsilon$-arithmetic. (That $F$ satisfies (3.1-1) follows from corollary 5.3-1 in chapter 5.) Define $y_\varepsilon = \Theta(\varepsilon)$. Then these $y_\varepsilon$'s satisfy (3.1-4) for $x$ being $0$, but $F(\varepsilon; 0, y_\varepsilon) = 0$ for all $\varepsilon$, while $\lim_{y \to 0} f(0, y) = \frac{d}{dt}(t+1)\big|_{t=0} = 1$. Hence $F$ is unstable at $0$.

## 3.2 A Geometric Characterization

F is unstable at x if and only if there is a reasonable at x stopping criterion whose $y_\varepsilon$'s satisfy

$$(3.2\text{-}1) \qquad \lim_{\varepsilon \to 0} |F(\varepsilon;\ x,\ y_\varepsilon) - f(x,\ y_\varepsilon)\ | > 0\ .$$

Interpreting this geometrically, we find that the graph of $F(\varepsilon;\ x,\ Y)$ versus finite $Y \in \mathcal{R}(\varepsilon)$ acts like an ε-wave. This is pictured in figure 3.2-1 for $x = \infty$ and $\varepsilon = \varepsilon_1, \varepsilon_2$ .

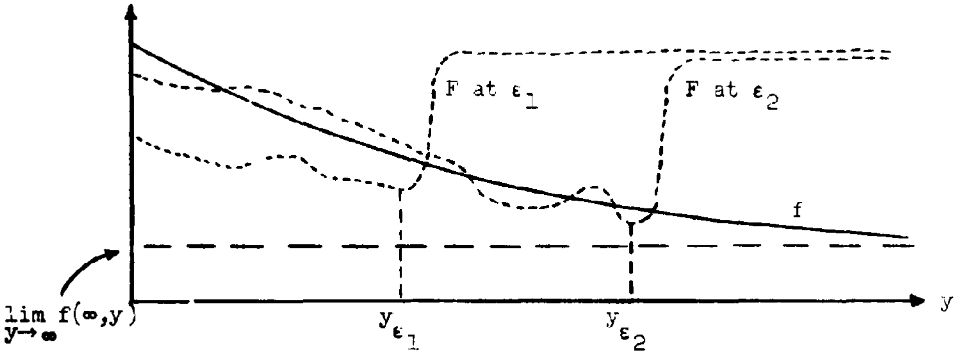


Figure 3.2-1 Instability

As $\varepsilon \to 0$ , the ε-wave moves towards x . The crest of the ε-wave stays uniformly away from $\lim_{y \to x} f(x,\ y)$ . (See example 3.1-2).

Two usual stopping criteria are

(1) choose $y_\varepsilon$ to be the first value of Y for which $TF(\varepsilon;\ x,\ Y) \leq RF(\varepsilon;\ x,\ Y)$ , as $Y \to x$ via some fixed approach, and

(2) choose $y_\varepsilon$ to be the first value of Y (as $Y \to x$ via some fixed approach) such that $F(\varepsilon;\ x,\ Y)$ and the previous four values of F are equal, within some tolerance.

The trouble with such stopping criteria is that they can make $F(\varepsilon;\ x,\ y_\varepsilon)$ ride the crest of the ε-wave out to x, thereby destroying convergence.

## 3.3 A Stopping Criterion That Works

The question arises, is there any stopping criterion (effectively computable or not) which yields convergence even when $F$ is unstable at $x$ ? The answer is given in the affirmative by the following definition and theorem.

DEFINITION 3.3-1: *Stopping criterion* $S.\,C.^{\#}$ *selects* $y_\varepsilon$ *to be that value of* $Y \in \mathcal{R}(\varepsilon)$ *closest to* $x$ *at which* $|F(\varepsilon;\, x,\, Y) - \lim_{y \to x} f(x,\, y)|$ *assumes its minimum over all* $Y \in \mathcal{R}(\varepsilon)$ *(taking the smaller value for* $y_\varepsilon$ *in case of a tie).*

We call $(y_\varepsilon,\, F(\varepsilon;\, x,\, y_\varepsilon))$ *the base of the* $\varepsilon$*-wave of* $F$ *at* $x$ . Of course $y_\varepsilon$ cannot be effectively computed from $F$ and $x$ .

THEOREM 3.3-1: *If* $\mathfrak{F} \sim f(\{x\} \times (\mathfrak{M} - \{x\}))$ *and* $\lim_{y \to x} f(x,\, y)$ *is finite, then* $S.\,C.^{\#}$ *works at* $x$ .

Note that $S.\,C.^{\#}$ works whether $F$ is stable or not. Thus there is some desirable behavior even in the presence of instability.

Proof: Let $\ell = \lim_{y \to x} f(x,\, y)$ . For any real input $y$ , we have

$$(3.3\text{-}1)\quad |F(\varepsilon;\, x,\, y) - \ell| \le |F(\varepsilon;\, x,\, y) - f(x,\, y)| + |f(x,\, y) - \ell| .$$

Let an $\eta > 0$ be given. By choosing $y$ sufficiently close to $x$ , keeping $y \in \mathfrak{M} - \{x\}$ , the second term on the right side of (3.3-1) becomes $< \eta/2$ . Then by making $\varepsilon$ sufficiently small, this value or $y$ is in $\mathcal{R}(\varepsilon)$ and the first term on the right side of (3.3-1) becomes $< \eta/2$ (the nesting property III of $\mathcal{R}(\varepsilon)$ allows this). For such $y$ and $\varepsilon$ , the left side of (3.3-1) is $< \eta$ . The left side bounds the distance of the $F(\varepsilon; x, y_{\varepsilon})$ of $S. C.^{\#}$ from $\ell$, and so $S. C.^{\#}$ works. This completes the proof.

Thus the height of the base of the $\varepsilon$-wave of $F$ an $x$ approaches $\lim_{y \to x} f(x, y)$ as $\varepsilon \to 0$.

## 3.4 An Algorithm for Overcoming Instability

We derive an effective analog to S. C.$^{\#}$ as follows. Suppose $\mathfrak{F} \approx f(\{x\} \times (\mathcal{M} - \{x\}))$ and $\lim_{y \to x} f(x, y)$ exists.

DEFINITION 3.4-1: S. C$^{\#\#}$ selects $y_\varepsilon$ to be the smallest value of $Y$ in $\mathbf{R}(\varepsilon)$ for which $(RF \hat{+} TF)(\varepsilon; x, Y)$ assumes its minimum over all $Y$ in $\mathbf{R}(\varepsilon)$ .

This is finitely computable because $\mathbf{R}(\varepsilon)$ is a finite set. This stopping criterion keeps us close enough to the base of the $\varepsilon$-wave of $F$ at $x$ that we get convergence even in the unstable case, provided only that

there is a sequence, $y_1, y_2, \ldots,$ with each $y_i \in \mathcal{M} - \{x\}$ ,

(3.4-1) such that

$$\lim_{i \to \infty} \limsup_{\varepsilon \to 0} TF(\varepsilon; x, y_i) = 0 .$$

THEOREM 3.4-1: Suppose that

(1) $\lim_{y \to x} f(x, y) \neq \omega$ ,

(2) $\mathfrak{F} \approx f(\{x\} \times \mathcal{M} - \{x\}))$ , and

(3) TF satisfies (3.4-1).

Then S. C.$^{\#\#}$ works at $x$ .

Proof: This proof in essentially the same as that of theorem 3.3-1. We will prove that for every $\eta > 0$ there is a $\delta > 0$ with

$$(RF \hat{+} TF)(\varepsilon; x, y_\varepsilon) < \eta \qquad \text{for all } \varepsilon \leq \delta ,$$

where $y_\varepsilon$ is chosen by S. C.$^{\#\#}$

Let an $\eta > 0$ be given. By assumption, there is a $Y$ in some $\mathbf{R}(\delta_1) - \{x\}$ with

$$\limsup_{\varepsilon \to 0} TF(\varepsilon; x, Y) < \eta/4$$

For this $Y$ there are $\delta_2, \delta_3 \in \mathcal{E}$ with

$$TF(\varepsilon; x, Y) < \eta/2 \qquad \text{for all } \varepsilon \leq \delta_2 ,$$
$$RF(\varepsilon; x, Y) < \eta/4 \qquad \text{for all } \varepsilon \leq \delta_3 .$$

There is a $\delta_4$ such that $\mathcal{R}(\delta_4)$ contains an $\eta_1$ and an $\eta_2$ with $3\eta/4 \leq \eta_1 < \eta_2 < \eta$ . Let $\delta = \min(\delta_1, \delta_2, \delta_3, \delta_4)$ . We have

$$(RF \overset{\frown}{+} TF)\ (\varepsilon; x, y_\varepsilon) \leq (RF \overset{\frown}{+} TF)\ (\varepsilon; x, Y) < \eta \quad \text{for all } \varepsilon \leq \delta$$

The first inequality makes use of the nesting property III of $\mathcal{R}$ . The second inequality uses the facts that

(1) $(RF \overset{\frown}{+} TF)\ (\varepsilon; x, Y) \in N_\varepsilon\ ((RF + TF)\ (\varepsilon; x, Y))$ ,

(2) $(RF + TF)\ (\varepsilon; x, Y) < 3\ \eta/4$ , and

(3) $\eta_1, \eta_2 \in \mathcal{R}(\varepsilon)$ (by property III of $\mathcal{R}$ ) .

This completes the proof.

Thus S. C.## is a (totally inefficient) algorithm for overcoming instability. It should be possible to find a more efficient algorithm for which theorem 3.4-1 holds because

(1) we do not need to find the exact minimum of $(RF \overset{\frown}{+} TF)$ $(\varepsilon; x, Y)$ over $Y \in \mathcal{R}(\varepsilon)$ ; we only need to stay "sufficiently close" to it, and

(2) in particular cases, it should be possible to localize the search for $y_\varepsilon$ .

On the other hand, it should be possible to show that any stopping criterion, for which theorem 3.4-1 holds, must require so many evaluations of $RF + TF$ to choose $y_{\varepsilon}$ that it cannot be very efficient.

In chapter 4, we present an efficient algorithm that almost satisfies this theorem (it has a somewhat more stringent third hypothesis, regarding $TF$).

## 3.5 Applications

We have proven that numerical instability is not an essential limitation of finite computations. Convergent roundoff-error and truncation-error bounds can be combined with an unstable subroutine to form a convergent algorithm for computing the associated limit. Here, we consider applications of this result to the initial-value problem for ordinary differential equations Instabilities can generally be classified as

(I) those due to the particular method of solution used, and

(II) those due to the problem being solved.

We will give an example of each and we will show that the instabilities in both of these examples can be overcome by S. C.## . Of course, the best way to overcome instabilities of type I is to find a stable method of solution.

EXAMPLE 3.5-1: Consider solving the initial-value problem $y' = -y$ , $y(0) = 1$ , by the corrector formula,

$$y_{n+1} = y_{n-1} - \frac{h}{3}(y_{n+1} + 4y_n + y_{n-1}) , \tag{3.5-1}$$

from Milne's method (see [R1, p. 182]). This is a well-known unstable formula. Since $y_n$ depends on $h$ , let us write $y_n(h)$ . Taking $y_0(h) = 1$ and $y_1(h) = e^{-h}$ , we find the solution of the above difference equation to be

$$y_n(h) = A(h)\; r_+(h)^n + B(h)\; r_-(h)^n ,$$

where

$$r_{\pm}(h) = \frac{-2h + \sqrt{3(h^2 + 3)}}{h + 3}$$

$$A(h) = \frac{e^{-h}(h + 3) + 2h}{2\sqrt{3(h^2 + 3)}} + \frac{1}{2} = 1 - B(h) \ .$$

For fixed $h > 0$ , we have $|r_+(h)| < 1 < |r_-(h)|$ and $B(h) \neq 0$ , so that $|y_n(h)| \to \infty$ as $n \to \infty$ , whereas $y(nh) \to 0$ . However, for any finite $x$ we have

$$(3.5\text{-}2) \qquad \lim_{n \to \infty} y_n(\tfrac{x}{n}) = e^{-x} \ .$$

Let $f_x(\infty, y) = y_n(\frac{x}{n})$ when $n = [y] \geq 1$ , and let it be $\omega$ otherwise. Let $F_x(\varepsilon; \infty, y)$ approximate $f_x(\infty, y)$ by evaluating (3.5-1) in some form of $\varepsilon$-arithmetic (see sec. 5.3), where the approximations used for the initial values converge to the correct values, $y_0(\frac{x}{n}) = 1$ and $y_1(\frac{x}{n}) = e^{-x/n}$ , as $\varepsilon \to 0$ . It follows from corollary 5.3-1 that $F_x$ satisfies (3.1-1). $RF_x$ can be defined as in sections 5.3 and 5.6, and $TF_x$ can be defined so that

$$TF_x(\varepsilon; \infty, n) \geq |y_n(\tfrac{x}{n}) - e^{-x}| \ .$$

It follows that S. C.## works when it is applied to $F_x$ .

Because of its extreme lack of efficiency, S. C.## could not be used in practice. But the $\varepsilon$-limit defined in section 4.1 could be applied to this $F_x$ with reasonable efficiency.

__EXAMPLE 3.5-2__: Consider solving the system,

$$(3.5\text{-}3) \qquad \begin{aligned} y' &= z \quad , \quad y(0) = 1, \\ z' &= y \quad , \quad z(0) = -1 \ , \end{aligned}$$

by the Newton-Cotes closed formula,

$$(3.5\text{-}4) \qquad \begin{pmatrix} y_{n+1} \\ z_{n+1} \end{pmatrix} = \begin{pmatrix} y_n \\ z_n \end{pmatrix} + \frac{h}{2} \begin{pmatrix} z_{n+1} + z_n \\ y_{n+1} + h_n \end{pmatrix} .$$

Again, we will write $y_n(h)$ and $z_n(h)$ . The general solution (3.5-3) is $y = Ae^{x} + Be^{-x}$ , and the initial values give $y = e^{-x}$ . The general solution to (3.5-4) for $y_n$ is

$$y_n(h) = A(h)\left(\frac{2-h}{2+h}\right)^n + B(h)\left(\frac{2+h}{2-h}\right)^n .$$

Taking $y_0 = -z_0 = 1$ yields $A(h) = 1$ and $B(h) = 0$ . But rounding errors in computing (3.5-4) in $\varepsilon$-arithmetic will build up so that $|\tilde{y}_n(h)|$ becomes large as $n$ does, even though $B(h)$ should have been zero. We computed $\tilde{y}_n(.05)$ and $\tilde{y}_n(\frac{5}{n})$ for $n = 20, 30, \ldots, 1000$, on an IBM360/65 computer in short and long-precision; $\tilde{y}_n(05)$ is graphed in figure 3.5-1 and all the data is given in tables 3.5-1 and 3.5-2.

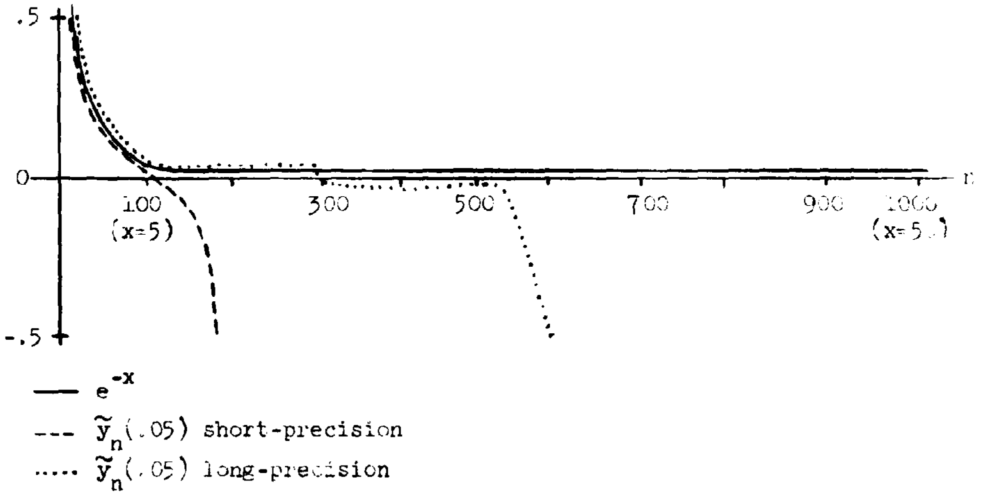


Figure 3.5-1. $\tilde{y}_n(.05)$

We again have (3.5-2). The rest follows as in the last example, except the initial values for $F_x$ are to be $y_0(h) = -z_0(h) = 1$ .

It is possible to construct examples where $\lim_{n \to \infty} y_n(\frac{x}{n}) = y(x)$ does not hold, even when the initial values are assumed exact. The methods of this chapter cannot be used to overcome such instabilities.

TABLE 3.5-1

Data from example 3.5-2

|  | $\tilde{y}_n(.05)$ |  |  |  |
|---|---|---|---|---|
| n | double-precision | single-precision | $y_n(.05)$ |  |
| 20 | 3.67802778856628'-01 | 3.677294'-01 | 3.67802778856713'-01 | 3.67[illegible] |
| 30 | 2.23060416715512'-01 | 2.229016'-01 | 2.23060416715708'-01 | 2.2[illegible] |
| 40 | 1.35278884134338'-01 | 1.349900'-01 | 1.35278884134720'-01 | 1.3[illegible] |
| 50 | 8.20422411202895'-02 | 8.154595'-02 | 8.20422411205736'-02 | 8.20649986272988'-02 |
| 60 | 4.97559495042088'-02 | 4.892212'-02 | 4.97559495053852'-02 | 4.97870683678640'-02 |
| 70 | 3.01753642659534'-02 | 2.878845'-02 | 3.01753642679265'-02 | 3.01973834223185'-02 |
| 80 | 1.83003764894594'-02 | 1.600527'-02 | 1.83003764927349'-02 | 1.83156388887342'-02 |
| 90 | 1.10985828253390'-02 | 7.308926'-03 | 1.10985828307569'-02 | 1.11089965382423'-02 |
| 100 | 6.73092931920775'-03 | 4.787731'-04 | 6.73092932815196'-03 | 6.73794699908547'-03 |
| 110 | 4.08208959176835'-03 | -6.229091'-03 | 4.08208960652379'-03 | 4.08677143846407'-03 |
| 120 | 2.47565448684730'-03 | -1.452738'-02 | 2.47565451118244'-03 | 2.47575217666636'-03 |
| 130 | 1.50140386065233'-03 | -2.653506'-02 | 1.50140390082156'-03 | 1.50343919297757'-03 |
| 140 | 9.10552542531220'-04 | -4.531748'-02 | 9.10552608762057'-04 | 9.11881965554516'-04 |
| 150 | 5.52220417798506'-04 | -7.566941'-02 | 5.52220526908476'-04 | 5.53084370147834'-04 |
| 160 | 3.34903599664287'-04 | -1.253403'-01 | 3.34903779775845'-04 | 3.35462627902512'-04 |
| 170 | 2.03107947683137'-04 | -2.070141'-01 | 2.03108244338655'-04 | 2.03468369010644'-04 |
| 180 | 1.23178051697556'-04 | -3.415465'-01 | 1.23178540851172'-04 | 1.23409804086680'-04 |
| 190 | 7.47025701142081'-05 | -5.632893'-01 | 7.47037766764656'-05 | 7.48518298877006'-05 |
| 200 | 4.53040796854943'-05 | -9.288276'-01 | 4.53054096205762'-05 | 4.53999297624848'-05 |
| 210 | 2.74740637318367'-05 | -1.531482'+00 | 2.74762566526953'-05 | 2.75364493497471'-05 |
| 220 | 1.66598396635851'-05 | -2.525161'+00 | 1.66634555556896'-05 | 1.67017007902456'-05 |
| 230 | 1.00998813299105'-05 | -4.163610'+00 | 1.01058435494416'-05 | 1.01300935986307'-05 |
| 240 | 6.11903419615134'-06 | -6.865200'+00 | 6.12886525873795'-06 | 6.14421235332822'-06 |
| 250 | 3.70074696903402'-06 | -1.131960'+01 | 3.71695734017579'-06 | 3.72665317207868'-06 |
| 260 | 2.22748450449694'-06 | -1.866377'+01 | 2.25421367340218'-06 | 2.26032540698106'-06 |
| 270 | 1.32933696862291'-06 | -3.077290'+01 | 1.36710723860851'-06 | 1.37095908638409'-06 |
| 280 | 7.56433494856405'-07 | -5.073912'+01 | 8.29106053214121'-07 | 8.31528719103573'-07 |
| 290 | 3.82996562343330'-07 | -8.366090'+01 | 5.02825841355337'-07 | 5.04347662367891'-07 |
| 300 | 1.07361840688759'-07 | -1.379449'+02 | 3.04947510339075'-07 | 3.05902320501827'-07 |
| 310 | -1.40856902826300'-07 | -2.274450'+02 | 1.84940741731457'-07 | 1.85539136261594'-07 |
| 320 | -4.25544941809680'-07 | -3.750098'+02 | 1.12160541708148'-07 | 1.12535174719260'-07 |
| 330 | -8.17772688723809'-07 | -6.183263'+02 | 6.80217187326516'-08 | 6.82560337633490'-08 |
| 340 | -1.41932731505055'-06 | -1.019509'+03 | 4.12529589183310'-08 | 4.13993771878519'-08 |
| 350 | -2.38332213795170'-06 | -1.681017'+03 | 2.50185771724790'-08 | 2.51099915574399'-08 |
| 360 | -3.95592340069216'-06 | -2.771725'+03 | 1.51729529262239'-08 | 1.52299797447127'-08 |
| 370 | -6.53871141864886'-06 | -4.570008'+03 | 9.20190223707887'-09 | 9.23744966197063'-09 |
| 380 | -1.07912783474700'-05 | -7.535027'+03 | 5.58065424972725'-09 | 5.60279643753729'-09 |
| 390 | -1.77993992273311'-05 | -1.242391'+04 | 3.38448520253133'-09 | 3.39826781949500'-09 |
| 400 | -2.93528343929222'-05 | -2.048504'+04 | 2.05258014088820'-09 | 2.06115362243957'-09 |
| 410 | -4.84018253436919'-05 | -3.377685'+04 | 1.24482306249044'-09 | 1.25015286638675'-09 |
| 420 | -7.98107327887827'-05 | -5.569184'+04 | 7.54944679644784'-10 | 7.58256042791193'-10 |
| 430 | -1.31600152448349'-04 | -9.182394'+04 | 4.57849381568909'-10 | 4.59905537865233'-10 |
| 440 | -2.16995058747146'-04 | -1.514001'+05 | 2.77470751056442'-10 | 2.78946809286893'-10 |
| 450 | -3.57801941490[illegible]'-04 | -2.496332'+05 | 1.68398274838872'-10 | 1.69189792261514'-10 |
| 460 | -5.89977322981009'-04 | -4.116075'+05 | 1.02128073845790'-10 | 1.02618796317019'-10 |
| 470 | -9.72809635151949'-04 | -6.786766'+05 | 6.19373534404134'-11 | 6.22414462290782'-11 |
| 480 | -1.60405914584733'-03 | -1.118995'+06 | 3.75629893597650'-11 | 3.77513454427912'-11 |
| 490 | -2.64492207409147'-03 | -1.844997'+06 | 2.27807307104144'-11 | 2.28973484564556'-11 |
| 500 | -4.36119374441567'-03 | -3.042077'+06 | 1.38157718686867'-11 | 1.38879438649641'-11 |
| 510 | -7.19114222576428'-03 | -5.015910'+06 | 8.37881605967698'-12 | 8.42346375446869'-12 |
| 520 | -1.18574247043968'-02 | -8.270530'+06 | 5.08147928535336'-12 | 5.10908902906335'-12 |
| 530 | -1.95516256167532'-02 | -1.363651'+07 | 3.08175183027841'-12 | 3.09881912872184'-12 |
| 540 | -3.22385403004472'-02 | -2.248355'+07 | 1.86898220185579'-12 | 1.87952881653909'-12 |
| 550 | -5.31579062023125'-02 | -3.707096'+07 | 1.13347688692316'-12 | 1.13999185304436'-12 |
| 560 | -8.76517039981834'-02 | -6.112374'+07 | 6.87416847476296'-13 | 6.91440010694023'-13 |
| 570 | -1.44528288689162'-01 | -1.007839'+08 | 4.16895948780193'-13 | 4.19379565837956'-13 |
| 580 | -2.38311695934463'-01 | -1.661779'+08 | 2.52833826734703'-13 | 2.54366564737603'-13 |
| 590 | -3.92950507711614'-01 | -2.739960'+08 | 1.53335488455461'-13 | 1.54281120319189'-13 |
| 600 | -6.47933375260[illegible]'-01 | -4.517663'+08 | 9.29929840620002'-14 | 9.35762290884020'-14 |

TABLE 3.5-1 (con't)

Data from example 3.5-2

| | $\tilde{y}_n(.05)$ | | | |
|---|---|---|---|---|
| n | double-precision | single-precision | $y_n(.05)$ | $e^{-.05n}$ |
| 610 | -1.06837286257908'+00 | -7.448829'+08 | 5.63972187512698'-14 | 5.67568523263274'-14 |
| 620 | -1.76163762625151'+00 | -1.228193'+09 | 3.42630779521816'-14 | 3.44247710846999'-14 |
| 630 | -2.90474385729172'+00 | -2.025111'+09 | 2.07430537765069'-14 | 2.08796791164594'-14 |
| 640 | -4.78961206254894'+00 | -3.339055'+09 | 1.25799871162652'-14 | 1.26641655490942'-14 |
| 650 | -7.89755821399810'+00 | -5.505450'+09 | 7.62935282097347'-15 | 7.68120468520213'-15 |
| 660 | -1.30222291344191'+01 | -9.077445'+09 | 4.62695421934397'-15 | 4.65888614510342'-15 |
| 670 | -2.14722635825262'+01 | -1.496717'+10 | 2.80659716843234'-15 | 2.82575728711563'-15 |
| 680 | -3.54054669594812'+01 | -2.467860'+10 | 1.70180661951750'-15 | 1.71390843154202'-15 |
| 690 | -5.83798296719404'+01 | -4.069136'+10 | 1.03209033629137'-15 | 1.03953801167023'-15 |
| 700 | -9.62620975010799'+01 | -6.709189'+10 | 6.25929203735285'-16 | 6.30511676014702'-16 |
| 710 | -1.58725907002114'+02 | -1.106208'+11 | 3.79605693719124'-16 | 3.82424662809715'-16 |
| 720 | -2.61722050606276'+02 | -1.823934'+11 | 2.30218500501407'-16 | 2.31952283024358'-16 |
| 730 | -4.31551679667754'+02 | -3.007370'+11 | 1.39620029019724'-16 | 1.40686171244615'-16 |
| 740 | -7.11582580805257'+02 | -4.958709'+11 | 8.46750042286430'-17 | 8.53304762574410'-17 |
| 750 | -1.17332359752443'+03 | -8.176138'+11 | 5.13526346575092'-17 | 5.17555500580189'-17 |
| 760 | -1.93468516746121'+03 | -1.348101'+12 | 3.11437018549998'-17 | 3.13913279204804'-17 |
| 770 | -3.19008899598695'+03 | -2.222778'+12 | 1.88876417286454'-17 | 1.90398028328646'-17 |
| 780 | -5.26011568883386'+03 | -3.664986'+12 | 1.14547400961535'-17 | 1.15482241730158'-17 |
| 790 | -8.67336939026156'+03 | -6.043004'+12 | 6.94692711384579'-18 | 7.00435202616867'-18 |
| 800 | -1.43014571700165'+04 | -9.964057'+12 | 4.21308523476864'-18 | 4.24835425529160'-18 |
| 810 | -2.35815738456857'+04 | -1.642879'+13 | 2.55509909698752'-13 | 2.57675710915499'-18 |
| 820 | -3.98834940683940'+04 | -2.708767'+13 | 1.54958445690809'-18 | 1.56288218933500'-18 |
| 830 | -6.41147245243432'+04 | -4.466251'+13 | 9.39772548126288'-19 | 9.47935965350479'-19 |
| 840 | -1.05718325971469'+05 | -7.364124'+13 | 5.69941469323965'-19 | 5.74952226429359'-19 |
| 850 | -1.74318216745455'+05 | -1.214235'+14 | 3.45650954694102'-19 | 3.48726153199446'-19 |
| 860 | -2.87432102335001'+05 | -2.002089'+14 | 2.09626056203032'-19 | 2.11513103759109'-19 |
| 870 | -4.73944806200956'+05 | -3.301066'+14 | 1.27131381650966'-19 | 1.28289182360879'-19 |
| 880 | -7.81484314034862'+05 | -5.442843'+14 | 7.71010459922486'-20 | 7.78113224113383'-20 |
| 890 | -1.28858408213801'+06 | -8.974361'+14 | 4.67592754511187'-20 | 4.71949527152614'-20 |
| 900 | -2.12473738361600'+06 | -1.479734'+15 | 2.83579789687082'-20 | 2.86251858054941'-20 |
| 910 | -3.50346478116109'+06 | -2.439871'+15 | 1.71981914482479'-20 | 1.73620528310030'-20 |
| 920 | -5.77683885429049'+06 | -4.022938'+15 | 1.04301434674511'-20 | 1.05306173575539'-20 |
| 930 | -9.52538964509908'+06 | -6.633027'+15 | 6.32554260597533'-21 | 6.38714229305844'-21 |
| 940 | -1.57063491261443'+07 | -1.093664'+16 | 3.83623575120270'-21 | 3.87399762868720'-21 |
| 950 | -2.58980904785613'+07 | -1.803272'+16 | 2.32655214825426'-21 | 2.34969833745282'-21 |
| 960 | -4.27031823276659'+07 | -2.973332'+16 | 1.41097816964182'-21 | 1.42516408274094'-21 |
| 970 | -7.04129820852805'+07 | -4.902548'+16 | 8.55712345282970'-22 | 8.64405711303612'-22 |
| 980 | -1.16103479316812'+08 | -8.083218'+16 | 5.18961691709419'-22 | 5.24288566336349'-22 |
| 990 | -1.91442224292436'+08 | -1.332748'+17 | 3.14733378497071'-22 | 3.17997090019776'-22 |
| 1000 | -3.15667759981840'+08 | -2.197452'+17 | 1.90875552327595'-22 | 1.92874984796393'-22 |

TABLE 3.5-2

Data from example 3.5-2

(Note: $y(5) = e^{-5} = 6.7379469990855'\text{-}03$)

|  | $\tilde{y}_n(\frac{5}{n})$ |  |  |
|---|---|---|---|
| n | single-precision | double-precision | $y_n(\frac{5}{n})$ |
| 20 | 6.259691'-03 | 6.56312402779648'-03 | 6.56312402790868'-03 |
| 30 | 5.460650'-03 | 6.66008827465367'-03 | 6.66008827578716'-03 |
| 40 | 5.009264'-03 | 6.69412021149203'-03 | 6.69412021215067'-03 |
| 50 | 5.949687'-03 | 6.70988861449936'-03 | 6.70988861592713'-03 |
| 60 | 1.785394'-03 | 6.71845852303765'-03 | 6.71845852888167'-03 |
| 70 | -1.430578'-03 | 6.72362739157639'-03 | 6.72362739775014'-03 |
| 80 | 2.511602'-03 | 6.72698278674873'-03 | 6.72698278853473'-03 |
| 90 | -9.466864'-03 | 6.72928349316458'-03 | 6.72928350441466'-03 |
| 100 | 4.787731'-04 | 6.73092931920775'-03 | 6.73092932815196'-03 |
| 110 | -1.192247'-02 | 6.73214710541409'-03 | 6.73214712415119'-03 |
| 120 | -9.010211'-03 | 6.73307338681753'-03 | 6.73307339919532'-03 |
| 130 | -1.851527'-02 | 6.73379425406474'-03 | 6.73379428368321'-03 |
| 140 | -3.144952'-02 | 6.73436628389642'-03 | 6.73436629883609'-03 |
| 150 | -2.433643'-02 | 6.73482774281800'-03 | 6.73482778138931'-03 |
| 160 | -2.318940'-02 | 6.73520547117676'-03 | 6.73520547763337'-03 |
| 170 | -2.624966'-02 | 6.73551848763167'-03 | 6.73551850753462'-03 |
| 180 | -2.789574'-02 | 6.73578078464965'-03 | 6.73578083277416'-03 |
| 190 | -3.652655'-02 | 6.73600280280934'-03 | 6.73600284084923'-03 |
| 200 | -4.467228'-02 | 6.73619235639365'-03 | 6.73619238936501'-03 |
| 210 | -6.243775'-02 | 6.73635547974731'-03 | 6.73635551095046'-03 |
| 220 | -7.417661'-02 | 6.73649684680107'-03 | 6.73649689923037'-03 |
| 230 | -7.697296'-02 | 6.73662016382665'-03 | 6.73662025025945'-03 |
| 240 | -5.590025'-02 | 6.73672842720946'-03 | 6.73672850653685'-03 |
| 250 | -7.729518'-02 | 6.73682396844148'-03 | 6.73682403412359'-03 |
| 260 | -1.126839'-01 | 6.73690867314930'-03 | 6.73690875302489'-03 |
| 270 | -1.092784'-01 | 6.73698417848589'-03 | 6.73698423442461'-03 |
| 280 | -8.893842'-02 | 6.73705172709532'-03 | 6.73705177405624'-03 |
| 290 | -1.420971'-01 | 6.73711236526989'-03 | 6.73711244818440'-03 |
| 300 | -9.052908'-02 | 6.73716704114330'-03 | 6.73716715674576'-03 |
| 310 | -1.554755'-01 | 6.73721657511228'-03 | 6.73721665691298'-03 |
| 320 | -6.847292'-02 | 6.73726155046883'-03 | 6.73726158543395'-03 |
| 330 | -1.843618'-01 | 6.73730241728272'-03 | 6.73730249947136'-03 |
| 340 | -2.111134'-01 | 6.73733975117423'-03 | 6.73733985321659'-03 |
| 350 | -8.065218'-02 | 6.73737395472379'-03 | 6.73737405122982'-03 |
| 360 | -1.433600'-01 | 6.73740522053359'-03 | 6.73740543922362'-03 |
| 370 | -1.938782'-01 | 6.73743410804914'-03 | 6.73743431683434'-03 |
| 380 | -2.166646'-01 | 6.73746083027843'-03 | 6.73746094479951'-03 |
| 390 | -1.720492'-01 | 6.73748531402137'-03 | 6.73748555086108'-03 |
| 400 | -3.021251'-01 | 6.73750823211830'-03 | 6.73750833465946'-03 |
| 410 | -3.281339'-01 | 6.73752922285242'-03 | 6.73752947178680'-03 |
| 420 | -2.661970'-01 | 6.73754901743251'-03 | 6.73754911718278'-03 |

```
 430   -3.216861'-01    6.73756717392245'-03    6.73756740798635'-03
 440   -2.298895'-01    6.73758422354465'-03    6.73758446592876'-03
 450   -2.621343'-01    6.73760017847728'-03    6.73760039936977'-03
 460   -4.247934'-01    6.73761493986216'-03    6.73761530502086'-03
 470   -3.659283'-01    6.73762889376321'-03    6.73762926941390'-03
 480   -4.369898'-01    6.73764202839905'-03    6.73764237016332'-03
 490   -2.354091'-01    6.73765426435703'-03    6.73765467704661'-03
 500   -4.692361'-01    6.73766612594913'-03    6.73766625293218'-03
 510   -4.564425'-01    6.73767672532680'-03    6.73767715458583'-03
 520   -5.697100'-01    6.73768716786513'-03    6.73768743337044'-03
 530   -3.301094'-01    6.73769694177762'-03    6.73769713584793'-03
 540   -5.685695'-01    6.73770594070581'-03    6.73770630430939'-03
 550   -4.983457'-01    6.73771446165973'-03    6.73771497723588'-03
 560   -4.773043'-01    6.73772287717993'-03    6.73772318970571'-03
 570   -5.245698'-01    6.73773079293770'-03    6.73773097374647'-03
 580   -6.700377'-01    6.73773792119116'-03    6.73773835864671'-03
 590   -3.889994'-01    6.73774510040816'-03    6.73774537123760'-03
 600   -6.671242'-01    6.73775166255439'-03    6.73775203613112'-03
 610   -4.950620'-01    6.73775782487857'-03    6.73775837593928'-03
 620   -3.840306'-01    6.73776493269523'-03    6.73776441146410'-03
 630   -3.033572'-01    6.73776967083937'-03    6.73777016187205'-03
 640   -2.756428'-01    6.73777552133174'-03    6.73777564484225'-03
 650   -3.408707'-01    6.73778042609231'-03    6.73778097670528'-03
 660   -3.852482'-01    6.73778549849570'-03    6.73778587256356'-03
 670   -5.888407'-01    6.73779006283562'-03    6.73779064640175'-03
 680   -7.872203'-01    6.73779471105827'-03    6.73779521118206'-03
 690   -3.430081'-01    6.73779892498558'-03    6.73779957893728'-03
 700   -6.937112'-01    6.73780350997523'-03    6.73780376084265'-03
 710   -3.669302'-01    6.73780712959780'-03    6.73780776729641'-03
 720   -7.802510'-01    6.73781081512993'-03    6.73781160797738'-03
 730   -5.162380'-01    6.73781432097763'-03    6.73781529190711'-03
 740   -1.016203'+00    6.73781841643391'-03    6.73781882749937'-03
 750   -8.151451'-01    6.73782126475786'-03    6.73782222261299'-03
 760   -7.357829'-01    6.73782467537295'-03    6.73782548459333'-03
 770   -6.082242'-01    6.73782753336580'-03    6.73782862031199'-03
 780   -5.016716'-01    6.73783071650218'-03    6.73783163619979'-03
 790   -4.737896'-01    6.73783354780333'-03    6.73783453828859'-03
 800   -1.252605'+00    6.73783652995367'-03    6.73783733222987'-03
 810   -1.300713'+00    6.73783944088058'-03    6.73784002333290'-03
 820   -1.331237'+00    6.73784140013634'-03    6.73784261658259'-03
 830   -5.401559'-01    6.73784431744590'-03    6.73784511666591'-03
 840   -6.525357'-01    6.73784715451264'-03    6.73784752799305'-03
 850   -8.088494'-01    6.73784877379256'-03    6.73784985471704'-03
 860   -9.591714'-01    6.73785099926720'-03    6.73785210075035'-03
 870   -1.185460'+00    6.73785365970621'-03    6.73785426977910'-03
 880   -1.424537'+00    6.73785540927053'-03    6.73785636528441'-03
 890   -6.967568'-01    6.73785762185479'-03    6.73785839055333'-03
 900   -9.581641'-01    6.73785952038243'-03    6.73786034868853'-03
 910   -1.313843'+00    6.73786126569431'-03    6.73786224262577'-03
 920   -1.630921'+00    6.73786333796671'-03    6.73786407514100'-03
 930   -9.874165'-01    6.73786484051156'-03    6.73786584886044'-03
 940   -1.385051'+00    6.73786580478379'-03    6.73786756627456'-03
 950   -7.232992'-01    6.73786855901075'-03    6.73786922974096'-03
 960   -1.191267'+00    6.73786999140930'-03    6.73787084149377'-03
 970   -1.731914'+00    6.73787132726719'-03    6.73787240365646'-03
 980   -1.096796'+00    6.73787322673937'-03    6.73787391824294'-03
 990   -1.628775'+00    6.73787457394860'-03    6.73787538716515'-03
1000   -9.935147'-01    6.73787547487356'-03    6.73787681224056'-03
```

_REMARKS_: Our definitions of subroutine and stability (def. 2.4-1 and 3.1-2) depend on the machine number system $(\mathcal{R}, \mathcal{E})$ being considered. We can eliminate this dependence by defining _an algorithm $\mathcal{A}$ for an ideal function_ $f$ to be a constructive mapping from {the set of all $(\mathcal{R}, \mathcal{E})$} into {the set of all subroutines}, such that $F \equiv \mathcal{A}(\mathcal{R}, \mathcal{E})$ is a subroutine relative to $(\mathcal{R}, \mathcal{E})$ and $F$ and $f$ satisfy (3.1-1). Thus $\mathcal{A}$ is a recursive operator (see sec. 1.5), mapping any determiner $(\alpha_{\mathcal{R}}, \ell_{\mathcal{R}})$ of $(\mathcal{R}, \mathcal{E})$ into a determiner $\gamma$ of such an $F \equiv \mathcal{A}(\mathcal{R}, \mathcal{E})$. Roughly speaking, Algol procedures and Fortran subroutines are examples of such algorithms. We would then say _$\mathcal{A}$ is stable relative to $(\mathcal{R}, \mathcal{E})$ at $x$_ if $\mathcal{A}(\mathcal{R}, \mathcal{E})$ is stable at $x$. Note that $x$ is a real input, and therefore $x$ depends on $(\mathcal{R}, \mathcal{E})$. We would say _$\mathcal{A}$ is stable at $c$_ (a numeric constant) if, for any $(\mathcal{R}, \mathcal{E})$ and any real input $x = c$, $\mathcal{A}(\mathcal{R}, \mathcal{E})$ is stable at $x$. If we allow $\mathcal{A}$ to take more arguments, say a list of algorithms of the above type (as well as $(\mathcal{R}, \mathcal{E})$), then we get stronger and more general concepts of stability, analogous to those found in the literature on the numerical solution of ordinary differential equations.

BLANK PAGE

## Chapter 4: ε-Limit, ε-Comparison, ε-Convergence, and ε-Continuity

### 4.1 ε-Limit and Truncation-Error Bounds

Define an operator, $\phi_{\mathrm{lim}}$, over the set $S_{\mathrm{lim}}$ of ideal functions of two variables, by

$$(4.1\text{-}1) \qquad \phi_{\mathrm{lim}}(f)(x) = \lim_{y \to x} f(x,y) \quad .$$

Thus $\phi_{\mathrm{lim}}$ maps an ideal function of two variables into an ideal function of one variable. This operator represents a notion of limit. (We consider limits of the forms $\lim_{\overline{y}_m \to \overline{x}_m} g(\overline{x_m}, \overline{y_m})$ and $\lim_{y \to x_m} h(\overline{x_m}, y)$ at the end of section 4.3 and in section 5.4.) Having $x$ in the argument list of $f$ considerably simplifies notation because the TF part of $\mathfrak{F}$ depends upon where $y$ is going; if we are interested in $\lim_{y \to x} g(y)$, we will simply form $f$, with $f(k,y) = g(y)$ when $k \neq \omega$, and then consider $\lim_{y \to x} f(x,y)$ .

To formulate an ε-notion of ε-limit corresponding to $\phi_{\mathrm{lim}}$, we must define a set $S_{\mathrm{lim}}$ of ε-functions and an ε-operator $(\Phi_{\mathrm{lim}}, Q_{\mathrm{lim}})$ over $S_{\mathrm{lim}}$ such that if $\mathfrak{F} \in S_{\mathrm{lim}}$ and $\mathfrak{F} \approx f(P)$ then $\Phi_{\mathrm{lim}}(\mathfrak{F}) \approx \phi_{\mathrm{lim}}(f)(Q_{\mathrm{lim}}(\mathfrak{F}, f, P))$ . One way to define $\Phi_{\mathrm{lim}}$ is to select some effective stopping criterion (see ch. 3) and define

$$(4.1\text{-}2) \quad \Phi_{\mathrm{lim}}(\mathfrak{F})(\varepsilon;x) \equiv (F(\varepsilon;\, x,\, y_\varepsilon),\ (RF \mathbin{\hat{+}} TF)(\varepsilon;\, x,\, y_\varepsilon),\ \omega) \quad ,$$

where $y_\varepsilon$ is the value chosen by the stopping criterion when it is given $\varepsilon$, $x$ and $\mathfrak{F}$ . If we selected the S. C.$^{\#\#}$ of definition 3.4-1, we would have a *totally* inefficient $\Phi_{\text{lim}}$ with good accuracy, a large $S_{\text{lim}}$ and a good $Q_{\text{lim}}$. But its total lack of efficiency rules out this $\Phi_{\text{lim}}$.

Another method is suggested by the proofs of theorems 3.3-1 and 3.4-1. Roughly, this method proceeds at precision $\varepsilon$ by

(1) finding a $\delta \leq \varepsilon$ and a $Y \in \mathcal{R}(\delta)$ such that the truncation-error bound, $TF(\delta;\, x,\, Y)$, is $\leq \varepsilon$,

(2) finding an $\eta \leq \delta$ such that $RF(\eta;\, x,\, Y) \leq \varepsilon$, and

(3) defining $\Phi_{\text{lim}}(\mathfrak{F})(\varepsilon;\, x)$ to be (approximately) $(F(\eta;\, x,\, Y),\, 2\varepsilon,\, \omega)$.

Of course, these steps will have to be modified and $S_{\text{lim}}$ will have to be defined so that this process halts for each $\mathfrak{F} \in S_{\text{lim}}$, any $\varepsilon$ and any real input $x$. As we shall see, the only stability requirements needed to insure that this method converges concern $TF$ (and not $F$ or $F \hat{+} RF$).

*DEFINITION 4.1-1: Suppose $\mathfrak{F} \approx f(P)$ for some $P$. We say $TF$ is stably convergent at $x$ relative to $f$ precisely when*

$$\left[\lim_{y \to x} f(x,y) \neq \omega\right] \Rightarrow \lim_{\varepsilon \to 0} TF(\varepsilon;\, x,\, y_\varepsilon) = 0 \quad ,$$

*as long as the $y_\varepsilon$'s are chosen by any reasonable at $x$ stopping criterion.*

This is a stability requirement on $TF$ because, if we assume that $\lim_{\varepsilon \to 0} TF(\varepsilon;\, x,\, Y)$ always exists, then it is equivalent to requiring that $\lim_{y \to x} f(x,y) \neq \omega$ should imply the existence of a $tf$ such that

(i) $\lim_{\varepsilon\to 0} \mathrm{TF}(\varepsilon;\, x,\, Y) = \mathrm{tf}(x,Y)$ for all $Y \in \mathfrak{M}$ at which $\mathrm{tf}(x,Y) \neq \omega$ ,

(ii) TF is stable at $x$ (under def. 3.1-2), and

(iii) $\lim_{y\to x} \mathrm{tf}(x,y) = 0$ .

Let $S_{\mathrm{lim}}$ be the set of all $\varepsilon$-functions $\mathfrak{F}$ of two variables such that for each $x \in \tilde{R}$ and each $\varepsilon$,

(1) $[Y \in \mathcal{R}(\varepsilon)$ and $\mathrm{RF}(\varepsilon;\, x,\, Y) \neq \omega]$ implies $\lim_{\varepsilon\to 0} \mathrm{RF}(\varepsilon;\, x,\, Y) = 0$, and

(2) $[\mathrm{TF}(\varepsilon;\, x,\, Y) \neq \omega$ for some $Y \in \mathcal{R}(\varepsilon)]$ implies $[\lim_{\varepsilon\to 0} \mathrm{TF}(\varepsilon;\, x,\, y_\varepsilon) = 0$ as long as the $y_\varepsilon$'s are chosen by any reasonable at $x$ stopping criterion].

For the following, we assume that an effective, reasonable at any $x \neq \omega$, stopping criterion, S.C., is given. We also assume that a $\lambda : \mathcal{E} \to \mathfrak{M}$ is given which satisfies

(i) $\lambda(\varepsilon) \in \mathcal{R}(\varepsilon)$ and $\lambda(\varepsilon) > 0$ for all $\varepsilon$,

(ii) $\lim_{\varepsilon\to 0} \lambda(\varepsilon) = 0$, and

(iii) $\lambda(\varepsilon_i) = < \gamma(i, \cdot) >$ for some recursive function, $\gamma$ .

We define $\Phi_{\mathrm{lim}}$ in terms of S.C., $\lambda$ and the $\overset{\frown}{I}$ and $\overset{\smile}{I}$ of section 2.8 by

*DEFINITION 4.1-2: Let* $\mathfrak{F} \in S_{\mathrm{lim}}$, $x$ *and* $\varepsilon$ *be given. Let* $y_{\varepsilon_1}, y_{\varepsilon_2}, \ldots$ *be the values selected by* S.C. *for* TF *and* $x$ . *If* $\mathrm{TF}(\varepsilon;\, x,\, y_\varepsilon) = \omega$ *then define* $\Phi_{\mathrm{lim}}(\mathfrak{F})(\varepsilon;\, x) \equiv (\omega, \omega, \omega)$ . *Otherwise, let* $\delta$ *be the largest member of* $\mathcal{E}$ *such that* $\delta \leq \varepsilon$ *and* $\mathrm{TF}(\delta;\, x,\, y_\delta) \leq \lambda(\varepsilon)$ . *If* $\mathrm{RF}(\delta;\, x,\, y_\delta) = \omega$ *then define*

$\Phi_{\lim}(\mathfrak{F})(\varepsilon; x) \equiv (\omega, \omega, \omega)$ . *Otherwise let* $j$ *be the smallest integer such that* $\varepsilon_j \leq \delta$ *and* $RF(\varepsilon_j; x, y_\delta) \leq \lambda(\varepsilon)$ . *Suppose* $F(\varepsilon_j; x, y_\delta)$ *is* $< \alpha_R(j, k, \cdot) >$ *and let* $\beta_\varepsilon(\cdot)$ be $\alpha_R(j, k, \cdot)$ .

*Define*

(4.1-3)
$$\Phi_{\lim}(\mathfrak{F})(\varepsilon; x) \equiv (\hat{I}(\varepsilon, \beta_\varepsilon), 2 \hat{\times} \lambda(\varepsilon) \hat{+} |\hat{I}(\varepsilon, B_\varepsilon) \hat{-} \check{I}(\varepsilon, \beta_\varepsilon)|, \omega) \quad .$$

For $\mathfrak{F} \sim f(P)$, define

(4.1-4)
$$Q_{\lim}(\mathfrak{F}, f, P) \equiv \{x: \{x\} \times (\mathfrak{M} \cap \{\text{some neighborhood of } x\}) \subset P$$
$$\text{and } TF \text{ is stably convergent at } x \text{ relative to } f\} .$$

*THEOREM 4.1-1: We have*

$$(\Phi_{\lim}, Q_{\lim}) \sim \phi_{\lim}(S_{\lim}) \quad .$$

*Proof*: Suppose $\mathfrak{F} \sim f(P)$ and $\mathfrak{F} \in S_{\lim}$ . Let $x \in Q_{\lim}(\mathfrak{F}, f, P)$ be such that $\ell \equiv \lim_{y \to x} f(x,y) \neq \omega$ . Then for sufficiently small $\varepsilon$, $TF(\varepsilon; x, y_\varepsilon) \neq \omega$ and we can find a $\delta$ with $TF(\delta; x, y_\delta) \leq \lambda(\varepsilon)$ . Let $\delta$ denote the largest such value $\leq \varepsilon$ . This means that $|f(x, y_\delta) - \ell| \leq \lambda(\varepsilon)$ . If $\delta$ is sufficiently small (it will be, if $\varepsilon$ was) then $RF(\delta; x, y_\delta) \neq \omega$ and we can find an $\eta \leq \delta$ with $RF(\eta; x, y_\delta) \leq \lambda(\varepsilon)$ . Let $\varepsilon_j$ be the largest such $\eta$ . This means that

$$|F(\varepsilon_j; x, y_\delta) - f(x, y_\delta)| \leq \lambda(\varepsilon) \quad .$$

Thus we have

$$|F(\varepsilon_j; x, y_\delta) - \ell| \leq 2\lambda(\varepsilon) \quad . \tag{4.1-5}$$

For each $\varepsilon$, the corresponding $F(\varepsilon_{j(\varepsilon)}; x, y_{\delta(\varepsilon)})$ equals some $< \alpha_{R}(j(\varepsilon), k(\varepsilon), \cdot) >$ . Let $\beta_{\varepsilon}(\cdot)$ be $\alpha_{R}(j(\varepsilon), k(\varepsilon), \cdot)$ . From (4.1-5) we know that

$$\lim_{\varepsilon \to 0} < \beta_{\varepsilon} > = \ell \quad , \tag{4.1-6}$$

if $\ell = \pm \infty$ then $< \beta_{\varepsilon} > = \ell$ for all sufficiently small $\varepsilon$ .

It follows that

$$\lim_{\varepsilon \to 0} \hat{I}(\varepsilon, \beta_{\varepsilon}) = \lim_{\varepsilon \to 0} \check{I}(\varepsilon, \beta_{\varepsilon}) = \ell \quad , \tag{4.1-7}$$

because $\hat{I}(\varepsilon, \beta_{\varepsilon})$ and $\check{I}(\varepsilon, \beta_{\varepsilon})$ are both in $N_{\varepsilon}(< \beta_{\varepsilon} >)$ . Further, by the triangle inequality and (4.1-5) we have

$$|\hat{I}(\varepsilon, \beta_{\varepsilon}) - \ell| \leq 2 \hat{\times} \lambda(\varepsilon) \hat{+} |\hat{I}(\varepsilon, \beta_{\varepsilon}) \hat{-} \check{I}(\varepsilon, \beta_{\varepsilon})| \quad . \tag{4.1-8}$$

By (4.1-7) and theorem 2.8-1, the right' side of (4.1-8) approaches 0 as $\varepsilon \to 0$, which means that

$$\lim_{\varepsilon \to 0} \Phi_{\text{lim}}(\mathcal{F})(\varepsilon; x) \equiv (\lim_{y \to x} f(x,y), 0, \omega) \quad .$$

This completes the proof.

$\Phi_{\text{lim}}$ overcomes instabilities in F and/or $F \hat{+} RF$ by using a stably convergent TF . First it picks a place $(y_{\delta(\varepsilon)})$ at which to evaluate the $\varepsilon$-limit, and then it increases the precision until the crest of the $\varepsilon_{j(\varepsilon)}$-wave has moved past $y_{\delta(\varepsilon)}$ . The efficiency of $\Phi_{\text{lim}}$ will depend on

(1) how closely TF and RF approximate the errors that they bound, how difficult they are to evaluate, and

(2) how judiciously S.C. chooses its $y_\varepsilon$'s; if they are unnecessarily close to $x$ then $j$ may have to be made very large (an expensive enterprise) before $RF(\varepsilon_j;\, x,\, y_\delta) \leq \lambda(\varepsilon)$, especially when $F$ or $F \mathbin{\hat{+}} RF$ is unstable at $x$.

Thus we say that $\Phi_{\mathrm{lim}}$ offers a <u>potentially</u> efficient algorithm for overcoming instability.

We will use $\underset{y \to x}{\mathrm{LIM}}\, \mathfrak{F}(\varepsilon;\, x,\, y)$ to denote $\Phi_{\mathrm{lim}}(\mathfrak{F})(\varepsilon;\, x)$ and we call this <u>the $\varepsilon$-limit of $\mathfrak{F}$ at $x$</u>.

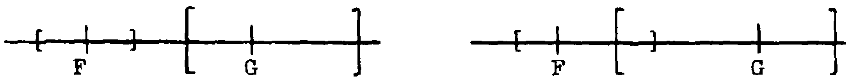
## 4.2 ε-Comparison Relations, $<_\varepsilon$ and $=_\varepsilon$

In the following sections, we will need an ε-less-than relation, $<_\varepsilon$, and an ε-equality relation, $=_\varepsilon$ . We will define these relations so that

(1) $\mathfrak{F}(\varepsilon; \bar{x}_m) <_\varepsilon \mathfrak{G}(\varepsilon; \bar{y}_n)$ is true when, based only on the information given by $\mathfrak{F}(\varepsilon; \bar{x}_m)$ and $\mathfrak{G}(\varepsilon; \bar{y}_n)$, $f(\bar{x}_m)$ must be less than $g(\bar{y}_n)$, and

(2) $\mathfrak{F}(\varepsilon; \bar{x}_m) =_\varepsilon \mathfrak{G}(\varepsilon; \bar{y}_n)$ is true when, based only on the information given by $\mathfrak{F}(\varepsilon; \bar{x}_m)$ and $\mathfrak{G}(\varepsilon; \bar{y}_n)$, $f(\bar{x}_m)$ might be equal to $g(\bar{y}_n)$ .

Essentially, the ε-less-than relationship holds when the interval $[F - RF, F + RF]$ lies entirely to the left of the interval $[G - RG, G + RG]$, and ε-equality holds when these intervals overlap (see figure 4.2-1). Of course $=_\varepsilon$ will not be an equivalence relation because it will not be transitive.



$\mathfrak{F} <_\varepsilon \mathfrak{G}$ $\quad$ $\mathfrak{F} =_\varepsilon \mathfrak{G}$

ε-Comparison

(a) FIGURE 4.2-1 (b)

DEFINITION 4.2-1: Let x and y be poor real inputs. For $\varepsilon \in \mathcal{E}$, define $x =_\varepsilon y$ to be true (and $x \neq_\varepsilon y$ to be false) precisely when

$$|X(\varepsilon) \dot{-} Y(\varepsilon)| \leq RX(\varepsilon) \dot{+} RY(\varepsilon) \quad . \tag{4.2-1}$$

Define $x <_\varepsilon y$ to be true (and $x \not<_\varepsilon y$ to be false) precisely when $x \neq_\varepsilon y$ and $X(\varepsilon) < Y(\varepsilon)$ .

For $a_1, a_2 \geq 0$, $a_3 \in \mathcal{R}(\varepsilon)$, when the triple $(\bar{a}_3)$ appears in an $\varepsilon$-comparison ($\varepsilon$ fixed) this triple is to be understood to denote the poor real input $a \equiv (A, RA)$ defined by

$$(A(\delta), RA(\delta)) \equiv \begin{cases} (\omega, \omega) & \delta > \varepsilon \\ (a_1, a_2) & \delta \leq \varepsilon \end{cases} .$$

This convention allows us to $\varepsilon$-compare $\varepsilon$-function values directly. At all times precisely one of $\mathcal{F}(\varepsilon; \bar{x}_m) =_\varepsilon \mathcal{H}(\varepsilon; \bar{y}_n)$, $\mathcal{F}(\varepsilon; \bar{x}_m) <_\varepsilon \mathcal{H}(\varepsilon; \bar{y}_n)$ and $\mathcal{H}(\varepsilon; \bar{y}_n) <_\varepsilon \mathcal{F}(\varepsilon; \bar{x}_m)$ holds.

THEOREM 4.2-1: Let $x$ and $y$ be real inputs. Then

$$x = y \Leftrightarrow [x =_\varepsilon y \text{ for all } \varepsilon]$$
$$x < y \Rightarrow [x <_\varepsilon y \text{ for all sufficiently small } \varepsilon] .$$
$$x < y \Leftarrow [x <_\varepsilon y \text{ for some } \varepsilon] .$$

Proof: If $x = y$ then we have, for all $\varepsilon$,

$$|X(\varepsilon) \underset{\smile}{-} Y(\varepsilon)| \leq |X(\varepsilon) - Y(\varepsilon)| \leq RY(\varepsilon) + RY(\varepsilon) \leq RX(\varepsilon) \overset{\wedge}{+} RY(\varepsilon) \quad ,$$

and so $x =_\varepsilon y$ for all $\varepsilon$ . If $x =_\varepsilon y$ for all $\varepsilon$, then applying theorem 2.8-1 and taking the limit of (4.2-1) as $\varepsilon \to 0$ yields $|x - y| \leq 0$, so $x = y$ .

If $x < y$ then, for all sufficiently small $\varepsilon$,

$$(4.2\text{-}2) \qquad X(\varepsilon) < Y(\varepsilon), \qquad |X(\varepsilon) \underset{\smile}{-} Y(\varepsilon)| > RX(\varepsilon) \overset{\wedge}{+} RY(\varepsilon) \quad ,$$

the second inequality holding because, by theorem 2.8-1, $|X(\varepsilon) \underset{\smile}{-} Y(\varepsilon)| \rightarrow |x-y| > 0$ whereas $RX(\varepsilon) \overset{\frown}{+} RY(\varepsilon) \rightarrow 0$. Thus $x <_{\varepsilon} y$ for all sufficiently small $\varepsilon$. If $x <_{\varepsilon} y$ (some $\varepsilon$) then (4.2-2) holds and so

$$|X(\varepsilon) - Y(\varepsilon)| > RX(\varepsilon) + RY(\varepsilon) \quad ,$$

implying $x < y$. This completes the proof.

Let bool[statement] be 1 if the statement is true and 0 if it is false. The notions of comparison given by

$$\phi_{*}(\bar{f}_2)(\bar{x}_m) = \begin{cases} \omega & \text{if } f_i(\bar{x}_m) = \omega \ (i = 1 \text{ or } 2) \\ \text{bool } [f_1(\bar{x}_m) * f_2(\bar{x}_m)] & \text{otherwise,} \end{cases}$$

for $*$ being $=$ and $<$, can be $\varepsilon$-ized easily, to yield <u>weak</u> $\varepsilon$-operators, $(\Phi_{*}, Q_{*})$. The weakness of these $\varepsilon$-operators is due to the fact that the information given by $\mathcal{F}_i(\varepsilon; \bar{x}_m)(i = 1, 2)$ may never (for any $\varepsilon$) be sufficient to determine that $f_i(\bar{x}_m)$ <u>must</u> equal $f_2(\bar{x}_m)$. See Bishop [B1, p. 24] and Aberth [A1, pp. 287-8] for similar considerations.

Techniques from interval analysis can be formalized in the $\varepsilon$-calculus to yield a weak $\varepsilon$-operator corresponding to the operator,

$$\phi_{pos}(f)(a, b) = \begin{cases} \omega & \text{if } a \text{ or } b \text{ is in } \{-\infty, \infty, \omega\} \\ \text{bool } [f(x) > 0 \text{ for } a \leq x \leq b] & \text{otherwise} \quad . \end{cases}$$

We leave this to the reader.

## 4.3 $\varepsilon$-Convergence and $\varepsilon$-Continuity: Pointwise

In this context, we say *$f$ converges at $x$* precisely when $\lim_{y \to x} f(x, y) \neq \omega$ ; i.e., precisely when the limit exists in the usual sense. Otherwise, we say *$f$ diverges at $x$* .

*DEFINITION 4.3-1*: *Fix* $x$ *and* $\varepsilon$ . *We say* ***$\mathfrak{F}$ $\varepsilon$-converges at $x$*** *precisely when*

$$\lim_{y \to x} \mathfrak{F}(\varepsilon;\, x,\, y) \neq_{\varepsilon} \omega \quad . \tag{4.3-1}$$

*Otherwise we say* ***$\mathfrak{F}$ $\varepsilon$-diverges at $x$*** .

We say *$f$ is continuous at $x$* precisely when

$$f(x,\, x) = \lim_{y \to x} f(x,\, y) \neq \omega \quad . \tag{4.3-2}$$

Otherwise we say *$f$ is discontinuous at $x$* . Note that (4.3-2) uses the transitivity relation $a = b \neq c \Rightarrow a \neq c$ to insure that $f(x, x) \neq \omega$ . Since $=_{\varepsilon}$ and $\neq_{\varepsilon}$ do not satisfy such a transitivity relation, we must explicitly insure this in

*DEFINITION 4.3-2*: *Fix* $x$ *and* $\varepsilon$ . *We say* ***$\mathfrak{F}$ is $\varepsilon$-continuous at $x$*** *precisely when*

$$\mathfrak{F}(\varepsilon;\, x,\, x) =_{\varepsilon} \lim_{y \to x} \mathfrak{F}(\varepsilon;\, x,\, y) \neq_{\varepsilon} \omega \ \ \underline{\text{and}} \ \ \mathfrak{F}(\varepsilon;\, x,\, x) \neq_{\varepsilon} \omega \quad .$$

*Otherwise we say* ***$\mathfrak{F}$ is $\varepsilon$-discontinuous at $x$*** . *We say* ***$\mathfrak{F}$ is strongly $\varepsilon$-discontinuous at $x$*** *precisely when*

$$\mathfrak{F}(\varepsilon;\ x,\ x) \neq_{\varepsilon} \underset{y \to x}{\mathrm{LIM}}\, \mathfrak{F}(\varepsilon;\ x,\ y) \quad .$$

If we are interested in the continuity of $g(y)$ at $y = x$, then we simply form $f$, with $f(k, y) = g(y)$ when $k \neq \omega$, and investigate the $\varepsilon$-continuity at $x$ of some $\mathfrak{F}$ corresponding to $f$. Let $\varepsilon_j$ be as in definition 4.1-2, when a value for $\varepsilon_j$ is found (i.e., when $TF(\varepsilon;\ x,\ y_{\varepsilon}) \neq \omega$); otherwise let $\varepsilon_j$ be $\varepsilon$. Let $\mathfrak{F}_{\varepsilon}$ denote the finite subset of $\mathfrak{M}^{(3)}$ given by

$$\mathfrak{F}_{\varepsilon} \equiv \{\mathfrak{F}(\eta;\ x,\ Y):\ Y \in \mathfrak{R}(\eta) \text{ and } \varepsilon_j \leq \eta \leq \varepsilon\} \quad .$$

From the definitions of $\varepsilon$-function, of $\varepsilon$-limit and of $\varepsilon$-equality, it follows that

(1) $\mathfrak{F}$ $\varepsilon$-converges at $x$ when, based only on information contained in $\mathfrak{F}_{\varepsilon}$, $f$ <u>must</u> converge at $x$, and

(2) $\mathfrak{F}$ $\varepsilon$-diverges at $x$ when, based only on information contained in $\mathfrak{F}_{\varepsilon}$, $f$ <u>might</u> diverge at $x$ .

Let $\mathfrak{F}'_{\varepsilon}$ denote the finite subset of $\mathfrak{M}^{(3)}$ given by

$$\mathfrak{F}'_{\varepsilon} \equiv \{\mathfrak{F}(\eta;\ x,\ x):\ \varepsilon_j \leq \eta \leq \varepsilon\} \quad .$$

As above, we have

(1) $\mathfrak{F}$ is $\varepsilon$-continuous at $x$ when, based only on $\mathfrak{F}_{\varepsilon} \cup \mathfrak{F}'_{\varepsilon}$, $f$ <u>might</u> be continuous at $x$, and

(2) $\mathfrak{F}$ is strongly $\varepsilon$-discontinuous at $x$ when, based only on $\mathfrak{F}_{\varepsilon} \cup \mathfrak{F}'_{\varepsilon}$, $f$ <u>must</u> be discontinuous at $x$ .

These definitions can be expressed in operator, $\varepsilon$-operator form as follows. Let $S_{\mathrm{lim}}$ be as in section 4.1. Define operators

$\phi_{\text{conv}}$, $\phi_{\text{cont}}$, each over $S_{\text{lim}}$, by

$$\phi_{\text{conv}}(f)(x) = \text{bool}\ [f \ \text{converges at} \ x] \quad ,$$

$$\phi_{\text{cont}}(f)(x) = \text{bool}\ [f \ \text{is continuous at} \ x] \quad ,$$

so long as $x \neq \omega$ . Of course, $\phi_{\text{conv}}(f)(\omega) = \phi_{\text{cont}}(f)(\omega) = \omega$ . Define $\varepsilon$-operators corresponding weakly to the above by equating, for $x \neq_{\varepsilon} \omega$,

$$(4.3\text{-}3) \qquad \begin{aligned} \Phi_{\text{conv}}(\mathfrak{F})(\varepsilon;\, x) \equiv (&\text{bool}\ [\mathfrak{F} \ \varepsilon\text{-converges at} \ x] \quad , \\ &\text{bool}\ [\mathfrak{F} \ \varepsilon\text{-diverges at} \ x],\ \omega) \quad , \end{aligned}$$

$$(4.3\text{-}4) \qquad \begin{aligned} \Phi_{\text{cont}}(\mathfrak{F})(\varepsilon;\, x) \equiv (&\text{bool}\ [\mathfrak{F} \ \text{is} \ \varepsilon\text{-continuous at} \ x] \quad , \\ &1\text{-bool}\ [\mathfrak{F} \ \text{is strongly} \ \varepsilon\text{-discontinuous at} \ x],\ \omega) \quad , \end{aligned}$$

$$(4.3\text{-}5) \qquad Q_{\text{cont}}(\mathfrak{F},\, f,\, P) \equiv Q_{\text{lim}}(\mathfrak{F},\, f,\, P) \cap \{x\colon\ (x,\, x) \in P\} \quad .$$

THEOREM 4.3-1: *We have*

$$(4.3\text{-}6) \qquad (\Phi_{\text{conv}},\, Q_{\text{lim}}) \sim \phi_{\text{conv}}(S_{\text{lim}}) \quad ,$$

$$(4.3\text{-}7) \qquad (\Phi_{\text{cont}},\, Q_{\text{cont}}) \sim \phi_{\text{cont}}(S_{\text{lim}}) \quad .$$

*Further, if* $f$ *converges at* $x$ *for all* $x \in Q_{\text{lim}}(\mathfrak{F},\, f,\, P)$ *then* $\Phi_{\text{conv}}(\mathfrak{F})$ *is not weak. If* $f$ *is discontinuous at* $x$ *and* $f(x,\, x) \neq \omega$ *and* $\lim_{y \to x} f(x,\, y) \neq \omega$ *for all* $x \in Q_{\text{cont}}(\mathfrak{F},\, f,\, P)$, *then* $\Phi_{\text{cont}}(\mathfrak{F})$ *is not weak.*

*Proof*: Consider $\varepsilon$-convergence first. Suppose $\mathfrak{F} \approx f(P)$ and $x$ is in $Q_{\text{lim}}(\mathfrak{F},\, f,\, P)$ . If $f$ converges at $x$ then by theorem 4.1-1,

$\lim_{y \to x} \mathfrak{F}(\varepsilon; x, y) \to (\lim_{y \to x} f(x, y), 0, \omega)$ as $\varepsilon \to 0$ and so $\lim_{y \to x} \mathfrak{F}(\varepsilon, x, y)$ must be $\neq_\varepsilon \omega$ for all sufficiently small $\varepsilon$. If $f$ diverges at $x$, then $\lim_{y \to x} f(x, y) = \omega$ and so $\lim_{y \to x} \mathfrak{F}(\varepsilon; x, y) =_\varepsilon \omega$ must hold for all $\varepsilon$. This and (4.3-3) yield (4.3-6) and the first remark after (4.3-7).

Consider $\varepsilon$-continuity. Suppose $\mathfrak{F} \approx f(P)$ and $x$ is in $Q_{\text{cont}}(\mathfrak{F}, f, P)$. If $f$ is continuous at $x$ then theorems 4.1-1, 4.1-2 and the fact that $(x, x) \in P$ give us $\varepsilon$-continuity for all sufficiently small $\varepsilon$. If $f$ is discontinuous at $x$ then $f(x, x) = \omega$ or $\lim_{y \to x} f(x, y) = \omega$ or else $f(x, x) \neq \lim_{y \to x} f(x, y)$. In the first two cases we have $\mathfrak{F}(\varepsilon; x, x) =_\varepsilon \omega$ or $\operatorname{LIM}_{y \to x} \mathfrak{F}(\varepsilon; x, y) =_\varepsilon \omega$ for all $\varepsilon$, and so $\mathfrak{F}$ is always $\varepsilon$-discontinuous at $x$. In the third case theorems 4.1-1 and 4.2-1 imply that $\mathfrak{F}(\varepsilon; x, y) \neq_\varepsilon \operatorname{LIM}_{y \to x} \mathfrak{F}(\varepsilon; x, y)$ for all sufficiently small $\varepsilon$, so $\mathfrak{F}$ is $\varepsilon$-discontinuous at $x$ for all sufficiently small $\varepsilon$. This and (4.3-4) yield (4.3-7) and the last remark. This completes the proof.

Let $f$ be an ideal function of $m + 1$ variables and $g$ an ideal function of $2m$ variables $(m \geq 1)$. We can easily discretize "$f$ converges at $\bar{x}_m$" (true if $\lim_{y \to \bar{x}_m} f(\bar{x}_m, y) \neq \omega$) and "$f$ is continuous at $\bar{x}_m$" (true if $f(\bar{x}_m, x_m) = \lim_{y \to \bar{x}_m} f(\bar{x}_m, y) \neq \omega$). However, with our present setup we cannot discretize "$g$ converges at $\bar{x}_m$" (true if $\lim_{\bar{y}_m \to \bar{x}_m} g(\bar{x}_m, \bar{y}_m) \neq \omega$) and "$g$ is continuous at $\bar{x}_m$" (true if $g(\bar{x}_m, \bar{x}_m) = \lim_{\bar{y}_m \to \bar{x}_m} g(\bar{x}_m, \bar{y}_m) \neq \omega$) because our truncation-error bounds do not give the necessary local information about all the possible approaches of $\bar{y}_m$ to $\bar{x}_m$ . (The $\bar{x}_m$ appearing in $g(\bar{x}_m, \bar{y}_m)$ may just be dummy variables telling where $\bar{y}_m$ is to go.) We could have done this latter discretization if we had assumed truncation-error bounds, $TG(\varepsilon; \bar{x}_m, \bar{y}_m)$, for limits of the form $\lim\limits_{\bar{y}_m \to \bar{x}_m} g(\bar{x}_m, \bar{y}_m)$ .

We would then have defined an $\varepsilon$-limit of the form $\operatorname*{LIM}\limits_{\bar{y}_m \to \bar{x}_m} \mathcal{J}(\varepsilon; \bar{x}_m, \bar{y}_m)$ .

Of course this $\varepsilon$-limit would not be more powerful, computationally, than $\operatorname*{LIM}\limits_{y \to x_m} \mathfrak{F}(\varepsilon; \bar{x}_m, y)$, because $m$ successive applications of the latter $\varepsilon$-limit are essentially as good as one application of the former; the difference between these two $\varepsilon$-limits is that the latter one will approach $\bar{x}_m$ along the m-dimensional axes whereas the former one may take any approach. This follows from the relation,

$$\left[\lim_{\bar{y}_m \to \bar{x}_m} g(\bar{x}_m, \bar{y}_m) \neq \omega\right] \Rightarrow \lim_{\bar{y}_m \to \bar{x}_m} g(\bar{x}_m, \bar{y}_m) = \lim_{y_1 \to x_1} \cdots \lim_{y_m \to x_m} g(\bar{x}_m, \bar{y}_m) \quad .$$

Thus, if the limit exists, the domain set of $\mathcal{J}$ is large enough and the truncation-error bounds involved are stably convergent, then both these $\varepsilon$-limits will work. We have avoided the more complicated form of $\varepsilon$-limit in order to simplify notation.

## 4.4 Discontinuities

In order to discretize convergence and continuity over intervals, we must know more about the kinds of discontinuities $f$ can have in $P$ while there still exists an $\varepsilon$-function corresponding to $f$ over $P$. Consider the ideal function $f$ defined for finite $x$ and $y$ by

$$f(x,\ y) = \text{bool}\ [x < y] \quad .$$

Define an $\varepsilon$-function $\mathfrak{F}$ by

$$\mathfrak{F}(\varepsilon; x,\ y) \equiv (\text{bool}\ [x <_{\eta} y \ \text{ for some } \ \eta \geq \varepsilon],\ \text{bool}\ [x \neq_{\eta} y \ \text{ or } \ RX(\eta) = RY(\eta) = 0 \ \text{ for some } \ \eta \geq \varepsilon],\ \omega),$$

so long as $x$ and $y$ are $\neq_{\varepsilon} -\infty,\ \infty,\ \omega$. In this case we have

$$\mathfrak{F} \approx f(\{(x,\ y):\ \ x \neq y \ \text{ or } \ x = y \in \mathfrak{M}\}) \quad .$$

However, we only have

$$\mathfrak{F} \sim f(\tilde{R}^{(2)}) \quad ,$$

the correspondence being weak because $RX(\eta) \neq 0$ for $\eta \geq \varepsilon$ implies $\mathfrak{F}(\varepsilon;\ x,\ x) \equiv (1,\ 1,\ \omega)$ and so $\lim_{\varepsilon \to 0} RF(\varepsilon;\ x,\ x) \neq 0$ for any $x \notin \mathfrak{M}$. Let $\mathfrak{F}'$ be any $\varepsilon$-function weakly corresponding to $f$ over $\tilde{R}^{(2)}$. Then for $x = y \notin \mathfrak{M}$, $RF'(\varepsilon; x,\ y)$ cannot go to $0$ with $\varepsilon$ because the inputted values of $x$ and $y$ will always be inexact, and so $\mathfrak{F}'$ will never have enough information to decide for sure that $x = y$.

Many variations on this basic theme are possible. The underlying principle is given by

*THEOREM 4.4-1*: *Suppose* $\mathfrak{F} \approx f(P)$ *and* $\bar{x}_m \in P$ *is a point of discontinuity of* $f$ (i.e., $f(\bar{x}_m) = \omega$ *or* $\lim\limits_{\bar{y}_m \to \bar{x}_m} f(\bar{y}_m) \neq f(\bar{x}_m)$) *and* $f(\bar{x}_m) \neq \omega$. *Then at least one* $x_j \in \mathcal{M}$.

*Proof*: Suppose $\mathfrak{F}$, $f$, $P$ and $\bar{x}_m$ satisfy the hypotheses, but $|x_i| \neq \infty$ and $RX_i(\varepsilon) \neq 0$ for all $\varepsilon$ and $i = 1,2,\dots, m$. We will prove that this implies $\lim\limits_{\varepsilon\to 0} RF(\varepsilon; \bar{x}_m) \neq 0$, a contradiction. Suppose $\gamma_1$ and $\gamma_2$ are the given determiners of $F$ and $RF$ and that $F$ and $RF$ involve respectively $r_1$ and $r_2$ subroutine constants. For $k = 1,2$ let $n_k(\varepsilon_i)$ be the least value of $n$ such that $\gamma_k(i, GN_n(\bar{x}_{m+r_k})) \neq 0$, and define

$$n(\varepsilon_i) = \max(n_1(\varepsilon_i), n_2(\varepsilon_i)) \quad .$$

Let $\sigma(\varepsilon)$ be the $m$-dimensional rectangle of real inputs,

$$\sigma(\varepsilon) \equiv \{\bar{y}_m\colon \ GN_{n(\varepsilon)}(\bar{y}_m) = GN_{n(\varepsilon)}(\bar{x}_m)\} \quad .$$

All the sides of $\sigma(\varepsilon)$ have positive length. Let $\ell_+$ and $\ell_-$ be the limit superior and limit inferior of $f(\bar{y}_m)$ as $\bar{y}_m \to \bar{x}_m$, and define

$$K = \begin{cases} \frac{1}{2}(\ell_+ - \ell_-) & \text{if } \ell_+ \neq \ell_- \\ \frac{1}{2}|\ell_+ - f(\bar{x}_m)| & \text{otherwise} \end{cases} ,$$

$$h(\varepsilon, c) = \sup_{\bar{y}_m \in \sigma(\varepsilon)} |c - f(\bar{y}_m)| \qquad \text{for } c \in \tilde{R} \quad .$$

$K > 0$ because $f$ is discontinuous at $\bar{x}_m$. There are $\bar{y}_m \in \sigma(\varepsilon)$ which make $f(\bar{y}_m)$ arbitrarily close to the one of $\ell_+$, $\ell_-$, $f(\bar{x}_m)$

which is furthest from $c$ . Thus we have

$$h(\varepsilon, c) \geq K \quad \text{for any} \quad c \in \tilde{R} \quad \text{and any} \quad \varepsilon \, .$$

For any real inputs $\bar{y}_m$ we have

$$RF(\varepsilon; \bar{y}_m) \geq |F(\varepsilon; \bar{y}_m) - f(\bar{y}_m)| \quad .$$

For $\bar{y}_m \in \sigma(\varepsilon)$ we have $F(\varepsilon; \bar{y}_m) = F(\varepsilon; \bar{x}_m)$ and $RF(\varepsilon; \bar{y}_m) = RF(\varepsilon; \bar{x}_m)$ , yielding

$$RF(\varepsilon; \bar{x}_m) \geq \sup_{\bar{y}_m \in \sigma(\varepsilon)} |F(\varepsilon; \bar{x}_m) - f(\bar{y}_m)| = h(\varepsilon, F(\varepsilon; \bar{x}_m)) \quad .$$

Thus $RF(\varepsilon; \bar{x}_m) \geq K > 0$ for any $\varepsilon$, the desired contradiction. This completes the proof.

*COROLLARY 4.4-1*: *Suppose* $\mathcal{F}$, $f$, $P$ *and* $\bar{x}_m$ *satisfy the hypotheses of the above theorem. Then for each* $j = 1,2,\ldots, m,$ *either* $x_j \in \mathcal{M}$ *or the function* $g(y) = f(x_1, \ldots, x_{j-1}, y, x_{j+1}, \ldots, x_m)$ *is discontinuous at* $y = x_j$ .

*Proof*: Define $\mathcal{G}$ by setting $T\mathcal{G} \equiv \omega$ and

$$G(\varepsilon; y) = F(\varepsilon; x_1,\ldots, x_{j-1}, y, x_{j+1},\ldots, x_m) \quad ,$$

$$RG(\varepsilon; y) = RF(\varepsilon; x_1,\ldots, x_{j-1}, y, x_{j+1}, \ldots, x_m) \quad .$$

Then $\mathcal{G} \approx g(\{x_j\})$ and $g(x_j) = f(\bar{x}_m) \neq \omega$ . If $g$ is discontinuous at $x_j$ then, by the above theorem, $x_j \in \mathcal{M}$ . This completes the proof.

COROLLARY 4.4-2: *If* $\mathcal{F} \approx f(P)$, $(\bar{x}_m, x_m) \in P$, $f(\bar{x}_m, x_m) \neq \omega$ *and* f *is discontinuous at* $\bar{x}_m$, *then* $x_m \in \mathcal{M}$ .

Proof: Under the given assumptions, $(\bar{x}_m, x_m)$ is a point of discontinuity of f, so, by corollary 4.4-2, either $x_m \in \mathcal{M}$ or $g(y) = f(\bar{x}_m, y)$ is continuous in y at $y = x_m$ . The latter alternative is ruled out by assumption, so we have $x_m \in \mathcal{M}$ . This completes the proof.

COROLLARY 4.4-3: *Let* P *be a set of m-tuples of numbers and suppose* $\mathcal{F} \approx f(P)$ . *Then* f *is continuous at every* $\bar{x}_m \in P$ *with* $f(\bar{x}_m) \neq \omega$ *and* $|x_i| \neq \infty$ $(i = 1,\ldots, m)$ .

The second use of P is as a set of real inputs (see sec. 2.3). For example, P might be $\tilde{R}^{(m)}$ .

Proof: Suppose $\mathcal{F}$, f and P satisfy the hypotheses. Assume that $\bar{x}_m \in P$ with $f(\bar{x}_m) \neq \omega$ (and hence $x_i \neq \omega$ for all i) and $|x_i| \neq \infty$ for $i = 1,\ldots, m$ . Define $\bar{y}_m$ by

$$Y_i(\varepsilon) = X_i(\varepsilon) \quad ,$$

$$RY_i(\varepsilon) = \max(\Theta(\varepsilon), RX_i(\varepsilon)) \quad ,$$

for $i = 1,\ldots, m$ and all $\varepsilon$, where $\Theta(\varepsilon)$ is the smallest positive number in $\mathbf{R}(\varepsilon)$ . Then each $y_i$ is a real input and $\bar{y}_m \in P$ . From theorem 4.4-1, we know that f cannot be discontinuous at $\bar{y}_m$, and hence not at $\bar{x}_m$ . This completes the proof.

## 4.5 ε-Convergence and ε-Continuity: Over Intervals

For simplicity, we consider open intervals. Let $o(a, b)$ denote the open interval between $a$ and $b (a, b \in \tilde{R})$. Define *the open ε-interval between a and b*, $\mathfrak{O}(\varepsilon; a, b)$, by

$$\mathfrak{O}(\varepsilon; a, b) \equiv \{Y: Y \in \mathfrak{R}(\varepsilon) \cap o(a, b), Y \neq_{\varepsilon} a, Y \neq_{\varepsilon} b\} .$$

For $Y \in \mathfrak{R}(\varepsilon)$, the decision $Y \in \mathfrak{O}(\varepsilon; a, b)$ is effective, given real inputs $a$ and $b$, and we have

$$\mathfrak{O}(\varepsilon; a, b) \subset o(a, b) \quad \text{for all } \varepsilon ,$$

(4.5-1)

$$\bigcup_{i \geq 1} \mathfrak{O}(\varepsilon_i; a, b) \equiv o(a, b) \cap \mathfrak{M} .$$

We say *f converges over o(a, b)* precisely when $f$ converges at all $x \in o(a, b)$. Otherwise, we say *f diverges in o(a, b)*.

DEFINITION 4.5-1: *We say $\mathfrak{F}$ ε-converges over $\mathfrak{O}(\varepsilon; a, b)$ precisely when $\mathfrak{F}$ ε-converges at all $x \in \mathfrak{O}(\varepsilon; a, b)$. Otherwise we say $\mathfrak{F}$ ε-diverges in $\mathfrak{O}(\varepsilon; a, b)$.*

We say *f is continuous over o(a, b)* precisely when $f$ is continuous at all $x \in o(a, b)$. Otherwise we say *f is discontinuous in o(a, b)*.

DEFINITION 4.5-2: *We say $\mathfrak{F}$ is ε-continuous over $\mathfrak{O}(\varepsilon; a, b)$ precisely when $\mathfrak{F}$ is ε-continuous at all $x \in \mathfrak{O}(\varepsilon; a, b)$. Otherwise we say $\mathfrak{F}$ is ε-discontinuous in $\mathfrak{O}(\varepsilon; a, b)$. We say $\mathfrak{F}$ is strongly ε-discontinuous in $\mathfrak{O}(\varepsilon; a, b)$ precisely when there* *is an* $x \in \mathfrak{O}(\varepsilon;\, a, b)$ *such that* $\mathfrak{F}$ *is strongly* *$\varepsilon$-discontinuous at* $x$ .

We express this in operator, $\varepsilon$-operator form by defining, for $a \neq \omega$ and $b \neq \omega$ and $f \in S_{\mathrm{lim}}$ ,

$$\phi_{\mathrm{convo}}(f)(a, b) = \mathrm{bool}\ [f \text{ converges over } o(a, b)]$$

$$\phi_{\mathrm{conto}}(f)(a, b) = \mathrm{bool}\ [f \text{ is continuous over } o(a, b)] \quad ,$$

and defining, for $a \neq_{\varepsilon} \omega$ and $b \neq_{\varepsilon} \omega$ ,

$$\Phi_{\mathrm{convo}}(\mathfrak{F})(\varepsilon;\, a, b) \equiv (\mathrm{bool}\ [\mathfrak{F} \text{ } \varepsilon\text{-converges over } \mathfrak{O}(\varepsilon;\, a, b)], \omega, \omega) \quad ,$$

$$\Phi_{\mathrm{conto}}(\mathfrak{F})(\varepsilon;\, a, b) \equiv (\mathrm{bool}\ [\mathfrak{F} \text{ } \varepsilon\text{-continuous over } \mathfrak{O}(\varepsilon;\, a, b)] \quad ,$$

$$1\text{-}\mathrm{bool}\ [\mathfrak{F} \text{ is strongly } \varepsilon\text{-discontinuous in } \mathfrak{O}(\varepsilon;\, a, b)], \omega)$$

Since the evaluation of $\Phi_{\mathrm{conto}}(\mathfrak{F})(\varepsilon;\, a, b)$ and $\Phi_{\mathrm{convo}}(\mathfrak{F})(\varepsilon;\, a, b)$ involves the evaluation of $\Phi_{\mathrm{lim}}(\mathfrak{F})$ only at $\varepsilon$-points $(\varepsilon;\, x)$ for which $x \in \mathfrak{R}(\varepsilon)$, the set $S_0$ of $\varepsilon$-functions to which $\Phi_{\mathrm{convo}}$ and $\Phi_{\mathrm{conto}}$ may be applied is defined as follows. Let $S_0$ be the set of all $\varepsilon$-functions $\mathfrak{F}$ of two variables such that for each $\varepsilon$ and each $x \in \mathfrak{R}(\varepsilon)$ ,

(1) $[Y \in \mathfrak{R}(\varepsilon)$ and $RF(\varepsilon;\, x, Y) \neq \omega]$ implies $\lim_{\varepsilon \to 0} RF(\varepsilon;\, x, Y) = 0$,
and

(2) $[TF(\varepsilon;\, x, Y) \neq \omega$ for some $Y \in \mathfrak{R}(\varepsilon)]$ implies
$[\lim_{\varepsilon \to 0} TF(\varepsilon;\, x, y_{\varepsilon}) = 0$ as long as the $y_{\varepsilon}$'s are chosen by any reasonable at $x$ stopping criterion].

In order for $\Phi_{\text{convo}}$ and $\Phi_{\text{conto}}$ to work well on $\mathfrak{F} \in \mathbf{S}_{\omega}$, $\mathfrak{F}$ will have to approximate its $f$ uniformly over $o(a, b) \cap \mathfrak{M}$ in a sense to be defined. Otherwise, for example, $f$ may converge over $o(a, b)$, but for each $i$, $\mathfrak{F}$ may $\varepsilon_i$-diverge at $x$ for each $x \in \Theta(\varepsilon_i; a, b) - \Theta(\varepsilon_{i-1}; a, b)$. We would then have

(1) $x \in \mathfrak{M} \cap o(a, b) \Rightarrow [\mathfrak{F}$ $\varepsilon$-converges at $x$ for all sufficiently small $\varepsilon]$, and

(2) $\mathfrak{F}$ $\varepsilon$-diverges in $\Theta(\varepsilon; a, b)$ for all $\varepsilon$.

DEFINITION 4.5-3: We say $\mathfrak{F}$ approximates $f$ uniformly at $a, b$, written $\mathfrak{F} \overset{u}{\approx} f[a, b]$, precisely when there is a $\delta > 0$ such that for each $\varepsilon \leq \delta$ and every $(x, y) \in \Theta(\varepsilon; a, b)^{(2)}$ we have

(1) $f(x, y) \neq \omega \Rightarrow \mathfrak{F}(\varepsilon; x, y) \neq_{\varepsilon} \omega$, and

(2) $\lim\limits_{z \to x} f(x, z) \neq \omega \Rightarrow TF(\varepsilon; x, y) \neq \omega$.

Let $P$ be a set of pairs of real inputs. We say $P$ covers $o(a, b)^{(2)}$ precisely when, for each pair of numbers $(c, d) \in o(a, b)^{(2)}$ there is a pair $(x, y) \in P$ with $x = c$ and $y = d$. Let $Q$ be a set of real inputs. We say $Q$ covers $o(a, b) \cap \mathfrak{M}$ precisely when, for each $\varepsilon$ and each number $c \in o(a, b) \cap \mathbf{R}(\varepsilon)$ there is a real input $x \in Q$ with $X(\delta) = c$ and $RX(\delta) = 0$ for $\delta \leq \varepsilon$. Define

$$Q_o(\mathfrak{F}, f, P) \equiv \{(a, b): \ a \neq b, P \text{ covers } o(a, b)^{(2)}, \mathfrak{F} \overset{u}{\approx} f[a, b],$$

$$f(x, y) \neq \omega \text{ for } (x, y) \in o(a, b)^{(2)}, Q_{\lim}(\mathfrak{F}, f, P) \text{ covers } o(a, b) \cap \mathfrak{M}\}.$$

THEOREM 4.5-1: We have

$$(\Phi_{\text{convo}}, Q_o) \sim \phi_{\text{convo}}(S_o) \quad ,$$

$$(\Phi_{\text{conto}}, Q_o) \sim \phi_{\text{conto}}(S_o) \quad .$$

Proof: Suppose $\mathfrak{F} \approx f(P)$, $\mathfrak{F} \in S_o$ and $(a, b) \in Q_o(\mathfrak{F}, f, P)$ . Then corollary 4.4-2 and the fact that $f(x, y) \neq \omega$ for $(x, y) \in o(a, b)^{(2)}$ imply that $f$ converges over $o(a, b) - \mathfrak{M}$ . Suppose $f$ converges over the rest of $o(a, b)$ . Then there is a $\delta > 0$ such that, for each $\varepsilon \leq \delta$ and every $(x, y) \in \mathfrak{O}(\varepsilon; a, b)^{(2)}$, none of $F$, $RF$ and $TF$ equals $\omega$ at $(\varepsilon; x, y)$ . This means that $\mathfrak{F}$ $\varepsilon$-converges over $\mathfrak{O}(\varepsilon; a, b)$ for all $\varepsilon \leq \delta$ and so

$$(4.5\text{-}2) \qquad \lim_{\varepsilon\to 0} \Phi_{\text{convo}}(\mathfrak{F})(\varepsilon; a, b) = (\phi_{\text{convo}}(f)(a, b), \omega, \omega) \quad .$$

On the other hand, suppose there is an $x_o \in \mathfrak{O}(a, b) \cap \mathfrak{M}$ such that $f$ diverges at $x_o$. For all sufficiently small $\varepsilon$, $x_o \in \mathfrak{O}(\varepsilon; a, b)$ and $TF(\varepsilon; x_o, y) = \omega$ for all $y$ . This means that $\mathfrak{F}$ $\varepsilon$-diverges in $\mathfrak{O}(\varepsilon; a, b)$ for all sufficiently small $\varepsilon$, again implying (4.5-2).

Consider continuity. Corollary 4.4-2 implies that $f$ is continuous over $o(a, b) - \mathfrak{M}$ . Suppose $f$ is continuous over $o(a, b) \cap \mathfrak{M}$ also. By the uniformity assumption, for all $\varepsilon \leq \delta$ we have $\mathfrak{F}(\varepsilon; x, x) \neq_\varepsilon \omega$, $\underset{y\to x}{\mathrm{LIM}}\, \mathfrak{F}(\varepsilon; x, y) \neq_\varepsilon \omega$ and $\mathfrak{F}(\varepsilon; x, x) =_\varepsilon$

$\underset{y\to x}{\mathrm{LIM}}\, \mathfrak{F}(\varepsilon; x, y)$ for all $x \in \mathfrak{O}(\varepsilon; a, b)$, i.e., that $\mathfrak{F}$ is $\varepsilon$-continuous over $\mathfrak{O}(\varepsilon; a, b)$ . This implies

$$(4.5\text{-}3) \qquad \lim_{\varepsilon\to 0} (\Phi_{\text{conto}}(\mathfrak{F})(\varepsilon; a, b))_1 = \phi_{\text{conto}}(f)(a, b) \quad .$$

On the other hand, suppose there is an $x_o \in o(a, b) \cap \mathfrak{M}$ such that $f$ is discontinuous at $x_o$. This $x_o$ is in $\mathfrak{S}(\varepsilon; a, b)$ for all sufficiently small $\varepsilon$. By theorem 4.3-1, $\mathfrak{F}$ is $\varepsilon$-discontinuous at $x_o$ for all sufficiently small $\varepsilon$. Thus $\mathfrak{F}$ is $\varepsilon$-discontinuous in $\mathfrak{S}(\varepsilon; a, b)$ for all sufficiently small $\varepsilon$, again yielding (4.5-3). This completes the proof.

It is not difficult to generalize this to half closed and closed intervals, and to a definition of "$\mathfrak{F}$ $\varepsilon$-converges for $x_m \in \mathfrak{S}(\varepsilon; a, b)$ at $\bar{x}_{m-1}$" and "$\mathfrak{F}$ is $\varepsilon$-continuous for $x_m \in \mathfrak{S}(\varepsilon; a, b)$ at $\bar{x}_{m-1}$" for $\mathfrak{F}$ of $m + 1$ variables.

*REMARKS*: In the "$\varepsilon$-calculus of stable $\varepsilon$-functions" mentioned in the remarks at the end of chapter 2, an $\varepsilon$-limit $\varepsilon$-operator could be defined by using a particular, reasonable at any $x \neq \omega$, stopping criterion to define

$$\Phi'_{\lim}(F)(\varepsilon;\, x) = F(\varepsilon;\, x,\, y_{\varepsilon}) \quad ,$$

$$Q'_{\lim}(F,\, f,\, P) \equiv \{x\colon \lim_{y \to x} f(x,\, y) \neq \omega \Rightarrow F \text{ is stable at } x\} \quad .$$

However, it would not be possible to define "$\varepsilon$-comparison" relations, $<_{\varepsilon}$ and $=_{\varepsilon}$, satisfying theorem 4.2-1. It would not be possible to define "$\varepsilon$-convergence" for reasons mentioned in section 2.5. Due to the lack of "$\varepsilon$-comparison" and "$\varepsilon$-convergence", it would not be possible to define "$\varepsilon$-continuity" either. A better name for this "$\varepsilon$-calculus" would be "a model of scientific computation" because the model would still be strong enough to do basic computation, but the reliability of results would have to be checked outside of the model, by physical tests or by an error analysis.

It is interesting that theorem 4.4-1 would no longer hold in this model. When "$\mathcal{F} \approx f(P)$", any $\bar{x}_m \in P$ could be a point of discontinuity of $f$, provided $P$ does not contain all "real inputs" $\bar{y}_m$ equal in value to $\bar{x}_m$ or $P$ does not contain a neighborhood of $\bar{x}_m$. However, we would have

*THEOREM*: *Suppose* $I_1, \ldots, I_m$ $(m \geq 1)$ *are intervals contained in* $R$ . *Let* $I \equiv I_1 \times \ldots \times I_m$ . *Suppose* "$\mathcal{F} \approx f(I)$" *and* $f(\bar{x}_m)$ *is finite for* $\bar{x}_m \in I$ . *Then* $f$ *is continuous in* $I$ .

The second use of $I$ is as a set of $m$-tuples of "real inputs", under the convention in section 2.3. By continuous in $I$ we mean continuous with respect to limits taken from the interior of $I$.

<u>Proof</u>: For simplicity, we consider only the case $m = 1$. Let $y_1, y_2, \ldots$ be arbitrary numbers in $I$ approaching $x \in I$. Let $z_1$ be a "real input" with $z_1 = y_1$. Let $\eta_1$ be the largest value of $\eta$ such that $|Z_1(\varepsilon) - y_1| \leq 1$ for $\varepsilon \leq \eta$ and $|F(\eta; z_1) - f(y_1)| \leq 1$. Suppose $F$ uses only $Z_1(\varepsilon_1), Z_1(\varepsilon_2), \ldots, Z_1(\gamma_1)$ in evaluating $F(\eta_1; z_1)$ (see sec. 2.4). For $i = 2, 3, \ldots$ define $z_i$, $\eta_i$ and $\gamma_i$ as follows.
Let $z_i$ be a "real input" with $z_i = y_i$ and $Z_i(\varepsilon) = Z_{i-1}(\varepsilon)$ for $\gamma_{i-1} \leq \varepsilon$. Let $\eta_i$ be the largest value of $\eta$ such that

$\eta < \gamma_{i-1}$, $|Z_i(\varepsilon) - y_i| \leq i/i$ for $\varepsilon \leq \eta$, and $|F(\eta; z_i) - f(y_i)| \leq 1/i$. Suppose $F$ uses only $Z_i(\varepsilon_1), Z_i(\varepsilon_2), \ldots, Z_i(\gamma_i)$ in evaluating $F(\eta_i; z_i)$. Note that $\gamma_i \leq \eta_i < \gamma_{i-1}$. Define $W: \mathcal{E} \to \mathcal{M}$ by $W(\varepsilon) = Z_i(\varepsilon)$ where $i$ is such that $\eta_{i+1} < \varepsilon \leq \eta_i$ (or, if $\eta_1 \leq \varepsilon$ then $i = 1$).
For $\eta_{i+1} < \varepsilon \leq \eta_i$ we have

$$\begin{aligned} |W(\varepsilon) - x| = |Z_i(\varepsilon) - x| &\leq |Z_i(\varepsilon) - y_i| + |y_i - x| \\ &\leq 1/i + |y_i - x| . \end{aligned}$$

As $\varepsilon \to 0$ we have $|y_i - x| \to 0$ and so $|W(\varepsilon) - x| \to 0$. Hence $w \equiv W$ is a "real input" with value $x$. We have

$$\begin{aligned} |f(x) - f(y_i)| &\leq |f(x) - F(\eta_i; z_i)| + |F(\eta_i; z_i) - f(y_i)| \\ &\leq |f(x) - F(\eta_i; z_i)| + 1/i . \end{aligned}$$

But $Z_i(\epsilon) = W(\epsilon)$ for $\epsilon \geq \gamma_i$, so $F(\eta_i; z_i) = F(\eta_i; w)$, yielding

$$|f(x) - f(y_i)| \leq |f(x) - F(\eta_i; w)| + 1/i \quad .$$

As $i \to \infty$ we have $\eta_i \to 0$, yielding $f(y_i) \to f(x)$. This completes the proof.

# Chapter 5: More $\varepsilon$-Operators

Our final tasks are to define $\varepsilon$-derivative, $\frac{D}{Dx}(\mathcal{F})$, $\varepsilon$-integral, $\int Dt(\mathcal{F})$, and to prove the fundamental theorem of the $\varepsilon$-calculus, essentially that

$$\int Dt\left(\frac{D}{Dt}(\mathcal{F})\right)(\varepsilon;\ a,\ b) =_{\varepsilon} \mathcal{F}(\varepsilon;\ b) - \mathcal{F}(\varepsilon;\ a) \quad ,$$

where "-" here denotes $\varepsilon$-subtraction of $\varepsilon$-functions, and is defined below. Our definitions will be based on

$$\frac{d}{dx}f(x) = \lim_{y \to x} (f(x) - f(y))/(x-y) \quad ,$$

$$\int_a^b f(t)dt = \lim_{n \to \infty} \frac{b-a}{n} \sum_{j=1}^{n} f(a + j\frac{(b-a)}{n}) \quad .$$

For this, we will need $\varepsilon$-operators for <u>$\varepsilon$-arithmetic</u> $(\Phi_+, \Phi_-, \ldots)$, <u>$\varepsilon$-limit</u> $(\Phi_{lim})$, <u>$\varepsilon$-composition</u> $(\Phi^n_{comp})$, and <u>$\varepsilon$-recursion</u> $(\Phi_{rec})$. It is interesting to note that all these $\varepsilon$-operators except $\Phi_{lim}$ will work in a fixed precision; i.e., when $\Phi(\mathcal{F})(\varepsilon;\ \bar{x}_m)$ is being evaluated, only values of $\mathcal{F}$ at points $(\varepsilon;\ \bar{y}_m)$ are required by $\Phi$. If we were to define $\Phi_{lim}$ in terms of the S.C.## of section 3.4, then $\Phi_{lim}$ would have this property also. We will also need two input $\varepsilon$-functions, $\mathcal{J}^m_j$ and $C^m_k$, the identity and the constant $\varepsilon$-functions. As mentioned earlier (see sec. 2.7), we will only be able to give partial definitions of these $\varepsilon$-operators because we have no automatic procedure for generating stably convergent truncation-error bounds; probably no such procedure exists. (However, it may be possible to generate such bounds from a definition of the ideal function expressed in terms of the operators and initial functions of this chapter; this is a worthwhile research project.) We will assume truncation-error bounds to be given; for completeness, unspecified truncation-error bounds may be taken to be identically $\omega$ . The roundoff-error bounds which we generate will be of the per step variety (as seen in interval analysis). Such bounds are notoriously inefficient in real situations. If better bounds are available, they can be used in place of our automatically generated bounds. (It may be possible to automatically improve such automatically generated bounds if we are given a definition of the ideal function under consideration in terms of the operators and initial functions of this chapter.)

For the following sections, we need definition 4.1-1 for TF of $m + 1 \geq 2$ variables (it was stated for TF of 2 variables).

*DEFINITION 5-1*: *Suppose* $\mathfrak{F} \approx f(P)$ . *We say TF is stably convergent at* $\bar{x}_m$ *relative to f* $(m \geq 1)$ *precisely when*

$$\lim_{y \to \bar{x}_m} f(\bar{x}_m, y) \neq \omega \Rightarrow \lim_{\varepsilon \to 0} TF(\varepsilon; \bar{x}_m, y_\varepsilon) = 0 \quad ,$$

*as long as the* $y_\varepsilon$*'s* *are chosen by a reasonable at* $x_m$ *stopping criterion. We always say* TF *of one variable are stably convergent at* $\bar{x}_0$ *relative to* f .

We will use the notation,

$$TF \downarrow \bar{x}_m(f) \quad \text{or} \quad TF \downarrow P(f) \quad ,$$

to denote that TF is stably convergent at $\bar{x}_m$, or at all $\bar{x}_m \in P$, relative to f .

In the following, we will need the mapping, $V: \mathcal{M}^{(3)} \rightarrow \mathcal{M}^{(2)}$, given by

$$V[(a, b, c)] \equiv (a, b) \quad .$$

## 5.1 Identity $\varepsilon$-Functions

For $1 \leq j \leq m$ and $x_i \neq \omega$ $(i = 1, \ldots, m)$, define the identity (ideal) functions of $m$ variables by

$$i_j^m(\bar{x}_m) = x_j \quad .$$

For $1 \leq j \leq m$, define

$$\mathcal{J}_j^m(\varepsilon;\, \bar{x}_m) \equiv (X_j(\varepsilon),\, RX_j(\varepsilon),\, TI_j^m(\varepsilon;\, \bar{x}_m)) \quad ,$$

as long as no $x_i =_\varepsilon \omega$, where $TI_j^m$ is to be defined. Of course $TI_1^1$ is identically $\omega$. For $1 \leq j < m$, $TI_j^m(\varepsilon;\, \bar{x}_m) = 0$ so long as no $x_i =_\varepsilon \omega$. For $m \geq 2$, define $TI_m^m$ by

$$TI_m^m(\varepsilon;\, \bar{x}_m) = (RI_{m-1}^m \mathbin{\hat{+}} RI_m^m \mathbin{\hat{+}} |I_{m-1}^m \mathbin{\overset{\frown}{-}} I_m^m|)(\varepsilon;\, \bar{x}_m) \quad .$$

THEOREM 5.1-1: $\quad \mathcal{J}_j^m \approx i_j^m(\tilde{R}^{(m)})$ and $TI_j^m \downarrow \tilde{R}^{(m-1)}(i_j^m) \quad .$

Proof: The only thing requiring proof is that $TI_m^m (m \geq 2)$ is stably convergent. But this follows immediately from theorem 2.8-1, completing the proof.

## 5.2 Constant $\varepsilon$-Functions

For any real input $k$ define the <u>constant $k$ ideal function of $m$ variables</u> $(m \geq 1)$ by

$$c_k^m(\bar{x}_m) = k \quad ,$$

so long as no $x_i = \omega$ . Define $\varepsilon$-functions $\mathbf{C}_k^m$ by

$$\mathbf{C}_k^1(\varepsilon;\, x_1) \equiv (I_1^2(\varepsilon;\, k,\, x_1),\, RI_1^2(\varepsilon;\, k,\, x_1),\, \omega) \quad ,$$

$$\mathbf{C}_k^m(\varepsilon;\, \bar{x}_m) \equiv \mathcal{S}_1^{m+1}(\varepsilon;\, k,\, \bar{x}_m) \qquad \text{for} \quad m \geq 2 \quad .$$

THEOREM 5.2-1: $\quad \mathbf{C}_k^m \approx c_k^m(\tilde{R}^{(m)})$ and $TC_k^m \downarrow \tilde{R}^{(m-1)}(c_k^m)$ .

This follows immediately from theorem 5.1-1.

## 5.3 ε-Arithmetic

For $*$ being $+$, $-$, $\times$, $\div$, define operators $\not{b}_*$ by

$$\not{b}_*(f, g)(\bar{x}_m) = f(\bar{x}_m) * g(\bar{x}_m) \qquad (m \geq 1) \quad .$$

We define ε-operators corresponding to these by first defining ε-arithmetic for machine numbers and corresponding ε-arithmetic roundoff-error bounds.

ε-Arithmetic subroutines, $FL_*$ for $*$ being $+$, $-$, $\times$, $\div$, are subroutines of two variables which approximate ideal arithmetic. Let $N_\varepsilon(a)$ be the ε-neighborhood of $a \in \tilde{R}$, as defined in section 2.8. The $FL_*$ must satisfy

(1) for $x \neq_\varepsilon \omega$ and $y \neq_\varepsilon \omega$, $FL_*(\varepsilon;\, x, y) = FL_*(\varepsilon;\, X(\varepsilon), Y(\varepsilon))$,

and

(2) for any $a, b \in \mathbf{R}(\varepsilon)$, $FL_*(\varepsilon;\, a, b)$ is in $N_\varepsilon(a * b)$ .

For example, the rounding subroutines $A_{n,*}$ of section 2.8 satisfy these. Condition (1) states that the $FL_*$ do not use the inputted error bounds. Condition (2) requires that, when $FL_*$ operates at ε-precision on members of $\mathbf{R}(\varepsilon)$, it must get an answer within two machine numbers from the correct answer, unless the correct answer is in $\{-\infty, \infty, \omega\}$, in which case $FL_*$ must get the correct answer.

Define a function, $w$: $\{(\varepsilon, X)\colon \varepsilon \in \mathcal{E}, X$ finite and in $\mathbf{R}(\varepsilon)\} \to \mathfrak{M}$, by

$$w(\varepsilon, X) = \max(|X - Y_1|, |X - Y_2|) \quad ,$$

where $Y_1$ is the second member of $\mathbf{R}(\varepsilon)$ below $X$ (or the first below $X$ if there is only one) and $Y_2$ is the second member of $\mathbf{R}(\varepsilon) - \{\omega\}$

above X (or the first above X if there is only one). For $X \in \mathfrak{R}(\varepsilon)$ , let $\overset{\frown}{|X|}$ denote $|X \hat{-} 0|$ and $\underset{\smile}{|X|}$ denote $|X \underset{\smile}{-} 0|$ . Define error bounds, $RFL_*$, for the $FL_*$, by

$$(5.3\text{-}1) \qquad R_{\pm}(\varepsilon; x, y) = (RX \hat{+} RY)(\varepsilon) ,$$

$$(5.3\text{-}2) \qquad R_{\times}(\varepsilon; x, y) = ( \overset{\frown}{|X|} \hat{\times} RY \hat{+} \overset{\frown}{|Y|} \hat{\times} RX \hat{+} RX \hat{\times} RY)(\varepsilon) ,$$

$$(5.3\text{-}3) \qquad R_{\div}(\varepsilon; x, y) = \begin{cases} \omega & \text{if } (\underset{\smile}{|Y|} \underset{\smile}{-} RY)(\varepsilon) \leq 0 , \\ ((RY \hat{\times} \overset{\frown}{|X|} \hat{\div} \underset{\smile}{|Y|} \hat{+} RX) \hat{\div} (\underset{\smile}{|Y|} \underset{\smile}{-} RY))(\varepsilon) & \text{otherwise,} \end{cases}$$

$$(5.3\text{-}4) \quad RFL_*(\varepsilon; x, y) = \begin{cases} \omega & \text{if } FL_*(\varepsilon; x, y) = \omega \text{ or } RX(\varepsilon) = \omega \\ & \text{or } RY(\varepsilon) = \infty \text{ or } [FL_*(\varepsilon; x, y) = \pm\infty \\ & \text{and } -\infty <_{\varepsilon} \frac{x}{y} <_{\varepsilon} \infty] \\ 0 & \text{if } FL_*(\varepsilon; x, y) = \pm\infty \\ w(\varepsilon, FL_*(\varepsilon, x, y)) \hat{+} R_*(\varepsilon; x, y) & \text{otherwise .} \end{cases}$$

We define $\varepsilon$-arithmetic $\varepsilon$-operators from the $FL_*$ and $RFL_*$ as follows. Let $S_{\text{arith}}$ be the set of pairs of $\varepsilon$-functions both of $m \geq 1$ variables. When $\mathfrak{F} \approx f(P)$ , let $f(\bar{x}_m)$ denote (as well as its numeric value) the poor real input, $(F(\cdot; \bar{x}_m), RF(\cdot; \bar{x}_m))$ . Suppose $\bar{\mathfrak{F}}_2 \in S_{\text{arith}}$ and $\bar{\mathfrak{F}}_2 \approx \bar{f}_2(\bar{P}_2)$ . Define

$$V[\Phi_*(\bar{\mathfrak{F}}_2)(\varepsilon; \bar{x}_m)] = (FL_*, RFL_*)(\varepsilon; f_1(\bar{x}_m), f_2(\bar{x}_m)) \quad .$$

For $m \geq 2$ , we assume the third part of $\Phi_*(\bar{\mathcal{F}}_2)$ to be given. We will abbreviate $\phi_*(f, g)$ by $f*g$ and $\Phi_*(\mathcal{F}, \mathcal{G})$ by $\mathcal{F}*\mathcal{G}$ . Also, we let $-\mathcal{F}$ denote $C_0^m - \mathcal{F}$ . Define

$$Q_{arith}(\bar{\mathcal{F}}_2, \bar{f}_2, \bar{P}_2) \equiv P_1 \cap P_2 \quad .$$

<u>THEOREM 5.3-1</u>: <u>For</u> * <u>being</u> +, -, ×, ÷ , <u>we</u> <u>have</u>

$$(\Phi_*, Q_{arith}) \approx \phi_*(S_{arith}) \quad .$$

<u>Proof:</u> It suffices to prove that, for each $\varepsilon$ and any real inputs $x$ and $y$ ,

$$RFL_*(\varepsilon; x, y) \geq |FL_*(\varepsilon; x, y) - (x * y)| \quad , \tag{5.3-5}$$

and that

$$x * y \neq \omega \Rightarrow \lim_{\varepsilon \to 0} RFL_*(\varepsilon; x, y) = 0 \quad . \tag{5.3-6}$$

Let $\varepsilon$, $x$ and $y$ be given. If $RX(\varepsilon)$ or $RY(\varepsilon)$ is $\infty$ or $FL_*(\varepsilon; x, y) = \omega$ or $[FL_*(\varepsilon; x, y) = \pm\infty$ and $-\infty <_\varepsilon x <_\varepsilon \infty$ and $-\infty <_\varepsilon y <_\varepsilon \infty]$ then (5.3-5) holds because $RFL_*(\varepsilon; x, y) = \omega$ . If $RX(\varepsilon)$ and $RY(\varepsilon)$ are finite, $|x| = \infty$ or $|y| = \infty$ , and $FL_*(\varepsilon; x, y) = \pm\infty$ then $FL_*(\varepsilon; x, y) = X(\varepsilon) * Y(\varepsilon) = x * y$ by conditions (1) and (2) on the $FL_*$ ; in this case (5.3-5) reduces to $0 \geq 0$ , which is true. Suppose $FL_*(\varepsilon; x, y)$ , $RX(\varepsilon)$ and $RY(\varepsilon)$ are finite. Then either $x, y$ and $x * y$ are finite or $x * y \equiv x \div \infty = 0$ . By the triangle inequality, we have

$$|FL_*(\varepsilon; x, y) - (x * y)| \leq |FL_*(\varepsilon; x, y) - (X(\varepsilon) * Y(\varepsilon))| + |X(\varepsilon) * Y(\varepsilon) - (x * y)| \quad . \tag{5.3-7}$$

By conditions (1) and (2) on $FL_*$ , we have

$$|FL_*(\varepsilon; x, y) - (X(\varepsilon) * Y(\varepsilon))| \leq w(\varepsilon, FL_*(\varepsilon; x, y)) \quad .$$

For the second term on the right side of (5.3-7), we have

$$|X(\varepsilon) \pm Y(\varepsilon) - (x \pm y)| \leq RX(\varepsilon) + RY(\varepsilon) \leq R_{\pm}(\varepsilon; x, y) \quad ,$$

$$|X(\varepsilon) \times Y(\varepsilon) - x \times y| \leq |X(\varepsilon) \times Y(\varepsilon) - X(\varepsilon) \times y| + |X(\varepsilon) \times y - x \times y|$$

$$\leq |X(\varepsilon)| \times RY(\varepsilon) + RX(\varepsilon) \times |y| \leq R_{\times}(\varepsilon; x, y) \quad .$$

If $*$ is $\div$ and $|y| = \infty$ then the above assumptions imply that $x$ is finite and $x \div y = 0$ ; in this case, $R_{\div}(\varepsilon; x, y)$ is either $\infty \div \infty = \omega$ or it is (some finite number) $\div \infty = 0$ , taking the latter value for all sufficiently small $\varepsilon$ , so (5.3-5) holds in this case. Suppose $y$ is finite. Then

$$|X(\varepsilon) \div Y(\varepsilon) - x \div y| = |(y \times X(\varepsilon) \div Y(\varepsilon) - x) \div y|$$

$$= |(y - Y(\varepsilon)) \times X(\varepsilon) \div Y(\varepsilon) + X(\varepsilon) - x| \div |y|$$

$$\leq (RY(\varepsilon) \times |X(\varepsilon) \div Y(\varepsilon)| + RX(\varepsilon)) \div |y| \leq R_{\div}(\varepsilon; x, y) \quad .$$

Thus (5.3-5) holds in all cases.

As in theorem 2.8-1, it follows that, for $x * y \neq \omega$ ,

$$\lim_{\varepsilon \to 0} FL_*(\varepsilon; x, y) = x * y \quad ,$$

(5.3-8)

$$|x * y| = \infty \Rightarrow [FL_*(\varepsilon; x, y) = x * y \text{ for all sufficiently small } \varepsilon] \quad .$$

This implies that, for $x * y$ being finite,

$$\lim_{\varepsilon\to 0} w(\varepsilon,\ FL_*(\varepsilon;\ x,\ y)) = 0 \quad .$$

For such $x * y$ , it follows from theorem 2.8-1 that

$$\lim_{\varepsilon\to 0} R_*(\varepsilon;\ x,\ y) = 0 \quad ,$$

and (5.3-6) follows. When $|x * y| = \infty$ , (5.3-6) follows from (5.3-4) and (5.3-8). This completes the proof.

We say that *f is rational* if it can be defined from the $i_j^m$ and $c_k^m$ by a finite number of arithmetic operations. We say that *$\mathfrak{F}$ is rational* if it can be defined from the $\mathcal{I}_j^m$ and $\mathbf{C}_k^m$ by a finite number of $\varepsilon$-arithmetic $\varepsilon$-operations.

COROLLARY 5.3-1: *Let $\mathfrak{F}$ be the rational $\varepsilon$-function whose definition corresponds to that of the rational function f . Then*

$$\mathfrak{F} \approx f(\tilde{R}^{(m)}) \quad .$$

This follows from theorems 5.1-1, 5.2-1 and 5.3-1 by a simple induction argument, which we omit. For example, this means that $\mathfrak{F} \equiv (\mathcal{I}_1^2 + \mathbf{C}_1^2 - (\mathcal{I}_2^2 + \mathbf{C}_1^2)) \div (\mathcal{I}_1^2 - \mathcal{I}_2^2)$ corresponds to the f of example 3.1-2 over $\tilde{R}^{(2)}$ .

## 5.4 ε-Limit

Here we generalize the definitions of section 4.1. Let $S_{lim}$ be the set of all ideal functions of 2,3,... variables. Define an operator $\phi_{lim}$ over $S_{lim}$ by

$$\phi_{lim}(f)(\bar{x}_m) = \lim_{y \to x_m} f(\bar{x}_m, y) \quad .$$

Assume a $\lambda(\cdot)$ as in section 4.1 and an effective, reasonable at any $x \neq \omega$ , stopping criterion, S.C., have been given. Let $S_{lim}$ be the set of all ε-functions of 2,3,... variables such that if $\mathfrak{F} \in S_{lim}$ then for each $\bar{x}_m \in \bar{R}^{(m)}$ and each $\varepsilon$

(1) $[Y \in \mathcal{R}(\varepsilon)$ and $RF(\varepsilon; \bar{x}_m, Y) \neq \omega]$ implies $\lim_{\varepsilon \to 0} RF(\varepsilon; \bar{x}_m, Y) = 0$ ,

and

(2) $[TF(\varepsilon; \bar{x}_m, Y) \neq \omega$ for some $Y \in \mathcal{R}(\varepsilon)]$ implies

$[\lim_{\varepsilon \to 0} TF(\varepsilon; \bar{x}_m, y_\varepsilon) = 0$ , as long as the $y_\varepsilon$ 's are chosen by a reasonable at $x_m$ stopping criterion] .

Define $\Phi_{lim}$ by

DEFINITION 5.4-1: Let $\mathfrak{F} \in S_{lim}$ , $\bar{x}_m$ and $\varepsilon$ be given. Let $y_{\varepsilon_1}, y_{\varepsilon_2}, \ldots$ denote the values selected by S.C. for TF and $\bar{x}_m$ . If $TF(\varepsilon; \bar{x}_m, y_\varepsilon) = \omega$ , define

$$V[\Phi_{lim}(\mathfrak{F})(\varepsilon; \bar{x}_m)] \equiv (\omega, \omega) \quad . \tag{5.4-1}$$

Otherwise, let $\delta$ be the largest member of $\mathcal{E}$ such that $\delta \leq \varepsilon$ and $TF(\delta; \bar{x}_m, y_\delta) \leq \lambda(\varepsilon)$ . If $RF(\delta; \bar{x}_m, y_\delta) = \omega$ , apply (5.4-1).

*Otherwise let* $j$ *be the smallest integer such that* $\varepsilon_j \leq \delta$ *and* $RF(\varepsilon_j; \bar{x}_m, y_\delta) \leq \lambda(\varepsilon)$ . *Suppose* $F(\varepsilon_j; \bar{x}_m, y_\delta)$ *is* $\langle\alpha_{\mathbb{R}}(j, k, \cdot)\rangle$ . *Let* $\beta_\varepsilon(\cdot)$ *be* $\alpha_{\mathbb{R}}(j, k, \cdot)$ . *Define*

$$V[\Phi_{\text{lim}}(\mathcal{F})(\varepsilon;\bar{x}_m)] \equiv (\hat{I}(\varepsilon,\beta_\varepsilon), 2 \,\hat{\times}\, \lambda(\varepsilon) \,\hat{+}\, |\hat{I}(\varepsilon,\beta_\varepsilon) \,\hat{-}\, \check{I}(\varepsilon,\beta_\varepsilon)|) \quad .$$

For $m \geq 2$ we assume the third part of $\Phi_{\text{lim}}(\mathcal{F})$ to be given. Define $Q_{\text{lim}}$ by

$$Q_{\text{lim}}(\mathcal{F},f,P) \equiv \{\bar{x}_m : \{\bar{x}_m\} \times (\mathcal{M} \cap \{\text{some neighborhood of } x_m\} \subset P \text{ and } TF \downarrow \bar{x}_m(f)\} \quad .$$

Theorem 4.1-1 generalizes immediately to

*THEOREM 5.4-1: We have*

$$(\Phi_{\text{lim}}, Q_{\text{lim}}) \approx \phi_{\text{lim}}(S_{\text{lim}}) \quad .$$

## 5.5 $\varepsilon$-Composition

For $n = 1, 2, \ldots,$ let $S^n_{\text{comp}}$ be the set of all $(n+1)$ - tuples $(f, \bar{g}_n)$, of ideal functions, where $f$ takes $n$ variables and each $g_i$ takes $m$ variables (some $m \geq 1$) . Define a composition operator $\phi^n_{\text{comp}}$ over $S^n_{\text{comp}}$ by

$$\phi^n_{\text{comp}}(f, \bar{g}_n)(\bar{x}_m) = \overline{f(g_n(\bar{x}_m))} \quad .$$

We abbreviate $\phi^n_{\text{comp}}(f, \bar{g}_n)$ by $f(\overline{g_n})$ . (In context, it will be clear whether the $g_i$ are functions or variables.) We will also use $\phi_+(f, g)(\bar{g}_n)$ or $(f + g)(\bar{g}_n)$ interchangeably with $\phi^n_{\text{comp}}(f + g, \bar{g}_n)$ , etc.

For the present, let $g_i(\bar{x}_m)$ denote (together with its numeric value) the poor real input, $(G_i(\cdot\,;\bar{x}_m)$ , $RG_i(\cdot\,;\bar{x}_m))$ . Define

$$V[\Phi^n_{\text{comp}}(\mathcal{F}, \bar{\mathcal{G}}_n)(\varepsilon;\bar{x}_m)] \equiv V[\mathcal{F}(\varepsilon;\, \overline{g_n(\bar{x}_m)})] \quad .$$

Let $\mathbf{S}^n_{\text{comp}}$ be the set of all $(n + 1)$ - tuples $(\mathcal{F}, \bar{\mathcal{G}}_n)$ , of $\varepsilon$-functions such that $(\mathcal{F}, \bar{\mathcal{G}}_n) \approx (f, \bar{g}_n)(P, \bar{P}_n)$ for some $(f, \bar{g}_n) \in S^n_{\text{comp}}$ and some $P, \bar{P}_n$ , and such that the computation of $\mathcal{F}(\varepsilon;\overline{g_n(\bar{x}_m)})$ via the determiners of $\mathcal{F}$ and $\bar{\mathcal{G}}_n$ halts for any real inputs $\bar{x}_m$ (see section 2.4). We assume the third part of $\Phi^n_{\text{comp}}(\mathcal{F}, \bar{\mathcal{G}}_n)$ to be given. We will abbreviate $\Phi^n_{\text{comp}}(\mathcal{F}, \bar{\mathcal{G}}_n)$ by $\mathcal{F}(\bar{\mathcal{G}}_n)$ . Define $Q^n_{\text{comp}}$ by

$$Q^n_{\text{comp}}(\mathcal{F}, \bar{\mathcal{G}}_n, f, \bar{g}_n, P, \bar{P}_n) \equiv \{\bar{x}_m : \bar{x}_m \in \bigcap_{i=1}^{n} P_i \text{ and } \overline{g_n(\bar{x}_m)} \in P\}$$

(Note that $\overline{g_n(\bar{x}_m)}$ again denotes poor real inputs, as explained above. And $\overline{g_n(\bar{x}_m)}$ cannot be in $P$ unless each $g_i(\bar{x}_m)$ is a real input.)

THEOREM 5.5-1: We have

$$(\Phi^n_{\text{comp}}, Q^n_{\text{comp}}) \approx \phi^n_{\text{comp}}(S^n_{\text{comp}}) \quad .$$

## 5.6 ε-Recursion

We are interested in the following form of recursion. Let $g_\infty, g_0$ and $h$ be given ideal functions of $m$, $m$ and $m+1$ variables $(m \geq 1)$. We define $f$ by recursion as follows:

$$(5.6\text{-}1) \quad f(\bar{x}_m) = \begin{cases} g_\infty(\bar{x}_m) & \text{if } x_m = \infty \\ g_0(\bar{x}_m) & \text{if } x_m < 1 \\ h(\bar{x}_m, f(\bar{x}_{m-1}, x_m - 1)) & \text{otherwise .} \end{cases}$$

We put this in operator form by defining

$$\phi_{rec}(g_\infty, g_0, h) = f \quad ,$$

where $f$ is as in (5.6-1), and defining $S_{rec}$ to be the corresponding set of $(g_\infty, g_0, h)$. The following example illustrates the use of this recursion. The operators of this example will be used later.

EXAMPLE 5.6-1: Define the operators $\sigma_+$ and $\sigma_\times$ over the $S_{lim}$ of section 5.4 by

$$\sigma_+(g)(\bar{x}_m) = \sum_{i=0}^{[x_m - 1]} g(\bar{x}_m, x_m - i) \quad ,$$

$$\sigma_\times(g)(\bar{x}_m) = \prod_{i=0}^{[x_m - 1]} g(\bar{x}_m, x_m - i) \quad ,$$

where the empty sum is defined to be $c_0^m(\bar{x}_m)$, the empty product is $c_1^m(\bar{x}_m)$, and $\sum_{i=0}^{\infty}$ and $\prod_{i=0}^{\infty}$ are $\equiv \omega$. Let $g_+$ be $c_0^{m+1}$ and $g_\times$ be $c_1^{m+1}$. For $*$ being $+$ and $\times$, we have

$$\sigma_*(g) \equiv \phi_{rec}(C_\omega^{m+1}, g_* g(\overline{I_{m+1}^{m+2}}) * I_{m+2}^{m+2})(\overline{I_m^m}, I_m^m) \quad .$$

Both $\sigma_+$ and $\sigma_x$ will be used in section 5.7, where we define an $\varepsilon$-function which corresponds to $e^x$ over $\tilde{R}$ .

Let $\mathscr{S}_\infty, \mathscr{S}_0, \mathscr{N}$ be given, and define the F part $\mathcal{F} \equiv \Phi_{rec}(\mathscr{S}_\infty, \mathscr{S}_0, \mathscr{N})$ by letting $\mathscr{S}_1$ denote $\mathscr{N}(\mathscr{S}^m_m, \mathcal{F}(\overline{\mathscr{S}^m_{m-1}}), \mathscr{S}^m_m - C^m_1))$ and equating

$$(5.6\text{-}2) \qquad F(\varepsilon;\bar{x}_m) = \begin{cases} G_\infty(\varepsilon;\bar{x}_m) & \text{if} \quad \mathscr{S}^m_m(\varepsilon;\bar{x}_m) =_\varepsilon \infty \\ G_0(\varepsilon;\bar{x}_m) & \text{if} \quad \mathscr{S}^m_m(\varepsilon,\bar{x}_m) <_\varepsilon 1 \\ G_1(\varepsilon;\bar{x}_m) & \text{if} \quad (\mathscr{S}^m_m - C^m_1)(\varepsilon;\bar{x}_m) <_\varepsilon \mathscr{S}^m_m(\varepsilon;\bar{x}_m) \\ \omega & \text{otherwise} \quad . \end{cases}$$

The third test is needed because it can happen that $\mathscr{S}^m_m(\varepsilon,\bar{x}_m) <_\varepsilon \infty$ but $(\mathscr{S}^m_m - C^m_1 - C^m_1 - \ldots - C^m_1)(\varepsilon;\bar{x}_m) \not<_\varepsilon 1$ no matter how many $C^m_1$ 's are $\varepsilon$-subtracted from $\mathscr{S}^m_m$ . Thus the evaluation of $F(\varepsilon,\bar{x}_m)$ via (5.6-2) with the third test replaced by "otherwise" (and the fourth alternative removed) would not halt for certain $\varepsilon$ and $\bar{x}_m$ . However, evaluation of $F(\varepsilon;\bar{x}_m)$ via (5.6-2) will always halt.

For the following definition of RF , we will need an $\varepsilon$-comparison operator, $\leq_\varepsilon$ . We want $S(\varepsilon;\bar{x}_m) \leq T(\varepsilon;\bar{y}_n)$ to hold when, based only on information given by $S(\varepsilon;\bar{x}_m)$ and $T(\varepsilon;\bar{y}_n)$ , $s(\bar{x}_m)$ _must_ be $\leq t(\bar{y}_n)$ .

_DEFINITION 5.6-1._ _Let_ x _and_ y _be poor real inputs. Define_ $x \leq_\varepsilon y$ _to be true (and_ $x \not\leq_\varepsilon y$ _to be false) precisely when_ $X(\varepsilon) \leq Y(\varepsilon)$ , $RX(\varepsilon)$ _and_ $RY(\varepsilon)$ _are finite, and_

$$|X(\varepsilon) \mathbin{\tilde{-}} Y(\varepsilon)| \geq RX(\varepsilon) \mathbin{\hat{+}} RY(\varepsilon) \quad .$$

We adopt the convention in section 4.2 concerning the use of triples, $\bar{a}_3$, with $\underline{<}_\varepsilon$ .

Define the subroutine ER by

$$(5.6\text{-}3) \qquad \mathrm{ER}(\varepsilon;\bar{x}_m) = (\mathrm{RG}_0 \hat{+} \mathrm{RG}_1 \hat{+} |G_0 \hat{-} G_1|)(\varepsilon;\bar{x}_m) \ .$$

ER bounds the error caused by using $G_j$ in place of $g_k$ for $(j,k) \in \{(0,0), (0,1), (1,0), (1,1)\}$ . RF is defined by

$$(5.6\text{-}4) \quad \mathrm{RF}(\varepsilon;\bar{x}_m) = \begin{cases} \mathrm{RG}_\infty(\varepsilon;\bar{x}_m) & \text{if } \infty \underline{<}_\varepsilon \mathcal{J}^m_m(\varepsilon;\bar{x}_m) \\ \mathrm{RG}_0(\varepsilon;\bar{x}_m) & \text{if } \mathcal{J}^m_m(\varepsilon;\bar{x}_m) <_\varepsilon 1 \\ \mathrm{RG}_1(\varepsilon;\bar{x}_m) & \text{if } 1 \underline{<}_\varepsilon \mathcal{J}^m_m(\varepsilon;\bar{x}_m) \text{ and} \\ & (\mathcal{J}^m_m - C^m_1)(\varepsilon;\bar{x}_m) <_\varepsilon \mathcal{J}^m_m(\varepsilon,\bar{x}_m) \\ \mathrm{ER}(\varepsilon;\bar{x}_m) & \text{if } (\mathcal{J}^m_m - C^m_1)(\varepsilon;\bar{x}_m) <_\varepsilon \mathcal{J}^m_m(\varepsilon;\bar{x}_m) \\ \omega & \text{otherwise .} \end{cases}$$

As usual, we assume TF to be given. Let $N_m$ be given by

$$N_m \equiv \{\bar{x}_m: \ x_m \text{ is a positive integer and no } x_i = \omega\} \ .$$

If $m = 1$ , let $S \equiv \{\bar{x}_0\}$ ; otherwise let $S$ be some subset of $\tilde{R}^{(m-1)}$ . Let $T \equiv S \times (\tilde{R} - \{\infty\})$ , where $\{\bar{x}_0\} \times P$ is defined to be $P$ , for any set $P$ .

For $P_1 \equiv S \times \{\infty\}$ and $P_2 \equiv T$ define $Q_{rec}$ by

$$Q_{rec}(\bar{\mathcal{G}}_3, \bar{g}_3, \bar{P}_3) \equiv \begin{cases} S \times \tilde{R} - N_m & \text{if } P_3 \equiv (T - N_m) \times \tilde{R} \\ S \times \tilde{R} & \text{if } P_3 \equiv T \times \tilde{R} \text{ and } g_0(\bar{x}_{m-1}, 1) = \\ & \quad h(\bar{x}_{m-1}, 1, y) \text{ for any } \bar{x}_{m-1} \in S \text{ and} \\ & \qquad \text{any } y \neq \omega \\ \{\,\} & \text{otherwise} . \end{cases}$$

For $P_1$ and $P_2$ not of this form, let $Q_{rec}$ be $\{\,\}$.

Let $S_{rec}$ be the set of all $\bar{\mathcal{F}}_3 \approx \bar{f}_3(\bar{P}_3)$, for some $\bar{f}_3 \in S_{rec}$ and some $\bar{P}_3$, such that the computation of $\Phi_{rec}(\bar{\mathcal{F}}_3)(\varepsilon;\bar{x}_m)$ via the determiners of $\bar{\mathcal{F}}_3$ halts for any real inputs $\bar{x}_m$ (see section [illegible]).

THEOREM 5.6-1: We have

$$(\Phi_{rec}, Q_{rec}) \approx \rho_{rec}(S_{rec})$$

Proof: Suppose $T \equiv S \times (\tilde{R} - \{\infty\})$, $\mathcal{G}_\infty \approx g_\infty(S \times \{\infty\})$, $\mathcal{G}_0 \approx g_0(1)$, $\mathcal{H} \approx h((T - N_m) \times \tilde{R})$, $(\mathcal{G}_\infty, \mathcal{G}_0, \mathcal{H}) \in S_{rec}$ and $(g_\infty, g_0, h) \in S_{rec}$. Let $f \equiv \rho_{rec}(g_\infty, g_0, h)$ and $\mathcal{F} \equiv \Phi_{rec}(\mathcal{G}_\infty, \mathcal{G}_0, \mathcal{H})$. Suppose $\bar{x}_m \in S \times \bar{R} - N_m$ and $f(\bar{x}_m) \neq \omega$. If $x_m < 1$ then $f(\bar{x}_m) = g_0(\bar{x}_m)$ and $V[\mathcal{F}(\varepsilon;\bar{x}_m)] \equiv V[\mathcal{G}_0(\varepsilon;\bar{x}_m)]$ for all sufficiently small $\varepsilon$, yielding convergence. If $x_m = \infty$ then $f(\bar{x}_m) = g_\infty(\bar{x}_m)$ and $V[\mathcal{F}(\varepsilon;\bar{x}_m)] \equiv V[\mathcal{G}_\infty(\varepsilon;\bar{x}_m)]$ for all sufficiently small $\varepsilon$, again yielding convergence. Suppose $1 < x_m < \infty$. Let $\mathcal{I}_{x_m}$ and $\mathcal{G}_{x_m}$ be given by

$$\mathcal{J}_{x_m} \equiv \mathcal{J}^m_m - \underbrace{C^m_1 - \ldots - C^m_1}_{[x_m-1]} \quad ,$$

$$\mathscr{J}_{x_m} \equiv \mathscr{K}(\overline{\mathcal{J}^m_{m-1}}, \mathcal{J}_{x_m}, \mathscr{J}_C(\overline{\mathcal{J}^m_{m-1}}, \mathcal{J}_{x_m} - C^m_1)) \quad .$$

For all sufficiently small $\varepsilon$ we have

$$\mathcal{J}_{x_m}(\varepsilon;\bar{x}_m) \not<_\varepsilon 1 \ , \quad (\mathcal{J}_{x_m} - C^m_1)(\varepsilon;\bar{x}_m) <_\varepsilon 1 \quad ,$$

and so

$$(5.6\text{-}5) \quad V[\mathscr{F}(\varepsilon;\bar{x}_m)] \equiv V[\mathscr{K}(\bar{\mathcal{J}}^m_m, \mathscr{K}(\bar{\mathcal{J}}^m_{m-1}, \mathcal{J}^m_m - C^m_1, \ldots \mathscr{J}_{x_m} \ldots))(\varepsilon;\bar{x}_m)] \quad .$$

We also have

$$(5.6\text{-}6) \quad f(\bar{x}_m) = h(\bar{x}_m, h(\bar{x}_{m-1}, x_m-1, \ldots h(\bar{x}_{m-1}, x_m - [x_m-1], g_0(\bar{x}_{m-1}, x_m - [x_m])) \ldots )) \quad .$$

Successive applications of corollary 5.3-1 and theorem 5.5-1, working from the inside to the outside on (5.6-5), give us convergence.

In addition to the above hypotheses, suppose $\mathscr{K} \approx h(T \times \tilde{R})$ and $g_0(\bar{x}_{m-1},1) = h(\bar{x}_{m-1},1,y)$ for any $\bar{x}_{m-1} \in S$ and any $y \neq \omega$ . From the above, we have $\mathscr{F} \approx f(S \times \tilde{R} - N_m)$ . Suppose $\bar{x}_m \in S \times \tilde{R} \cap N_m$ and $f(\bar{x}_m) \neq \omega$ . Define $\mathscr{J}'_{x_m}$ by

$$\mathscr{J}'_{x_m}(\varepsilon;\bar{y}_m) \equiv (G_{x_m}(\varepsilon;\bar{y}_m), EF(\varepsilon;\bar{y}_{m-1},z),\omega) \quad \text{for all } \varepsilon \text{ and } \bar{y}_m ;$$

where $z$ is the real input $(I_{x_m}(\cdot;\bar{y}_m),RI_{x_m}(\cdot;\bar{y}_m))$ . (When $\bar{y}_m$ is $\bar{x}_m$ , the value of $z$ is $1$ .) Equations (5.6-5) and (5.6-6) become

$$(5.6\text{-}7) \quad V[\mathcal{F}(\varepsilon;\bar{x}_m)] \equiv V[\boldsymbol{\mathcal{N}}(\mathcal{S}^{\overline{m}}_m, \boldsymbol{\mathcal{N}}(\mathcal{S}^{\overline{m}}_{m-1}, \mathcal{S}^m_m - C^m_1, \ldots \boldsymbol{\mathcal{S}}'_{x_m} \ldots))(\varepsilon;\bar{x}_m)] \quad .$$

$$(5.6\text{-}8) \quad f(\bar{x}_m) = h(\bar{x}_m, h(\bar{x}_{m-1}, x_m-1, \ldots h(\bar{x}_{m-1}, 1, g_0(\bar{x}_{m-1}, 0))\ldots)) \quad .$$

If $x_m = 1$ , these last equations are

$$V[\mathcal{F}(\varepsilon;\bar{x}_m)] \equiv V[\boldsymbol{\mathcal{S}}'_{x_m}(\varepsilon;\bar{x}_m)] \quad ,$$

$$f(\bar{x}_{m-1}, 1) = h(\bar{x}_{m-1}, 1, g_0(\bar{x}_{m-1}, 0)) \quad ,$$

and $z$ is just $x_m$ . In general, we have

$$ER(\varepsilon;\bar{x}_{m-1}, z) = (RG_0 \hat{+} RG_1 \hat{+} |G_0 \hat{-} G_1|)(\varepsilon;\bar{x}_{m-1},z) \quad ,$$

$$G_0(\varepsilon;\bar{x}_{m-1}, z) = (\boldsymbol{\mathcal{S}}_0(\mathcal{S}^{\overline{m}}_{m-1}, \mathcal{S}_{x_m})(\varepsilon;\bar{x}_m))_1 \quad ,$$

$$G_1(\varepsilon;\bar{x}_{m-1}, z) = G_{x_m}(\varepsilon;\bar{x}_m) \quad ,$$

$$RG_1(\varepsilon;\bar{x}_{m-1}, z) = RG_{x_m}(\varepsilon;\bar{x}_m) \quad ,$$

$$g_0(\bar{x}_{m-1}, z) = g_0(\bar{x}_{m-1}, 1) \quad ,$$

$$g_1(\bar{x}_{m-1}, z) = h(\bar{x}_{m-1}, 1, g_0(\bar{x}_{m-1}, 0)) = g_0(\bar{x}_{m-1}, 1) \quad ,$$

the last equality following from our additional hypotheses. Further, by theorem 5.5-1 and corollary 5.3-1, we have

$$\lim_{\varepsilon\to 0} G_0(\varepsilon;\bar{x}_{m-1}, z) = g_0(\bar{x}_{m-1}, 1) \quad ,$$

$$\lim_{\varepsilon\to 0} G_1(\varepsilon;\bar{x}_{m-1}, z) = g_0(\bar{x}_{m-1}, 1) \quad ,$$

$$\lim_{\varepsilon\to 0} RG_0(\varepsilon;\bar{x}_{m-1},z) = \lim_{\varepsilon\to 0} RG_1(\varepsilon;\bar{x}_{m-1},z) = 0 \quad .$$

By this and theorem 2.8-1, we have

$$\lim_{\varepsilon\to 0} ER(\varepsilon;\bar{x}_{m-1},z) = 0 \quad .$$

From this and (5.6-7) and (5.6-8), it follows that

$$\lim_{\varepsilon\to 0} V[\mathfrak{F}(\varepsilon;\bar{x}_m)] \equiv (f(\bar{x}_m), 0) \quad .$$

This completes the proof.

Let $\mathscr{S}_+$ be $C_0^{m+1}$ and $\mathscr{S}_\times$ be $C_1^{m+1}$ . Let $\mathscr{S}$ be an $\varepsilon$-function of $m+1$ variables. For $*$ being $+$ and $\times$ , define

$$\Sigma_*(\mathscr{S}) \equiv \Phi_{rec}(C_\omega^{m+1}, \mathscr{S}_*, \mathscr{S}(\mathscr{I}^{\overline{m+2}}_{m+1}) * \mathscr{I}^{m+2}_{m+2})(\overline{\mathscr{I}^m_m}, \mathscr{I}^m_m) \quad ,$$

$$Q_{sum}(\mathscr{S},g,P) \equiv \begin{cases} S \times \tilde{R} - N_m & \text{if } P \equiv S \times (\tilde{R} - N_1 - \{\infty\})^{(2)} \\ \{\ \} & \text{otherwise .} \end{cases}$$

Let $S^*_{sum}$ be the set of all $\mathscr{S}$ of $m+1 = 2,3,\ldots$ variables such that the computation of $\Sigma_*(\mathscr{S})(\varepsilon;\bar{x}_m)$ via the determiner of $\mathscr{S}$ halts for any real inputs $\bar{x}_m$ .

COROLLARY 5.6-1: *For* $*$ *being* $+$ *and* $\times$ , *we have*

$$(\Sigma_*, Q_{sum}) \approx \sigma_*(S^*_{sum}) \quad .$$

Proof: Suppose $\mathcal{J} \approx g(S \times (\tilde{R} - N_1 - \{\infty\})^{(2)})$ . Then

$$\mathcal{J}(\overline{\mathcal{S}^{m+2}_{m+1}}) * \mathcal{S}^{m+2}_{m+2} \approx g(\overline{i^{m+2}_{m+1}}) * i^{m+2}_{m+2}(S \times (\tilde{R} - N_1 - \{\infty\})^{(2)} \times \tilde{R}) \quad \text{and}$$

theorems 5.6-1 and 5.5-1 yield $\Sigma_*(\mathcal{J}) \approx \sigma_*(g)(S \times \tilde{R} - N_m)$ . This completes the proof.

## 5.7 An $\varepsilon$-Function Corresponding to $e^x$ Over $\tilde{\mathfrak{R}}$

First we mention that $e^{\omega} = \omega$ , $e^{-\infty} = 0$ and $e^{\infty} = \infty$ . Let $f_{exp}(x) = e^x$ as in section 2.6. Define two basic $\varepsilon$-functions, $\mathfrak{P}$ and $\mathfrak{A}$ , by

$$(5.7\text{-}1) \qquad \mathfrak{P} \equiv \Sigma_+(C^2_1) \quad ,$$

$$(5.7\text{-}2) \qquad \mathfrak{A} \equiv \Phi_{rec}(C^1_\infty,\ C^1_1 - \mathcal{S}^1_1,\ \mathcal{S}^2_1 - C^2_1)(\mathcal{S}^1_1 + C^1_1) \quad .$$

By corollary 5.6-1, $\mathfrak{P} \approx \sigma_+(c^2_1)(\tilde{\mathfrak{R}} - N_1)$ ; $p \equiv \sigma_+(c^2_1)$ is essentially an entier for positive reals, because

$$(5.7\text{-}3) \qquad p(x) = \begin{cases} \omega & \text{if } x = \omega, \infty \\ 0 & \text{if } x < 1 \\ [x] & \text{otherwise} \end{cases} \quad .$$

By theorem 5.6-1, $\mathfrak{A} \approx a(\tilde{\mathfrak{R}})$ , where $a(x) = |x|$ . We use $\mathfrak{P}$ and $\mathfrak{A}$ to define $\mathfrak{F}_{exp}$ as follows:

$$(5.7\text{-}4) \qquad \mathfrak{T} \equiv \Sigma_x(\mathcal{S}^4_1 \div \mathfrak{P}(\mathcal{S}^4_4)) \quad ,$$

$$(5.7\text{-}5) \qquad \mathfrak{F} \equiv \Sigma_+(\mathfrak{T}(\mathcal{S}^3_1,\ \mathcal{S}^3_2,\ \mathcal{S}^3_3 - C^3_1))(\ -\ \mathfrak{A}(\mathcal{S}^3_1),\ \mathcal{S}^3_3) \quad ,$$

$$(5.7\text{-}6) \qquad \mathcal{L} \equiv \Phi_{rec}(C^1_0,\ \Phi_{lim}(\mathfrak{F})(\mathcal{S}^1_1,\ C^1_\infty),\ \Phi_{lim}(\mathfrak{F})(\mathcal{S}^2_1,\ C^2_\infty)) \quad ,$$

$$(5.7\text{-}7) \qquad \mathfrak{F}_{exp} \equiv \Phi_{rec}(C^1_\infty,\ \mathcal{L}(C^1_1 - \mathcal{S}^1_1),\ C^2_1 \div \mathcal{L}(\mathcal{S}^2_1 - C^2_1))(\mathcal{S}^1_1 + C^1_1) \quad .$$

$\mathfrak{T}$ forms terms, $x^{n-1}/(n-1)!$ . $\mathfrak{F}$ forms sums, $\sum_{n=1}^{N}(-|x|)^{n-1}/(n-1)!$ .

$\mathcal{J}$ checks for $x = -\infty$ , to insure that $\mathcal{F}_{\exp}(\varepsilon;-\infty) = 0$ , and otherwise it approximates $\sum_{n=0}^{\infty}(-|x|)^n/n!$ . $\mathcal{F}_{\exp}$ checks for $x = \infty$ , to insure that $\mathcal{F}_{\exp}(\varepsilon;\infty) = \infty$ , and otherwise it computes $\mathcal{K}(\varepsilon;x)$ and reciprocates this value when $x \geq 0$ . The only essential part of this definition that is missing is a definition of the TF part of $\mathcal{F}$ , because TF is the only truncation-error bound used here. In the following discussion, we will define TF and use our previous theorems to prove that $\mathcal{F}_{\exp} \approx f_{\exp}(\tilde{R})$. The only nontrivial part of this development is the definition of a stably convergent TF .

Theorems 5.1-1, 5.5-1 and 5.3-1 yield that $\mathcal{I}_1^4 \div \rho(\mathcal{I}_4^4) \approx i_1^4 \div p(i_4^4)(\tilde{R}^{(4)} - N_4)$ . This and corollary 5.6-1 imply that $\mathcal{T} \approx t(\tilde{R}^{(3)} - N_3)$ , where

$$t(x,m,n) = \begin{cases} \omega & \text{if } x = \omega \text{ , } m = \omega \text{ , or } n \geq \infty \\ 1 & \text{if } n < 1 \\ x^{[n]}/[n]! & \text{otherwise .} \end{cases}$$

Thus $\mathcal{F} \approx f(\tilde{R}^{(3)} - N_3)$ , where $f$ is defined in section 2.6. For the next step, we need the TF part of $\mathcal{F}$ . We could define TF from $\mathcal{TF}' \equiv \mathcal{T}(-\mathcal{A}(\mathcal{I}_1^3), \mathcal{I}_2^3, \mathcal{I}_3^3)$ without making use of any special properties of $(\mathcal{R}, \mathcal{E})$ , but a considerable derivation is required to insure that the resulting TF is stably convergent at all $(x, \infty)$ with $x$ finite. For simplicity, we instead sketch a definition of TF for the case where $(\mathcal{R}, \mathcal{E})$ is a floating-point number system and $\varepsilon$-arithmetic satisfies the usual relations needed for an error analysis in the style of Wilkinson [W2] . We work from the tf of section 2.6. Let a finite $x$ be given and let

$$Y \equiv Y(\varepsilon) = (|X(\varepsilon)| \mathbin{\hat{+}} RX(\varepsilon)) \mathbin{\hat{\div}} ((1 \underset{\smile}{-} \varepsilon) \underset{\smile}{\times} (1 \underset{\smile}{-} \varepsilon) \underset{\smile}{\times} (1 \underset{\smile}{-} \varepsilon)) \quad .$$

For integers $n > Y$ let $fl_\varepsilon(Y^n/n!)$ denote the product $Y \times (Y/2) \times (Y/3) \times \ldots \times (Y/m)$ evaluated in floating-point $\varepsilon$-arithmetic, where $m$ is the largest integer with $Y < m \leq n$ such that overflow and underflow do not occur (or $m$ is $\omega$, if there is no such integer). Using this $m$ instead of $n$ is two-thirds of the trick needed to define TF for an arbitrary $(\mathcal{R}, \mathcal{E})$. Assume that the value $I$ used in place of $i$ in $(Y/i)$ satisfies $I = i \times (1+\eta_i(\varepsilon))$, where $|\eta_i(\varepsilon)| \leq \varepsilon$. Then for some $|\eta_j'(\varepsilon)| \leq \varepsilon$ $(j=1,\ldots, 3m)$, we have

$$\begin{aligned} fl_\varepsilon(Y^n/n!) &= (Y^m/m!) \times \prod_{j=1}^{3m} (1 + \eta_j'(\varepsilon)) \\ &\geq ((|X(\varepsilon)| + RX(\varepsilon))^m/m!) \times \prod_{j=1}^{3m} \frac{(1+\eta_j'(\varepsilon))}{(1-\varepsilon)} \\ &\geq |x|^m/m! \quad . \end{aligned}$$

Further, we have

$$\lim_{\varepsilon \to 0} fl_\varepsilon(Y^n/n!) = |x|^n/n! \quad ,$$

so we define

$$TF(\varepsilon;x,k,y) = \begin{cases} fl_\varepsilon(Y^n/n!) & \text{if } k = \infty \text{ and } |x| + 1 <_\varepsilon y <_\varepsilon \infty \\ \omega & \text{otherwise} \end{cases} \quad ,$$

where $n$ is the largest integer $<_\varepsilon y$. We prove that TF is stably convergent at $(x, \infty)$ for finite $x$ as follows. Let $y_{\varepsilon_1}, y_{\varepsilon_2}, \ldots$

satisfy $y_\varepsilon \in \mathcal{R}(\varepsilon) - \{\infty\}$ and $\lim_{\varepsilon \to 0} y_\varepsilon = \infty$. For some $|\eta_j''(\varepsilon)| \leq 2\varepsilon$ we have

$$Y(\varepsilon) = (|X(\varepsilon)| + RX(\varepsilon))(1-\varepsilon)^{-3} \times \prod_{j=1}^{7} (1 + \eta''(\varepsilon)) \quad ,$$

$$TF(\varepsilon;x,\infty,y_\varepsilon) \leq ((|X(\varepsilon)| + RX(\varepsilon))(\frac{1+\varepsilon}{1-\varepsilon})^3(1 + 2\varepsilon)^7)^m/m! \quad ,$$

for all sufficiently small $\varepsilon$. As $\varepsilon \to 0$, $m \to \infty$ and the right side above goes to $0$. Thus $TF$ is stably convergent at all $(x, \infty)$ with $x$ finite.

Thus $\mathcal{F} \equiv (F, RF, TF)$ is in $S_{\lim}$ and $\Phi_{\lim}(\mathcal{F})(\mathcal{I}_1^m, \mathcal{I}_\infty^m) \approx \phi_{\lim}(f)(i_1^m, c_\infty^m)(R \times \tilde{R}^{(m-1)})$. This and theorem 5.6-1 yield $\mathcal{G} \approx g(\tilde{R})$, where $g$ is defined in section 2.6. This implies that $\mathcal{F}_{\exp} \approx f_{\exp}(\tilde{R})$.

REMARKS: It is easy to define initial $\varepsilon$-functions and $\varepsilon$-operators analogous to those of this chapter for the "$\varepsilon$-calculus of stable $\varepsilon$-functions" discussed at the end of chapters 2 and 4. However, our example in section 5.7 would have to be changed, because the subtractions in $\Sigma(-|x|)^n/n!$ makes $F$ unstable at $(x,\infty)$ . This can be remedied by defining $F$ in terms of $\Sigma|x|^n/n!$ .

# Chapter 6. ε-Derivative and ε-Integral

## 6.1 ε-Differentiability and ε-Derivative

Define a difference operator, d, over the set $S_d$ of ideal functions of one variable by

$$d(f) \equiv (f(i_1^2) - f(i_2^2)) \div (i_1^2 - i_2^2) .$$

We say f is differentiable at x precisely when d(f) converges at x . Otherwise, we say f is nondifferentiable at x . Define a difference ε-operator, $(D, Q_D)$, over the set $S_D$ of ε-functions of one variable by

$$D(\mathcal{F}) \equiv (\mathcal{F}(\mathcal{I}_1^2) - \mathcal{F}(\mathcal{I}_2^2)) \div (\mathcal{I}_1^2 - \mathcal{I}_2^2) ,$$

$$Q_D(\mathcal{F}, f, P) \equiv P \times P .$$

By theorems 5.1-1, 5.5-1 and corollary 5.3-1, we have

$$(D, Q_D) \approx d(S_D) .$$

DEFINITION 6.1-1: We say $\mathcal{F}$ is ε-differentiable at x precisely when $D(\mathcal{F})$ ε-converges at x . Otherwise, we say $\mathcal{F}$ is ε-nondifferentiable at x .

Our previous analysis of ε-convergence at x carries over immediately to ε-differentiability at x , so we will not bother to express this in operator, ε-operator form.

Define a derivative operator, $\frac{d}{dt}$ , over $S_d$ by

$$\frac{d}{dt}(f) \equiv \phi_{\lim}(d(f)) .$$

Let $S_{d/dt}$ be the set of all $\varepsilon$-functions, $\mathcal{F}$, of one variable such that $D(\mathcal{F}) \in S_{lim}$. Define an $\varepsilon$-derivative $\varepsilon$-operator, $(\frac{D}{Dt}, Q_{d/dt})$, by

$$\frac{D}{Dt}(\mathcal{F}) \equiv \Phi_{lim}(D(\mathcal{F})) ,$$

$$Q_{d/dt}(\mathcal{F}, f, P) \equiv Q_{lim}(D(\mathcal{F}), d(f), P \times P).$$

We call $\frac{D}{Dt}(\mathcal{F})(\varepsilon; x)$ *the $\varepsilon$-derivative of $\mathcal{F}$ at $x$*. By theorem 4.1-1 we have

$$(\frac{D}{Dt}, Q_{d/dt}) \approx (S_{d/dt}) .$$

Thus, under the usual conditions, the $\varepsilon$-derivative at $x$ of an $\varepsilon$-function approaches the derivative at $x$ of its corresponding ideal function as $\varepsilon \to 0$.

It deserves mention here that there is a function, $f$, such that

(1) $f(x)$ is finite for all $x \in R$,

(2) there is a $\mathcal{F} \approx f(\tilde{R})$, and

(3) $f$ is nondifferentiable at every point in $R$.

See Grzegorczyk [Gl, pp. 199-201].

BLANK PAGE

## 6.2 $\varepsilon$-Integrability and $\varepsilon$-Integral

Let $C$ denote a finite closed interval of numbers. For $P$ being a set of real inputs, we say P covers C almost everywhere precisely when $C$ has a subset $C'$ of Lebesgue measure zero such that for any $c \in C - C'$ there is an $x \in P$ with $x = c$. As is usual, we write "a.e." for "almost everywhere." Let "over $C$" be implicit in the statements "f is continuous a.e., bounded or integrable. From analysis, we know that the bounded, Rieman integrable functions are precisely those that are bounded and continuous a.e. (See Royden [R3, p. 70].)

Suppose $f$ is a bounded ideal function and that $\mathfrak{F} \approx f(P)$. In order for the information contained in the set $\{\mathfrak{F}(\varepsilon; x): \varepsilon \in \mathcal{E}, x \in C \cap P\}$ to determine whether $f$ is continuous a.e., $P$ will have to cover $C$ a.e. (Remember that [illegible] gives no information because [illegible]'s of $\mathfrak{F}$'s of one variable are $= \omega$.) Similarly, the set,

$$\{f: \mathfrak{F} \approx f(P) \text{ and } f \text{ is bounded}\},$$

will contain both integrable and nonintegrable ideal functions unless $P$ covers $C$ a.e.

Now, suppose we have a definition of "$\mathfrak{F}$ is $\varepsilon$-integrable (over C)" which is based only on the values of $\mathfrak{F}$. Then, the weakest requirement on the size of $P$ which might make the conditions,

(1) $\mathfrak{F} \approx f(P)$,

(2) $f$ is bounded, and

(3) $\mathfrak{F}$ is $\varepsilon$-integrable for all sufficiently small $\varepsilon$,

equivalent ot {$f$ is integrable} is

(4) $P$ covers $C$ a.e.

However, by theorem 4.4-1, we know that conditions 1, 2 and 4 by themselves inply that $f$ is integrable, i.e., that $f$ is continuous everywhere in $C$ except possibly at points in $\mathfrak{M} \cap C$, a set of measure zero. Hence essentially the only definition of "$\mathfrak{F}$ is $\varepsilon$-integrable", which uses only the values of $\mathfrak{F}$ and for which we have

$$(1) - (4) \text{ hold } \Leftrightarrow f \text{ is integrable,}$$

is "$\mathfrak{F}$ is $\varepsilon$-integrable precisely when $1 = 1$."

Consider basing our definition on the values of $\mathfrak{F}' \equiv \mathfrak{F}(\mathcal{J}_2^2)$, an $\varepsilon$-function which may have a worthwhile truncation-error bound. (This is reasonable because it is only by a fluke of notation that $\mathfrak{F}$ of one variable have no worthwhile truncation-error bound.) Then we have the additional information given by $TF'$, which satisfies

$$TF'(\varepsilon;\, x,\, Y) \geq \left| f(Y) - \lim_{y \to x} f(y) \right| .$$

Again, let us assume that $f$ is bounded. Let $z$ be a poor real input such that, for each $\varepsilon$, $|Z(\varepsilon) - c| \leq RZ(\varepsilon)$ for all $c \in C$.
If $TF'(\varepsilon;\, z,\, z) < \omega$ then we will know that $\lim_{y \to c} f(y)$ exists for all $c \in C$. And this implies that $f$ is integrable, by

> <u>THEOREM 6.2-1</u>: <u>Suppose</u> $f$ <u>is bounded over</u> $C$ <u>and</u> $\lim_{y \to c} f(y)$ <u>exists for all</u> $c \in C$. <u>Then</u> $f$ <u>has at most a countable number of discontinuities in</u> $C$.

So far as we know, this is a new result.

<u>Proof</u>: (This was proved independently by Bill Glassmire and Paul Rosenthal.) Define

$$g(c) = \lim_{y \to c} f(y) \qquad \text{for } c \in C .$$

First we prove that $g$ is continuous. Suppose $y_i \to c$ as $i \to \infty$. For each $i$ there is an $x_i \neq c$ with

$$|g(y_i) - f(x_i)| < 1/i, \quad |y_i - x_i| < 1/i \quad ,$$

because $g(y_i) = \lim_{x \to y_i} f(x)$. Thus

$$\lim_{i \to \infty} x_i = c, \quad \lim_{i \to \infty} g(y_i) = g(c) \quad .$$

Since $y_i$ was an arbitrary approach, this means that $g$ is continuous in $C$. Next we prove that $g \neq f$ only on a countable set. Suppose not. We have

$$\{x: \; g(x) > f(x)\} \equiv \bigcup_{n=1}^{\infty} \{x: \; g(x) > f(x) + 1/n\} \quad .$$

If this set is uncountable then at least one of the sets on the right is uncountable; suppose $E_n \equiv \{x: \; g(x) > f(x) + 1/n\}$ is. Then its members have a cluster point, $x_0$, and there is a sequence, $x_1, x_2, \ldots$, from $E_n$ and approaching $x_0$ such that

$$g(x_m) > f(x_m) + 1/n \qquad \text{for } m = 1, 2, \ldots ,$$

$$\lim_{m \to \infty} g(x_m) = g(x_0) \geq g(x_0) + 1/n \quad ,$$

a contradiction. Similarly, the set $\{x: \; g(x) < f(x)\}$ is countable. Therefore, the set $\{x: \; g(x) \neq f(x)\}$ is countable. This completes the proof.

Thus we can define $\varepsilon$-integrability as follows. For finite a and b, let c[a, b] denote the closed interval between a and b. For given, finite real inputs a and b, let $z \equiv z(a, b)$ be a poor real input such that

(1) for each $\varepsilon$, $|Z(\varepsilon) - x| \leq RZ(\varepsilon)$ for all $x \in c[a, b]$,

(2) $\bigcap_{i \geq 1} \{x: |Z(\varepsilon_i) - x| \leq RZ(\varepsilon_i)\} \equiv c[a, b]$, and

(3) $Z(\cdot)$ and $RZ(\cdot)$ are effectively computable from a and b.

It is easy to verify that such z exist.

DEFINITION 6.2-1: Suppose $\mathfrak{F} \approx f(P)$ for some P. Let $\mathfrak{F}' \equiv \mathfrak{F}(\mathcal{J}_2^2)$. Let $z \equiv z(a, b)$ be as above. For finite a and b, we say TF' is good relative to f, a and b precisely when [f is integrable over c[a, b]] $\Rightarrow$ [TF'($\varepsilon$; z, z) < $\omega$ for all sufficiently small $\varepsilon$].

Let TF' and $z \equiv z(a, b)$ be as above. We put these results in operator, $\varepsilon$-operator form by defining

$S_{int} \equiv$ {ideal functions of one variable, bounded over $o(-\infty, \infty)$},

$\not{p}_{int}(f)(a, b) =$ bool [f is integrable over c[a, b]],

provided a and b are finite, and

$S_{int}$ {$\mathfrak{F}$: $\mathfrak{F} \approx f(P)$ for some $f \in S_{int}$ and some P, and computation of TF'($\varepsilon$; x, x) via the determiner of TF' halts for any poor real input x (see sec. 2.4)},

$\Phi_{\mathrm{int}}(\mathcal{F})(\varepsilon;\ a,\ b) \equiv (\mathrm{bool}\ [\mathrm{TF}'(\varepsilon;\ x,\ z) < \omega],\ \ \mathrm{bool}\ [\mathrm{TF}'(\varepsilon;\ z,\ z) = \omega],\omega)$,

provided $-\infty <_{\varepsilon} \begin{matrix} a \\ b \end{matrix} <_{\varepsilon} \omega$, and

$Q_{\mathrm{int}}(\mathcal{F},\ f,\ P) \equiv \{(a,\ b)\text{:}\ \ \mathrm{TF}'\ \text{ is good relative to } f, a \text{ and } b\}$ .

An immediate consequence of the above analysis is

THEOREM 6.2-2: *We have*

$$(\Phi_{\mathrm{int}},\ Q_{\mathrm{int}}) \sim \phi_{\mathrm{int}}(S_{\mathrm{int}}) \ .$$

*Further, if* $f$ *is integrable over* $c[a,\ b]$ *for all* $(a,\ b) \in Q_{\mathrm{int}}(\mathcal{F},\ f,\ P)$, *then* $\Phi_{\mathrm{int}}(\mathcal{F})$ *is not weak, and vice versa.*

Now for the integral. Using the notation of chapter 5, we define a partial sum and an integral operator over $S_{\mathrm{int}}$ by

$$h^m \equiv (i_2^m - i_1^m) \div p(i_4^m) \qquad (m = 4,\ 5) \quad ,$$

$$t \equiv i_1^5 + p(i_5^5) \times h^5 \quad ,$$

$$\phi_{\mathrm{psum}}(f) \equiv h^4 \times \sigma_+(f(t)) \quad ,$$

$$\int \mathrm{dt}(f)(a,\ b) = \begin{cases} \phi_{\mathrm{lim}}(\phi_{\mathrm{psum}}(f))(i_1^2,\ i_2^2,\ c_\infty^2)(a,\ b) \\ \qquad\qquad \text{if } \ \phi_{\mathrm{int}}(f)(a,\ b) = 1 \\ \omega \qquad\qquad\qquad\qquad \text{otherwise} \end{cases} .$$

We define $\varepsilon$-operators for these by

$$\mathcal{K}^m \equiv (\mathcal{J}_2^m - \mathcal{J}_1^m) + \wp(\mathcal{J}_4^m) \qquad (m = 4, 5) \quad ,$$

$$\mathcal{T} \equiv \mathcal{J}_1^5 + \wp(\mathcal{J}_5^5) \times \mathcal{K}^5 \quad ,$$

$$\Phi_{psum}(\mathfrak{F}) \equiv \mathcal{K}^4 \times \Sigma_+(\mathfrak{F}(\mathcal{T})) \quad ,$$

$$Q_{psum}(\mathfrak{F}, f, P) \equiv Q_{sum}(\mathfrak{F}(\mathcal{T}\ ), f(t), Q^1_{comp}(\mathfrak{F}, \mathcal{T}, f, t, P,$$

$$\widetilde{R}^{(3)} \times (\widetilde{R} - N_1)^{(2)})) - N_4 \quad ,$$

$$\int Dt(\mathfrak{F})(\varepsilon;\ a, b) = \begin{cases} \Phi_{lim}(\Phi_{psum}(\mathfrak{F}))(\mathcal{J}_1^2, \mathcal{J}_2^2, C_\infty^2)(\varepsilon;\ a, b) & \\ \qquad \text{if} \quad \Phi_{int}(\mathfrak{F})(\varepsilon;\ a, b) \equiv (1, 0, \omega) & \\ (\omega, \omega, \omega) & \text{otherwise} \end{cases} \quad ,$$

$$Q_{integral}(\mathfrak{F}, f, P) \equiv Q_{int}(\mathfrak{F}, f, P) \cap$$

$$Q^3_{comp}(\Phi_{lim}(\Phi_{psum}(\mathfrak{F})), \mathcal{J}_1^2, \mathcal{J}_2^2, C_\infty^2, \ldots) \quad ,$$

$$S_{integral} \equiv \{\mathfrak{F}:\ \mathfrak{F} \in S_{int} \text{ and } \Phi_{psum}(\mathfrak{F}) \in S_{lim}\} \quad .$$

The theorems of chapter 5 and theorem 6.2-2 yield

<u>THEOREM 6.2-3</u>: <u>We have</u>

$$(\int Dt, Q_{integral}) \sim \int dt(S_{integral}) \quad .$$

<u>Further,</u> $\int Dt(\mathfrak{F})$ <u>is weak if and only if</u> $\Phi_{int}(\mathfrak{F})$ <u>is weak.</u>

## 6.3 The Fundamental Theorem of the $\varepsilon$-Calculus

Following is the $\varepsilon$-calculus analog to the fundamental theorem of the calculus.

<u>THEOREM 6.3-1</u>: <u>Fix</u> $a, b \in R$ <u>and assume that</u>

(1) $\frac{d}{dt}(f)$ <u>is bounded and integrable over</u> $c[a, b]$, <u>and</u>

(2) $\mathcal{F} \approx f(\{\ \})$ .

<u>Then, for any</u> $\varepsilon$ <u>we have</u>

$$(6.3\text{-}1) \qquad \int Dt(\frac{D}{Dt}(\mathcal{F}))(\varepsilon;\, a,\, b) =_{\varepsilon} (\mathcal{F}(\mathcal{S}_2^2) - \mathcal{F}(\mathcal{S}_1^2))(\varepsilon;\, a,\, b) \quad .$$

<u>Proof</u>: For any $a_1, a_2 \geq 0$, $a_3 \in R(\varepsilon)$ let $\rho(\bar{a}_3)$ be

$$\rho(\bar{a}_3) \equiv \{x: \ |a_1 - x| \leq a_2\} \quad .$$

For any $\varepsilon$-function $\mathcal{G}$ of $m$ variables and any $(\varepsilon;\, \bar{x}_m)$ we have $[\mathcal{G} \approx g(\{\ \})$ implies $g(\bar{x}_m) \in \rho(\mathcal{G}(\varepsilon;\, \bar{x}_m))]$ . This means that

$$A = \int dt(\frac{d}{dt}(f))(a,\, b) \in \rho(\int Dt(\frac{D}{Dt}(\mathcal{F}))(\varepsilon;\, a,\, b)) \quad ,$$

$$B = f(b) - f(a) \in \rho((\mathcal{F}(\mathcal{S}_2^2) - \mathcal{F}(\mathcal{S}_1^2))(\varepsilon;\, a,\, b)) \quad .$$

The fundamental theorem of the calculus tells us that $A = B$, yielding (6.3-1). This completes the proof.

BLANK PAGE

## Chapter 7: Computable Real Functions and Completeness

### 7.1 Computable Real Functions

We say that a set of real inputs <u>P covers $\tilde{R}^{(m)}$</u> $(m \geq 1)$ precisely when each value in $\tilde{R}^{(m)}$ is taken on by some member of P. We say <u>P covers $\tilde{R}^{(0)}$</u> precisely when $P \equiv \{\bar{x}_0\}$. Let $\mathcal{K}_1$ be the class of all ideal functions, f, such that there is a P covering $\tilde{R}^{(m)}$ $(m \geq 0)$ and an $\mathcal{F}$ with $\mathcal{F} \approx f(P)$. We say <u>f is computable</u>$_1$ precisely when $f \in \mathcal{K}_1$. $\mathcal{K}_1$ depends on $(\mathcal{R}, \mathcal{C})$ and it contains many functions with discontinuities. We will not consider $\mathcal{K}_1$ further.

By <u>specialization</u> of an ideal function f of $m \geq 1$ variables, let us mean the replacement of a variable by a numeric constant, yielding an ideal function of m-1 variables. Let $\mathcal{K}_2'$ be the class of all ideal functions f such that there is an $\mathcal{F} \approx f(\tilde{R}^{(m)})$ and F and RF are subroutines of m variables and <u>no</u> constants. Let $\mathcal{K}_2$ be the smallest class of ideal functions containing $\mathcal{K}_2'$ and closed under specialization. We say <u>f is computable</u>$_2$ precisely when $f \in \mathcal{K}_2$. If $f \in \mathcal{K}_2$ then there is an $\mathcal{F} \approx f(\tilde{R}^{(m)})$ such that F and RF are subroutines of m variables and $n \geq 0$ constants. (We do not know whether the reverse is true.) As we shall see, $\mathcal{K}_2$ is independent of $(\mathcal{R}, \mathcal{C})$. By theorem 4.4-1, we know that any $f \in \mathcal{K}_2$ is continuous at all $\bar{x}_m \in R^{(m)}$ with $f(\bar{x}_m) \neq \omega$.

Let $\theta$ be as in section 2.2 and $\mathbf{n}$ as in 1.5. Let us say $\alpha_1, \alpha_2 \colon \mathbf{n} \to \mathbf{n}$ give $a \in \tilde{R}$ precisely when

(1) for each $n \geq 1$, either $\alpha_1(n) = \alpha_2(n) = 3$ or

$$|a - \theta(\alpha_1(n), \alpha_2(n))/n| \leq 1/n \quad ,$$

(2) if $a \neq \omega$ then, for all sufficiently large $n$, either

$$\alpha_1(n) \neq 3 \quad \text{or} \quad \alpha_2(n) \neq 3 \quad .$$

For $m \geq 1$, we say $\underline{\overline{\alpha_{2m}} \text{ give } \bar{x}_m}$ precisely when $\alpha_{2i-1}, \alpha_{2i}$ give $x_i$ $(i = 1,2,\ldots, m)$ . Also, we say $\underline{\bar{\alpha}_0 \text{ give } \bar{x}_0}$ . We say recursive operators $\underline{\psi_1, \psi_2 \text{ give } f}$ precisely when for any $\overline{\alpha_{2m}}$ and $\bar{x}_m \in \tilde{R}^{(m)}$ ,

$$[\overline{\alpha_{2m}} \text{ give } \bar{x}_m] \Rightarrow [\psi_1(\overline{\alpha_{2m}}), \psi_2(\overline{\alpha_{2m}}) \text{ give } f(\bar{x}_m)] \quad .$$

Let $\mathcal{K}_3'$ be the class of all ideal functions $f$ for which there exist recursive operators $\psi_1, \psi_2$ which give $f$ . Let $\mathcal{K}_3$ be the smallest class of ideal functions which contains $\mathcal{K}_3'$ and which is closed under specialization. We say $\underline{f \text{ is computable}_3}$ precisely when $f \in \mathcal{K}_3$ . This is analogous to Grzegorczyk's definition of computable continuous real functions of one variable [G2] . $\mathcal{K}_3$ obviously does not depend on $(\mathcal{R}, \mathcal{C})$ and we have

<u>THEOREM 7.1-1</u>: $\mathcal{K}_2 \equiv \mathcal{K}_3$ .

<u>Proof</u>: We prove this by proving $\mathcal{K}_2' \equiv \mathcal{K}_3'$ . Our proof is based on two transformation functions, $t_1$ and $t_2$ . Suppose $\alpha_1, \alpha_2$ give $a \in \tilde{R}$ and define the poor real input $t_1(\alpha_1, \alpha_2) \equiv x$ by

$$\alpha(\cdot) \equiv \theta(\alpha_1(\cdot), \alpha_2(\cdot)) \quad ,$$

$$X(\epsilon) = \hat{I}(\epsilon, \alpha) \quad ,$$

$$RX(\epsilon) = \begin{cases} \omega & \text{if } X(\epsilon) = \omega \ , \\ |\hat{I}(\epsilon, \alpha) \overset{\frown}{-} \check{I}(\epsilon, \alpha)| & \text{otherwise} \ . \end{cases}$$

If $a \neq \omega$ then this $x$ is a real input. Let $y$ be a poor real input and define $t_2(y) \equiv (\beta_1, \beta_2)$ as follows. Let an $n \geq 1$ be given. If $RY(\epsilon_n) = \omega$ or $\limsup_{\epsilon \to 0} RY(\epsilon) > 0$ then define $\beta_1(n) = \beta_2(n) = 3$ . Otherwise let $j(n)$ be the smallest value of $j$ such that $RY(\epsilon_j) \leq \check{I}(\epsilon_j, \gamma_{3n})$, where $\gamma_{3n}(k) = [k/3n]$ so that $< \gamma_{3n} > = 1/3n$ . Then

$$|Y(\epsilon_{j(n)}) - y| \leq 1/3n \ .$$

Suppose $Y(\epsilon_{j(n)})$ is $< \alpha_{\mathbb{R}}(j(n), k(n), \cdot) >$ and let $\delta_n(\cdot)$ be $\alpha_{\mathbb{R}}(j(n), k(n), \cdot)$ . If $< \delta_n > = -\infty, \infty$ or $\omega$ then $< \delta_n > = y$ ; in this case, define $\beta_1(n) = \beta_2(n) = 1, 2$ or $3$ respectively. Suppose $< \delta_n >$ is finite. For any integer $i$, define

$$r(i) = \begin{cases} i/3 & \text{if } i = 3[i/3] \\ (i-1)/3 & \text{if } i = 3[i/3] + 1 \\ (i+1)/3 & \text{otherwise} \ . \end{cases}$$

Note that $r(i)$ is always an integer. Define

$$\beta_1(n) = |r(\delta_n(3n))| \ ,$$

$$\beta_2(n) = |r(\delta_n(3n))| - r(\delta_n(3n)) \ .$$

In this case we have

$$\beta_1(n) - \beta_2(n) = r(\delta_n(3n)) \quad ,$$

$$|y - r(\delta_n(3n))/n|$$

$$\leq |y - Y(\varepsilon_{j(n)})| + |Y(\varepsilon_{j(n)}) - \delta_n(3n)/3n| + 1/3n \leq 1/n \quad .$$

Thus if $y$ is a real input then $\beta_1, \beta_2$ give $y$. Further, if $y$ is a real input or if $RY(\cdot) = \omega$ then $\beta_1$ and $\beta_2$ are computable from $(Y, RY)$.

We prove $K'_2 \subset K'_3$ as follows. Suppose $f \in K'_2$. Then there is an $\mathfrak{F} \approx f(\widetilde{R}^{(m)})$ such that $F$ and $RF$ are subroutines of no constants (see sec. 2.4). We will construct recursive operators $\phi_1, \phi_2$ which give $f$. Suppose $\overline{\alpha_{2m}}$ give $\bar{x}_m$. Define poor real inputs $\bar{y}_m$ by

$$y_i = t_1(\alpha_{2i-1}, \alpha_{2i}) \quad .$$

Let $f(\bar{y}_m)$ denote the poor real input $(F(\cdot\,; \bar{y}_m), RF(\cdot\,; \bar{y}_m))$ and define $\phi_1$ and $\phi_2$ by

$$(\phi_1(\overline{\alpha_{2m}}), \phi_2(\overline{\alpha_{2m}})) = t_2(f(\bar{y}_m)) \quad .$$

Then $\phi_1$ and $\phi_2$ are recursive operators and they give $f$. Thus $f \in K'_3$.

We prove $K'_3 \subset K'_2$ as follows. Suppose $f \in K'_3$. Then there are recursive operators $\phi_1, \phi_2$ giving $f$. We will define an $\mathfrak{F} \approx f(\widetilde{R}^{(m)})$ such that $F$ and $RF$ are subroutines of no constants. Suppose $\bar{x}_m$ are poor real inputs. Define $\overline{\alpha_{2m}}$ by

$$(\alpha_{2i-1}, \alpha_{2i}) = t_2(x_i) \quad .$$

Define F and RF by

$$(F(\cdot;\ \bar{x}_m),\ RF(\cdot;\ \bar{x}_m)) \equiv t_1(\psi_1(\overline{\alpha_{2m}}),\ \psi_2(\overline{\alpha_{2m}})) \quad ,$$

and set $TF \equiv \omega$ . Then $\mathfrak{F} \equiv (F, RF, TF)$ is the desired $\mathcal{E}$-function. This completes the proof.

Let $\mathcal{K}$ be the class of all $f\colon R \to R$ such that there is an $f' \in \mathcal{K}_3'$ (or $\mathcal{K}_2'$) with $f(x) = f'(x)$ for all $x \in R$ . $\mathcal{K}$ is precisely Grzegorczyk's class of computable continuous real functions [G2]. In [G2] Grzegorczyk proves $\mathcal{K}$ to be equivalent to several other classes of computable real functions which have appeared in the literature. In [G2, p. 192] he proves that the $f \in \mathcal{K}$ are computably uniformly continuous in any segment. He also constructs an $f \in \mathcal{K}$ which is not differentiable at any point [G2, p. 199].

## 7.2 Completeness

Let $\mathcal{K}^*$ be the class of all $f\colon R^{(m)} \to R$ $(m \geq 0)$ such that there is an $f' \in \mathcal{K}_3$ with $f'(\bar{x}_m) = f(\bar{x}_m)$ for all $\bar{x}_m \in R^{(m)}$. Let $\mathcal{K}^{**}$ be the class of all $f\colon R^{(m)} \to R$ $(m \geq 0)$ which can be defined exclusively in terms of the $c^n_k$, $i^n_j$, $\phi_+$, $\phi_\times$, $\phi_\div$, $\phi_{\lim}$, $\phi^n_{\text{comp}}$, $\phi_{\text{rec}}$ from chapter 5.

THEOREM 7.2-1: $\mathcal{K}^* \subset \mathcal{K}^{**}$ .

In this sense, the initial functions and operators of chapter 5 are complete.

Proof: We only sketch the proof. The Stone-Wierstrauss theorem (see Bishop [B1, pp. 97, 100]) shows that we can construct an arbitrarily close (in the sup norm) polynomial approximation to a continuous function over a compact set if we are given

(1) access to any finite number of (arbitrarily close approximations to) values of the function over the compact set, and

(2) the modulus of continuity of $f$ .

Grzegorczyk [G1, p. 192] has shown that every $f \in \mathcal{K}^*$ of one variable has a computable modulus of continuity, and his proof generalizes to $f$ of any number of variables. Thus any $f \in \mathcal{K}^*$ can be written as a polynomial in $m$ variables:

$$f(\bar{x}_m) = \lim_{n\to\infty} \sum_{\bar{j}_m} c(n,\ p_1^{j_1} \times \ldots \times p_m^{j_m}) \times x_1^{j_1} \times \ldots \times x_m^{j_m} \quad ,$$

where $p_i$ denotes the $i^{th}$ prime number, $c(n, k)$ is the $k^{th}$ coefficient of the $n^{th}$ polynomial, the sum is taken over all $\bar{j}_m$

such that $0 \leq j_i \leq n$ $(i = 1,2,\ldots, m)$, and where the $n^{th}$ polynomial approximates f over the m-dimensional square, $[-n, n]^{(m)}$, with a maximum error less than $1/n$. Further, we can assume that each $c(n, k)$ is rational. Thus, in order to show that $K^* \subset K^{**}$ we need only show that $K^{**}$ includes all computable rational functions, $c(n, k)$. But, since division is one of the closure operations of $K^{**}$, we need only show that $K^{**}$ contains all recursive rational functions, $b(n, k)$. It is obvious that the initial functions and operations, except the effective minimum, used to define the recursive functions in [M1, p. 120-1] can be simulated by the operators and initial functions of $K^{**}$. That the effective minimum operator can also be simulated in this way follow from the following:

$$t^n \equiv i^n_m - i^n_{m+1} - c^n_1 + f(i^n_{m-1}, i^n_m - c^n_1) \quad \text{for} \quad n = m+1, m+2 \quad ,$$

$$h \equiv \phi_{rec}(c^{m+1}_\omega , i^{m+1}_m, t^{m+2})(\overline{i^{m+1}_m}, t^{m+1}) \quad ,$$

$$g \equiv \phi_{rec}(c^m_\omega, c^m_0, h)(\overline{i^m_{m-1}}, i^m_m + c^m_1) \quad ,$$

$$\mu y(f(\overline{i^m_{m-1}}, y) \neq 0) \equiv \phi_{lim}(g(\overline{i^{m+1}_{m-1}}, i^{m+1}_{m+1}))(\overline{i^m_{m-1}}, c^m_\infty) \quad .$$

This deserves some explanation. For $i > j \geq 0$, if $i = j+1$ and $f(\overline{x}_{m-1}, j) = 0$ then $h(\overline{x}_{m-1}, i, j) = i$, or otherwise $h(\overline{x}_{m-1}, i, j) = j$. If $f(\overline{x}_{m-1}, k) = 0$ for $k = 0, 1,\ldots, \ell$ then $g(\overline{x}_{m-1}, \ell) = \ell+1$; otherwise, $g(\overline{x}_{m-1}, \ell)$ is the least value of $n$ such that $f(\overline{x}_{m-1},n) \neq 0$. This completes the proof.

BLANK PAGE

## Summary and Conclusions

We have developed a theory of numerical computation based on recursive function theory, with a flavor of interval analysis. This theory concerns itself with a general class of variable-precision computations and the finite-precision (or intermediate) results arising in such computations. For example, the floating-point computations of modern digital computers are in this class. Our main goal was to form a realistic model of such computations. This was done by developing the concepts of

(1) a machine number system $(\mathcal{R}, \mathcal{E})$ (sec. 2.2),

(2) a real input $\mathbf{x} \equiv (X, RX)$ (sec. 2.3),

(3) a subroutine $F$ (sec. 2.4),

(4) an $\varepsilon$-function $\mathfrak{F} \equiv (F, RF, TF)$ (sec. 2.5), and

(5) $\varepsilon$-arithmetic (sec. 2.8 and 5.3).

If this model had been our only goal, we would probably have dispensed with roundoff-error and truncation-error bounds (the RX, RF and TF indicated above) because such bounds are usually not computed on the computer. (We discuss the removal of these bounds in the remarks at the end of chapters 2, 4 and 5.) However, our secondary goal necessitated the incorporation of these bounds. This secondary goal was to find out how concepts from the calculus such as convergence, continuity, differentiability and integrability apply, at each fixed level of precision, to numerically computed functions which, after all, can be viewed at a fixed precision as a discrete set of points on a graph. This secondary goal was achieved by associating the numerically computed function, F, with its underlying mathematical (or ideal) function, f, through the use of roundoff-error bounds, RF, and truncation-error bounds, TF . Thus we defined an $\varepsilon$-function $\mathfrak{F}$ to be a triple (F, RF, TF) .

In trying to apply convergence and continuity to $\varepsilon$-functions, we were lead to an investigation of stopping criteria and stability (ch. 3). Out of this came a new and simple definition of stability, the concept of an $\varepsilon$-wave, and a proof that instability can be overcome, given the requisite error bounds.

As presented in chapters 2 and 3, the concepts of subroutine $\varepsilon$-function and stability are machine dependent because they are defined in terms of a fixed machine number system. In the remarks at the end of chapter 3, we show how these concepts can be made machine independent.

The part of the $\varepsilon$-calculus dealing with notions from the calculus is of definitional interest only. For example, one may have wondered whether there is a definition of $\varepsilon$-continuity which satisfies the following: for each fixed precision $\varepsilon$, many numerically computed functions which look possibly continuous at a point $x$, but whose corresponding ideal function is discontinuous at $x$, may be accepted as $\varepsilon$-continuous at $x$; but, as $\varepsilon \to 0$ these functions should be weeded out as $\varepsilon$-discontinuous at $x$ . We found (in sec. 4.3) that it is possible to form such a definition by making use of computable information about (i.e., bounds on) truncation and roundoff errors. We do not expect such definitions to be of practical importance.

On the other hand, the part of the $\varepsilon$-calculus which models scientific computation should have practical implications. Our work on stopping criteria and stability tends in this direction. But we as yet have no concrete applications.

# REFERENCES

[A1] Aberth, Oliver. Analysis in the computable number field. J. Assoc. Comput. Mach. V 15, No. 2 (April 1968) 275-299.

[B1] Bishop, Errett. Foundations of Constructive Analysis. McGraw-Hill. New York. 1967.

[G1] Grzegorczyk, Andrzej. Computable functionals. Fund. Math. 42 (1955) 168-202.

[G2] ————————. On definitions of computable real continuous functions. Fund. Math. 44 (1957) 61-71.

[G3] ————————. Some approaches to constructive analysis. Constructivity in Mathematics. North-Holland. Amsterdam (1959) 43-61.

[K1] Klaua, Dieter. Konstructive Analysis. (German) Deutscher Verlag der Wissenschaften. Berlin. 1961.

[K2] Kreisel, Georg. Interpretations of analysis by means of functionals of finite types. Constructivity in Mathematics. North-Holland. Amsterdam (1959) 101-128.

[M1] Mazur, S. Computable Analysis. Panstwowe Wydawnictwo Naukowe. Warszawa. 1963.

[M2] Mendelson, Elliott. Introduction to Mathematical Logic. Van Nostrand. Princeton, N. J. 1964.

[M3] Moore, Ramon E. Interval Analysis. Printice-Hall. Englewood Cliffs, N. J. 1966.

[M4] ——————. The automatic analysis and control of errors in digital computing based on the use of interval numbers. Error in Digital Computation. John Wiley and Sons. New York. V 1 (1964) 61-130.

[N-Z] Niven, Ivan M., Zuckerman, Herbert S. An Introduction to the Theory of Numbers. John Wiley and Sons. New York. 1966.

[R1] Ralston, Anthony. A First Course in Numerical Analysis. McGraw-Hill. New York. 1965.

[R2] Riesel, Hans. A case of numerical divergence. BIT (1961) 130-1.

[R3] Royden, H. L. Real Analysis. Macmillan. New York. 1964.

[S1] Scott, Dana. Some definitional suggestions for automata theory. Journal of Computer and System Sciences. V 1, No. 2 (August 1967) 187-212.

[S2] Shoenfield, Joseph R. Mathematical Logic. Addison-Wesley. Reading, Mass. 1967.

[W1] van Wijngaarden, A. Numerical analysis as an independent science BIT 6 (1966) 66-81.

[W2] Wilkinson, J. H. Rounding Errors in Algebraic Processes. Prentice-Hall. Englewood Cliffs, N. J. 1963.

**DOCUMENT CONTROL DATA - R & D**

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

| | |
|---|---|
| 1. ORIGINATING ACTIVITY (Corporate author)<br>Computer Science Department<br>Stanford University<br>Stanford, California 94305 | 2a. REPORT SECURITY CLASSIFICATION<br>Unclassified<br>2b. GROUP<br>-L- |

3. REPORT TITLE

ε-CALCULUS

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Manuscript for Publication (Technical Report)

5. AUTHOR(S) (First name, middle initial, last name)

Richman, Paul L.

| | | |
|---|---|---|
| 6. REPORT DATE<br>August 16, 1968 | 7a. TOTAL NO. OF PAGES<br>138 | 7b. NO. OF REFS<br>19 |

| | |
|---|---|
| 8a. CONTRACT OR GRANT NO.<br>N00014-67-A-0112-0029<br>b. PROJECT NO. | 9a. ORIGINATOR'S REPORT NUMBER(S)<br>CS 105 |
| c.<br>d. | 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)<br>none |

10. DISTRIBUTION STATEMENT

Releasable without limitations on dissemination.

| | |
|---|---|
| 11. SUPPLEMENTARY NOTES<br>--- | 12. SPONSORING MILITARY ACTIVITY<br>Office of Naval Research |

13. ABSTRACT

We use recursive function theory to lay the basis for a partially constructive theory of calculus, which we call the ε-calculus. This theory differs from other theories that have grown out of recursive function theory in that

(1) it is directly related to the variable-precision computations used in scientific computation today, and

(2) it deals explicitly with intermediate results rather than ideal answers.

As $\varepsilon \to 0$, intermediate results in the ε-calculus approach their corresponding answers in the calculus. Thus we say "the ε-calculus approaches the calculus, as $\varepsilon \to 0$." It is hoped that investigations in the ε-calculus will lead to a better understanding of numerical analysis. Several new results in this direction are presented, concerning instability and also machine numbers. Discrete notions of limit, convergence, continuity, arithmetic, derivative and integral are also presented and analyzed.

| 14. KEY WORDS | LINK A |  | LINK B |  | LINK C |  |
|---|---|---|---|---|---|---|
|  | ROLE | WT | ROLE | WT | ROLE | WT |
| recursive |  |  |  |  |  |  |
| computable |  |  |  |  |  |  |
| numerical analysis |  |  |  |  |  |  |
| automatic analysis |  |  |  |  |  |  |