INFORMATION TO USERS

This was produced from a copy of a document sent to us for microfilming. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help you understand markings or notations which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure you of complete continuity.

2. When an image on the film is obliterated with a round black mark it is an indication that the film inspector noticed either blurred copy because of movement during exposure, or duplicate copy. Unless we meant to delete copyrighted materials that should not have been filmed, you will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed the photographer has followed a definite method in "sectioning" the material. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For any illustrations that cannot be reproduced satisfactorily by xerography, photographic prints can be purchased at additional cost and tipped into your xerographic copy. Requests can be made to our Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases we have filmed the best available copy.

HOEFER, MAX PETER

NUMERICAL SOLUTIONS OF THE KORTEWEG-DEVRIES EQUATION

City University of New York PH.D. 1981

NUMERICAL SOLUTIONS OF THE

KORTEWEG-DEVRIES EQUATION

by

MAX PETER HOEFER

A dissertation submitted to the Graduate Faculty in Mathematics in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1981

This manuscript has been read and accepted for the Graduate Faculty in Mathematics in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

1/28/81

Date — Chairman of Examining Committee

Date — Executive Officer

Supervisory Committee

The City University of New York

Acknowledgements

I wish to thank all the people who, because of their time, thoughts and understanding enabled me to complete this research. Among many are: The late Professor Harry E. Rauch, who started me on this project; Professors Richard Sacksteder and Stanley Kaplan, who helped me through a very bleak period; and, most of all, Professor Linda Keen, who provided the guidance that led to the successful completion of this paper.

I also wish to thank the City University of New York, which so generously provided me with the use of their facilities at the University Computer Center.

Finally, I wish to express sincere gratitude to my wife and companion Rosalba who, with her quiet confidence, helped me nuture this paper from infancy to full maturity.

Table of Contents

List of Tables

List of Figures

CHAPTER I

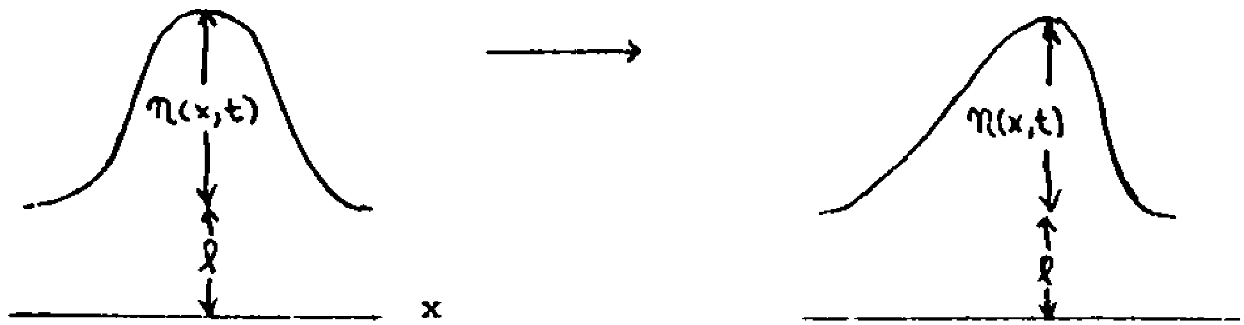
INTRODUCTION

I. HISTORY

In a meeting of the British Association for the Advancement of Science in 1844, J. Scott Russell issued what was in effect a challenge to mathematicians to describe certain types of wave phenomena. In particular, in his "Report on Waves", Russell was interested in obtaining a quantitative description of what he had personally observed as a "great wave of translation" (now popularly called "solitary wave") moving in a shallow channel [1]. Until that time it was believed that all waves moving in shallow rectangular canals would steepen along their forward side and slacken along their backward side (see figure 1.1). The wave that Russell observed, however, did not change shape as it moved down the canal (see figure 1.2) [2].

FIGURE 1.1

Direction of travel



time $t = 0$ — 1.1(a)

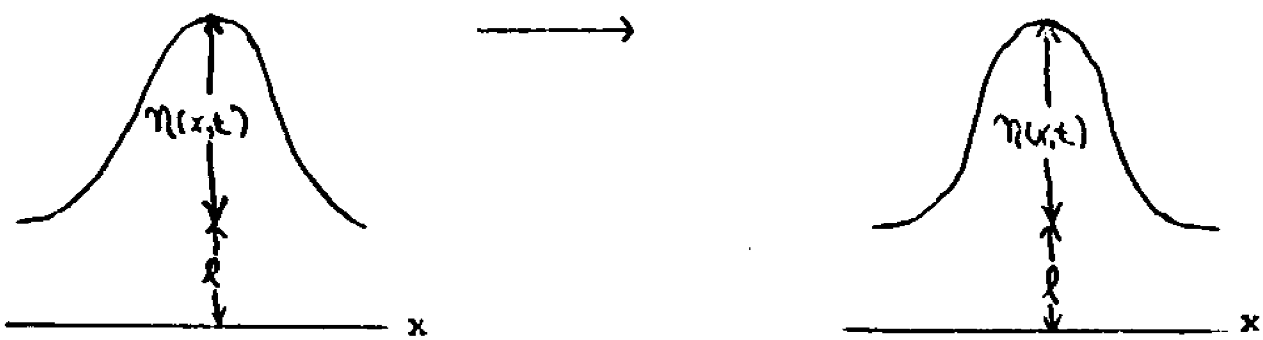
time $t > 0$ — 1.1(b)

Legend: η is the height of the wave above the undisturbed medium,

and ℓ is the depth of the medium.

FIGURE 1.2

Direction of travel



time $t = 0$ — 1.2(a)

time $t > 0$ — 1.2(b)

Legend: Same as Figure 1.1.

In 1895 Korteweg and deVries met the challenge by discovering the equation bearing their name (and hereafter abbreviated "KdV"). They formulated the equation as

$$(1.1)\quad \frac{\partial \eta}{\partial t} = \frac{3}{2}\left(\frac{g}{\ell}\right)^{1/2}\frac{\partial}{\partial x}\left(\frac{1}{2}\eta^2 + \frac{2}{3}\gamma\eta + \frac{\sigma}{3}\frac{\partial^2 \eta}{\partial x^2}\right),$$

$$-\infty < x < \infty,\quad t > 0$$

with

ℓ = equilibrium level of the water,

$\eta(x,t)$ = the surface elevation above ℓ of a shallow water wave in a channel of rectangular cross section moving towards increasing x,

γ = (small) arbitrary constant related to the uniform motion of the liquid,

g = universal gravitational constant, and

$\sigma = \frac{\ell^3}{3} - \frac{T\ell}{\rho g}$ where

T = surface capillary tension and

ρ = density of the water (see figures 1.1,1.2) [2].

In studying their equation, Korteweg and deVries discovered a class of solutions U in terms of the Jacobi elliptic function, cn(Θ). Because of the symbolism, and because a class of intensively studied waves had previously been termed "sinusoidal", these solutions were termed "cnoidal" and the waves were designated "cnoidal waves". Under the assumption $U \to 0$ as $|x| \to \infty$ a specific solution is the single solitary wave (originally mentioned by Russell) in terms of the function sech^2 (see Chapter I, Section II).

From the turn of the century until 1960, no new important application of KdV was discovered. In 1960 Gardner and Morikawa realized that an important equation from plasma physics is, after a change of variables, KdV [3][4]. Since then, many other problems were found that gave rise to KdV and KdV-like equations (see Chapter I, Section III for a brief discussion of a few important papers). This resulted in intensive activity in the study of KdV, activity that has continued until today. It is this activity that provided the impetus for this paper.

II. DISCUSSION OF THE EQUATION

Equation (1.1) can be simplified. First rewrite it in the form

(1.2) $$\eta_t - 3\left(\frac{g}{\ell}\right)^{1/2}\left(\frac{\eta}{2} + \frac{\nu}{3}\right)\eta_x - \frac{\sigma}{2}\left(\frac{g}{\ell}\right)^{1/2}\eta_{xxx} = 0,$$

$-\infty < x < \infty$, $0 < t$.

Then set

(1.3a) $$u(x,t) = \frac{\eta}{2} + \frac{\nu}{3},$$

(1.3b) $t' = \frac{t}{2}$,

(1.3c) $x' = -\frac{1}{6}\left(\frac{\ell}{g}\right)^{1/2} x$, and

(1.3d) $\delta^2 = \frac{\sigma}{216}\left(\frac{\ell}{g}\right) = \frac{1}{216}\left(\frac{\ell^3}{3} - \frac{T\ell}{\rho g}\right)\left(\frac{\ell}{g}\right)$.

Finally substitute these equations into (1.2) and obtain, after dropping the primes from the t's and x's,

(1.4) $u_t + uu_x + \delta^2 u_{xxx} = 0, -\infty < x < +\infty,\ 0 < t.$

Equation (1.4) subject to the initial conditions

(1.4a) $u(x,0) = a \cdot \mathrm{sech}^2(bx)$

has solution

(1.5) $u(x,t) = a \cdot \mathrm{sech}^2(bx - ct)$

providing

(1.6) $b = \frac{1}{2\delta}\left(\frac{a}{3}\right)^{1/2}$

and

(1.7) $c = \frac{a}{6\delta}\left(\frac{a}{3}\right)^{1/2}$,

as is easily verifiable by direct substitution. This is, indeed, Russell's solitary wave moving to the right with velocity $v = \frac{c}{b} = \frac{a}{3}$. Our discussions throughout the remainder of this paper will assume KdV is written in the form (1.4).

KdV is one of the simplest examples of a nonlinear dispersive equation [5]. To understand "nonlinear dispersion", we must first define "linear dispersion". For one dimensional plane waves, one way to define a linear dispersive equation is to start with any linear equation admitting elementary solutions of the form

(1.8) $a \cdot \cos(kx-wt)$

(where

a = wave amplitude,

k = wave number,

x = space coordinate,

w = w(k) = angular frequency (a function of k) and

t = time coordinate)

for which the function w(k) has certain properties. The functional relationship w = w(k) is called the "dispersion relation"; it is the relationship between w and k that identically solves the equation. General solutions are combinations of elementary solutions. The most general solutions are given in terms of Fourier integrals,

$$u(x,t) = \int_0^\infty a(k)\cos(kx-wt)\, dk,$$

when they exist. The "spectrum function" a(k) is determined by initial or boundary conditions.

The "phase velocity" of a wave is defined to be $c_p = \frac{w(k)}{k}$; it is the velocity of each elementary wave train and is the velocity an observer must travel at to maintain observation of a particular "crest". The "group velocity" of a wave is defined to be $c_g = \frac{dw(k)}{dk}$; it is the "local velocity" which an observer must travel at to observe a constant wave number. If $c_p = \frac{w(k)}{k}$ is not constant, crests with different k will travel at different velocities; hence the term "dispersive". Therefore we define a linear dispersive equation to be any equation with solutions of the form (1.8) for which c_p is not constant.

The "linearized form" of a nonlinear equation is arrived at by neglecting all but the first order powers of the dependent variable and its partial derivatives [5]. The linearization of KdV (1.4) is therefore

(1.9) $u_t + u_x + \delta^2 u_{xxx} = 0.$

The dispersion relation is $w = k - \delta^2 k^3$; the phase velocity is $c_p = \frac{w}{k} = 1 - \delta^2 k^2$ and the group velocity is $\frac{dw}{dk} = 1 - 3\delta^2 k^2$. A nonlinear equation is called dispersive if its linearized form is dispersive. Since c_p for (1.9) is not constant, the linearized form of KdV is dispersive and, hence, KdV (1.4) is dispersive.

KdV will have a solitary wave solution provided that the nonlinear term (which tends to make a wave "break") is neutralized by the third degree ("dispersive") term, given appropriate initial conditions. If we ignore the third degree term (1.4) becomes

(1.10) $u_t + uu_x = 0.$

This is a nonlinear hyperbolic equation. Considering (1.10) subject to (1.4a), we can use the method of characteristics to see the wave steepening in front until at $t = t_B = (2ab \cdot \text{sech}^2(bx_B) \cdot \tanh(x_B))^{-1}$ (where $x_B = \text{sech}^{-1}((2b)^{\frac{1}{2}}\tanh(bx_B))/b)$ the wave breaks (the solution becomes non-singular at $x = x_B$). The larger the amplitude "a", the more unstable the wave and, therefore, the faster the wave breaks. The addition of the dispersive term has a dampening effect upon the nonlinear term. δ is often called the "coefficient of dispersion"; when δ is large enough, the third degree term prevents the solution from becoming non-singular. As we have seen, for given "a" and "δ" in (1.4) and (1.4a) there exist values of "b" and "c" that cause the initial profile to be preserved for all t. The result is the solitary wave.

We select the values $a = .9$ in (1.4a) and $\delta = .022$ in (1.4) as did others who have done numerical work on KdV (see Chapter I, Section III for details). The amplitude "a" can be selected arbitrarily, but the coefficient of dispersion δ cannot because it is a physical constant dependent upon surface tension, rest depth and density of the medium (recall (1.3d)). In an early numerical paper on KdV, Zabusky and Kruskal [6] (see also Chapter I, Section III) solved KdV using a finite differencing technique. To begin their method, they ignored the dispersive term for small values of t. To justify this, a small value of δ was required; they selected δ = .022. After inspecting the physical constants in the definition of δ, this assumption does not appear to be unreasonable. Since then, other authors have chosen the same value for δ (as we do) to facilitate comparing their various techniques as applied to KdV.

III. RECENT WORK ON KDV

Since Gardner and Morikawa demonstrated a relationship between KdV and plasma physics, many other problems have been found that give rise to KdV or KdV-like equations. Among the varied problems that yield KdV are ion-acoustic waves in a plasma, longitudinal waves propagating in a one-dimensional lattice of equal masses coupled by nonlinear springs, pressure waves in a liquid-gas mixture, and waves in elastic rods [2]. As was mentioned in Chapter I, Section II, KdV has been studied extensively because it is a standard example of a nonlinear dispersive wave [7].

We have seen what fields have caused interest in the study of KdV. Conversely, KdV has stimulated a wealth of research in theoretical mathematics, especially in fields allied to geometry. In [8], Adler and Möser are interested in a simpler construction of a recently discovered family of rational solutions u to KdV (and higher dimensional KdV-like equations) [9]. The emphasis of their investigation is on the algebraic properties of these solutions. It had previously been demonstrated that this family constitutes a manifold M [9]; moreover, M decomposes into countably many manifolds M_d for d = 1,2, ... By factoring differential operators determined from KdV, the authors construct classes of polynomials of d variables, P_d. They achieve their originally stated goal by demonstrating that each of the previously known manifolds M_d has a representation in terms of P_d.

In [10], Lax is interested in constructing a family of solutions to KdV which are periodic in x (in contrast to the previously mentioned paper). He states (without proof) that constructing solutions to KdV which are periodic in x is more difficult than constructing solutions to KdV assuming non-periodicity. Zabusky and Kruskal (op. cit.) suggested that with sinusoidal initial conditions the solution to KdV would break up into a complicated wave pattern and then recombine into shapes which were "similar" to the initial conditions. Lax cites this work as suggesting to him that he should seek solutions periodic in x and "almost periodic" in t. He constructs a family of solutions to KdV (possessing those properties) variationally by minimizing conserved functionals arising from KdV. He shows that his solutions lie on N-dimensional tori. In particular, the case $N = 1$ corresponds to the elliptic functional ("cnoidal") solutions of Korteweg and deVries.

Since our interest is in numerical solutions to KdV, let us discuss some important results from this area. In 1965, Zabusky and Kruskal (op. cit.) were interested in solving KdV numerically subject to periodic initial conditions. They hoped to show that a wave with sufficient amplitude would "break up" into a number of smaller waves. They solved KdV (3.4) subject to the initial conditions

(1.11) $u(x,0) = -\cos(\pi x), \quad -1 \leq x \leq 1$

(their choice of $\delta = .022$ has already been discussed). The method they used was a finite-differencing technique which gave them solutions discrete in space (x) and time (t). Their results showed the original wave steepen in front and then break up into 8 "wavelets". Since no known analytic solution exists for this initial-value problem, they compared their results with empirical data;

the comparison was favorable. Their results served to stimulate researchers seeking to construct classes of solutions to KdV because it presented properties of sought-after solutions. In 1975 Greig and Morris [11] solved (1.4) subject to initial conditions that cause a solitary wave solution. They substituted $a = .9$ into (1.4a); this makes $b = 12.44824$ and $c = 3.734472$ in (1.6) and (1.7), and (1.4a) becomes

$$(1.12)\quad u(x,0) = .9\operatorname{sech}^2(12.44824x).$$

They knew the analytic solution to KdV subject to (1.12) is

$$(1.13)\quad u(x,t) = .9\operatorname{sech}^2(12.44824x-3.734472t).$$

The method employed in their numerical work is a "hop-scotch" differencing technique giving solutions which are discrete in space and time. They compared their numerical results with the analytic solution (1.13) to demonstrate the validity of their technique. They claimed (without proof) that their "hopscotch" technique gave better results to (1.12) subject to (1.13) than did the method of Zabusky and Kruskal. To carry out the "hopscotch" technique they required more computer space because the method is dependent upon "hopping" from a partition of the spatial interval to points contained within subintervals of this partition. In 1978 Fornberg solved KdV using the "Fast Fourier Transform"[12] . His solution is continuous in space, but the requirement of an expansion in orthogonal functions can be rather severe. For $u(x,0) = .9\text{sech}^2(12.44824x)$ we found a set of non-orthogonal functions that converged much faster to these initial conditions than did other sets of orthogonal functions (see Chapter III). Fornberg also compared the known analytic solution to his numerical approximation to demonstrate that his technique worked.

Our paper presents a new technique for the numerical solution of KdV. Our technique has the advantage that it yields solutions which are continuous in the space variable (x) while not requiring orthogonality conditions for its approximating functions. As will be seen in the next Chapter, convergence of an approximation to a solution of KdV is dependent upon convergence of the approximation to the initial conditions. The non-requirement of orthogonal functions will enable us to select functions that converge more rapidly to the initial conditions.

We begin by outlining the technique that we will use to solve KdV. Since the main concern of this paper is the numerical solution to (1.4) subject to (1.12), we present numerical solutions to this initial value problem; we follow this with a detailed discussion of the results we obtain. We conclude with results obtained from solving (1.4) subject to (1.11) and indicate other techniques which may be used to solve KdV and similar nonlinear time dependent partial differential equations.

CHAPTER II

DEVELOPMENT OF THE ITERATIVE PROCEDURE

We present the iterative process that gives approximate numerical solutions to the Korteweg-deVries equation in this Chapter. The process is a refinement of a procedure suggested by Rosen [13] . We write KdV as

$$u_t = L[u] ,$$

where

$$L[u] = -uu_x - \delta^2 u_{xxx} \text{ and}$$

$$\delta = .022.$$

We assume $u \to 0$ as $|x| \to \infty$, so we approximate it by a function which is 0 (for computing machine accuracy - see Chapter III for a complete discussion) outside some finite interval $[a,b]$. Our process employs linear programming to minimize a measure of error $\varepsilon(t)$ over a set of points $P = \{a=x_0 < x_1 \ldots < x_m = b\}$; the solution is then interpolated to all $[a,b]$. We assume an initial condition $u(x,0) = f(x)$ and boundary conditions $u(a,t) = u(b,t) = 0$.

We approximate $u(x,t)$ in $[a,b]$ by

$$v(x,t) = \sum_{i=0}^{n} a_i(t)\phi_i(x).$$

Each $a_i(t)$ is an unknown real coefficient to be determined, the $\varphi_i(x)$'s are preselected functions (not necessarily orthogonal) that are at least three-times differentiable, and n+1 is the number of functions used in the approximation. Also, initially we choose the set of points $P \subseteq [a,b]$ mentioned above.

We begin at time t = 0. The real unknowns $a_i(0)$ are determined by minimizing the maximum error $\varepsilon(0)$ between $v(x,0)$ and $u(x,0) = f(x)$ over P; that is, by solving

$$(2.1)\quad \max_{x \in P} \left| \sum_{i=0}^{n} a_i(0)\, \varphi_i(x) - f(x) \right| = \min.$$

Reformulating (2.1) into standard linear programming form [14] [15] and then matching the boundary conditions at x = a and x = b requires a number of steps. (2.1) is equivalent to

(2.2) minimize $\varepsilon(0) \geqq 0$ subject to (the constraints)

$$-\varepsilon(0) \leqq \sum_{i=0}^{n} a_i(0)\, \varphi_i(x_0) - f(x_0) \leqq \varepsilon(0)$$

$$\vdots \qquad\qquad \vdots$$

$$-\varepsilon(0) \leqq \sum_{i=0}^{n} a_i(0)\, \varphi_i(x_m) - f(x_m) \leqq \varepsilon(0).$$

Splitting the inequalities and writing everything " $\geqq$ " we obtain

(2.3) minimize $\mathcal{E}(0) \geqq 0$ subject to

$$\sum_{i=0}^{n} a_i(0)\varphi_i(x_0) - f(x_0) \geqq -\mathcal{E}(0)$$

$$\vdots \qquad \vdots$$

$$\sum_{i=0}^{n} a_i(0)\varphi_i(x_m) - f(x_m) \geqq -\mathcal{E}(0)$$

$$-\sum_{i=0}^{n} a_i(0)\varphi_i(x_0) + f(x_0) \geqq -\mathcal{E}(0)$$

$$\vdots \qquad \vdots$$

$$-\sum_{i=0}^{n} a_i(0)\varphi_i(x_m) + f(x_m) \geqq -\mathcal{E}(0).$$

Collecting variables ($a_i(0)$ and $\mathcal{E}(0)$) on the left side and constants ($f(x_i)$) on the right side changes (2.3) to

(2.4) minimize $\mathcal{E}(0) \geqq 0$ subject to

$$\sum_{i=0}^{n} a_i(0)\varphi_i(x_0) + \mathcal{E}(0) \geqq f(x_0)$$

$$\vdots \qquad \vdots$$

$$\sum_{i=0}^{n} a_i(0)\varphi_i(x_m) + \mathcal{E}(0) \geqq f(x_m)$$

$$-\sum_{i=0}^{n} a_i(0)\varphi_i(x_0) + \mathcal{E}(0) \geqq -f(x_0)$$

$$\vdots \qquad \vdots$$

$$-\sum_{i=0}^{n} a_i(0)\varphi_i(x_m) + \mathcal{E}(0) \geqq -f(x_m).$$

Finally, we exactly match the boundary conditions of 0 at $a = x_0$ and $b = x_m$ by changing (2.4) at those points only;

(2.5) minimize $\mathcal{E}(0) \geqq 0$ subject to

$$\sum_{i=0}^{n} a_i(0)\varphi_i(x_0) \geqq 0$$

$$\sum_{i=0}^{n} a_i(0)\varphi_i(x_1) + \mathcal{E}(0) \geqq f(x_1)$$

$$\vdots \qquad \vdots$$

$$\sum_{i=0}^{n} a_i(0)\varphi_i(x_{m-1}) + \mathcal{E}(0) \geqq f(x_{m-1})$$

$$\sum_{i=0}^{n} a_i(0)\varphi_i(x_m) \geqq 0$$

$$-\sum_{i=0}^{n} a_i(0)\varphi_i(x_0) \geqq 0$$

$$-\sum_{i=0}^{n} a_i(0)\varphi_i(x_1) + \mathcal{E}(0) \geqq -f(x_1)$$

$$\vdots \qquad \vdots$$

$$-\sum_{i=0}^{n} a_i(0)\varphi_i(x_{m-1}) + \mathcal{E}(0) \geqq -f(x_{m-1})$$

$$-\sum_{i=0}^{n} a_i(0)\varphi_i(x_m) \geqq 0.$$

We put (2.5) into more convenient form by defining a (2m+2) X (n+2) matrix A:

$$A=\begin{bmatrix}
\varphi_0(x_0) & \varphi_1(x_0) & \cdots & \varphi_n(x_0) & 0\\
\varphi_0(x_1) & \varphi_1(x_1) & \cdots & \varphi_n(x_1) & 1\\
\cdot & \cdot & \cdots & \cdot & \cdot\\
\cdot & \cdot & \cdots & \cdot & \cdot\\
\cdot & \cdot & \cdots & \cdot & \cdot\\
\varphi_0(x_{m-1}) & \varphi_1(x_{m-1}) & \cdots & \varphi_n(x_{m-1}) & 1\\
\varphi_0(x_m) & \varphi_1(x_m) & \cdots & \varphi_n(x_m) & 0\\
-\varphi_0(x_0) & -\varphi_1(x_0) & \cdots & -\varphi_n(x_0) & 0\\
-\varphi_0(x_1) & -\varphi_1(x_1) & \cdots & -\varphi_n(x_1) & 1\\
\cdot & \cdot & \cdots & \cdot & \cdot\\
\cdot & \cdot & \cdots & \cdot & \cdot\\
\cdot & \cdot & \cdots & \cdot & \cdot\\
-\varphi_0(x_{m-1}) & -\varphi_1(x_{m-1}) & \cdots & -\varphi_n(x_{m-1}) & 1\\
-\varphi_0(x_m) & -\varphi_1(x_m) & \cdots & -\varphi_n(x_m) & 0
\end{bmatrix}$$

The column vectors a(0) and b each have n+2 elements,

$$a(0) = \begin{bmatrix} a_0(0)\\ a_1(0)\\ \cdot\\ \cdot\\ \cdot\\ a_n(0)\\ \varepsilon(0)\end{bmatrix}, \quad b = \begin{bmatrix} 0\\ 0\\ \cdot\\ \cdot\\ \cdot\\ 0\\ 1\end{bmatrix},$$

and the column vector c(0) has 2m+2 elements,

$$c(0) = \begin{bmatrix} 0 \\ f(x_1) \\ \cdot \\ \cdot \\ \cdot \\ f(x_{m-1}) \\ 0 \\ 0 \\ -f(x_1) \\ \cdot \\ \cdot \\ \cdot \\ -f(x_{m-1}) \\ 0 \end{bmatrix}$$

We write (2.5) as

(2.6) minimize $b^t \cdot a(0) \geqq 0$ subject to $\Lambda \cdot a(0) \geqq c(0)$.

Now we can solve (2.6) for a(0) using the simplex method of linear programming [14] [15].

Next, we extrapolate the approximate solution from $t = 0$ to $t = h, 2h, \ldots, Nh$, where $0 < h \ll 1$ is the time difference length and N is the number of time differencing steps. To do this, we need an approximation for v_t. For $N = 1$ we set

(2.7) $v_t \approx \frac{v(x,t+h) - v(x,t)}{h}$.

For $N \geq 2$, we set

(2.8) $v_t \approx \frac{v(x,t+h) - v(x,t-h)}{h}$.

The reason for this is that (2.7) is not very stable and causes unreasonably large errors after a number of iterations. We use the leapfrog time differencing scheme (2.8) after N = 1 to be able to iterate a large number of times with relatively small error.

We assume that the approximate solution is known at $t = (k-1)h$, $k \geq 2$, and show how to extrapolate it to $t = kh$, $k \geq 2$ (for k = 1 the formula is similar and uses (2.7)). Since $u(x,(k-1)h) \approx v(x,(k-1)h) = \sum_{i=0}^{N} a((k-1)h)\,\varphi_i(x)$, $L\left[u(x,(k-1)h)\right]$ will be approximated by $L\left[v(x,(k-1)h)\right]$. We determine $v(kh,x)$ by minimizing the maximum difference $\varepsilon(kh)$ between the approximation to v_t (2.8) and $L\left[v(x,(k-1)h)\right]$ over P; that is, by solving

(2.9) $\max_{x \in P} \left| \frac{v(x,kh) - v(x,(k-2)h)}{2h} - L\left[v(x,(k-1)h)\right] \right| = \min.$

Write L_j^{k-1} instead of $L\left[v(x_j,(k-1)h)\right]$ and v_j^{k-2} instead of $v(x_j,(k-2)h)$, simplify and then exactly match boundary conditions as we did above ((2.1) - (2.5)). Now we have

(2.10) minimize $\mathcal{E}(kh) \geqq 0$ subject to

$$
\begin{array}{lllll}
\sum_{i=0}^{N} a_i(kh)\,\varphi_i(x_0) & & & \geqq & 0 \\
\sum_{i=0}^{N} a_i(kh)\,\varphi_i(x_1) & + & \mathcal{E}(kh) & \geqq & v_1^{k-2} + 2hL_1^{k-1} \\
\cdot & & \cdot & \cdot & \cdot \\
\cdot & & \cdot & \cdot & \cdot \\
\cdot & & \cdot & \cdot & \cdot \\
\sum_{i=0}^{N} a_i(kh)\,\varphi_i(x_{m-1}) & + & \mathcal{E}(kh) & \geqq & v_{m-1}^{k-2} + 2hL_{m-1}^{k-1} \\
\sum_{i=0}^{N} a_i(kh)\,\varphi_i(x_m) & & & \geqq & 0 \\
-\sum_{i=0}^{N} a_i(kh)\,\varphi_i(x_0) & & & \geqq & 0 \\
-\sum_{i=0}^{N} a_i(kh)\,\varphi_i(x_1) & + & \mathcal{E}(kh) & \geqq & -(v_1^{k-2} + 2hL_1^{k-1}) \\
\cdot & & \cdot & \cdot & \cdot \\
\cdot & & \cdot & \cdot & \cdot \\
\cdot & & \cdot & \cdot & \cdot \\
-\sum_{i=0}^{N} a_i(kh)\,\varphi_i(x_{m-1}) & + & \mathcal{E}(kh) & \geqq & -(v_{m-1}^{k-2} + 2hL_{m-1}^{k-1}) \\
-\sum_{i=0}^{N} a_i(kh)\,\varphi_i(x_m) & & & \geqq & 0.
\end{array}
$$

Write A and b as above but modify a(kh) and c(kh) to be

$$a(kh) = \begin{bmatrix} a_0(kh) \\ \cdot \\ \cdot \\ \cdot \\ a_n(kh) \\ \varepsilon(kh) \end{bmatrix}, \quad c(kh) = \begin{bmatrix} 0 \\ v_1^{k-2} + 2hL_1^{k-1} \\ \cdot \\ \cdot \\ \cdot \\ v_{m-1}^{k-2} + 2hL_{m-1}^{k-1} \\ 0 \\ 0 \\ -(v_1^{k-2} + 2hL_1^{k-1}) \\ \cdot \\ \cdot \\ \cdot \\ -(v_{m-1}^{k-2} + 2hL_{m-1}^{k-1}) \\ 0 \end{bmatrix}.$$

Now rewrite (2.10) as

(2.11) minimize $b^t \cdot a(kh) \geqq 0$ subject to $A \cdot a(kh) \geqq c(kh)$.

This linear program is now solved in the same manner as (2.6).

Linear programs (2.6) and (2.11) completely describe the iterative process. We remark that, from one iteration to the next, the only physical changes occur in vector c.

In the actual numerical solutions to our linear programs, we solve the maximum duals to (2.6) and (2.11).

These can both be expressed by

$$\text{(2.12) maximize } d^t.c(kh) \text{ subject to } \begin{cases} d \geqq 0 \\ d^t.A = b^t \end{cases}, \ k \geqq 0.$$

Here d is an unknown vector with 2m+2 entries ; A, b and c are as defined above. The simplex method that we employ is a standard algorithm which yields solutions simultaneously to the maximum program (2.12) and its minimum dual (2.11). In computer programming it is more efficient to use an internally stored program than one which must be entered and stored. The University Computer Center of the City University of New York has available a subprogram which implements the simplex method in this form [16] , and we use this to solve our programs.

CHAPTER III

THE SOLITARY WAVE

I. NUMERICAL DATA

In this section we use the methods of the previous chapter to derive approximate numerical solutions to KdV (1.6). Recall that (1.6) with initial conditions $u(x,0) = a\cdot \mathrm{sech}^2(bx)$ will have the solution $u(x,t) = a\cdot \mathrm{sech}^2(bx - ct)$ if $b = \frac{1}{2\delta}\left(\frac{a}{3}\right)^{1/2}$ and $c = \frac{a}{6\delta}\left(\frac{a}{3}\right)^{1/2}$. Following [6] [11], we use $\delta = .022$ and $a = .9$; this implies $b = 12.44824$ and $c = 3.734472$.

We choose our parameters as follows. First, we take -1 and +1 as endpoints of the spatial interval. The remainder of the points in P are chosen between the endpoints at intervals of .04; that is, $P = \{-1, -.96, \ldots, .96, 1\}$. Thus our parameter m is 50. Next we choose the functions $\varphi_i(x)$ by setting $\varphi_i(x) = H_i(bx)$, where $H_i(y)$ is the i^{th} Hermite function. The Hermite functions are defined as follows:

$$H_0(y) = \exp(-y^2/2)$$

$$H_1(y) = 2y\cdot \exp(-y^2/2)$$

$$H_{i+1}(y) = 2yH_i(y) - 2iH_{i-1}(y),\quad i=1,2, \ldots , n-1.$$

Below we present two approximate numerical solutions, one obtained by choosing 10 functions (n = 9) and one obtained by choosing 20 functions (n = 19), and compare them with the analytic solution to KdV. We label the approximate solutions v_1 and v_2, respectively, and the analytic solution u. In each case a time differencing step of h = .001 is used.

For our solution with n = 9, Tables 3.1 - 3.3 give a sampling of computed values of the coefficients $a_i(t)$ (i = 0,1, ..., 9) for t = 0, t = .25 and t = .5. The numbers are rounded to three significant figures and listed in the computer version of scientific notation (thus, .213 x 10^{-3} will be seen as .213D-03).

TABLE 3.1

$a_i(0)$ i = 0, ..., 9

i	$a_i(0)$	i	$a_i(0)$
0	.770D 00	5	.381D-18
1	.239D-16	6	-.924D-04
2	-.390D-01	7	-.645D-19
3	-.658D-17	8	.378D-05
4	.258D-02	9	.163D-20

TABLE 3.2

$a_i(.25)$ $i = 0, \ldots, 9$

i	$a_i(.25)$	i	$a_i(.25)$
0	.582D 00	5	.479D-04
1	.335D 00	6	.116D-03
2	.746D-01	7	-.776D-05
3	.982D-03	8	-.288D-05
4	-.216D-02	9	-.125D-05

TABLE 3.3

$a_i(.5)$ $i = 0, \ldots, 9$

i	$a_i(.5)$	i	$a_i(.5)$
0	.279D 00	5	.943D-03
1	.294D 00	6	-.152D-04
2	.156D 00	7	-.776D-05
3	.505D-01	8	-.147D-05
4	.997D-02	9	.155D-05

Tables 3.4 - 3.6 compare the analytic solution $u(x,t)$ to the calculated solution $v_1(x,t)$ and show the absolute value of the difference between u and v_1 at selected points $x \in P$ for times $t = 0$, $t = .25$ and $t = .5$. At the end of each table we show the maximum error up to that time over all points $x \in P$ (we call that error $\bar{\varepsilon}(t)$ so not to confuse it with the iteration error $\varepsilon(t)$ - thus $\bar{\varepsilon}(t) = \max_{t' \leq t} \left\{ \max_{x \in P} |u(x,t') - v(x,t')| \right\}$).

TABLE 3.4

x	$v_1(x,0)$	$u(x,0)$	$\lvert v_1 - u \rvert$
-.40	.595D-03	.170D-03	.425D-03
-.36	.208D-02	.461D-03	.162D-02
-.32	.449D-02	.125D-02	.325D-02
-.28	.549D-02	.337D-02	.212D-02
-.24	.601D-02	.910D-02	.309D-02
-.20	.212D-01	.242D-01	.325D-02
-.16	.679D-01	.646D-01	.325D-02
-.12	.163D 00	.164D 00	.130D-02
-.08	.377D 00	.380D 00	.325D-02
-.04	.712D 00	.709D 00	.325D-02
0	.897D 00	.900D 00	.325D-02
.04	.712D 00	.709D 00	.325D-02
.08	.377D 00	.380D 00	.325D-02
.12	.163D 00	.164D 00	.130D-02
.16	.679D-01	.646D-01	.325D-02
.20	.212D-01	.242D-01	.325D-02
.24	.601D-02	.910D-02	.309D-02
.28	.549D-02	.337D-02	.212D-02
.32	.449D-02	.125D-02	.325D-02
.36	.208D-02	.461D-03	.162D-02
.40	.595D-03	.170D-03	.425D-03

Maximum error $\mathcal{E}(0)$ = .325D-02

TABLE 3.5

x	$v_1(x,.25)$	$u(x,.25)$	$\|v_1 - u\|$
-.40	.143D-02	.263D-04	.140D-02
-.36	.424D-02	.713D-04	.417D-02
-.32	.719D-02	.193D-03	.699D-02
-.28	.441D-02	.522D-03	.389D-02
-.24	-.391D-02	.141D-02	.531D-02
-.20	-.356D-02	.382D-02	.738D-02
-.16	.906D-02	.103D-01	.124D-02
-.12	.296D-01	.276D-01	.201D-02
-.08	.785D-01	.728D-01	.571D-02
-.04	.181D 00	.184D 00	.269D-02
0	.410D 00	.417D 00	.758D-02
.04	.755D 00	.749D 00	.661D-02
.08	.898D 00	.897D 00	.151D-02
.12	.664D 00	.668D 00	.401D-02
.16	.343D 00	.345D 00	.257D-02
.20	.159D 00	.147D 00	.112D-01
.24	.628D-01	.573D-01	.543D-02
.28	.718D-02	.216D-01	.144D-01
.32	-.101D-01	.804D-02	.182D-01
.36	-.712D-02	.298D-02	.101D-01
.40	-.241D-02	.110D-02	.351D-02

Maximum error $\overline{\mathcal{E}}(.25)$ = .188D-01

(occurred when t = .24)

TABLE 3.6

x	$v_1(x,.5)$	$u(x,.5)$	$\|v_1 - u\|$
-.40	-.282D-02	.407D-05	.282D-02
-.36	-.897D-02	.110D-04	.899D-02
-.32	-.169D-01	.298D-04	.169D-01
-.28	-.135D-01	.807D-04	.136D-01
-.24	.654D-02	.218D-03	.633D-02
-.20	.131D-01	.591D-03	.125D-01
-.16	-.528D-02	.160D-02	.688D-02
-.12	-.144D-02	.432D-02	.579D-02
-.08	.207D-01	.117D-01	.901D-02
-.04	.353D-01	.312D-01	.453D-02
0	.861D-01	.820D-01	.409D-02
.04	.200D 00	.205D 00	.517D-02
.08	.452D 00	.456D 00	.459D-02
.12	.789D 00	.785D 00	.373D-02
.16	.883D 00	.886D 00	.341D-02
.20	.633D 00	.625D 00	.780D-02
.24	.313D 00	.313D 00	.316D-04
.28	.120D 00	.131D 00	.107D-01
.32	.405D-01	.508D-01	.102D-01
.36	.121D-01	.191D-01	.698D-02
.40	.296D-02	.710D-02	.414D-02

Maximum error $\overline{\mathcal{E}}(.5)$ = .208D-01

(occurred when t = .47)

Appendices 1 - 3 depict $u(x,t)$ (solid line) and $v_1(x,t)$ (dashed line) for the three times given above, respectively. In some cases the solutions are indistinguishable on a graph.

Now we repeat the above procedure for $n = 19$ (20 functions). Tables 3.7 - 3.9 give a sampling of the computed values of the coefficients $a_i(t)$ ($i = 0, \ldots, 19$) for $t = 0$, $t = .5$ and $t = 1.0$.

TABLE 3.7

$a_i(0)$ $i = 0, \ldots, 19$

i	$a_i(0)$	i	$a_i(0)$
0	.771D 00	10	-.776D-07
1	.543D-17	11	-.505D-23
2	-.387D-01	12	.202D-08
3	-.907D-18	13	-.134D-24
4	.258D-02	14	-.400D-10
5	.112D-18	15	.846D-27
6	-.855D-04	16	.855D-12
7	-.109D-19	17	.204D-27
8	.309D-05	18	-.165D-13
9	.193D-21	19	-.334D-29

TABLE 3.8

$a_i(.5)$ i = 0, ..., 19

i	$a_i(.5)$	i	$a_i(.5)$
0	.278D 00	10	.179D-06
1	.296D 00	11	.308D-08
2	.156D 00	12	-.272D-08
3	.498D-01	13	.554D-11
4	.987D-02	14	.652D-10
5	.105D-02	15	.467D-11
6	-.118D-04	16	.516D-12
7	-.123D-04	17	-.246D-12
8	.323D-05	18	.340D-13
9	.130D-05	19	.913D-14

TABLE 3.9

$a_i(1.0)$ $i = 0, \ldots, 19$

i	$a_i(1.0)$	i	$a_i(1.0)$
0	.176D-01	10	.354D-05
1	.316D-01	11	.514D-06
2	.306D-01	12	.597D-07
3	.206D-01	13	.577D-08
4	.106D-01	14	.664D-09
5	.436D-02	15	.698D-10
6	.148D-02	16	.952D-11
7	.411D-03	17	.156D-11
8	.976D-04	18	.141D-12
9	.201D-04	19	.304D-13

Tables 3.10 - 3.12 compare $u(x,t)$ to $v_2(x,t)$ and show the absolute value of the difference between u and v_2 at selected points $x \in P$ for times $t = 0$, $t = .5$ and $t = 1.0$. Given at the end of each table is the maximum error up to that time, $\bar{\epsilon}(t)$.

TABLE 3.10

x	$v_2(x,0)$	$u(x,0)$	$\|v_2 - u\|$
-.40	.363D-03	.170D-03	.193D-03
-.36	.516D-03	.461D-03	.551D-04
-.32	.105D-02	.125D-02	.193D-03
-.28	.357D-02	.337D-02	.193D-03
-.24	.891D-02	.910D-02	.193D-03
-.20	.246D-01	.244D-01	.193D-03
-.16	.644D-01	.646D-01	.193D-03
-.12	.165D 00	.164D 00	.193D-03
-.08	.380D 00	.380D 00	.193D-03
-.04	.709D 00	.709D 00	.193D-03
0	.900D 00	.900D 00	.193D-03
.04	.709D 00	.709D 00	.193D-03
.08	.380D 00	.380D 00	.193D-03
.12	.165D 00	.164D 00	.193D-03
.16	.644D-01	.646D-01	.193D-03
.20	.246D-01	.244D-01	.193D-03
.24	.891D-02	.910D-02	.193D-03
.28	.357D-02	.337D-02	.193D-03
.32	.105D-02	.125D-02	.193D-03
.36	.516D-03	.461D-03	.551D-04
.40	.363D-03	.170D-03	.193D-03

Maximum error $\mathcal{E}(0)$ = .193D-03

TABLE 3.11

x	$v_2(x,.5)$	$u(x,.5)$	$\|v_2 - u\|$
-.40	.130D-02	.407D-05	.130D-02
-.36	-.177D-02	.110D-04	.177D-02
-.32	.161D-03	.298D-04	.131D-03
-.28	.348D-02	.807D-04	.340D-02
-.24	-.183D-02	.218D-03	.205D-02
-.20	-.997D-03	.591D-03	.159D-02
-.16	.122D-02	.160D-02	.379D-03
-.12	.819D-02	.432D-02	.387D-02
-.08	.112D-01	.117D-01	.408D-03
-.04	.281D-01	.312D-01	.307D-02
0	.812D-01	.820D-01	.739D-03
.04	.210D 00	.205D 00	.422D-02
.08	.453D 00	.456D 00	.349D-02
.12	.786D 00	.785D 00	.342D-03
.16	.887D 00	.886D 00	.108D-02
.20	.625D 00	.625D 00	.445D-04
.24	.312D 00	.313D 00	.623D-03
.28	.130D 00	.131D 00	.750D-03
.32	.529D-01	.508D-01	.212D-02
.36	.186D-01	.191D-01	.537D-03
.40	.537D-02	.710D-02	.173D-02

Maximum error $\overline{E}(.5)$ = .542D-02

(occurred when t = .39)

TABLE 3.12

x	$v_2(x,1.0)$	$u(x,1.0)$	$\lvert v_2 - u \rvert$
-.40	.443D-02	.972D-07	.443D-02
-.36	.273D-02	.263D-06	.273D-02
-.32	-.404D-02	.712D-06	.405D-02
-.28	.245D-02	.193D-05	.245D-02
-.24	.156D-02	.522D-05	.156D-02
-.20	-.677D-03	.141D-04	.691D-03
-.16	-.365D-03	.382D-04	.403D-03
-.12	-.909D-03	.103D-03	.101D-02
-.08	-.342D-03	.280D-03	.622D-03
-.04	.576D-02	.758D-03	.500D-02
0	-.615D-02	.205D-02	.820D-02
.04	.104D-01	.554D-02	.485D-02
.08	.195D-01	.149D-01	.456D-02
.12	.271D-01	.398D-01	.126D-01
.16	.106D 00	.104D 00	.254D-02
.20	.261D 00	.255D 00	.678D-02
.24	.540D 00	.539D 00	.131D-02
.28	.840D 00	.846D 00	.672D-02
.32	.842D 00	.846D 00	.443D-02
.36	.546D 00	.539D 00	.726D-02
.40	.258D 00	.255D 00	.362D-02

Maximum error $\bar{\mathcal{E}}(1.0)$ = .126D-01

Appendices 4 - 6 depict $u(x,t)$ (solid line) and $v_2(x,t)$ (dashed line) for $t = 0$, $t = .5$ and $t = 1.0$. As before, in some cases the two solutions are indistinguishable when shown on a graph.

II. DISCUSSION

In the choice of the $\varphi_i(x)$'s our first consideration was the following. For our method a necessary condition for convergence of the approximating numerical solutions to the actual solution of second order equations is that the maximum of the error at $t = 0$, $\varepsilon(0)$, go to 0 [13][17]. Though our equation is not second order, the convergence condition suggested we should choose functions which converge rapidly to the initial conditions. We tried various sets of functions. Since the use of orthogonal functions is an efficient way of spanning particular function spaces, and since much is known about the properties of these functions, our first thought was to use a set of orthogonal functions. Over $(-1,+1)$ the Chebyschev polynomials are orthogonal with weight $(1-x^2)^{-\frac{1}{2}}$ but they converged very slowly to our initial conditions. The set $\{1, \cos(k\pi x), \sin(k\pi x)\}$ $(k = 1, \ldots, n)$ did not converge much faster nor did even or odd expansions formed from this set. The speed of convergence of the approximating series to the initial conditions is necessary for the practical implementation of any algorithm on a digital computer. As the number of functions in the approximation increases, so does required computer time and space (usually this is a nonlinear relationship). Since the obvious orthogonal sets converged so slowly to $a \cdot \text{sech}^2(bx)$ that the computer time and space needed would have been unreasonably large, we were forced to look elsewhere. An important characteristic of the linear programming technique enabled us to do this; nowhere does this approach use an orthogonality property of its approximating functions, as do many methods currently being used. Thus we were able to try non-orthogonal sets of functions. Since the Hermite functions $H_i(x)$ have exponential factors, they looked like good candidates for approximating $a \cdot \text{sech}^2(bx)$. They are orthogonal over $(-\infty, +\infty)$, but not over $[-1, +1]$. The "almost orthogonal" behavior of this set, together with the fact that each function goes to 0 as $|x| \to \infty$ as do our initial conditions, suggested we try them. As expected, this set of functions converged much faster to $a \cdot \text{sech}^2(bx)$ than sets previously used. Finally, the "b" in the initial conditions suggested we try $H_i(bx)$ $(i=0, \ldots, n)$, and this set did indeed converge fastest of all sets that we considered to $a \cdot \text{sech}^2(bx)$.

The selection $\delta = .022$ was made so that we could compare our results with work done in the past using the same value for that parameter [6] [11] . We tried various other values for δ , and our work showed the maximum error varies very little (see Table 3.13; in it n = 9 and $0 \leq t \leq .5$).

TABLE 3.13

δ	$\max_{x,t} \lvert u(x,t) - v_1(x,t) \rvert$
.625D-03	.230D-01
.220D-01	.208D-01
.100D 00	.336D-01

This is not surprising since when both u and v are considered to be functions of δ , $u(x,t,\delta)$ and $v(x,t,\delta)$, they depend continuously on δ for $\delta > 0$. However, since "b" depends inversely upon δ , increasing the size of δ decreases the size of "b" which has the effect of flattening the wave. If "a" is held constant, this increases $a \cdot \text{sech}^2(bx)$ at all $x \neq 0$. This means that if δ is made so large that $a \cdot \text{sech}^2(bx)$ can no longer be considered aproximately 0 at -1 and +1, the interval would have to be increased until the value at its endpoints was approximately 0 for us to be able to employ our method.

The maximum error $\overline{E}(t)$ is calculated over a finite number of points $P \subseteq [-1,+1]$ but is a good estimate of the sup error over $[-1,+1]$. Consider the solution found for n = 9 and estimate the sup error by

$$(3.1)\quad \bar{\mathcal{E}}_T(t) = \max_{x\in P} \left\{ |u(x,t)-v_1(x,t)| + \Delta x\, |u_x(x,t)-v_{1x}(x,t)| \right\},$$

with $\Delta x = .04$ (the distance between adjacent points). For each t, $v_1(x,t)$ is differentiable with respect to x since it is the finite sum of functions which are differentiable with respect to x. Since

$$\frac{d}{dx} H_i(bx) = 2biH_{i-1}(bx) - b^2 x H_i(bx),$$

it follows that

$$v_{1x} = \frac{\partial}{\partial x} \sum_{i=0}^{9} a_i(t) H_i(bx)$$
$$= \sum_{i=0}^{9} a_i(t) \left[2iH_{i-1}(bx) - b^2 x H_i(bx) \right]$$

for all t. Substituting this result into (3.1), we calculated $\bar{\mathcal{E}}_T(t)$ for all values of t in the iterative procedure. Table 3.14 presents a sampling of these results, along with corresponding $\bar{\mathcal{E}}$'s for comparison.

TABLE 3.14

t	$\bar{\mathcal{E}}(t)$	$\bar{\mathcal{E}}_T(t)$
0	.523D-02	.118D-01
.25	.187D-01	.296D-01
.5	.208D-01	.329D-01

The ratio of the number of functions n to the number of points m is important. First let us fix the number of points. The two solutions given earlier (v_1 and v_2) have number of functions n to number of points m ratios R of $R \approx \frac{1}{5}$ and $R \approx \frac{2}{5}$, respectively. We observed that if we raised the ratio R to a number greater than $\frac{3}{5}$, the procedure became unstable after a small number of iterations. We can explain why this occurs by looking at the first iteration when t = 0. We note that at this time the analytic solution $u(x,0) = a \cdot \text{sech}^2(bx)$ has only one critical point (x = 0) and it changes direction there. The approximate solution v(x,0) is the product of an n^{th} degree polynomial with $\exp(-(bx)^2/2)$ which may have as many as n+1 critical points where it may change direction. Suppose we have found $a_i(0)$ (i = 0, ..., n) and $\mathcal{E}(0)$ which satisfy

$$\min_{a_i(0)} \left\{ \max_{x \in P} \left| u(x,0) - \sum_{i=0}^{n} a_i(0) \Phi_i(x) \right| \right\} = \mathcal{E}(0).$$

Select n^* such that $n < n^* < m$ and suppose a solution to the corresponding linear program when t = 0 is $a_i^*(0)$ (i = 0, ..., n^*) and $\mathcal{E}^*(0)$. Since the second solution has more functions, and therefore can more easily approximate the conditions $a \cdot \text{sech}^2(bx)$, $\mathcal{E}^*(0) \leq \mathcal{E}(0)$; practically, $\mathcal{E}^*(0) < \mathcal{E}(0)$. This is because the addition of higher degree polynomial terms (times $\exp(-(bx)^2/2)$) causes the maximum error measured at the grid points $x \in P$ to diminish. However, a polynomial of higher degree in general (times $\exp(-(bx)^2/2)$) has more critical points and therefore so does $v^*(x,0)$. This means we can expect larger errors in the derivative ($|u_x(x,0)-v_x(x,0)| < |u_x(x,0)-v^*_x(x,0)|$) for some $x \in I$ (see figure 3.1).

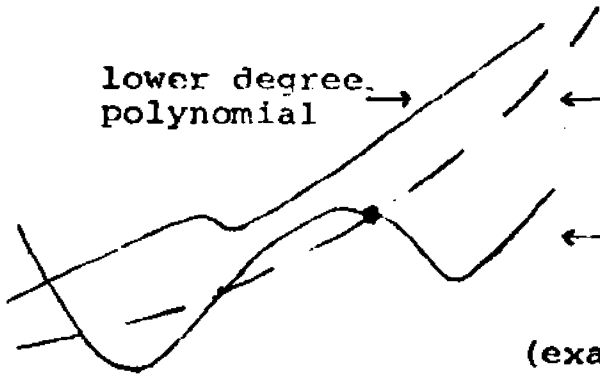


FIGURE 3.1

If n* is allowed to become too large, the supremum of the error over $[-1,+1]$ of $v^*(x,0)$ will be much greater than the error for $v(x,0)$.

This is precisely what was observed in our numerical work. Increasing the number of functions while holding the number of grid points fixed causes the approximation v^* to have a larger maximum error in its derivative for some $x \in P$ and hence a larger $\mathcal{E}^*_T$; this means we have $\mathcal{E}^*_T(0) > \mathcal{E}_T(0)$ even though $\mathcal{E}^*(0) < \mathcal{E}(0)$. As a specific example, let $v_3(x,t)$ be an approximate solution to KdV using 5 Hermite functions and 21 equally spaced points in $[-1,+1]$ (call this grid selection P_{21}) and let v^*_3 be the same except that it uses 10 functions. For $v^*_3(x,0)$ we have $\mathcal{E}^*(0) = .240D-07$ while for $v_3(x,0)$ we have

$\mathcal{E}(0)$ = .143D-02 ($\mathcal{E}^*(0) < \mathcal{E}(0)$). However, for $v_3(x,0)$ we have $\mathcal{E}_T(0)$ = .317D-01 while for $v^*_3(x,0)$ we have $\mathcal{E}^*_T(0)$ = .125D 00 ($\mathcal{E}_T(0) < \mathcal{E}^*_T(0)$). Our observations confirm that the error in the derivative of $v^*_3(x,0)$ is greater than the error in the derivative of $v_3(x,0)$ since

$$\max_{x \in P_{21}} |u_x(x,0) - v^*_{3_x}(x,0)| = 1.25$$

while

$$\max_{x \in P_{21}} |u_x(x,0) - v_{3_x}(x,0)| = .303.$$

Finally, we note that at intermediate points (between the 21 points in P_{21}) the size of the errors are consistent with what we observed with the $\mathcal{E}_T$'s. Using our original P we have $\mathcal{E}(0)$ = .264D-01 while $\mathcal{E}^*(0)$ = 1.74 and

$$\max_{x \in P} |u_x(x,0) - v_{3_x}(x,0)| = .907$$

while

$$\max_{x \in P} |u_x(x,0) - v^*_{3_x}(x,0)| = 2.67.$$

We conclude that as R gets large (approaches 1), $\mathcal{E}^*_T(0) \gg \mathcal{E}_T(0)$ and the iteration process may become unstable after a few iterations. This is especially true for mathematical models describing spatial movement such as KdV. From the work we did, the results show that the best ratio R will in general depend upon the number of points in P. For m = 51, it appears best for R to satisfy $\frac{1}{5} \leq R \leq \frac{1}{2}$.

CHAPTER IV

PERIODIC INITIAL CONDITIONS

In this chapter we solve KdV over $[-1,+1]$ with the periodic initial conditions (1.11) ($f(x) = -\cos(\pi x)$) using the technique developed in Chapter II. We select $\delta = .022$ and $P = \{-1, -.96, -.92, \ldots, +.96, +1\}$ with parameter $m = 50$ as we did in Chapter III. Because of the periodic initial conditions we choose periodic approximating functions. Specifically,

$$\phi_i(x) = \begin{cases} \cos(i\pi x) & 0 \leq i \leq 4 \\ \\ \sin(i\pi x) & 5 \leq i \leq 9, \end{cases}$$

with parameter $n = 9$. This set is orthogonal over $[-1,+1]$ in contrast with the set used in Chapter III. The time differencing step is $h = .001$. We label our approximate solution $v(x,t)$ (we know of no analytic solution). Because of the periodic initial conditions and because of our use of approximating functions with the same period, the artificial boundary conditions at $x = -1$ and $x = +1$ in Chapter III can be ignored. Thus we use system (2.4) in lieu of system (2.5) and matrix A becomes

$$A = \begin{bmatrix} \varphi_0(x_0) & \varphi_1(x_0) & \cdots & \varphi_n(x_0) & 1 \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & & \cdots & \cdot & \cdot \\ \varphi_0(x_m) & \varphi_1(x_m) & \cdots & \varphi_n(x_m) & 1 \\ -\varphi_0(x_0) & -\varphi_1(x_0) & \cdots & -\varphi_n(x_0) & 1 \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ -\varphi_0(x_m) & -\varphi_1(x_m) & \cdots & -\varphi_n(x_m) & 1 \end{bmatrix} .$$

The vectors c become

$$c(0) = \begin{bmatrix} f(x_0) \\ \cdot \\ \cdot \\ \cdot \\ f(x_m) \\ -f(x_0) \\ \cdot \\ \cdot \\ \cdot \\ -f(x_m) \end{bmatrix} , \quad c(kh) = \begin{bmatrix} v_0^{k-2} + 2hL_0^{k-1} \\ \cdot \\ \cdot \\ \cdot \\ v_m^{k-2} + 2hL_m^{k-1} \\ -(v_0^{k-2} + 2hL_0^{k-1}) \\ \cdot \\ \cdot \\ \cdot \\ -(v_m^{k-2} + 2hL_m^{k-1}) \end{bmatrix}$$

(the v's and L's are defined in Chapter II). Vectors "a" and "b" remain as in Chapter II.

Tables 4.1 - 4.3 give a sampling of the computed values of the coefficients $a_i(t)$ $(i = 0, \ldots, 9)$ for $t = 0$, $t = .25$ and $t = .5$.

TABLE 4.1

$a_i(0)$ $i = 0, \ldots, 9$

i	$a_i(0)$	i	$a_i(0)$
0	.268D-16	5	.485D-16
1	-.100D 01	6	.143D-15
2	.125D-16	7	.954D-16
3	.600D-16	8	.806D-16
4	-.519D-16	9	.401D-16

TABLE 4.2

$a_i(.25)$ $i = 0, \ldots, 9$

i	$a_i(.25)$	i	$a_i(.25)$
0	-.384D-03	5	.171D-02
1	-.983D 00	6	.316D 00
2	.395D-02	7	-.816D-02
3	.157D 00	8	-.162D 00
4	-.436D-02	9	.783D-02

TABLE 4.3

$a_i(.5)$ $i = 0, \ldots, 9$

i	$a_i(.5)$	i	$a_i(.5)$
0	.296D-01	5	.622D-01
1	-.432D 00	6	.259D 00
2	.447D-02	7	-.713D-01
3	.132D 00	8	-.104D 01
4	-.484D-01	9	.657D-01

Since $a_1(0) = -1$ and $a_i(0) = 0$ $(i \neq 1)$ (0 with respect to the machine accuracy of the IBM 370 which is approximately 10^{-15} [18]), $v(x,0)$ exactly matches the initial conditions. Tables 4.4 - 4.6 show the values of the approximate function $v(x,t)$ at all points $x \in P$ for times $t = 0$, $t = .25$ and $t = .5$.

TABLE 4.4

x	v(x,0)	x	v(x,0)	x	v(x,0)
-1.00	.100D 01	-.32	-.536D 00	.36	-.426D 00
-.96	.992D 00	-.28	-.637D 00	.40	-.309D 00
-.92	.969D 00	-.24	-.729D 00	.44	-.187D 00
-.88	.930D 00	-.20	-.809D 00	.48	-.628D-01
-.84	.876D 00	-.16	-.876D 00	.52	.628D-01
-.80	.809D 00	-.12	-.930D 00	.56	.187D 00
-.76	.729D 00	-.08	-.969D 00	.60	.309D 00
-.72	.637D 00	-.04	-.992D 00	.64	.426D 00
-.68	.536D 00	0	-.100D 01	.68	.526D 00
-.64	.426D 00	.04	-.992D 00	.72	.637D 00
-.60	.309D 00	.08	-.969D 00	.76	.729D 00
-.56	.187D 00	.12	-.930D 00	.80	.809D 00
-.52	.628D-01	.16	-.876D 00	.84	.876D 00
-.48	-.628D-01	.20	-.809D 00	.88	.930D 00
-.44	-.187D 00	.24	-.729D 00	.92	.969D 00
-.40	-.309D 00	.28	-.637D 00	.96	.992D 00
-.36	-.426D 00	.32	-.536D 00	1.00	.100D 01

TABLE 4.5

x	v(x,.25)	x	v(x,.25)	x	v(x,.25)
-1.00	.766D 00	-.32	-.408D 00	.36	-.469D 00
-.96	.848D 00	-.28	-.352D 00	.40	-.334D-01
-.92	.950D 00	-.24	-.508D 00	.44	.105D 00
-.88	.102D 01	-.20	-.762D 00	.48	.114D 00
-.84	.971D 00	-.16	-.951D 00	.52	-.100D 00
-.80	.765D 00	-.12	-.101D 01	.56	-.196D 00
-.76	.491D 00	-.08	-.948D 00	.60	-.100D-02
-.72	.344D 00	-.04	-.852D 00	.64	.436D 00
-.68	.418D 00	0	-.768D 00	.68	.884D 00
-.64	.650D 00	.04	-.689D 00	.72	.111D 01
-.60	.850D 00	.08	-.611D 00	.76	.105D 01
-.56	.695D 00	.12	-.571D 00	.80	.835D 00
-.52	.256D 00	.16	-.635D 00	.84	.635D 00
-.48	-.306D 00	.20	-.820D 00	.88	.567D 00
-.44	-.714D 00	.24	-.103D 01	.92	.612D 00
-.40	-.801D 00	.28	-.109D 01	.96	.692D 00
-.36	-.625D 00	.32	-.892D 00	1.00	.766D 00

TABLE 4.6

x	v(x,.5)	x	v(x,.5)	x	v(x,.5)
-1.00	.323D 00	-.32	.104D 01	.36	-.108D 00
-.96	-.581D 00	-.28	.636D 00	.40	.540D 00
-.92	-.549D 00	-.24	-.511D 00	.44	.130D 01
-.88	.419D 00	-.20	-.138D 01	.48	.844D 00
-.84	.139D 01	-.16	-.125D 01	.52	-.373D 00
-.80	.141D 01	-.12	-.321D 00	.56	-.120D 01
-.76	.432D 00	-.08	.514D 00	.60	-.834D 00
-.72	-.627D 00	-.04	.493D 00	.64	.444D 00
-.68	-.746D 00	0	-.327D 00	.68	.151D 01
-.64	.198D 00	.04	-.114D 01	.72	.149D 01
-.60	.130D 01	.08	-.117D 01	.76	.465D 00
-.56	.148D 01	.12	-.331D 00	.80	-.529D 00
-.52	.449D 00	.16	.601D 00	.84	-.556D 00
-.48	-.957D 00	.20	.735D 00	.88	.354D 00
-.44	-.157D 00	.24	-.123D 00	.92	.126D 01
-.40	-.949D 00	.28	-.125D 01	.96	.124D 01
-.36	.301D 00	.32	-.161D 01	1.00	.323D 00

Appendices 7 - 9 at the end of the paper depict the corresponding wave shapes. Since v matches the initial conditions exactly, $v(x,0) = -\cos(\pi x)$ $(-1 \leq x \leq +1)$ and we observe a wave with two crests (at x = -1 and at x = +1) in Table 4.4 and Appendix 7. As t increases the two initial waves break up and form numerous wavelets. When t = .25, Table 4.5 and Appendix 8 show a wave pattern with 6 peaks. By the time t = .5 we see a wave pattern which has stabilized with 8 peaks (Table 4.6 and Appendix 9). We have no analytic solution with which to compare our results, but we can compare our work with the numerical work of Zabusky and Kruskal mentioned in Chapter I, Section III. In it they show the original waves changing into a more complex pattern containing eight wavelets , agreeing with our results.

CHAPTER V

CONCLUSION

I. FINAL COMMENTS ON THE LINEAR PROGRAMMING TECHNIQUE

The linear programming approach is an effective, efficient method for solving KdV and should also prove to be so for other equations of the form $u_t = F[u]$ (F a not-necessarily linear differential operator, t not necessarily a time variable). It is superior to finite-differencing techniques in space and time with respect to the fact that it gives solutions which are continuous in the space variable; it has an advantage over the Fast-Fourier Transform because it does not require its set of approximating functions to be orthogonal. In particular, the latter property allows us to calculate accurate approximate solutions to the solitary wave problem using a small number of approximating functions.

Experiments where we used the linear programming approach to other time-dependent partial differential equations have indicated that the method does not only perform well on KdV. A simple linear example

$$u_t = u_{xx} \quad (0 < x < 1, \ t > 0)$$

subject to

$$u(x,0) = \sin(\pi x) \quad (0 < x < 1)$$

and

$$u(0,t) = u(1,t) = 0 \quad (0 \leq t)$$

(the heat equation) has as its Fourier Series solution

$$u(x,t) = \exp(-\pi^2 t)\sin(\pi x).$$

Using our technique with

$$v(x,t) = \sum_{n=1}^{5} a_n(t)\sin(n\pi x)$$

we find that

$$a_n(t) = \begin{cases} \exp(-\pi^2 t) & \text{(to within three significant figures) if } n = 1 \\ 0 & \text{if } n > 1 \end{cases}$$

(we use $\Delta x = .01$ and $h = .001$ up to $t = .5$; the above results are therefore for when $t = 0, .001, \ldots, .500$). Thus the linear programming solution is an excellent approximation to the solution for this problem using separation of variables, namely

$$u(x,t) = \exp(-\pi^2 t)\sin(\pi x).$$

II. SUGGESTIONS FOR FURTHER RESEARCH

Aesthetically, we would prefer an approximate solution to KdV continuous in both space and time which has not been calculated by a device such as interpolation. We will suggest two numerical techniques which may enable us to achieve this goal, and point out what we believe may be the main difficulties in attempting to apply the techniques. In both cases we consider the (possibly) nonlinear partial differential equation

$$(5.1) \quad K[u(x,t)] = 0 \quad (a < x < b,\ 0 < t)$$

subject to

$$(5.1a) \quad u(a,t) = u(b,t) = 0 \quad (t \geq 0)$$

and

$$(5.1b) \quad u(x,0) = f(x) \quad (a < x < b)$$

(K is a partial differential operator).

For the first method we select an approximate solution of the form

$$(5.2) \quad v(x,t) = \sum_{i=0}^{M} \sum_{j=0}^{N} a_{ij} X_i(x) T_j(t).$$

Here we have two sets of approximating functions, $\{X_i(x)\}$ for the space variable and $\{T_j(t)\}$ for the time variable. These functions must have certain properties which at first may seem extreme, but in actuality are not. First, select the $X_i(x)$ so that $X_i(a) = X_i(b) = 0$ $(i = 0, \ldots, M)$; then the boundary conditions will be satisfied identically, $v(a,t) = v(b,t) = 0$. Next, if we require $T_j(0) = 1$ ($j = 0, \ldots, N$) and the $X_i(x)$ to be orthonormal over (a,b) with weight function $w(x)$, we can approximate the initial conditions (5.1b):

$$\begin{aligned} v(x,0) &= \sum_{i=0}^{M}\sum_{j=0}^{N} a_{ij} X_i(x) T_j(0) \\ &= \sum_{i=0}^{M}\sum_{j=0}^{N} a_{ij} X_i(x) = f(x); \end{aligned}$$

multiplying both sides by $X_k(x)$, integrating over (a,b) with as a factor the weight function $w(x)$, and using the orthonormality of the $X_i(x)$'s yields

$$(5.3) \quad \sum_{j=0}^{N} a_{kj} = \int_a^b w(x) f(x) X_k(x)\, dx \quad (k = 0, \ldots, M).$$

We can calculate the integral in (5.3) numerically, if necessary. This is a system of N+1 linear algebraic equations in the $(M+1)(N+1)$ unknowns a_{ij}. Next we substitute (5.2) into (5.1) and have

$$(5.4) \quad K[v] = 0.$$

We further assume (5.4) can be rewritten as

$$(5.5) \quad \sum_{i=0}^{M}\sum_{j=0}^{N} F_{ij}[a_{kl}] X_i(x) T_j(t) = 0 \quad (k = 0, \ldots, M; \; l = 0, \ldots, N)$$

where each $F_{ij}[a_{kl}]$ is a function of the unknowns a_{kl}. We therefore can describe (5.4) by

(5.6) $F_{ij} = 0$ ($i = 0, \ldots, M$; $j = 0, \ldots, N$),

a system of (M+1)(N+1) nonlinear equations in the (M+1)(N+1) unknowns a_{ij}. Thus an approximate solution to (5.1) subject to (5.1a) and (5.1b) is (5.2) where the a_{ij}'s are defined by the overdetermined system (5.3) and (5.6) (with a total of (N+1)+(M+1)(N+1) = (M+2)(N+1) equations in (M+1)(N+1) unknowns). In particular, if (5.1) subject to (5.1a) and (5.1b) is KdV, the system (5.6) is quadratic. Herein lies the difficulty. Not only is the total system overdetermined, but in general it is extremely large when enough functions are chosen to accurately approximate the initial conditions; this is an obstacle in applying a Galerkin-type method, such as the one outlined above. However, the advent of new, large, high-speed computers and some recent results in the numerical solution of systems of nonlinear algebraic equations may make this a viable technique anyway.

Another approach that leads to solutions of KdV continuous in space and time combines the linear programming technique with the method outlined above. This "nonlinear programming" approach minimizes a linear function, $\mathcal{E}$, subject to the constraints

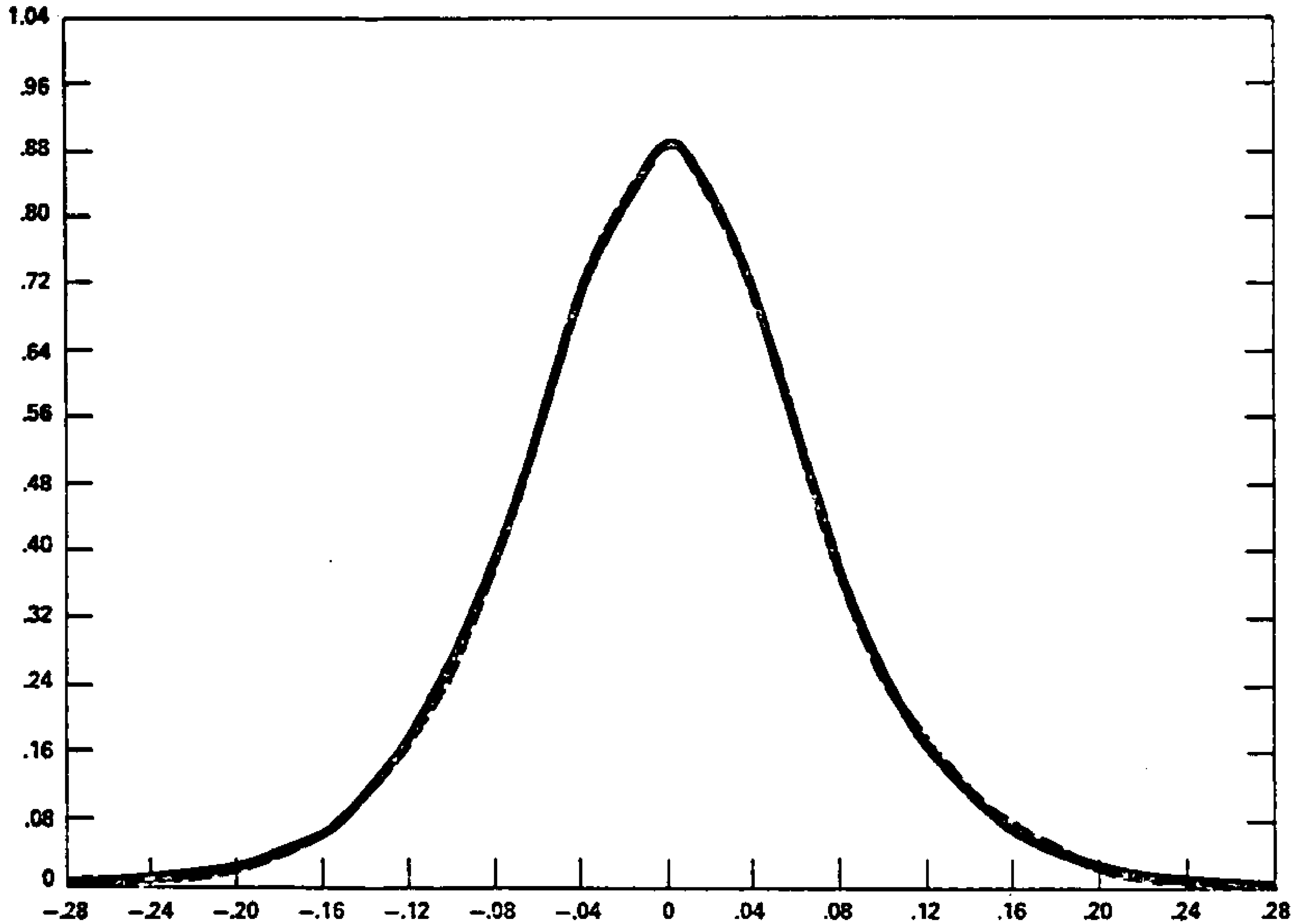
$$-\varepsilon \leq \sum_{j=0}^{N} a_{ij} - \int_a^b w(x) f(x) X_i(x) \, dx \leq \varepsilon \quad (i = 0, \ldots, M)$$

(5.7)

$$-\varepsilon \leq F_{ij}\left[a_{kl}\right] \leq \varepsilon \quad (i = 0, \ldots, M; \; j = 0, \ldots, N).$$

The variables are again a_{ij} ($i = 0, \ldots, M$; $j = 0, \ldots, M$). This can be put in the standard form of a minimum nonlinear programming problem with $(M+1)(N+1)+1$ variables and $2(M+2)(N+1)$ constraints. This eliminates the problem of the overdetermined system. The new difficulty is that the constraints (5.7) can no longer be assumed to define a convex set, and therefore the technique discussed in Chapter II must be modified.

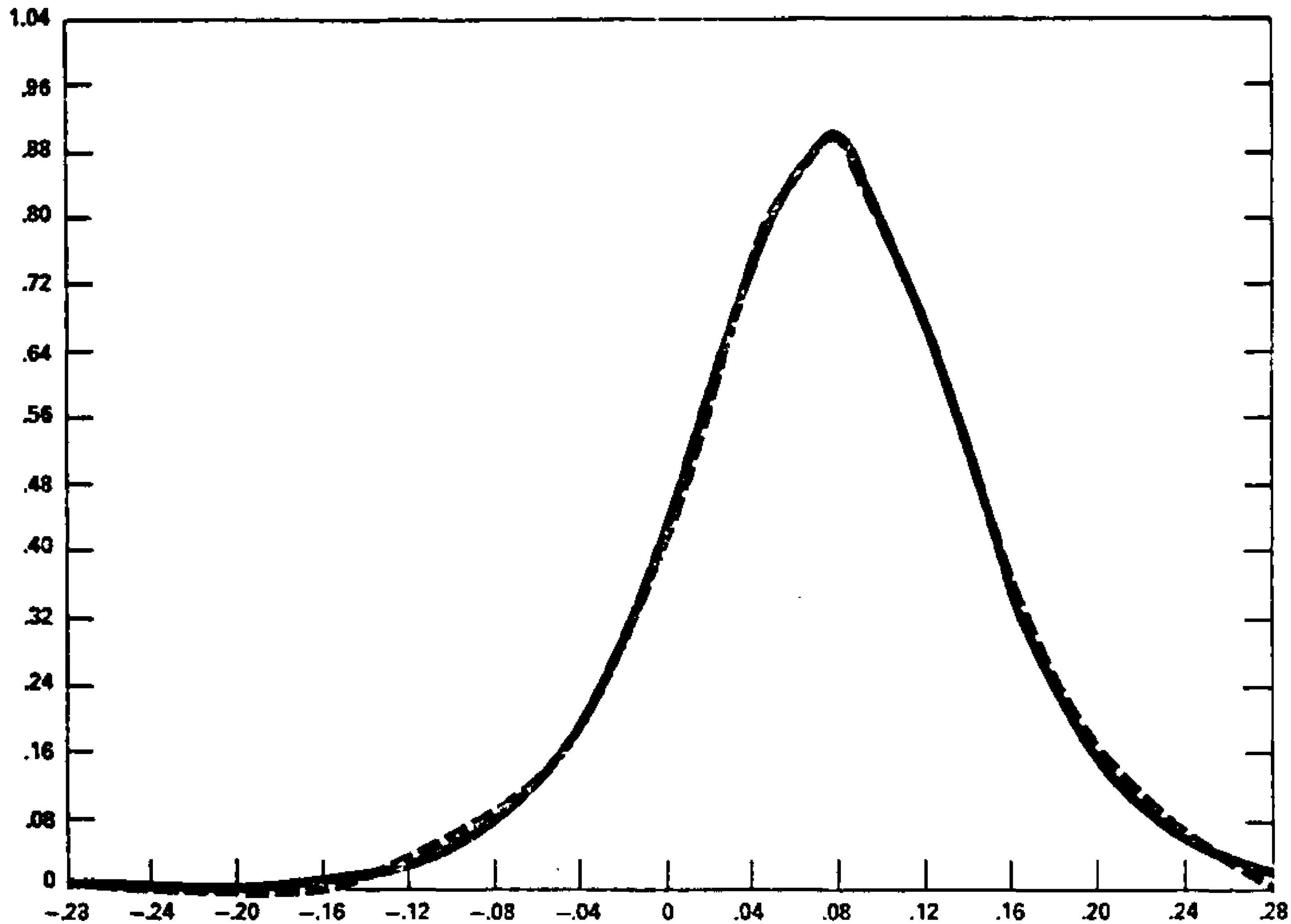
APPENDIX 1



U (SOLID LINE) AND V (DASHED LINE)

t=0

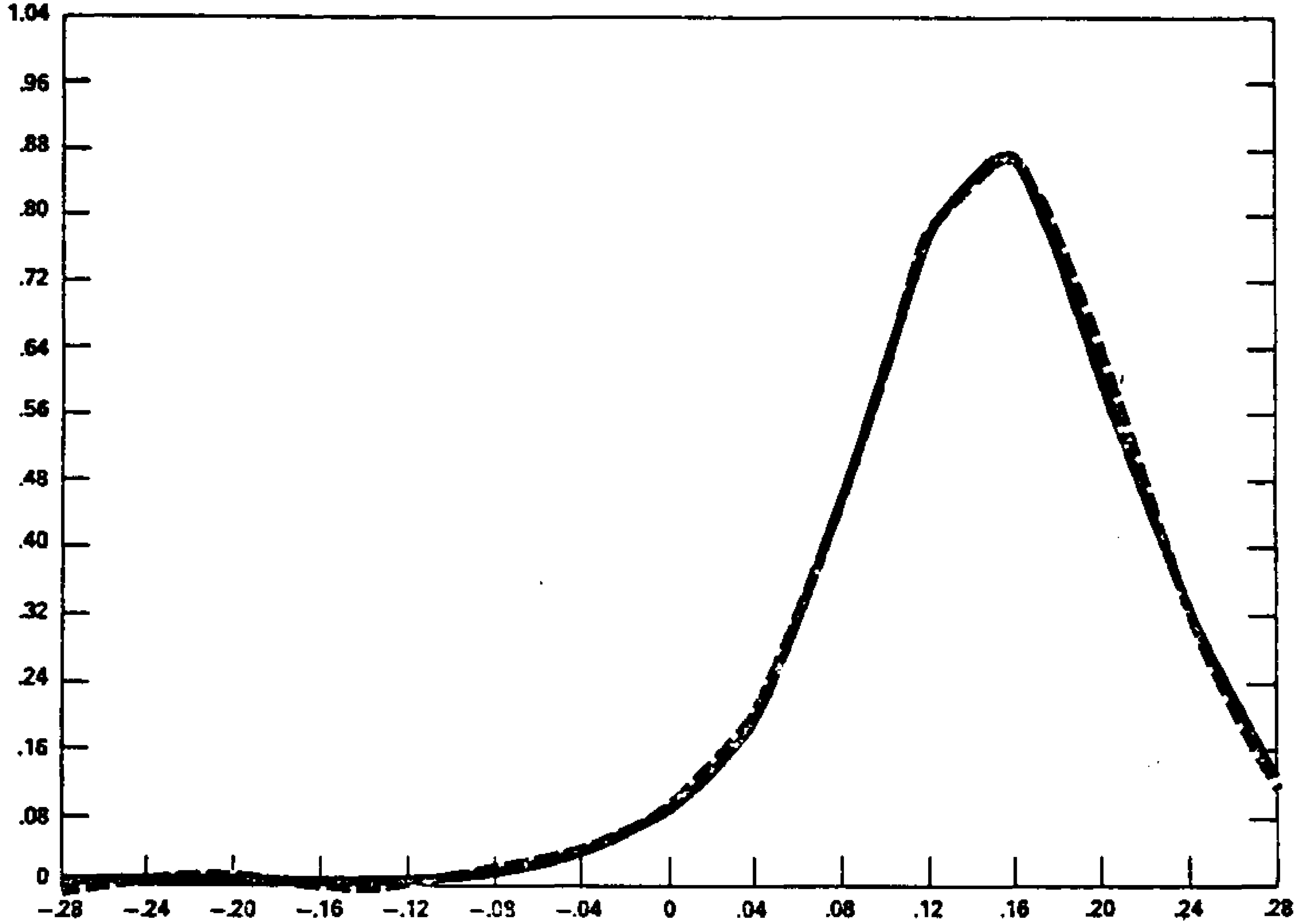
APPENDIX 2



U (SOLID LINE) AND V (DASHED LINE)

t=.25

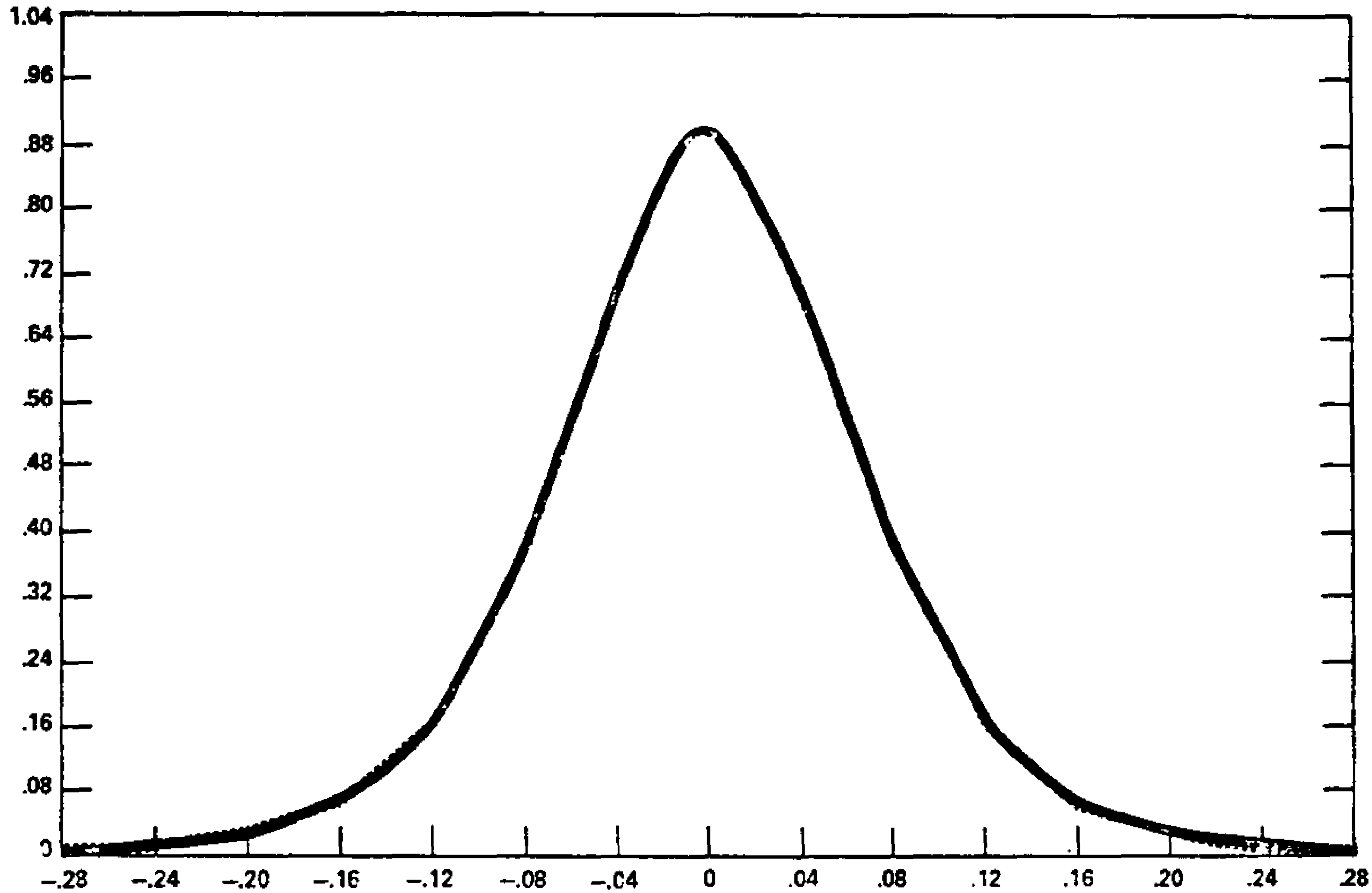
APPENDIX 3



U (SOLID LINE) AND V (DASHED LINE)

t=.50

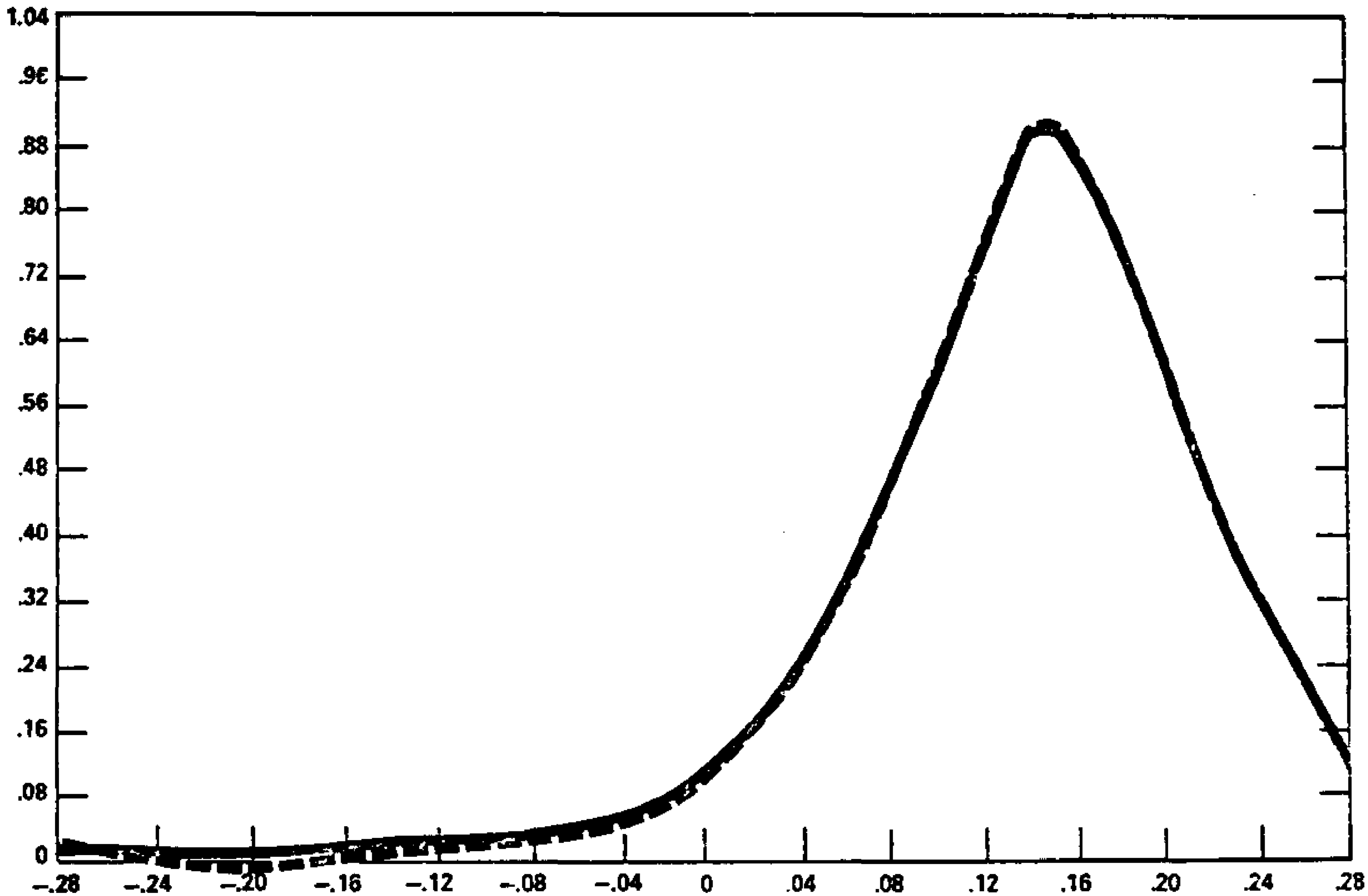
APPENDIX 4



U (SOLID LINE) AND V (DASHED LINE)

t=0

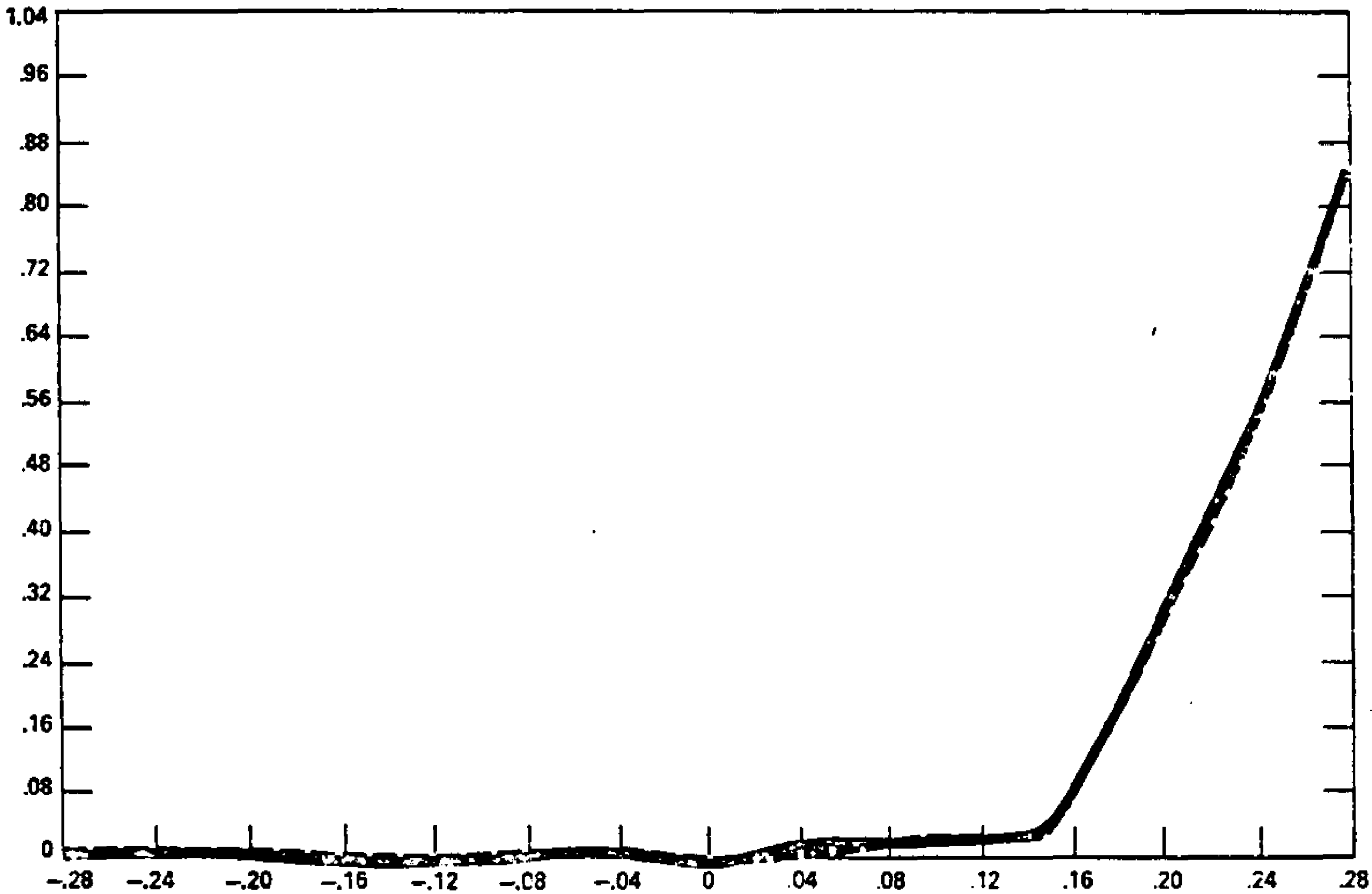
APPENDIX 5



U (SOLID LINE) AND V (DASHED LINE)

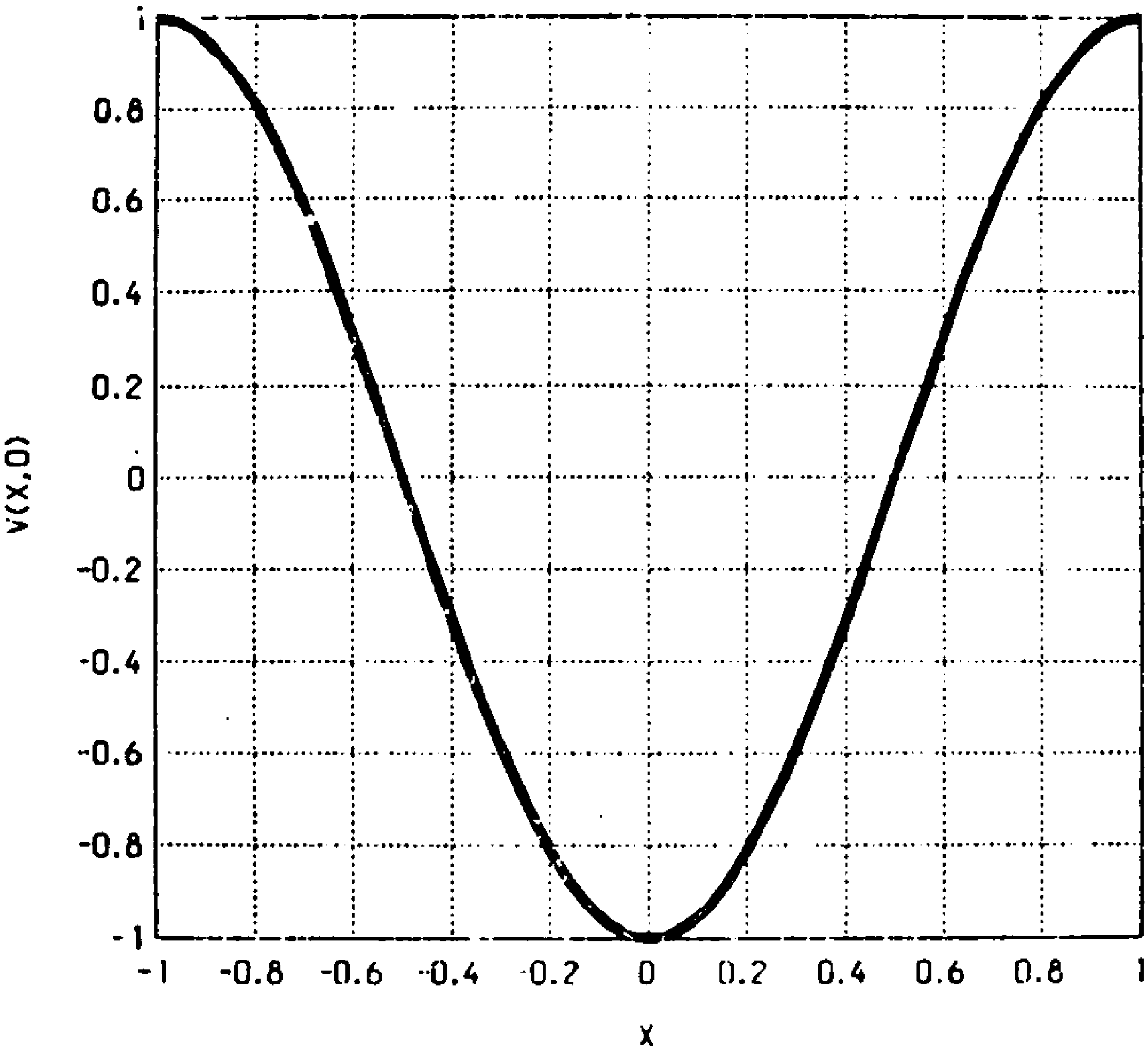
t=.50

APPENDIX 6

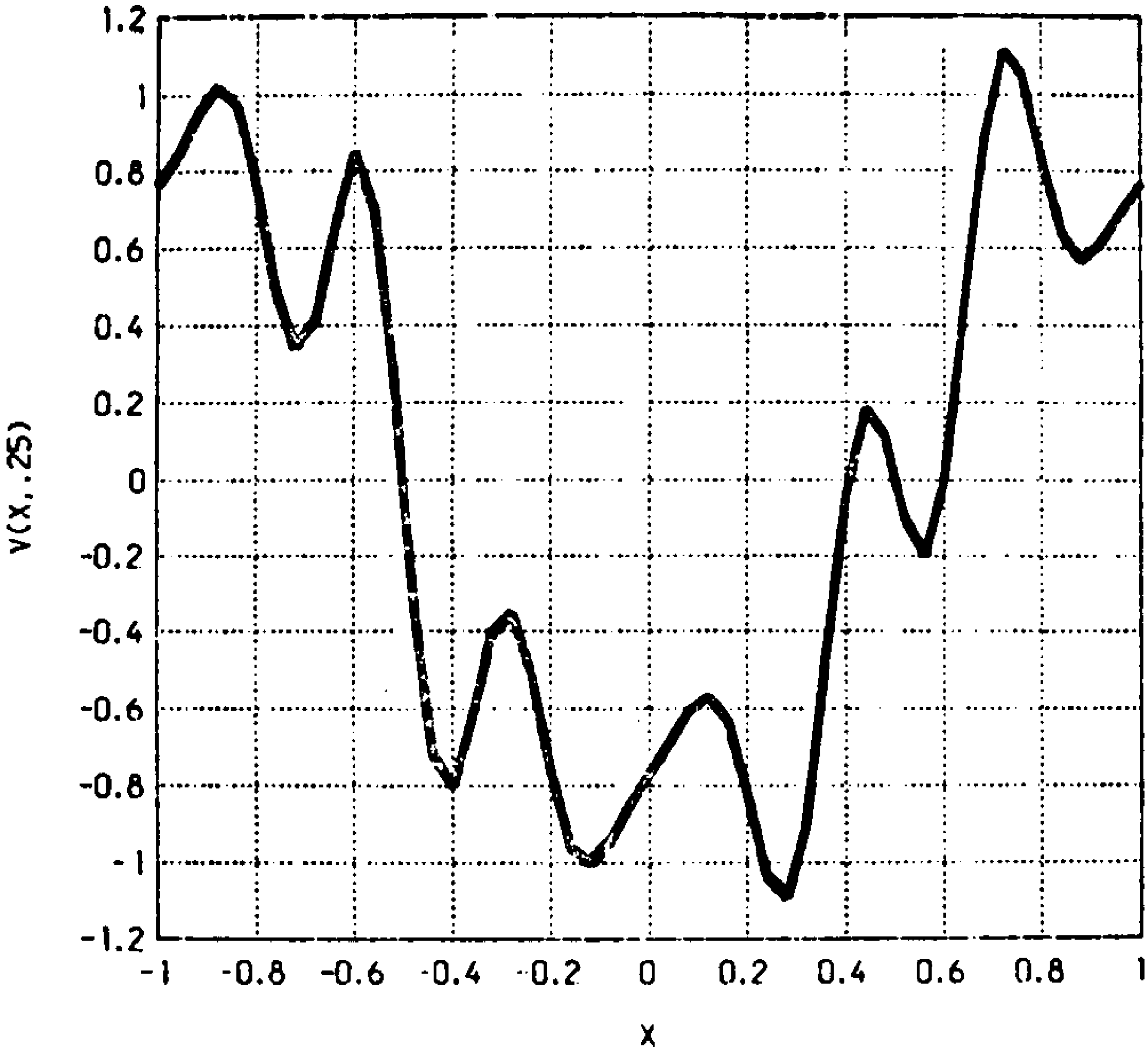


U (SOLID LINE) AND V (DASHED LINE)

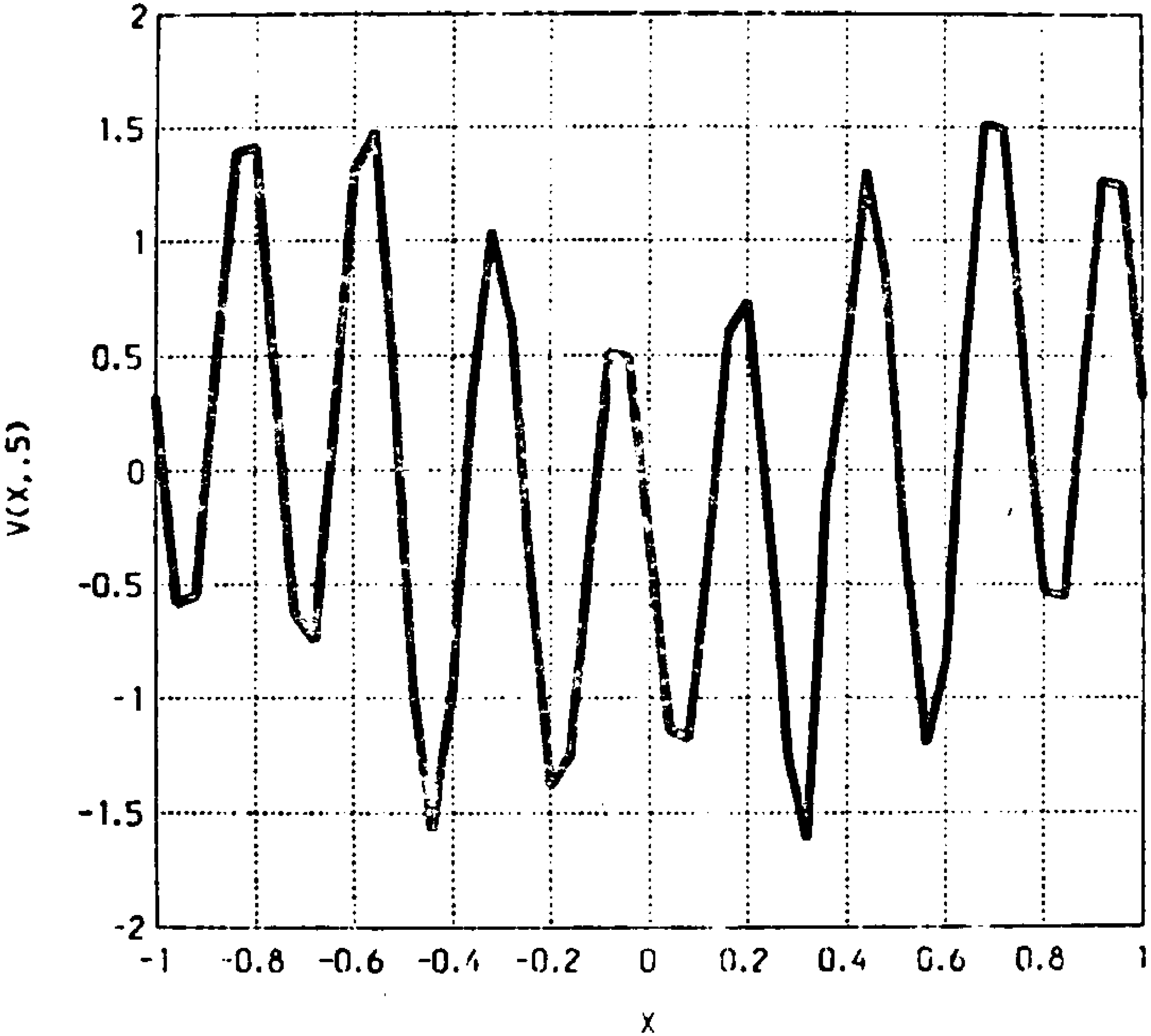
t= 1



APPENDIX 7



APPENDIX 8



APPENDIX 9

Bibliography

[1] J. S. Russell, "Report on Waves", Rep. 14th Meeting of the British Assoc. for the Advancement of Science, John Murray, London, 1844, pp. 311 - 390.

[2] R. M. Miura, "The Korteweg-DeVries Equation: A Survey of Results", S.I.A.M. Review, 1976, Vol.18, No. 3, pp. 412 - 459.

[3] C. S. Gardner and G. K. Morikawa, "Similarity in the Asymptotic Behavior of Collision Free Hydromagnetic Waves and Water Waves", NYU Courant Inst. Math. Sci. Res. Rep. NYO - 9082, 1960.

[4] S. Leibovich and A. R. Seebass, "Examples of Dissipative and Dispersive Systems Leading to the Burgers and Korteweg-DeVries Equations", Nonlinear Waves, Leibovich and Seebass, Eds., Cornell U. Press, Ithaca N.Y., 1974.

[5] G. B. Whitham, Linear and Nonlinear Waves, Wiley-Interscience, N.Y., 1974.

[6] N. J. Zabusky and M. D. Kruskal, "Interaction of Solitons in a Collisionless Plasma and the Recurrence of Initial States", Phys. Rev. Lett., 1965, Vol. 15, pp. 240 - 243.

[7] A. Jeffrey and T. Kakutani, "Weak Nonlinear Dispersive Waves: A Discussion Centered About the Korteweg-DeVries Equation", S.I.A.M. Review, 1972,

Vol. 14, No. 4, pp.582 - 643.

[8] K. Adler and J. Moser, "On a Class of Polynomials Connected with the Korteweg-DeVries Equation", Comm. Math. Phys., Springer-Verlag N.Y., 1978, Vol. 1, pp. 1 - 30.

[9] H. Aireault, H.P. McKean and J. Moser, "Rational and Elliptic Solutions of the Korteweg-DeVries Equation and a Related Many Body Problem", Comm. Pure Appl. Math., 1977, Vol. 30, pp. 95 - 148.

[10] P. D. Lax, "Periodic Solutions of the KdV Equation", Comm. Pure Appl. Math., 1975, Vol. 28, pp. 141-188.

[11] C. Greig and G. Morris, "A Hopscotch Method for the Korteweg-DeVries Equation", Proceedings of the 5th Manitoba Conference on Numerical Mathematics, 1975, Eds. B. L. Hartnell and H.C. Williams, Manitoba Utilitas Mathematica Publishing Inc., Winnepeg, 1976.

[12] B. Fornberg, "Pseudospectral Calculations on 2-D Turbulence and Nonlinear Waves", Computational Fluid Dynamics, S.I.A.M.-A.M.S. Proc., 1978, Vol. 11, pp. 1 - 17.

[13] J. B. Rosen, "Approximate Computational Solutions of Nonlinear Parabolic Partial Differential Equations Using Linear Programming", Numerical Solutions of Nonlinear Differential Equations, Ed. Greenspan, Wiley, N.Y., 1966.

[14] G. B. Dantzig, Linear Programming and Extentions, Princeton U. Press, Princeton N.J., 1963.

[15] G. Hadley, Linear Algebra, Addison-Wesley, Lex. Mass., 1961.

[16] Programmers Manual, International Mathematics and Statistics Library, 5th Ed., Houston Tex., 1977.

[17] P. C. Cheung, "Linear Programming Applied to the Solution of Differential Equations", S.I.A.M. Review, 1977, Vol. 19, No. 3, pp. 398-421.

[18] G. E. Forsythe, M. A. Malcolm and C. B. Moler, Computer Methods for Mathematical Computations, Prentice-Hall, Englewood Cliffs, N.J., 1977.