I. SEISMIC RAY-TRACING IN PIECEWISE HOMOGENEOUS MEDIA

II. ANALYSIS OF OPTIMAL STEP SIZE SELECTION IN HOMOTOPY AND CONTINUATION METHODS

Thesis by

David J. Perozzi

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

1980

(Submitted May 20, 1980)

ACKNOWLEDGEMENTS

I would like to express my gratitude to all professors of the Applied Mathematics Department for the many enlightening contributions to my education. In particular, my deepest thanks go to my advisor, Professor Herbert B. Keller, for his interest, encouragement, and infinite patience.

I wish to also thank my fellow graduate students for many stimulating discussions, both academic and otherwise. Special thanks are due to Dr. Bengt Fornberg and Dr. Gerald B. Whitham for their supervision of the writing of this thesis.

My research at Caltech was supported by a grant from the United States Geological Survey, to whom I wish to express my sincere appreciation. In addition, many thanks are due to the Institute for support in the form of Graduate Teaching Assistantships and Graduate Research Assistantships.

Finally, I wish to dedicate this thesis to my family in recognition of their constant faith and continuing support.

ABSTRACT

Part I

The general problem of the inversion of seismograms usually involves the solution of a nonlinear least squares system. The major component of any such system is the solution of the direct problem. That is, the tracing of a ray between two given end points, where all velocities and interface shapes are specified.

This problem is studied for piecewise constant velocities and fairly arbitrary interface shapes. An efficient computer code is developed for the solution of this problem yielding travel times, amplitudes, ray paths, phase shifts, and caustic locations.

The results are then extended to a wider class of velocity distributions. Conditions are given for the class of velocity distributions for which the problem may be studied completely algebraically.

A standard nonlinear least squares technique is then applied to invert for hypocenters, interface shapes, and elastic constants.

Part II

A brief historical survey of continuation methods is given, with particular emphasis on contributions after 1950.

The problem of selecting an "optimal" step is studied. Optimality here refers to work and storage required for the computation of the solution. The problem is first cast in its most general setting and a couple of trivial theorems are presented.

The problem is then dissected into its component parts, each of which is studied separately. Several combinations of components are also examined. For several specific iterative methods, theorems are presented which optimize an upper bound on the work.

Several computational procedures naturally arising from this theory are suggested.

TABLE OF CONTENTS

PART I

SEISMIC RAY-TRACING IN PIECEWISE HOMOGENOUS MEDIA

I. INTRODUCTION

There is presently an intense interest in the inversion of seismograms. Two major groups are working fairly independently on this problem: seismologists and oil exploration companies. The former deal primarily with large scale problems (in the sense that distances are on the order of tens to hundreds of kilometers) and are interested in source location, velocity distributions, and fault line location. The latter group works on smaller scale problems (the size of an oil field) and focus their attention on velocity distributions and detailed interface shapes (sources are known explosive charges).

Currently, most (though by no means all) inversion programs use only the travel time from source to receiver. In a large percentage of cases only the first travel time is used. Since it is frequently possible to identify what type of ray causes subsequent peaks, it would seem worthwhile to investigate the possibility of using this information. Presently, there are two major methods of attacking the problem: (1) directly via the partial differential equations of propagation; (2) the reduction of the equations in (1) to a set of nonlinear ordinary differential equations. The former has been employed extensively by Bamberger, Chavent, Hemon, and Lailly [1] among others. The latter is used by Červený, Molotkov, and Pšenčík [4], Pereyra, Keller, and Lee [13], Pereyra and Lee [15] and numerous others. All succeeding sections of this work shall deal with this latter case.

Approach (1) has been limited to fairly specific interfaces, primarily parallel planes or concentric circles. In [1], Bamberger et al. attempt to recover the acoustical impedance via the use of normal incidence seismograms. Parallel planes are used, where the travel time between planes is kept constant (i.e., $\frac{y_i - y_{i+1}}{v_i}$ = constant).

Červený et al. present an excellent overview of the ray method. They limit their consideration to the solution of the direct problem with planar interfaces, but allow fairly arbitrary velocity distributions. They use a standard "shooting" technique which seems to be very popular among seismologists. This consists of solving an initial value problem for a wide range of initial conditions (take-off angles). Presumably, if enough rays are traced, some will terminate in the vicinity of the known receiver locations. The relevant data are then interpolated for use at these points. There is a very brief section acknowledging the existence of programs for the solution of the problem via a boundary value problem, but this technique is not seriously discussed.

Pereyra et al. reduce the problem to a nonlinear two-point boundary value problem, for which they have powerful programs available. Arbitrary velocity distributions are permitted, as are arbitrarily shaped interfaces. Parallel shooting is used between each pair of of interface crossings, and then matching takes place to determine the intersections of the rays with the interfaces. This process is then iterated until a continuous ray satisfying the boundary conditions is obtained. In the inversion process, only first travel times are used, and amplitudes are not included.

The work which is, perhaps, closest to that which follows was done by Chander [5]. This is actually a three-dimensional program (although all examples presented are two-dimensional). The media are taken as piecewise constant. All interfaces are planar and non-intersecting in the region of interest.

The following material may be viewed as both an extension and a generalization of Chander's work, allowing for more general interface shapes and more complicated velocity distributions. We shall confine ourselves to the two-dimensional case, although the modifications required for three dimensions are fairly minor. For velocity distributions which do not fall within the class exhibited in section II.6, one may be able to employ an approximate velocity distribution to obtain a good initial guess for the code of Pereyra ѐt al. in a more efficient manner.

For further references on the ray method, one is referred to Červený et al. [4], which in itself contains an excellent exposition of the subject. More detailed references on the wave front method and normal incidence inversion may be found in the bibliography contained in Bamberger et al. [1].

II. DIRECT PROBLEM

II.1. Introduction to the Direct Problem

The major goal of this work is to develop fast, efficient, and accurate methods for determining large classes of seismic rays joining two arbitrary points, $\underset{\sim}{x}_I$ and $\underset{\sim}{x}_F$, for very general geometric configurations. Only two-dimensional situations are considered. The geological interfaces and free surfaces separating various regions of different homogeneous isotropic elastic material (i.e., differing constant wave speeds c_p and c_s) are allowed to be fairly arbitrary. For each ray which makes contact with N interfaces, there can be 2^{N+1} (or more) distinct seismic rays connecting $\underset{\sim}{x}_I$ to $\underset{\sim}{x}_F$. This occurs because, on contact with an interface, a seismic wave (compressive or shear) splits into a compressive and a shear wave. The procedures presented allow the easy determination of all such rays (when they exist). Travel time and amplitude variation are determined along each ray. Phase shifts may also be determined since the occurrence of every caustic on any given ray segment may be detected. From these data, compiled for some set of ray classes, we can construct peak seismograms.

Of course the motivating factor is the solution of geophysical inverse problems. Chapter III incorporates the methods of Chapter II to determine source location, media speeds, and interface shapes.

In section II.2, the ray problem is formulated for general piecewise constant configurations. It reduces to systems of coupled nonlinear equations. Also, the notion of a ray "signature" is introduced. This is a useful tool for devising simple computer codes to solve the nonlinear system.

Section II.3 describes the solution procedures employed (Newton's method and a continuation in the speeds). A special continuation (or homotopy) procedure is developed to obtain the initial ray of any given class.

In section II.4, the computation of travel time, amplitude variation, and phase is discussed. Section II.5 is a discussion of nonphysical, nonunique and diffracted rays.

Section II.6 covers the extensions to linearly varying elastic materials. The materials are allowed to be plane stratified with an arbitrary variation of speeds in the principle direction.

Section II.7 contains worked-out examples.

II.2. Formulation of the Problem and Classification of Rays

II.2.1. Problem Formulation

The structure of the earth is modelled by piecewise constant regions of arbitrary shape. All interfaces (and the free surface of the earth) are assumed to be smooth curves (piecewise C^3 and continuous is sufficient). These curves are represented by the formulae:

$$y = f_i(x) \quad , \quad i = 1,2, \dots , M \qquad (2.1)$$

We shall adopt the convention that $i = 1$ represents the free surface of the earth. The region of physical interest is thus confined to $y \leq f_1(x)$. For purposes of simplification, we shall not consider intersecting interfaces at present. The inclusion of intersecting interfaces poses no great difficulties, but it does complicate the notation somewhat (and the coding as well). So the configuration consists of layers, but the interfaces need not be planar nor simple geometric shapes (see Figure 1).

The medium between each successive pair of interfaces is assumed homogeneous,isotropic, and perfectly elastic. The elastic constants in each medium are assumed known: λ (Lamé constant), μ (shear modulus), ρ (density). In such media, at most two kinds of signals can propagate: compressive (P) waves with speed $C_P = \sqrt{\frac{\lambda + 2\mu}{\rho}}$ and shear (SV) waves with speed $C_S = \sqrt{\frac{\mu}{\rho}}$. These speeds are, of course, different in each medium. Since the speeds are constant, the rays are merely straight lines. Hence, no differential equations need be solved.

A ray is determined geometrically by specifying the initial point

$\underset{\sim}{x}_I$, the final point $\underset{\sim}{x}_F$, and each contact interface. On each segment $[\underset{\sim}{x}_{k-1}, \underset{\sim}{x}_k]$ its type of propagation must be known. To determine the ray path, only the speed of propagation is required in each medium. At each contact with an interface, Snell's law must hold. This supplies sufficient information to determine the contact nodes, as shown by the following.

Denote the speed on segment $[\underset{\sim}{x}_{k-1}, \underset{\sim}{x}_k]$ by v_k for each $k = 1,2, \ldots, N_2$. Let the tangent vector to the interface at node $\underset{\sim}{x}_k$ be denoted by $\underset{\sim}{\tau}_k$. The most general form of Snell's law then requires:

$$v_{k+1}\left(\underset{\sim}{\tau}_k \cdot \frac{\underset{\sim}{x}_k - \underset{\sim}{x}_{k-1}}{||\underset{\sim}{x}_k - \underset{\sim}{x}_{k-1}||}\right) = v_k\left(\underset{\sim}{\tau}_k \cdot \frac{\underset{\sim}{x}_{k+1} - \underset{\sim}{x}_k}{||\underset{\sim}{x}_{k+1} - \underset{\sim}{x}_k||}\right) \qquad (2.2)$$

In the above the vector notation $\underset{\sim}{x}_k \equiv (x_k, y_k)$ is employed and $(\underset{\sim}{\tau} \cdot \underset{\sim}{x})$ represents the usual scalar product of vectors. From the interface representation (2.1), when $\underset{\sim}{x}_k$ is on interface i_k we have

$$\text{a)} \quad \underset{\sim}{x}_k \equiv (x_k, y_k) = (x_k, f_{i_k}(x_k)) \qquad (2.3)$$

With $\frac{df}{dx}(x) = f'(x)$, a tangent to interface i_k at $\underset{\sim}{x}_k$ is given by:

$$\text{b)} \quad \underset{\sim}{\tau}_k = (1, f'_{i_k}(x_k)) \qquad (2.3)$$

Substituting (2.3) in (2.2) one obtains:

$$\phi_k = v_{k+1} \frac{(x_k - x_{k-1}) + f'_{i_k}(x_k)(f_{i_k}(x_k) - f_{i_{k-1}}(x_{k-1}))}{\left[(x_k - x_{k-1})^2 + (f_{i_k}(x_k) - f_{i_{k-1}}(x_{k-1}))^2\right]^{\frac{1}{2}}}$$

$$- v_k \frac{(x_{k+1} - x_k) + f'_{i_k}(x_k)(f_{i_{k+1}}(x_{k+1}) - f_{i_k}(x_k))}{\left[(x_{k+1} - x_k)^2 + (f_{i_{k+1}}(x_{k+1}) - f_{i_k}(x_k))^2\right]^{\frac{1}{2}}} = 0 \qquad (2.4)$$

Relation (2.4) must hold on each interface which is crossed between $\underset{\sim}{x}_I$ and $\underset{\sim}{x}_F$. Since the source location and receiver location are given, we obtain a system with N variables, where N is the number of interfaces which are crossed. The unknowns are the scalars $x_1, \dots, x_N$. For $k = 1$ and $k = N$ in (2.4), we require values for x_0 and x_{N+1}. By definition we use:

$$\text{a)} \qquad \begin{aligned} \underset{\sim}{x}_0 &= \underset{\sim}{x}_I = (x_0, y_0) = (x_I, f_{i_0}(x_I)) \\ \underset{\sim}{x}_{N+1} &= \underset{\sim}{x}_F = (x_{N+1}, y_{N+1}) = (x_F, f_{i_{N+1}}(x_{N+1}) \end{aligned} \qquad (2.5)$$

The use of f_{i_0} and $f_{i_{N+1}}$ is simply for a standardization of the computer code. In reality, these functions are defined by

$$\text{b)} \qquad \begin{aligned} f_{i_0}(x) &= y_I \\ f_{i_{N+1}}(x) &= y_F \end{aligned} \qquad (2.5)$$

Using (2.5) in (2.4), we are left with N nonlinear equations in N unknowns. By introducing the vector notation

$$X \equiv (x_1, x_2, \dots, x_N)^T , \; V \equiv (v_1, v_2, \dots, v_N)^T$$

$$\Phi\,(X,V) \equiv (\phi_1, \phi_2, \dots, \phi_N)^T \qquad (2.6)$$

then the system to be solved is given by:

$$\Phi\,(X,V) = 0 \qquad (2.7)$$

II.2.2. Ray Signatures, Ray Classes and Propagation Types

In order to implement the preceding formulation, it is necessary to specify on each desired ray the speed of propagation on each of its segments $[\underset{\sim}{x}_{k-1}, \underset{\sim}{x}_k]$ and the appropriate interface formula at the nodes $\underset{\sim}{x}_k$. This is done by assigning an integer label, say m, to each different material. The speeds in material m are then labelled as:

$$C_{P,m} \equiv v_{2m}, \ C_{s,m} \equiv v_{2m+1}, \ m = 1,2, \dots M. \tag{2.8}$$

Therefore, P waves are denoted by even subscripts and S waves by odd subscripts. Given $\underset{\sim}{x}_I$ and $\underset{\sim}{x}_F$, we may classify any desired ray by sequentially listing its speed on the first segment, the number of the first interface, speed on the second segment, number of second interface, ..., speed on the final segment. We shall call this listing a ray signature. This is equivalent to listing the subscripts of the speeds and the subscripts of the interfaces. Thus a signature is specified by giving an ordered set of integers

$$\mathcal{S} \equiv [j_1, i_1; \ j_2, i_2; \ \dots \ ; \ j_N, i_N; \ j_{N+1}] \tag{2.9}$$

Given two signatures $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$, we shall say they correspond to rays of the same <u>class</u> if $i_k^{(1)} = i_k^{(2)}$ for all $k = 1, \dots, N$. In general, there are 2^{N+1} ray types in each class (there may be more actual rays, since there may be more than one ray corresponding to a given signature).

Some simple examples will be illustrative. Figure 1 depicts a sketch of four layers, the first and third of which are composed of the same material. Two classes of rays are indicated, one with one internal node (I) and the other with two internal nodes (II). Classes I and II, with all possible signatures, are listed below:

Class I	Class II
1. [4,2;2] ∿ [P;P]	1. [4,3; 4,2;2] ∿ [P;P;P]
2. [4,2;3] ∿ [P;S]	2. [4,3; 4,2;3] ∿ [P,P;S]
3. [5,2;2] ∿ [S;P]	3. [4,3; 5,2;2] ∿ [P;S;P]
4. [5,2;3] ∿ [S;S]	4. [4,3; 5,2;3] ∿ [P;S;S]
	5. [5,3; 4,2;2] ∿ [S;P;P]
	6. [5,3; 4,2;3] ∿ [S;P;S]
	7. [5,3; 5,2;2] ∿ [S;S;P]
	8. [5,3; 5,3;3] ∿ [S;S;S]

Next to each ray signature is listed the sequence of propagation types. This is redundant information since it can be gleaned by simply observing the parity of the speed indices. However, it is often quite useful to display this simpler propagation signature.

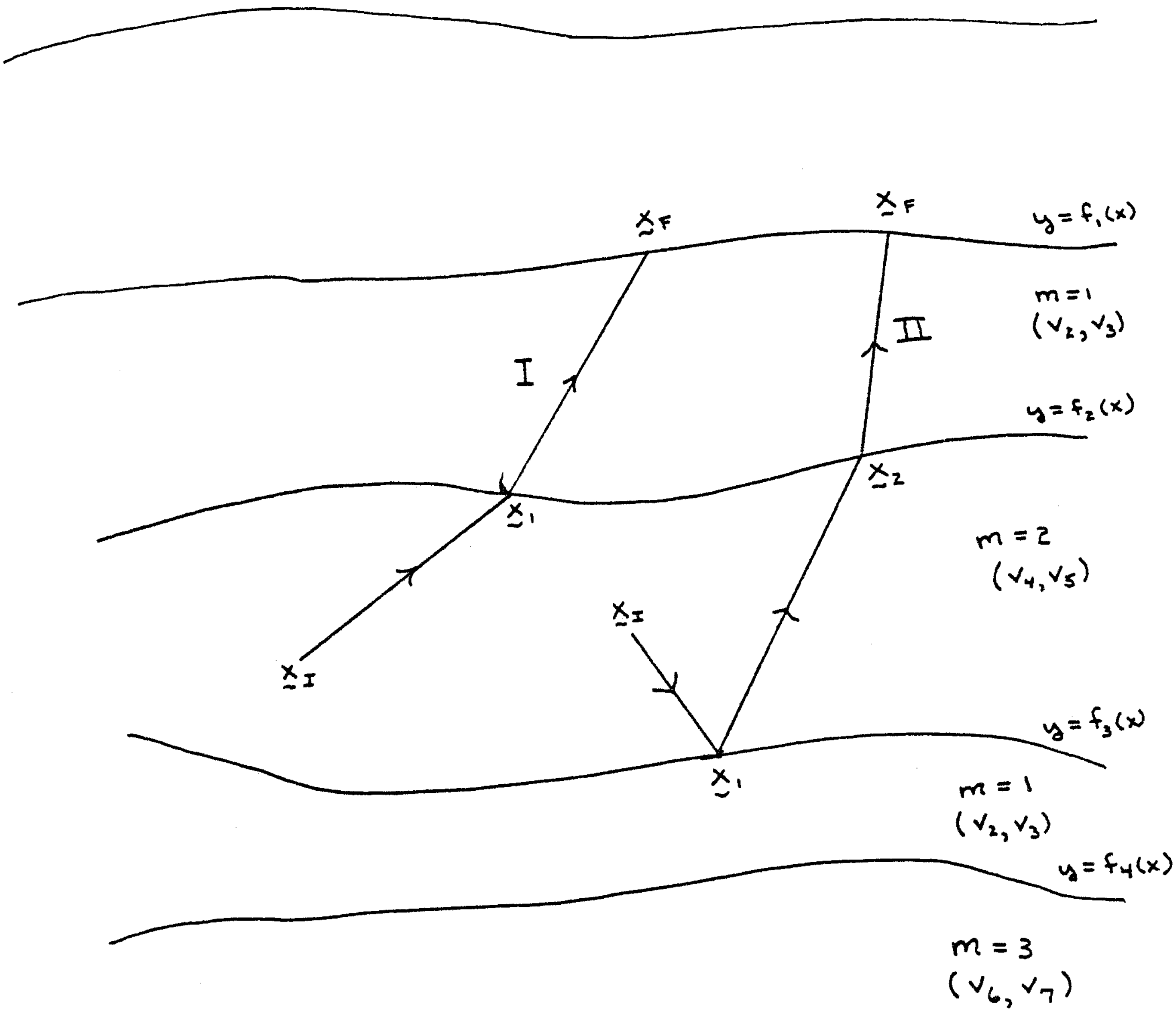


Figure 1. Four layered media with ray classes I and II depicted.

II.3. Solution Procedures

II.3.1. Newton's Method and Continuation for Rays of the Same Class

Newton's method and a continuation procedure are utilized to solve the system (2.7). In particular, given an approximation to the solution, say X^V, then an improved value is given by

$$X^{V+1} = X^V + \delta X^V \tag{3.1a}$$

where

$$J^V \delta X^V = -\Phi^V \tag{3.1b}$$

In (3.1) we have used:

$$X^V \equiv (x_1{}^V, x_2{}^V, \ldots, x_N{}^V), \; \Phi^V \equiv \Phi(X^V, V), \; J^V = J(X^V, V) \tag{3.2}$$

where $J(X,V)$ is the Jacobian matrix of the system

$$J(X,V) \equiv \frac{\partial \Phi(X,V)}{\partial X} \tag{3.3a}$$

From (2.4) - (2.6), we deduce that J is an NxN tridiagonal matrix, $J \equiv [b_k, a_k, c_k]$, where

$$b_k \equiv \frac{\partial \phi_k}{\partial x_{n-1}}, \quad a_k \equiv \frac{\partial \phi_k}{\partial x_k}, \quad c_k \equiv \frac{\partial \phi_k}{\partial x_{k+1}} \tag{3.3b}$$

Introducing the following notation:

$$\Delta x_j \equiv x_j - x_{j-1}, \; \Delta y_j \equiv y_j - y_{j-1}, \; y'_j = \frac{df_{i_j}(x_j)}{dx_j} \quad , \tag{3.3c}$$

$$D_j \equiv [(\Delta x_j)^2 + (\Delta y_j)^2]^{\frac{1}{2}}$$

then the components of J may be written explicitly as:

$$a_k = \frac{v_{k+1}}{D_k}\left[1 + y''_k \Delta y_k + (y'_k)^2 - \left(\frac{\Delta x_k + y'_k\ \Delta y_k}{D_k}\right)^2\right]$$

$$+ \frac{v_k}{D_k}\left[1 - y''_k \Delta y_{k+1} + (y'_k)^2 - \left(\frac{\Delta x_{k+1} + y'_k \Delta y_{k+1}}{D_{k+1}}\right)^2\right]$$

$$b_k = - \frac{v_{k+1}}{D_k}\left[1 + y'_{k-1} y'_k - \left(\frac{\Delta x_k + y'_{k-1}\Delta y_k}{D_k}\right)\left(\frac{\Delta x_k + y'_k \Delta y_k}{D_k}\right)\right] \quad (3.3d)$$

$$c_k = - \frac{v_k}{D_{k+1}}\left[1 + y'_k y'_{k+1} - \left(\frac{\Delta x_{k+1} + y'_k \Delta y_{k+1}}{D_{k+1}}\right)\left(\frac{\Delta x_{k+1} + y'_{k+1}\Delta y_{k+1}}{D_{k+1}}\right)\right]$$

Newton's method converges quadratically when X^o, the initial guess, is sufficiently close to a solution [if J is nonsingular at the root]. That is, we are assured that

$$||\delta X^v|| \leq K\ ||\delta X^{v-1}||^2, \quad v = 1,2,\ldots \qquad (3.4)$$

if X^o is close enough to a solution. To obtain such a X^o we use a continuation method.

We introduce a one parameter family of speeds.

$$w(\lambda) \equiv \lambda \hat{V} + (1-\lambda) V, \quad 0 \leq \lambda \leq 1 \qquad (3.5)$$

Clearly $w(0) = \hat{V}$ and $w(1) = V$. Thus if the solution of (2.7) using the speeds (3.5) is denoted by $X(\lambda)$, it follows that $X(0)$ is the solution using speeds $\hat{V}$ and $X(1)$ is the solution with speeds V. If we know the solution for any value of λ, then we can use $X^o(\lambda + \Delta\lambda)$ as the initial guess in Newton's method for the value $\lambda + \Delta\lambda$, where $X^o(\lambda + \Delta\lambda)$ is given by:

$$X^{o}(\lambda + \Delta\lambda) = X(\lambda) + \Delta\lambda\dot{X}(\lambda) \qquad (3.6)$$

This is accurate to order $(\Delta\lambda)^2$ if we know $\dot{X}(\lambda) \equiv \frac{dX(\lambda)}{d\lambda}$. We can obtain this derivative by substituting (3.5) in (2.7) and differentiating to obtain

$$J(X(\lambda),\, w(\lambda))\ \dot{X}(\lambda) = -\frac{\partial\Phi(X(\lambda), V(\lambda)}{\partial V}\ (V-\hat{V}) \qquad (3.7)$$

The matrix $B \equiv \frac{\partial\Phi}{\partial V}$ is Nx$(N+1)$ and is bi-diagonal. In particular, the k^{th} row of $B = (b_{k,j})$ (recalling (3.3c)) has the elements:

$$b_{k,k} \equiv \frac{\partial\phi_k}{\partial V_k} = -\left(\frac{\Delta x_{k+1} + y_k'\Delta y_{k+1}}{D_{k+1}}\right)$$

$$b_{k,k+1} \equiv \frac{\partial\phi_k}{\partial V_{k+1}} = \frac{\Delta x_k + y_k'\Delta y_k}{D_k}$$

$$b_{k,j} \equiv 0\ ,\quad j \neq k,k+1 \qquad (3.8)$$

Since we know $J(X(\lambda), w(\lambda))$ at λ, we can easily obtain $\dot{X}(\lambda)$ by solving (3.7).

Note that once we have any one ray in a given class, we may use this continuation technique to compute all desired rays in that class. If the rays of different propagation types are ordered appropriately, we can obtain vectors V and $\hat{V}$ which differ in only one component, i.e., for some k :

$$V - \hat{V} = (0,\ldots,0,\ v_k-\hat{v}_k,0,\ldots,0)^T. \qquad (3.9)$$

The right hand side of (3.6) has only two non-zero components in this case, the $(k-1)^{st}$ and the k^{th}. In more than 90% of all test examples, this continuation procedure succeeded for $\Delta\lambda = 1$. That is, in <u>one step</u>, we obtain an initial guess for which Newton's method converges for the new speeds ^{V} from (3.6)with $\lambda = 0$ and $\Delta\lambda = 1$.

<u>II.3.2. Obtaining the First Ray of a Class</u>

In II.3.1, it was shown how to compute all the rays of a given class after one ray of that class had been determined. A simple technique for determining some first ray in a class will now be exhibited. Usually we choose the pure compressive ray [P;P;...;P], but this is by no means necessary.

Assuming that $\underset{\sim}{x}_I$, $\underset{\sim}{x}_F$, and a signature have been specified, then an initial vector $\hat{X}$ is chosen arbitrarily. For example, we often choose

$$\hat{x}_k = x_I + k\ (\frac{x_F - x_I}{N+1})\ ,\ k = 0,1,\ldots,\ N+1 \qquad (3.10)$$

From the signature, we can now determine the nodes $\underset{\sim}{\hat{x}}_k = (\hat{x}_k, f_{i_k}(\hat{x}_k))$. For $\hat{v}_1$, the correct value of v_1 is chosen. Noting that (2.2) is linear in the speeds v_k, we now use the values of $\hat{x}_k$ and $\hat{v}$ to successively compute $\hat{v}_2$, ..., $\hat{v}_{N+1}$ to generate the speed vector $\hat{V}$. The speeds thus generated are generally unrelated to any physical materials (in fact, some may even be negative). Additionally, the ray in question may pass through the same material several times with a different propagation speed $\hat{v}$ on each segment. One may envision this as treating the earth as a Riemann sheeted material. Every time the ray traverses a medium, it travels on a different Riemann sheet. Regardless of the non-physical nature of the generated data, the procedure has generated vectors $\hat{X}$ and $\hat{V}$ such that

$$\Phi\,(\hat{X},\ \hat{V}) = 0.$$

Once again, we can employ the continuation method of II.3.1 to determine the solution for the true physical speeds V. The continuation process for this first ray is typically somewhat slower than for the subsequent rays. But, in most cases, it converges surprisingly quickly.

For certain exceptional cases, some segment $[\hat{x}_k, \hat{x}_{k+1}]$ may be tangent to an interface. A simple shift of some components of $\hat{X}$ can always be found so that this does not occur.

II.3.3. Smale's Boundary Conditions for the Simplest Case; Niceties of J

There is no guarantee that the continuation method given in the preceding sections must always yield a path leading to $\lambda = 1$. However, Smale has presented conditions under which the continuation method cannot fail. These conditions follow.

Suppose we wish to solve

$$f(u) = 0\ . \tag{3.11}$$

We use the continuation method with the parametrized problem:

$$G(u,\lambda) \equiv f(u) - \lambda f(u^o),\ u = u(s),\ \lambda = \lambda(s). \tag{3.12}$$

By differentiating, we obtain the initial value problem:

$$f'(u)\dot{u} - \dot{\lambda}\, f(u^o) = 0 \tag{3.13a}$$

$$||\, \dot{u}\, ||^2 + \dot{\lambda}^2 = 1 \tag{3.13b}$$

$$u(0) = u^o,\ \lambda(0) = 1 \tag{3.13c}$$

where b is simply a normalization which makes the parameter s mimic arc length.

Smale's boundary conditions guarantee that a path exists satisfying (3.12) such that $\lambda(s)$ has an odd number of zeroes on that path (i.e. the path passes through an odd number of solutions of (3.11)) for almost all choices of u^o.

Given $\Omega \subset R^N$ and $\partial\Omega$ smooth and connected, Smale's conditions become:

$$f'(u) \text{ is non-singular} \quad \forall\ u \in \partial\, \Omega \tag{3.14a}$$

$$\sigma(f'(u))^{-1}f(u) \text{ points in } \forall\ u \in \partial\, \Omega \tag{3.14b}$$

where σ is a constant equal to either 1 or -1.

We shall show that for the simplest case (simple reflection or transmission), that we can guarantee that Smale's conditions hold.

By defining $s_i = x_i - x_{i-1}$, we may formulate the problem for parallel plane layers in terms of the s_i's.

$$\frac{s_i}{v_i D_i} - \frac{s_{i+1}}{v_{i+1} D_{i+1}} = 0,\ i=1,\ldots,N \tag{3.15a}$$

$$\sum_{i=1}^{n+1} s_i - (x_F - x_I) = 0 \tag{3.15b}$$

For N=1, we will show that Smale's boundary conditions hold for

$$\Omega \equiv \{0 \leq s_1,\ 0 \leq s_2,\ s_1+s_2 \leq (x_F - x_I) + \varepsilon\} \tag{3.16}$$

where ε is a sufficiently small positive number.

The Jacobian for this case is

$$\hat{J} = \begin{pmatrix} \dfrac{k_i}{v_i D_i^{\,3}} & -\dfrac{k_{i+1}}{v_{i+1} D_{i+1}^{\,3}} \\ 1 & 1 \end{pmatrix} \equiv \begin{pmatrix} a_1 & -b_1 \\ 1 & 1 \end{pmatrix} \qquad (3.17)$$

where $k_i = |y_i - y_{i+1}|$.

Since both a_1 and b_1 are positive, J is non-singular (i.e., $\det J = a_1 + b_1 > 0$) for all finite values of $\underset{\sim}{s}$. Thus we may write

$$\hat{J}^{-1} = \frac{1}{a_1 + b_1} \begin{pmatrix} 1 & b_1 \\ -1 & a_1 \end{pmatrix} \qquad (3.18)$$

and we note that condition (3.14a) is satisfied on $\partial\Omega$.

To test condition (3.14b) we form the product

$$\hat{J}^{-1}\underset{\sim}{f} = \frac{1}{a_1 + b_1} \begin{bmatrix} f_1 + b_1 f_2 \\ -f_1 + a_1 f_2 \end{bmatrix} \qquad (3.19)$$

The outward normal to $s_1 + s_2 = 1 + \varepsilon$ is $\underset{\sim}{n}_1 = (1,1)^T$. The component of $\hat{J}^{-1}f$ in the $\underset{\sim}{n}_1$ direction is

$$c_1 = \underset{\sim}{n}_1 \cdot \hat{J}^{-1}\underset{\sim}{f} = f_2 = s_1 + s_2 - (x_F - x_I) \qquad (3.20)$$

On $s_1 + s_2 = (x_F - x_I) + \varepsilon$, we have $c_1 = \varepsilon > 0$. Hence $J^{-1}f$ points out on this portion of the boundary.

The outward normal to $s_1 = 0$ is $\underset{\sim}{n}_2 = (-1,0)^T$. Thus we have on $s_1 = 0$,

$$c_2 = \underset{\sim}{n}_2 \cdot \hat{J}^{-1}\underset{\sim}{f} = \frac{1}{a_1 + b_1}(f_1 + b_1 f_2)$$

$$= -\left[\frac{s_1}{v_1 D_1} - \frac{s_2}{v_2 D_2} + \frac{k_1}{v_2 D_2{}^3}\left[s_1 + s_2 - (x_F - x_I)\right]\right] \det \hat{J}$$

$$= \left(\frac{s_2}{v_2 D_2} - \frac{k_2}{v_2 D_2{}^3}\left[s_2 - (x_F - x_I)\right]\right)\det \hat{J} \qquad (3.21)$$

Since $0 \leq s_2 \leq (x_F - x_I) + \varepsilon_1$, we have $s_2 - (x_F - x_I) \leq \varepsilon_1$. Hence we can choose ε_1 small enough that c_2 is positive.

Similarly for $s_2 = 0$, $\underset{\sim}{n}_3 = (0,-1)^T$. On $s_2 = 0$ we obtain

$$c_3 = \underset{\sim}{n}_3 \cdot \hat{J}^{-1}\underset{\sim}{f} = -\frac{1}{a_1 + b_1}(-f_1 + a_1 f_2)$$

$$= \frac{s_1}{v_1 D_1} - \frac{k_1}{v_1 D_1{}^3}(s_1 - (x_F - x_I)) \qquad (3.22)$$

As before, we may choose an ε_2 such that for $0 \leq s_1 \leq x_F - x_I + \varepsilon_2$, $c_3 > 0$. We now take $\varepsilon = \min(\varepsilon_1, \varepsilon_2)$. Then $c_1 > 0$, $c_2 > 0$, $c_3 > 0$. Hence (3.14b) holds with $\sigma = -1$ on $\partial\Omega$. Thus Smale's boundary conditions are satisfied and the continuation method must produce a path which leads to a solution in Ω.

We return now to the matrix J of (3.3) for the case of parallel plane layers. For this case it is easy to show that J is symmetric and positive definite.

From (3.3) with $y' = y'' = 0$, we obtain the elements of J as

$$b_i = -\frac{k_i}{v_i D_i} < 0$$

$$a_i = \frac{k_i}{v_i D_i{}^3} + \frac{k_{i+1}}{v_{i+1} D_{i+1}{}^3} = -(b_i + c_i) > 0$$

$$c_i = -\frac{k_{i+1}}{v_{i+1} D_{i+1}{}^3} = b_{i+1} < 0 \qquad (3.23)$$

Since J is tridiagonal and $c_i = b_{i+1}$, J is symmetric. First let us show that any matrix of this type has a positive determinant. Since all principal submatrices are also of the form, they all will have positive determinant and hence the matrix must be positive definite. Thus we need only show that $\det J > 0$. If we decompose J into $L\,U$ factored form where L is lower triangular and U is upper triangular,

$$J = LU = \begin{bmatrix} \alpha_1 & & & & \\ b_2 & \alpha_2 & & & \\ & b_3 & \alpha_3 & & \\ & & \ddots & \ddots & \\ & & & \ddots & \ddots \\ & & & & \ddots \quad \ddots \\ & & & b_N & \alpha_N \end{bmatrix} \begin{bmatrix} 1 & \gamma_1 & & & \\ & 1 & \gamma_2 & & \\ & & 1 & \ddots & \\ & & & \ddots & \ddots \\ & & & & \ddots \quad \ddots \\ & & & & \gamma_{N-1} \\ & & & & 1 \end{bmatrix} \qquad (3.24)$$

Then we find [Isaacson and Keller [10], p. 56]

$$\begin{aligned} \alpha_1 &= a \quad , \gamma_1 = c_1/_{\alpha_1} & & \\ \alpha_i &= a_i - b_i\gamma_{i-1} \quad , & i &= 2,3,\ldots N \\ \gamma_i &= c_i/\alpha_i \quad , & i &= 2,3,\ldots N \end{aligned} \qquad (3.25)$$

By Theorem 5, Isaacson and Keller [10], p. 56, $\alpha_i > |a_i| - |b_i|$. However, we have from (3.23)

$$|a_i| - |b_i| = |c_i| > 0 \qquad (3.26)$$

Hence $\alpha_i > 0$ for all i. Now $\det J = \det L \cdot \det U = \det L = \prod_{i=1}^{N} \alpha_i > 0$. Thus $\det J > 0$. If we denote by $J^{(k)}$ the matrix composed of the first k columns and rows of J, then each $J^{(k)}$ has the form exhibited in (3.24). The above determinant condition thus applies to each $J^{(k)}$, $k = 1, \ldots, N$. Hence

$$\det J^{(k)} > 0 \ , \ k = 1, \ldots, N \qquad (3.27)$$

Thus by Theorem 1, p. 152, Franklin [6], J is positive definite.

II.4. Travel Time, Amplitude, Phase

II.4.1. Travel Time

After a ray path has been determined, we compute the time for a signal of the given propagation type to travel from $\underset{\sim}{x}_I$ to $\underset{\sim}{x}_F$. Since the speed on the k^{th} segment $[\underset{\sim}{x}_{k-1}, \underset{\sim}{x}_k]$ is v_k and the length of that segment is D_k, clearly the travel time is given by:

$$t = \sum_{k=1}^{N+1} \frac{D_k}{v_k} \; . \tag{4.1}$$

II.4.2. Amplitude Calculation

The amplitude along a ray is computed assuming that a source of unit strength is located at $\underset{\sim}{x}_I$. In a narrow tube of rays surrounding the ray in question, it is assumed that the energy carried by the wave is conserved. The change in energy along a ray is proportional to the normal cross-sectional area of the ray tube which is proportional to the Jacobian of the mapping induced by the rays. Also, at each interface a ray may split into two reflected and two transmitted rays. This must be taken into account for energy conservation. This yields the standard reflection and transmission coefficients at the interfaces.

The Jacobian of the mapping has been calculated by Cerveńy et al. [4]. This result follows.

With the following definitions

$$R_k = \frac{(1 + (f'_{i_k}(x_k))^2)^{3/2}}{|f''_{i_k}(x_k)|} \equiv \text{radius of curvature of interface } i_k \text{ at } x_k$$

$\theta_k \equiv$ incident angle

$\hat{\theta}_k \equiv$ angle of reflection/transmission

v_k, v_{k+1} $\equiv$ propagation speeds

r_k $\equiv$ radius of curvature of wavefront prior to incidence

$\hat{r}_k$ $\equiv$ radius of curvature of wavefront after incidence (4.2)

The geometric spreading factor G may now be computed as follows:

$$G = [(\hat{r}_{N+1} + D_{N+1}) \prod_{j=1}^{N} d_j]^{-\frac{1}{2}} \tag{4.3a}$$

$$d_j = \frac{v_{j+1} \cos^2\theta_j}{v_j \cos^2 \hat{\theta}_j} + \frac{r_j}{\sigma_j R_j \cos^2\hat{\theta}_j} [\frac{v_{j+1}}{v_j}\cos\theta_j + \hat{\sigma}_j \cos\hat{\theta}_j] \tag{4.3b}$$

$$\hat{r}_j = \frac{r_j}{d_j} \tag{4.3c}$$

$$r_j = \hat{r}_{j-1} + D_j \tag{4.3d}$$

$$r_0 = 0 \tag{4.3e}$$

where σ_j and $\hat{\sigma}_j$ are determined by the following rules. The tangent line to interface i_j at the point $\underset{\sim}{x}_j$ is given by (see Figure 2).

$$t_j(x) = f'_{i_j}(x_j)\ (x-x_j) + y_j\ . \tag{4.4}$$

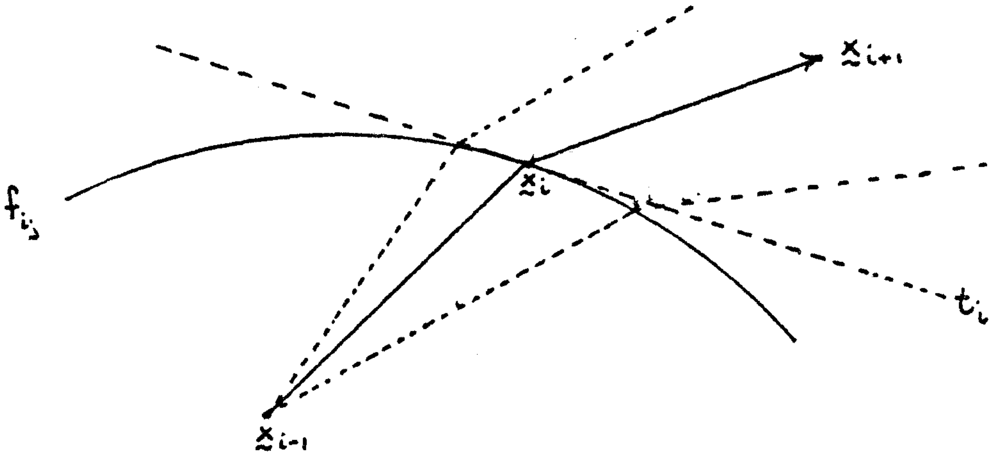


Figure 2. Tangent line t_i and a tube of rays

One now wishes to determine under what conditions spreading occurs, and under what conditions focussing occurs.

There are two cases to be considered. If the curvature is negative (with respect to the incident ray),then spreading occurs for a transmitted ray and focussing for a reflected ray. Similarly, if the curvature is positive (with respect to the incident ray), focussing occurs for the transmitted rays and spreading for the reflected rays. (Note that this implies that the amplitude is <u>not</u> independent of the direction of transversal!) Since $\sigma_j = 1$ corresponds to spreading, $\sigma_j = -1$ to focussing, we obtain the following determination.

$$\sigma_j = \operatorname{sgn} (f''_{i_j}(x_j) \; [t_j(x_{j-1}) - y_{j-1}]) \tag{4.5}$$

The factor $\hat{\sigma}_j$ merely signifies whether the ray is reflected or transmitted. This determination is easily made by examining the ray signature. Hence from (2.9), a ray is reflected at $\underset{\sim}{x}_k$ if $j_k - j_k \bmod 2 = j_{k+1} - j_{k+1} \bmod 2$. So the Jacobian is completely determined with

$$\hat{\sigma}_k = \begin{cases} 1 \; , & j_k - j_k \bmod 2 = j_{k+1} - j_{k+1} \bmod 2 \\ -1 \; , & \text{otherwise} \end{cases} \tag{4.6}$$

Next we must consider the amplitude change due to the splitting of rays at an interface. This requires the solution of a 4x4 linear system at each interface (2x2 at free surfaces). A derivation of this system may be found in Keller [12], pp. 371-381. Only the portions relevant to the functioning of a working computer code will be detailed here.

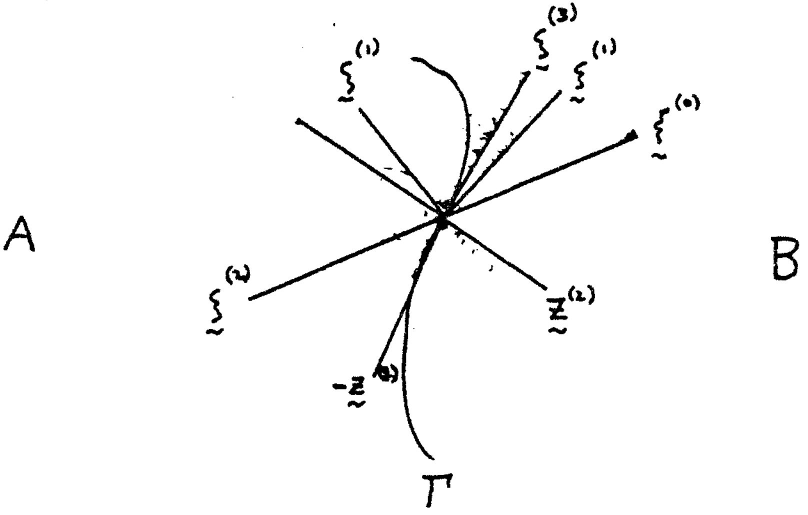
Let Γ be an interface separating medium A from medium B. We define the following unit vectors (see Figure 3) at the point $\underset{\sim}{x}_j$.

$$
\begin{aligned}
\underset{\sim}{\xi}_j^{(0)} &\equiv \text{incident ray} \\
\underset{\sim}{\xi}_j^{(1)} &\equiv \text{reflected P ray} \\
\underset{\sim}{\xi}_j^{(2)} &\equiv \text{reflected S ray} \\
\underset{\sim}{\xi}_j^{(3)} &\equiv \text{transmitted P ray} \\
\underset{\sim}{\xi}_j^{(4)} &\equiv \text{reflected S ray}
\end{aligned}
\tag{4.7}
$$

The unit normal, $\underset{\sim}{z}_j^{(1)}$, and unit tangent, $\underset{\sim}{z}_j^{(2)}$, to interface i_j at the point $\underset{\sim}{x}_j$ are defined by:

$$
\begin{aligned}
\underset{\sim}{z}_j^{(1)} \cdot \underset{\sim}{\xi}_j^{(0)} &> 0 \\
z_j^{(2)} \cdot \underset{\sim}{\xi}_j^{(0)} &\geq 0
\end{aligned}
\tag{4.8}
$$

(see Figure 3).



$\underset{\sim}{\xi}^{(1)}$
$\underset{\sim}{\xi}^{(3)}$
$\underset{\sim}{\xi}^{(1)}$
$\underset{\sim}{\xi}^{(0)}$
A
B
$\underset{\sim}{\xi}^{(2)}$
$\underset{\sim}{z}^{(2)}$
$-\underset{\sim}{z}^{(1)}$
Γ

Figure 3. Unit vectors $\underset{\sim}{z}^{(n)}$ and $\underset{\sim}{\xi}^{(n)}$

We now define the angles θ_v^j by

$$\cos\theta_v^j = \underset{\sim}{\xi}_j^{(v)} \cdot \underset{\sim}{z}_j^{(1)}$$

$$\sin\theta_v^j = \underset{\sim}{\xi}_j^{(v)} \cdot \underset{\sim}{z}_j^{(2)} \qquad (4.9)$$

and the elastic parameters by

$\mu_A^j \ (\mu_B^j) \equiv$ shear modulus in medium A (B)

$\rho_A^j \ (\rho_B^j) \equiv$ density of medium A (B)

$\lambda_A^j \ (\lambda_B^j) \equiv$ Lamé constant in medium A(B) (4.10)

Thus we can define the amplitudes and the appropriate speeds by

$\alpha_v^j \equiv$ amplitude of ray in direction $\underset{\sim}{\xi}_j^{(v)}$

$c_1^j = \sqrt{\dfrac{\lambda_A^j + 2\mu_A^j}{\rho_A^j}} \equiv$ speed of P ray in medium A

$c_2^j = \sqrt{\dfrac{\mu_A^j}{\rho_A^j}} \equiv$ speed of S ray in medium A

$c_3^j = \sqrt{\dfrac{\lambda_B^j + 2\mu_B^j}{\rho_B^j}} \equiv$ speed of P ray in medium B

$c_4^j = \sqrt{\dfrac{\mu_B^j}{\rho_B^j}} \equiv$ speed of S ray in medium B (4.11)

Using (4.7), (4.10) and (4.11) one can now write the system for the amplitudes as

$$Q_j \underset{\sim}{\alpha}_j = \alpha_0^j \underset{\sim}{K}_j \tag{4.12}$$

with:

$$Q_j = \begin{bmatrix} \cos\theta_1^j & \sin\theta_2^j & -\cos\theta_3^j & -\sin\theta_4^j \\ \sin\theta_1^j & -\cos\theta_2^j & -\sin\theta_3^j & \cos\theta_4^j \\ -\rho_A^j c_1^j \cos 2\theta_2^j & -\rho_A^j c_2^j \sin 2\theta_2^j & \rho_B^j c_3^j \cos 2\theta_4^j & \rho_B^j c_4^j \sin 2\theta_4^j \\ -\dfrac{\mu_A^j}{c_1^j}\sin 2\theta_1^j & \rho_A^j c_2^j \cos 2\theta_2^j & \dfrac{\mu_B^j}{c_3^j}\sin 2\theta_3^j & -\rho_B^j c_4^j \cos 2\theta_4^j \end{bmatrix} \tag{4.13}$$

$$\underset{\sim}{\alpha}_j = (\alpha_1^j, \alpha_2^j, \alpha_3^j, \alpha_4^j)^T$$

$$\underset{\sim}{K}_j = (k_1^j, k_2^j, k_3^j, k_4^j)^T .$$

The k_v^j are determined by the type of the incident ray as follows.

(Incident P ray)

$$k_1^j = \cos\theta_1^j,\ k_2^j = -\cos\theta_2^j,\ k_3^j = -\rho_A^j c_2^j \sin\theta_2^j,\ k_4^j = \frac{\mu_A^j}{c_1^j}\sin 2\theta_1^j \tag{4.14a}$$

(Incident S ray)

$$k_1^j = -\sin\theta_2^j,\ k_2^j = -\cos\theta_2^j,\ k_3^j = -\rho_A^j c_2^j \sin 2\theta_2^j,\ k_4^j = \frac{\mu_A^j}{c_1^j}\sin 2\theta_1^j \tag{4.14b}$$

For reflection from a free surface, the density on one side of the interface is zero. Thus no transmitted rays exist. One is left with a 2x2 system, which may be solved explicitly to yield:

(Incident P ray)

$$\alpha_1^j = -\alpha_0^j\left(\frac{(c_1^j)^2\cos^2 2\theta_2^j + (c_2^j)^2\sin 2\theta_2^j\ \sin 2\theta_0^j}{(c_1^j)^2\cos^2 2\theta_2^j - (c_2^j)^2\sin 2\theta_2^j\ \sin 2\theta_0^j}\right)$$

$$\alpha_2^j = \alpha_0^j\left(\frac{2c_1^j c_2^j\ \sin 2\theta_0^j\ \cos 2\theta_2^j}{(c_1^j)^2\cos^2 2\theta_2^j - (c_2^j)^2\sin 2\theta_2^j \sin 2\theta_0^j}\right) \qquad (4.15a)$$

(Incident S ray)

$$\alpha_1^j = -\alpha_0^j\left(\frac{2c_1^j c_2^j\ \cos 2\theta_0^j\ \sin 2\theta_0^j}{(c_1^j)\ \cos^2 2\theta_0^j + (c_2^j)^2\sin 2\theta_1^j\ \sin 2\theta_0^j}\right)$$

$$\alpha_2^j = \alpha_0^j\left(\frac{-(c_1^j)^2\cos^2 2\theta_0^j + (c_2^j)^2\sin 2\theta_1^j\ \sin 2\theta_0^j}{(c_1^j)^2\cos^2 2\theta_0^j + (c_2^j)^2\sin 2\theta_1^j\ \sin 2\theta_0^j}\right) \qquad (4.15b)$$

From the signature (2.9), we know which α_v^j to choose for each j, say $\alpha_{v_j}^j$. This yields the amplitude factor

$$E = \prod_{j=1}^{N} \alpha_{v_j}^j \qquad (4.16)$$

If α_I is the strength of the disturbance at $\underset{\sim}{x}_I$ and α_F the amplitude of the travel ray at $\underset{\sim}{x}_F$, then from (4.3) and (4.16) we obtain

$$\alpha_F = G\,E\,\alpha_I \qquad (4.17)$$

II.4.3. Phase Shifts

In order to produce a synthetic seismogram, it is necessary to input an initial impulse. There are standard convolutions which take the so-called "peak" seismograms (amplitude versus time for individual rays) as input and produce a continuous seismogram as output. Although we are not concerned with that problem here, we must supply all necessary information required as input in order to utilize the existing software. The preceding sections have shown how the amplitude and travel time may be computed. The only additional data necessary are the phase shifts.

There are three sources of phase shifts to consider: (1) reflections, (2) caustics, and (3) critical and supercritical rays.

Let us first deal with ordinary reflections, since this is the simplest case. Upon reflection, a phase change of π occurs. That is, the amplitude changes sign (since the wave reverses direction), hence a factor of $e^{\pi i}$ is introduced. We define the quantities γ_k by

$$\gamma_k = \begin{cases} 1 \;, & j_k - j_k \bmod 2 = j_{k+1} \bmod 2 \\ 0 \;, & \text{otherwise} \end{cases} \tag{4.18}$$

The total phase shift due to ordinary reflections is then expressed trivially by:

$$\psi_1 = \pi \sum_{k=1}^{N} \gamma_k \tag{4.19}$$

Secondly, passage through a caustic results in a phase shift of $\pi/2$. How to determine whether a caustic has been encountered is treated in II.4.4. It suffices here to state that since the rays are straight line segments, a maximum of one caustic can be encountered on each segment. The quantities β_k are used for this purpose.

$$\beta_k = \begin{cases} 1 \;, & \text{caustic crossed on segment } [\underset{\sim}{x}_k, \underset{\sim}{x}_{k+1}] \\ 0 \;, & \text{otherwise} \end{cases} \tag{4.20}$$

The total phase shift due to passage through caustics may be expressed as:

$$\psi_2 = \frac{\pi}{2} \sum_{k=1}^{N} \beta_k \tag{4.21}$$

The final source of phase shifts occurs when one or more of the splitting rays is supercritical. In this case the associated angles are imaginary. This may yield a complex solution to (4.13). If we represent the amplitudes by:

$$\alpha_n = \omega_n e^{i\phi_n}$$

$$\omega_n \geq 0$$

$$0 \leq \phi_n < 2\pi \tag{4.22}$$

then the phase shift is given by ϕ_{v_j}. Hence the total contribution due to supercritical rays is

$$\psi_3 = \sum_{k=1}^{N} \phi_{v_k} \tag{4.23}$$

Thus we arrive at the total phase shift ψ, from (4.19), (4.21), and (4.23):

$$\psi = \sum_{k=1}^{3} \psi_k \; . \tag{4.24}$$

II.4.4. Location of Caustics on Ray Segments

When a ray passes through a caustic, then the amplitude as given by geometrical optics is formally infinite. We can detect such a point by examining the factors d_j of (4.3b). If d_j is 0 at any point of

$[\underset{\sim}{x}_{j-1}, \underset{\sim}{x}_j]$, then that point lies on a caustic. Hence, if $d_j < 0$, we have passed through a caustic on $[\underset{\sim}{x}_{j-1}, \underset{\sim}{x}_j]$. Locating the point on $[\underset{\sim}{x}_{j-1}, \underset{\sim}{x}_j]$ where the caustic occurs is then quite simple. For infinite amplitude $(d_j = 0)$, we need (from (4.3b)):

$$\sigma_j R_j v_{j+1} \cos^2\theta_j + r_j v_{j+1} \cos\theta_j + \tilde{\sigma}_j v_j \cos\tilde{\theta}_j = 0 \qquad (4.25)$$

Replacing r_j by $\tilde{r}_{j-1} + \tilde{D}_j$ via (4.3d):

$$\sigma_j R_j v_{j+1} \cos^2\theta_j + (\tilde{r}_{j-1} + \tilde{D}_j) v_{j+1} \cos\theta_j + \tilde{\sigma}_j v_j \cos\tilde{\theta}_j = 0 \qquad (4.26)$$

Solving for $\tilde{D}_j$ we obtain

$$\tilde{D}_j = - \frac{\sigma_j R_j v_{j+1} \cos^2\theta_j + \tilde{\sigma}_j v_j \cos\tilde{\theta}_j + \tilde{r}_{j-1} v_{j+1} \cos\theta_j}{v_{j+1} \cos\theta_j}$$

$$= -\sigma_j R_j \cos\theta_j - \frac{v_j \tilde{\sigma}_j \cos\tilde{\theta}_j}{v_{j+1} \cos\theta_j} - \tilde{r}_{j-1} \qquad (4.27)$$

$\tilde{D}_j$ is the distance along the segment $[\underset{\sim}{x}_j, \underset{\sim}{x}_{j+1}]$ where a caustic would be located. If $\tilde{D}_j > 0$, then we can determine the crossing of the caustic $\underset{\sim}{x}_c = (x_c, y_c)$. (Note $\bar{D}_j < 0 \Rightarrow$ spreading is occurring, hence no caustics are possible.) If $\tilde{D}_j < D_j$((3.3c)), then $[\underset{\sim}{x}_{j-1}, \underset{\sim}{x}_j]$ passes through a caustic. For this case we obtain the following equation for x_c:

$$(x_c - x_{j-1})^2 + (m x_{j-1} + b - f_{i_{j-1}}(x_{j-1}))^2 = \tilde{D}_j^2 \qquad (4.28a)$$

where we have used

$$m = \frac{f_{i_j}(x_j) - f_{i_{j-1}}(x_{j-1})}{x_j - x_{j-1}}$$

$$b = y_{j-1} - m\, x_{j-1} \tag{4.28b}$$

The point x_c is then determined by the solution of this quadratic as:

$$x_c^{\pm} = x_{j-1} \quad \frac{1}{2u} \quad \sqrt{v^2 - 4uw} \tag{4.29a}$$

with the definitions

$$u = m^2 + 1, \; v = 2\left[m(b - y_{j-1}) - x_{j-1}\right]$$

$$w = x_{j-1}^2 + (b - y_{j-1})^2 - \widetilde{D}_j^{\,2} \tag{4.29b}$$

Finally $\underset{\sim}{x}_c$ is determined by

$$x_c = \begin{cases} x_c^{+} \;, & x_{j-1} < x_j \\ x_c^{-} \;, & x_{j-1} \geq x_j \end{cases} \tag{4.30a}$$

$$y_c = m\, x_c + b \tag{4.30b}$$

II.5. Diffracted Rays and the Problems of Existence and Uniqueness

II.5.1. Diffracted Rays

For planar interfaces, the path of a diffracted ray may be calculated in exactly the same manner as non-diffracted rays. One need only specify that the diffracted ray travels in the medium with the higher velocity. However, the amplitude cannot be determined by simple ray theory. Thus although we can determine the path and the travel time, we are unable to include these rays in an artificial seismogram.

For non-planar interfaces, simple modifications to the system (2.7) are required. The system decouples into n-1 independent systems, if n diffractions occur. Consider Figure 4.

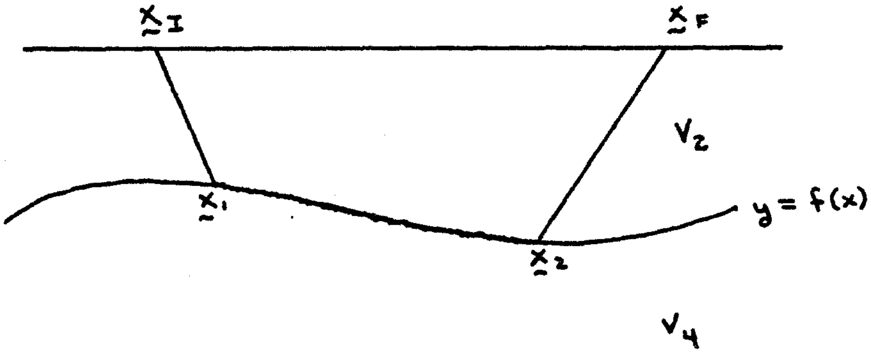


Figure 4. An example with a diffracted ray

The algebraic system is

$$v_4 \left[\frac{x_1 - x_I + f'(x_1)\,(f(x_1) - y_I)}{D_1} \right] + \delta_1 v_2 \sqrt{1 + f'(x_1)^2} = 0 \tag{5.1a}$$

$$v_4 \quad \frac{x_F - x_2 + f'(x_2)\ (y_F - f(x_2))}{D_3} \quad + \delta_2 v_2 \quad \sqrt{1 + f'(x_2)^2} \qquad (5.1b)$$

δ_i is chosen so that the angle between $[\underset{\sim}{x}_i, \underset{\sim}{x}_{i+1}]$ and $[\underset{\sim}{x}_{i+1}, \underset{\sim}{x}_{i+2}]$ is not acute. This difficulty does not arise if we use a parametric form of representation for the interfaces, i.e.,$[x_i = x_i(t),\ y_i = y_i(t)$ denotes the i^{th} interface$]$. Note that (5.1a) and (5.1b) are independent. For larger systems we obtain similar results. The 9x9 system in Figure 5 decouples into one 3x3, one 2x2, and one 4x4 system.

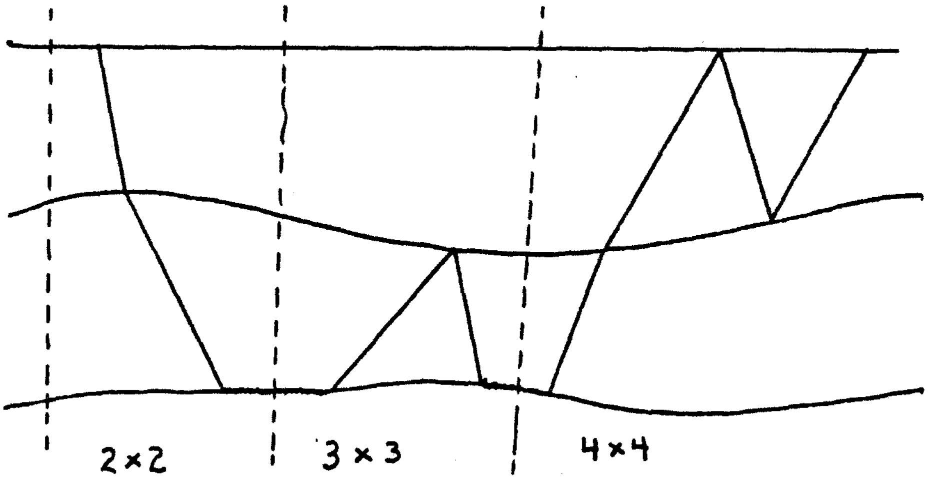


Figure 5. An example with two diffracted segments

Although calculation of the amplitudes is not possible in an elementary way, the travel time may be determined using any numerical integration routine. For example, in the system (5.1), once x_1 and x_2 have been determined, the travel time along the curve $f(x)$ from x_1 to x_2 is simply

$$t = \int_{x_1}^{x_2} \frac{\sqrt{1 + (f'(x))^2}}{v_4} \, dx \qquad (5.2)$$

Thus, we are able to identify the time of arrival of a diffracted ray, allowing us to determine the beginning of a head wave.

II.5.2. The Problem of Existence, Non-Physical Rays, and Non-Uniqueness

In the most general setting, we have no guarantee that a ray of any given type exists. More important, the solution which we obtain may not correspond to a physically meaningful ray. From II.3.3, we know that for parallel plane layers a solution exists (and is in fact unique), and thus for small perturbations this will remain true. (The perturbations must be small in the first two derivatives of f_{i_j}, as well as f_{i_j} itself.) Figure 6 illustrates the very simplest case in which a solution exists, but the ray is non-physical. In general, only examination of the solution allows us to determine if it is physically acceptable. A scope with graphing capabilities may be utilized to reject any rays which are unacceptable.

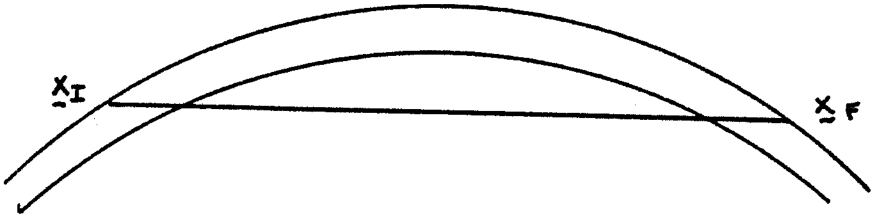


Figure 6. A ray path directly joining $\underset{\sim}{x}_I$ to $\underset{\sim}{x}_F$ exists, but is non-physical.

Similarly, the problem of non-unique rays is equally perplexing. For very general interface shapes the solution need not be unique. Indeed, an infinite number of solutions may exist. Figure 7 depicts such a case, where the source and receiver are located at the same point, the center of a circle.

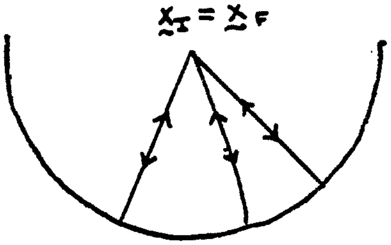


Figure 7. All radii of the circle are physical ray paths

A less contrived example of both non-existence and non-uniqueness appears in II.7.

II.6. Stratified Media with Non-Constant Speeds

The previous sections have all dealt with media having piecewise constant elastic properties. However, for certain non-constant velocity distributions, the problem may still be rendered completely algebraic with a tridiagonal Jacobian matrix.

Let us examine the travel time principle upon which geometrical optics is based. We desire to locally minimize the travel time between two specified end points. Hence, we must minimize the functional

$$I = \int_{x_1}^{x_2} \frac{\sqrt{1 + y'^2}}{v(x,y)}\, dx = \int_{t_1}^{t_2} \frac{\sqrt{\dot{x} + \dot{y}^2}}{v(x,y)}\, dt \tag{6.1}$$

where $\dot{a} = \frac{da}{dt}$ and $a' = \frac{da}{dx}$.

Setting $w = \frac{1}{v(x,y)}$, from the calculus of variations, we need the first variation to be zero, yielding [Gelfand and Fomin [8] , p. 19]:

$$w_y - w_x y' - w \frac{y''}{1 + y'^2} = 0 \tag{6.2a}$$

$$\begin{aligned} y(x_1) &= y_1 \\ y(x_2) &= y_2 \end{aligned} \tag{6.2b}$$

where the subscripts x and y refer to partial differentiation. In general, we cannot obtain even a first integral of this system. However, if we restrict ourselves to the class of velocity distributions given by the ansatz

$$v(x,y) = h(Ax + By + C) \tag{6.3}$$

then we are able to obtain a first integral in the following manner.

First, we make the change of variable

$$p = \frac{Ax + c}{A} \tag{6.4}$$

Substituting (6.4) into (6.1) and using (6.3) we obtain

$$I = \int_{t_1}^{t_2} \frac{\sqrt{\dot{x}^2 + \dot{y}^2}}{v(x,y)} \, dt = \int_{t_1}^{t_2} \frac{\sqrt{\dot{p}^2 + \dot{y}^2}}{h(Ap + By)} \, dt \tag{6.5}$$

We now perform a rotation of the p-y coordinate system.

Define X, Y, and θ via:

$$X = p \sin\theta + y \cos\theta$$

$$Y = -p \cos\theta + y \sin\theta$$

$$\sin \theta = \frac{A}{\sqrt{A^2 + B^2}}$$

$$\cos \theta = \frac{B}{\sqrt{A^2 + B^2}} \tag{6.6}$$

Using (6.6) in (6.5) yields

$$I = \int_{t_1}^{t_2} \frac{\sqrt{\dot{X}^2 + \dot{Y}^2}}{h(\sqrt{A^2+B^2}\ X)} \, dt \tag{6.7}$$

Finally, to obtain our final desired form, we define

$$a = \sqrt{A^2 + B^2}\ X \ , \quad b = \sqrt{A^2 + B^2}\ Y \tag{6.8}$$

Substituting (6.8) into (6.7) gives

$$I = \frac{1}{\sqrt{A^2 + B^2}} \int_{t_1}^{t_2} \frac{\sqrt{\dot{a}^2 + \dot{b}^2}}{h(a)} \, dt \tag{6.9}$$

We thus arrive at the problem which is of immediate interest.

$$\text{Minimize} \qquad I = \int_{t_1}^{t_2} \frac{\sqrt{\dot{a}^2 + \dot{b}^2}}{h(a)}\, dt$$

$$= \int_{a_1}^{a_2} \frac{\sqrt{1 + b'^2}}{h(a)}\, da \tag{6.10}$$

where $b' = \frac{db}{da}$. Note that h is now independent of b. Thus from (6.2) we must solve

$$b''h - b'(1 + b'^2)h' = 0 \tag{6.11}$$

This is a first order differential equation for $z = b'$,

$$z(1 + z\ )h' - hz' = 0 \tag{6.12}$$

In differential form, we have

$$\frac{dz}{z(1 + z^2)} = \frac{h'}{h}\, da \tag{6.13}$$

This can be integrated to get

$$\log z - \tfrac{1}{2}\log\,(1 + z^2) = \log h + c_1 \tag{6.14}$$

Solving this for z, we obtain the following expression for $\frac{db}{da}$,

$$z = \frac{db}{da} = c_2 \frac{h(a)}{\sqrt{1 - c_2^2\, h^2(a)}} \tag{6.15}$$

where $c_2 = e^{c_1}$.

Finally, to obtain b as a function of a, we integrate once more:

$$b(a) = c_2 \int_{a_1}^{a_2} \frac{h(c)}{\sqrt{1 - c_2^2\, h^2(c)}}\, dc + c_3 \tag{6.16}$$

with c_2 and c_3 determined by the initial conditions $b(a_1) = b_1$, $b(a_2) = b_2$. Thus, if we can integrate (6.16) explicitly, we are left with purely an algebraic problem [i.e., matching segments of curves at interfaces, involving no differential equations]. Even if we cannot perform the integration, a numerical integration can be applied to approximate $b(a)$. Since this has not, as yet, been implemented, it is not clear if this is more efficient than solving the system of o.d.e.'s, although one would tend to believe it should be.

In order to be useful, it is necessary that we can obtain a closed form for b for a wide variety of choices of the function h. The following list of integrals indicates that this is possible for many elementary functions. (Recall, however, that we have already performed a rotation,translation, and compression or expansion on the original coordinate system.) E and S represent integration constants in the following:

h(a)	b(a)
constant	$b = Ea + S$ (lines)
a	$(b - S)^2 + a^2 = E^2$ (circles)
e^a	$b = \arcsin E e^a + S$
$\sin a$	$b = \arcsin (E \cos a) + S$
$\sinh a$	$b = \arcsin (E \cosh a) + S$
$\cosh a$	$b = \arcsin (E \sinh a) + S$
$\frac{1}{a}$	$b = \log (a + \sqrt{a^2 - E}) + S$
sech a	$b = F(\arcsin (\text{sech } a), E) + S$ (F is the Legendre elliptic integral)
$\frac{1}{a^2}$	$b = \text{arcsec } (Ea^2) + S$
cosec a	$b = F(\arcsin (E \cos a, \frac{1}{E})) + S$
$\sec a$	$b = F(\arcsin(E \sin a, \frac{1}{E})) + S$
$a^{-\frac{1}{2}}$	$b = \sqrt{x - E} + S$
$(c_1a + c_2a^2)^{-\frac{1}{2}}$	$b = \log (2\sqrt{c_2(E + c_1a + c_2a^2)} + 2c_2a + c_1) + S$
$\cot c_1a$	$b = \arcsin (E \sin c_1a) + S$

II.7. Examples of the Direct Problem

II.7.1 Example 1 - Vertical Layers

The set-up for example 1 is shown in Figure 8.

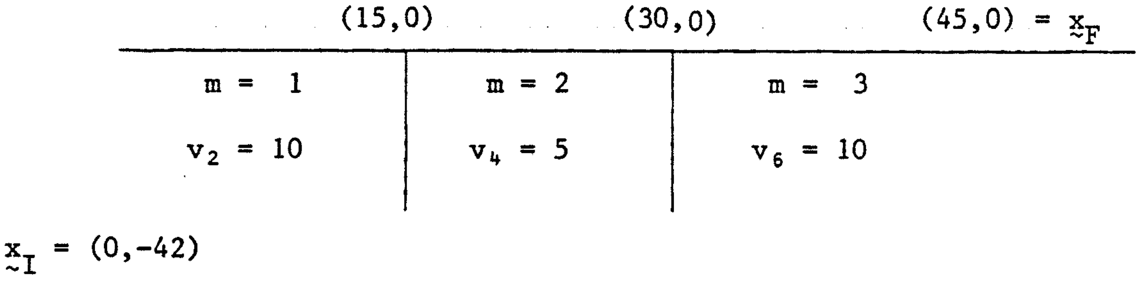

(15,0)
(30,0)
(45,0) = $\underset{\sim}{x}_F$
m = 1
m = 2
m = 3
v_2 = 10
v_4 = 5
v_6 = 10
$\underset{\sim}{x}_I$ = (0,-42)

Figure 8. Geometry of Example 1

This example was used to test whether results were correct. Since the geometry is so simple, the y-coordinates can be calculated easily by hand a priori. Also, Pereyra [15] was dealing with this geometry at the time the code was being developed, and thus this provided an additional test of the validity of the results. One interesting sidelight is that due to the symmetry of the geometry (45-30 = 15-0), and the fact $v_2 = v_6$, all rays which are strictly of type P (i.e. PPP, PPPPP, PPPPPPP, etc.) must pass through the point (22.5, -21). Plot #1 illustrates that this is indeed the case. Note that in this example, the interfaces are given as $x = f_i(y)$. The program can, in general, handle interfaces represented both as $x = f_i(y)$ and $y = g_i(x)$ simultaneously.

II.7.2. Example 2 - Layered Media with Oscillatory Interfaces

Three media consisting of three pieces separated by sinusoidal cracks are considered (Figure 9).

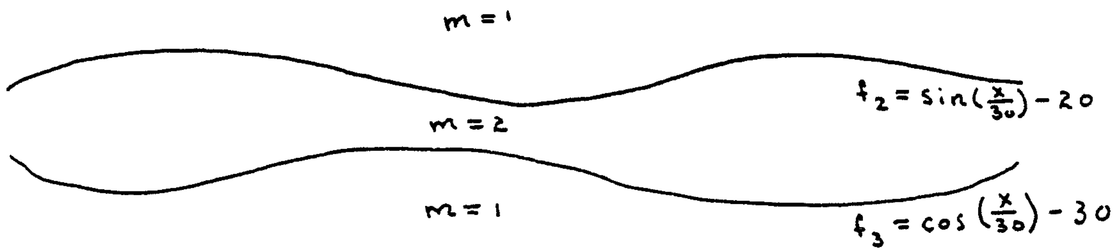

m = 1
$f_2 = \sin(\frac{x}{30}) - 20$
m = 2
m = 1
$f_3 = \cos(\frac{x}{30}) - 30$

Figure 9. Geometry of Example 2

The speeds are given by:

$$v_2 = 6.8,\ v_3 = 3.7$$
$$v_4 = 6.1,\ v_5 = 3.5 \tag{7.1}$$

This test example is included merely to illustrate the applicability to truly non-planar interfaces. Several families of rays are depicted in plots 2 - 5. The given disturbance was assumed to generate only P waves. The initial and final points are not shown on the plots, but all rays begin and end at the same points.

II.7.3. Example 3 - Parallel Plane Layers

A point seismogram is computed for each of three receivers. Time is plotted on the horizontal scale versus the base 10 logarithm of the amplitude on the vertical scale. The 20 most direct rays are computed (using only a P wave source). All P to S and S to P conversions are included (see Figure 10).

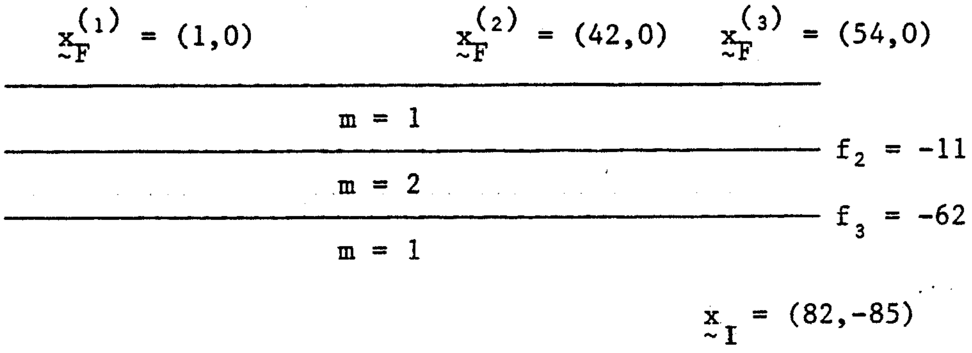
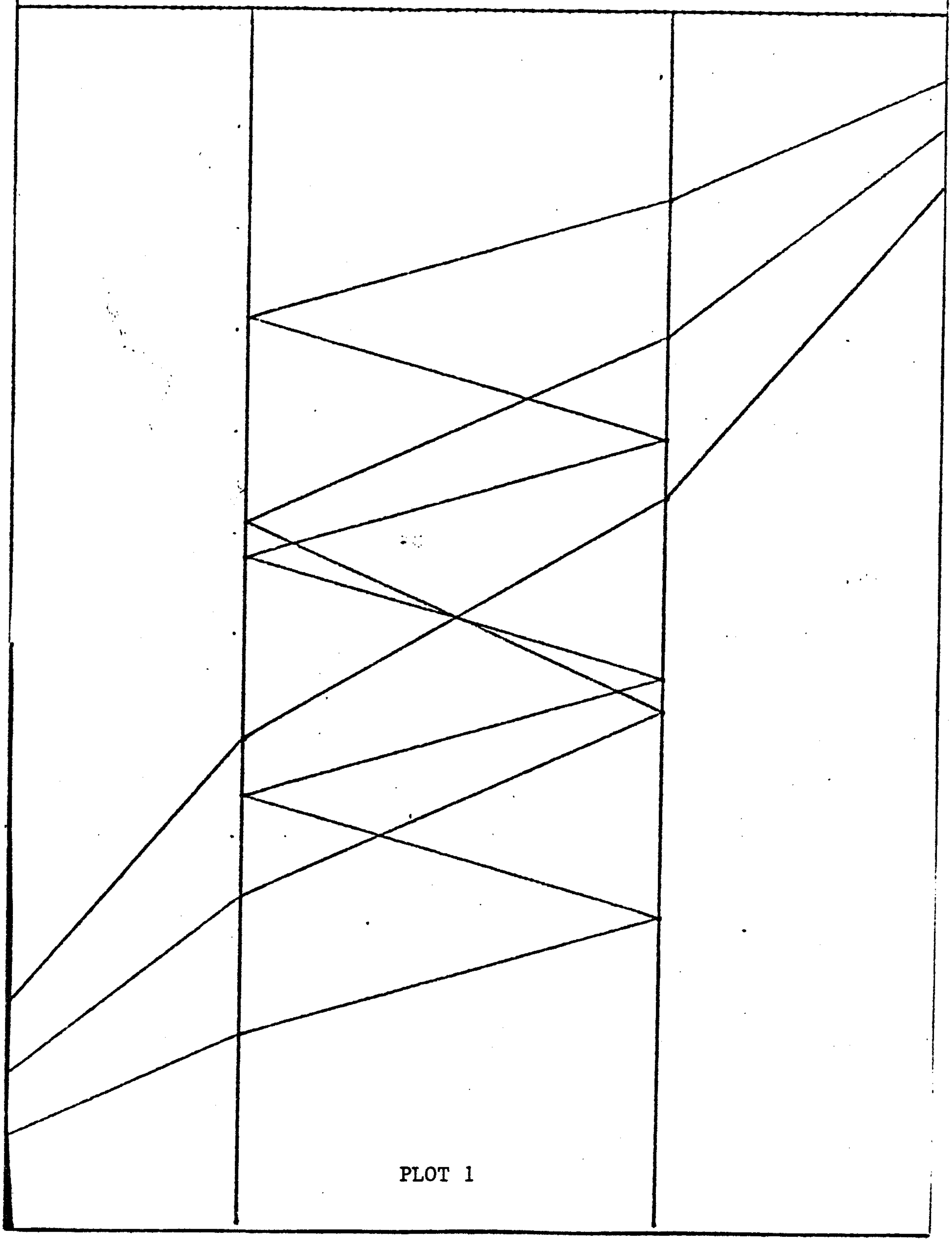
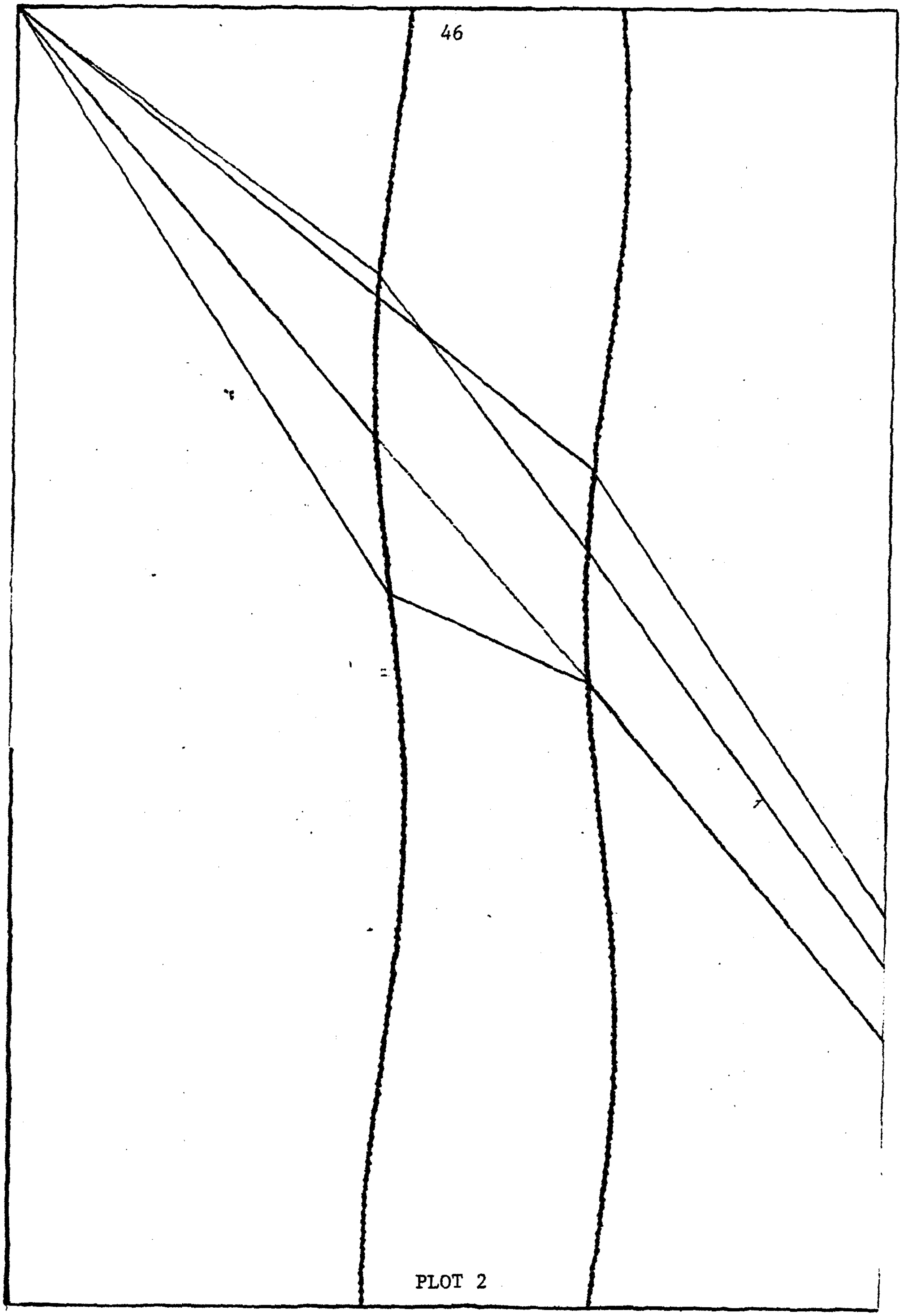

$\underset{\sim}{x}_F^{(1)} = (1,0)$
$\underset{\sim}{x}_F^{(2)} = (42,0)$
$\underset{\sim}{x}_F^{(3)} = (54,0)$
m = 1
$f_2 = -11$
m = 2
$f_3 = -62$
m = 1
$\underset{\sim}{x}_I = (82,-85)$

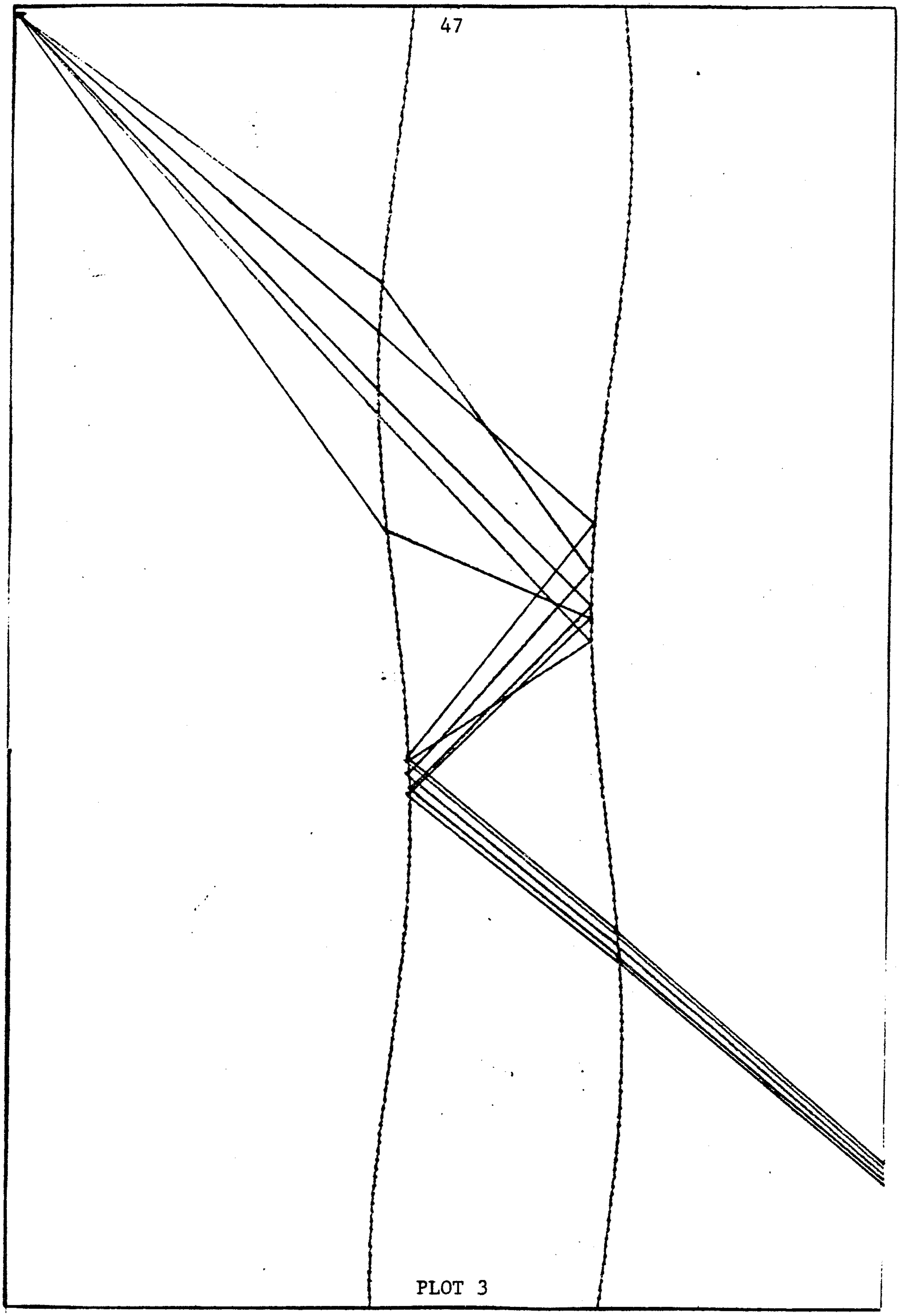
Figure 10. Geometry of Example 3



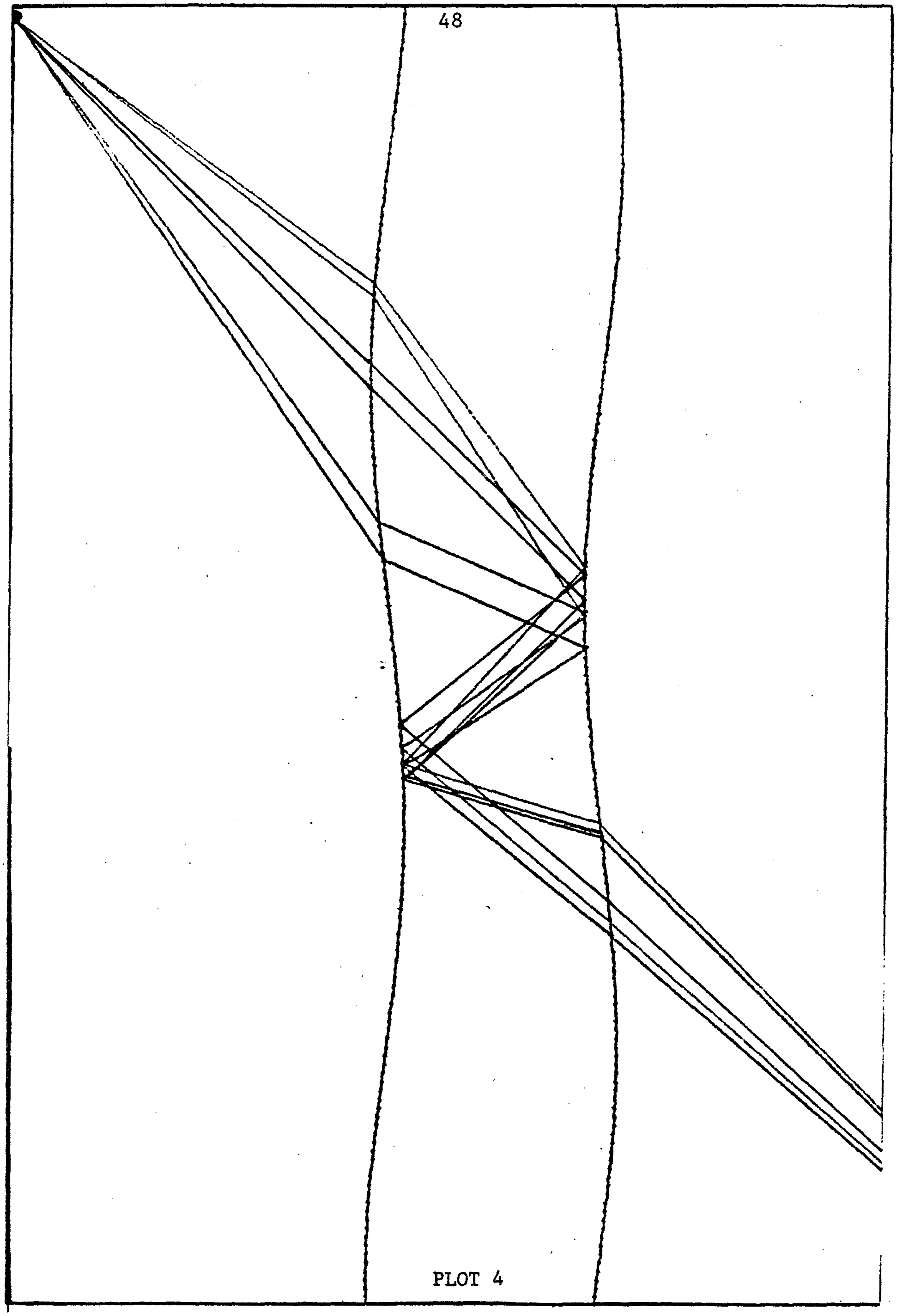
PLOT 1



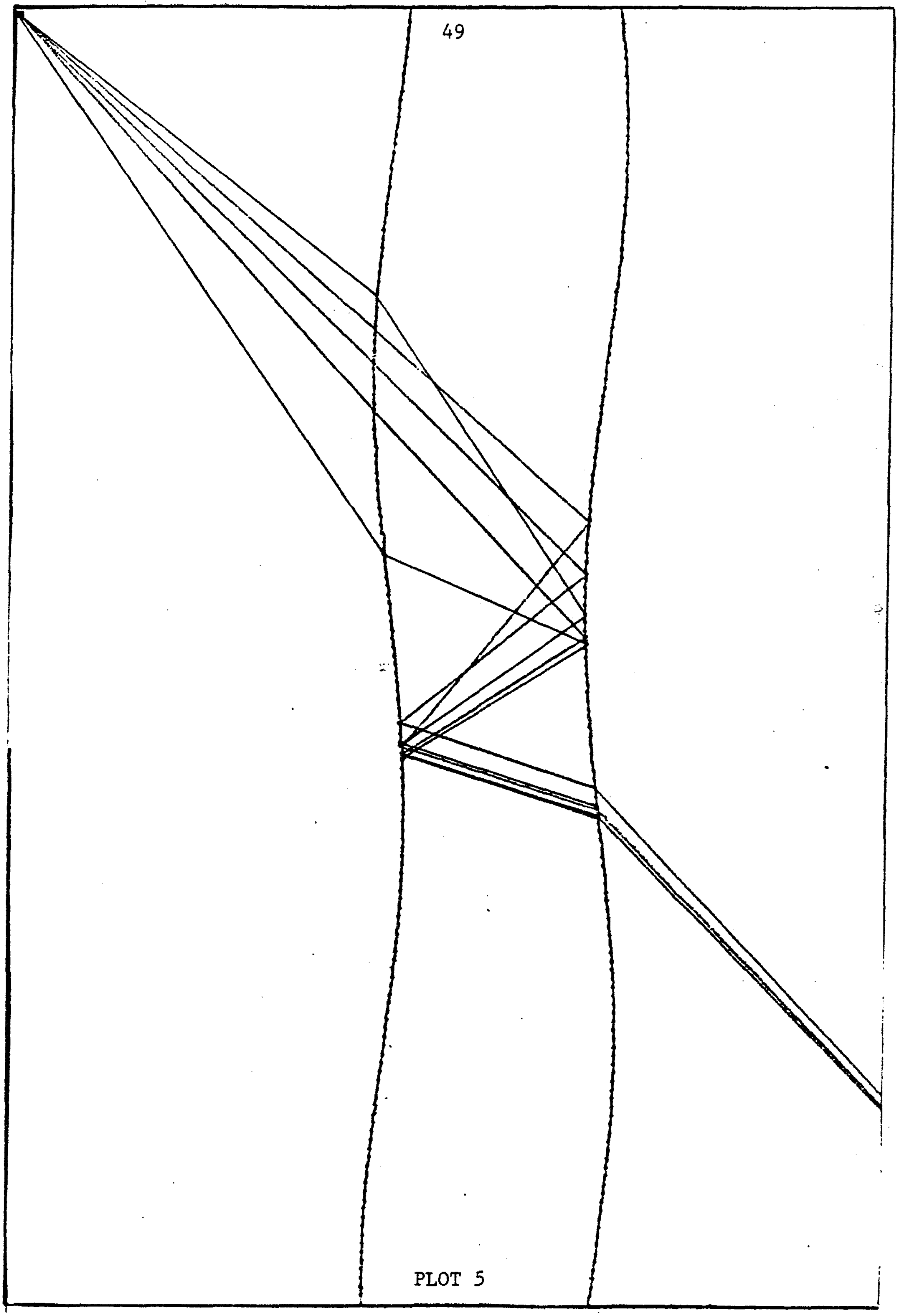
PLOT 2



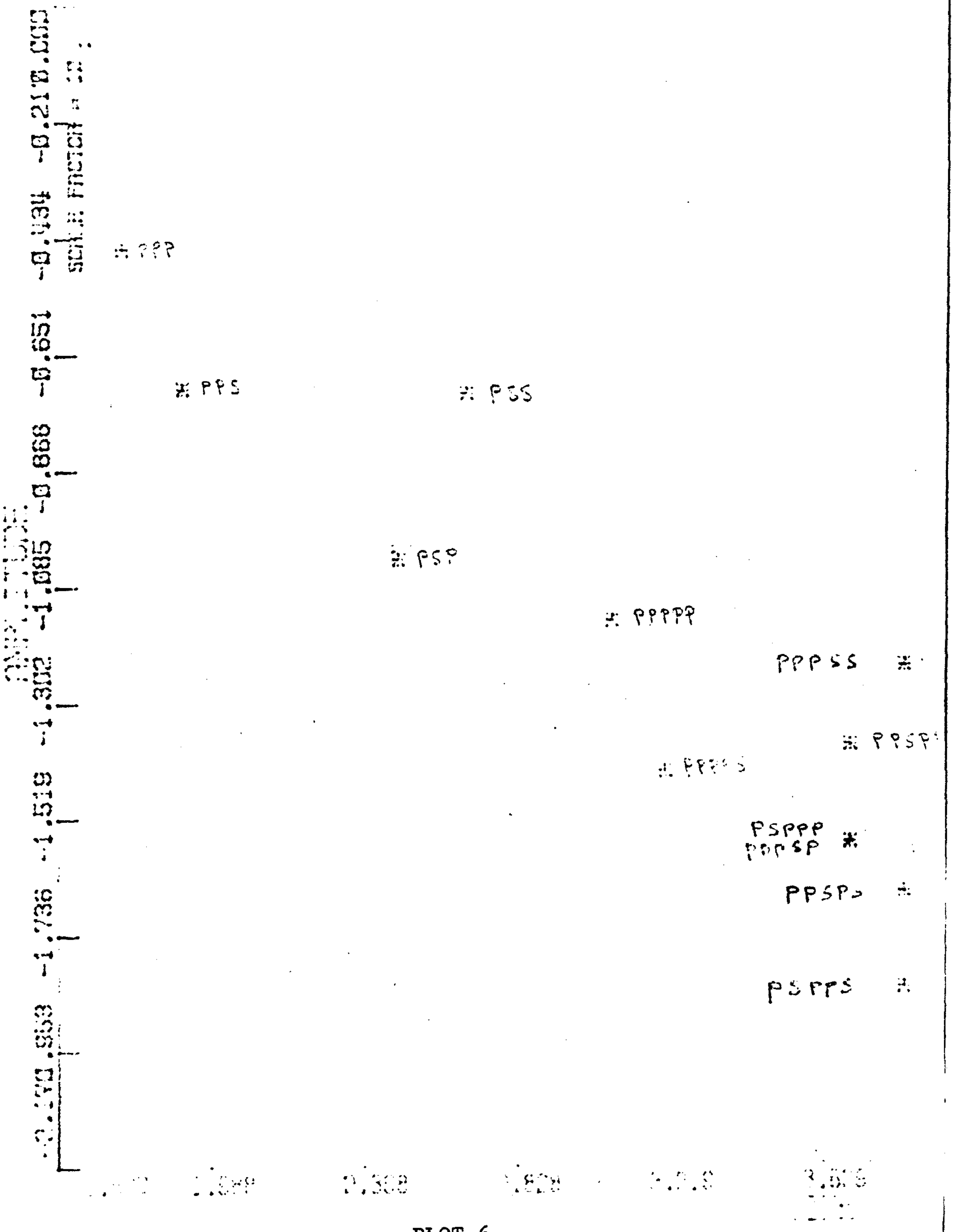
PLOT 3



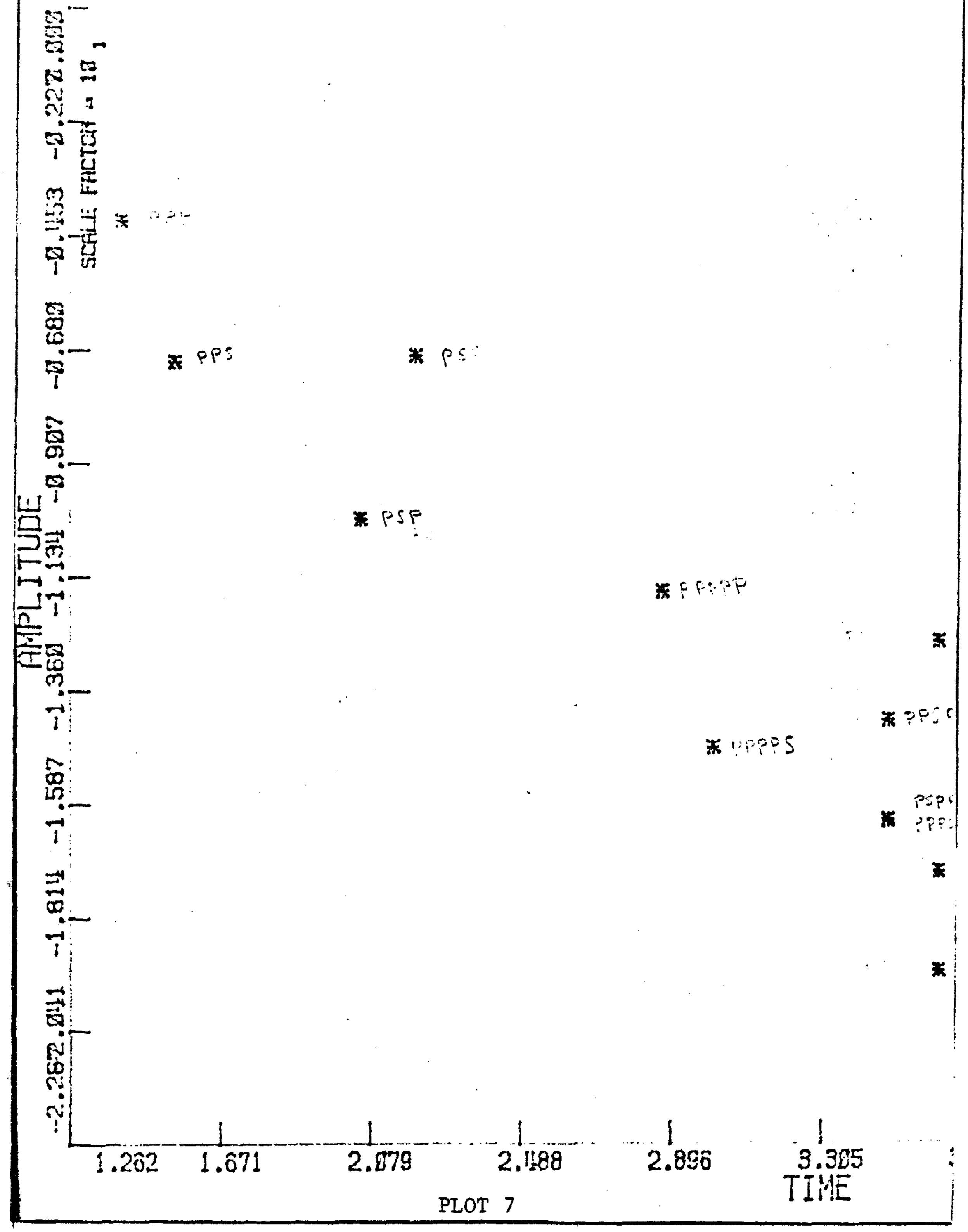
PLOT 4



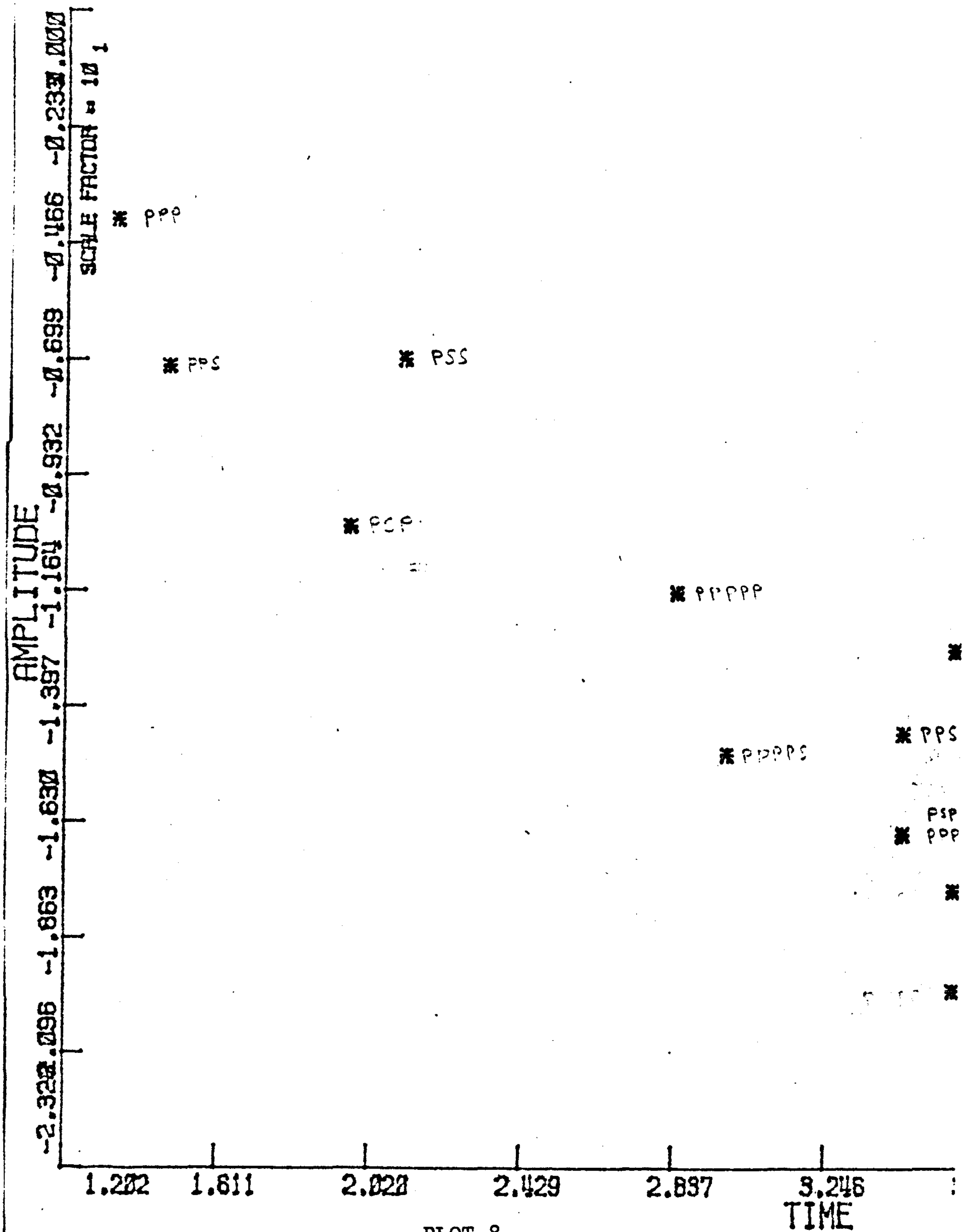
PLOT 5


-2.170 -1.953 -1.736 -1.519 -1.302 -1.085 -0.868 -0.651 -0.434 -0.217 0.000
SCALE FACTOR
* PPP
* PPS
* PSS
* PSP
* PPPPP
PPPSS *
* PPSP
* PPPPS
PSPPP *
PPPSP
PPSPP *
PSPPS *

PLOT 6


-2.262
-2.041
-1.814
-1.587
-1.360
-1.134
-0.907
-0.680
-0.453
-0.227
0.000
AMPLITUDE
SCALE FACTOR = 10 1
PPS
PS
PSP
PPPPP
PPPPS
PPS
PSP
PPP
1.262
1.671
2.079
2.488
2.896
3.305
TIME

PLOT 7


AMPLITUDE
SCALE FACTOR = 1Ø 1
-2.329
-2.Ø96
-1.863
-1.63Ø
-1.397
-1.164
-Ø.932
-Ø.699
-Ø.466
-Ø.233
Ø.ØØØ
PPP
PPS
PSS
PSP
PPPPP
PPPPS
PPS
PSP
PPP
1.2Ø2
1.611
2.Ø2Ø
2.429
2.837
3.246
TIME

PLOT 8

The same basic pattern is observed at each station with merely a time shift. This is precisely what one would expect from plane layers (see Plots 6 - 8).

This example serves to illustrate some of the output options available to the code user. Table 1 gives the maximum available output for each ray. The first six entries labeled SMAX refer to the maximum norm of the residual using the initial guess $\underset{\sim}{x}(1) = \underset{\sim}{x}(0) + \dot{\underset{\sim}{x}}(0)$ via continuation. Even though the iteration is converging, the program decides to try a smaller step. (A maximum of 6 iterations per step was imposed. If the residual, r, is not smaller than the tolerance, here set to 10^{-7}, a smaller continuation step is used. Also if r exceeds 10^4 , a smaller step is immediately implemented.) The blank line indicates that a smaller step is being tried. The following two values of SMAX indicate the residual using $\underset{\sim}{x}(1) = \underset{\sim}{x}(\frac{1}{2}) + \frac{1}{2}\dot{\underset{\sim}{x}}(\frac{1}{2})$ as initial guess (i.e., two steps of continuation lead to convergence). At the third iteration $r < 10^{-7}$. (The final residual is not printed.) The x and y coordinates of the intersections of the ray with the interfaces are then printed. Finally, the amplitude and travel time are output.

II.7.4. Modelling of a River Bed

Figure 11 shows the set-up for Example 4.

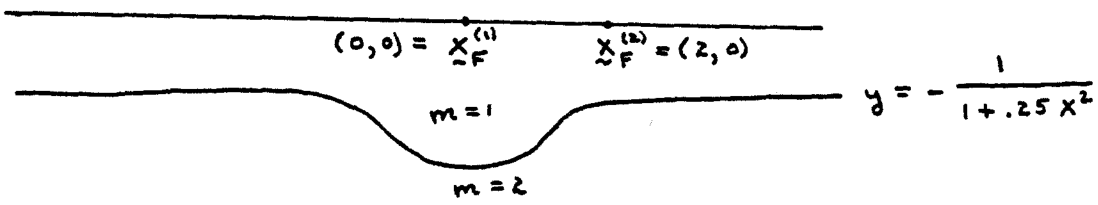

$(0,0) = \underset{\sim}{x}_F^{(1)}$
$\underset{\sim}{x}_F^{(2)} = (2,0)$
$m = 1$
$m = 2$
$y = -\frac{1}{1+.25x^2}$

Figure 11. Model of a River Bed

TABLE 1

```
ENTER INITIAL POINT
82.,-85.,

 SMAX    0.2694202E+02SMAX
 SMAX    0.8163634E+01SMAX
 SMAX    0.2294151E+01SMAX
 SMAX    0.3611835E+00SMAX
 SMAX    0.7807753E-02SMAX
 SMAX    0.2627497E-05SMAX

 SMAX    0.1742341E-01SMAX
 SMAX    0.2224505E-05SMAX
     0.8200000E+02      -0.8500000E+02
     0.5932782E+02      -0.6200000E+02
     0.1184322E+02      -0.1100000E+02
     0.1000000E+01       0.0000000E+00
THE AMPLITUDE FACTOR IS  0.2874527E-04
THE ARRIVAL TIME IS  0.1752660E+03           PPP
AMP=  0.2874527E-04TIME=  0.1752660E+03

 SMAX    0.5504957E+03SMAX
 SMAX    0.3798785E+05SMAX

 SMAX    0.8245828E+01SMAX
 SMAX    0.2322185E+01SMAX
 SMAX    0.2011011E+00SMAX
 SMAX    0.1913102E-02SMAX
 SMAX    0.4721957E-06SMAX
     0.8200000E+02      -0.8500000E+02
     0.5727913E+02      -0.6200000E+02
     0.5761724E+01      -0.1100000E+02
     0.1000000E+01       0.0000000E+00
THE AMPLITUDE FACTOR IS  0.6908614E-07
THE ARRIVAL TIME IS  0.1914058E+03           PPS
AMP=  0.6908614E-07TIME=  0.1914058E+03

 SMAX    0.3451305E+02SMAX
 SMAX    0.1710306E+02SMAX
 SMAX    0.2767163E+01SMAX
 SMAX    0.4368305E+00SMAX
 SMAX    0.7900561E-02SMAX
 SMAX    0.3283522E-05SMAX
     0.8200000E+02      -0.8500000E+02
     0.4515273E+02      -0.6200000E+02
     0.1862261E+02      -0.1100000E+02
     0.1000000E+01       0.0000000E+00
THE AMPLITUDE FACTOR IS  0.4369812E-10
THE ARRIVAL TIME IS  0.2490988E+03           PSP
AMP=  0.4369812E-10TIME=  0.2490988E+03

 SMAX    0.2107211E+02SMAX
 SMAX    0.3994740E+01SMAX
 SMAX    0.6219177E-01SMAX
```

This problem is of true physical interest due to the fact that many cities are located over former river beds. Table 2 gives an example of the minimum printed output (short of complete suppression). The integer, n, following these data, allows one to determine the number of continuation steps used. Since only uniform steps are currently employed, 2^{n-1} steps are taken across the interval. The amplitude is decomposed into vertical and horizontal components to reflect that aspect of the actual recording machinery. Note especially that for each of the stations, the most direct ray is not the strongest. There are several multiply reflected rays of greater or equal intensity (see Plots #9 - 12).

On this problem the first attempt to produce a gather was put forth. The source was located at (0,-8) with receivers at (i,0) for i = 1,2,3,4 (see Figure 12).

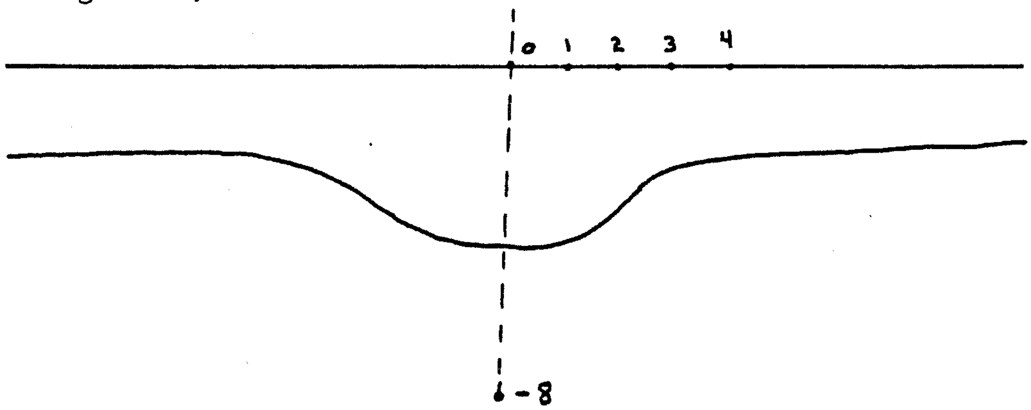


Figure 12. Source and Receivers for Gather

Graph 1 depicts the results. Both P and S waves are generated by the source. The 20 most direct rays at each station are included. The right half of the basin is clearly discernible in the gather, with an echoing effect attributed to the slower S wave arrivals.

II.7.5. Example of Non-Uniqueness and Non-Existence

The geometry of this example is depicted in Figure 13.

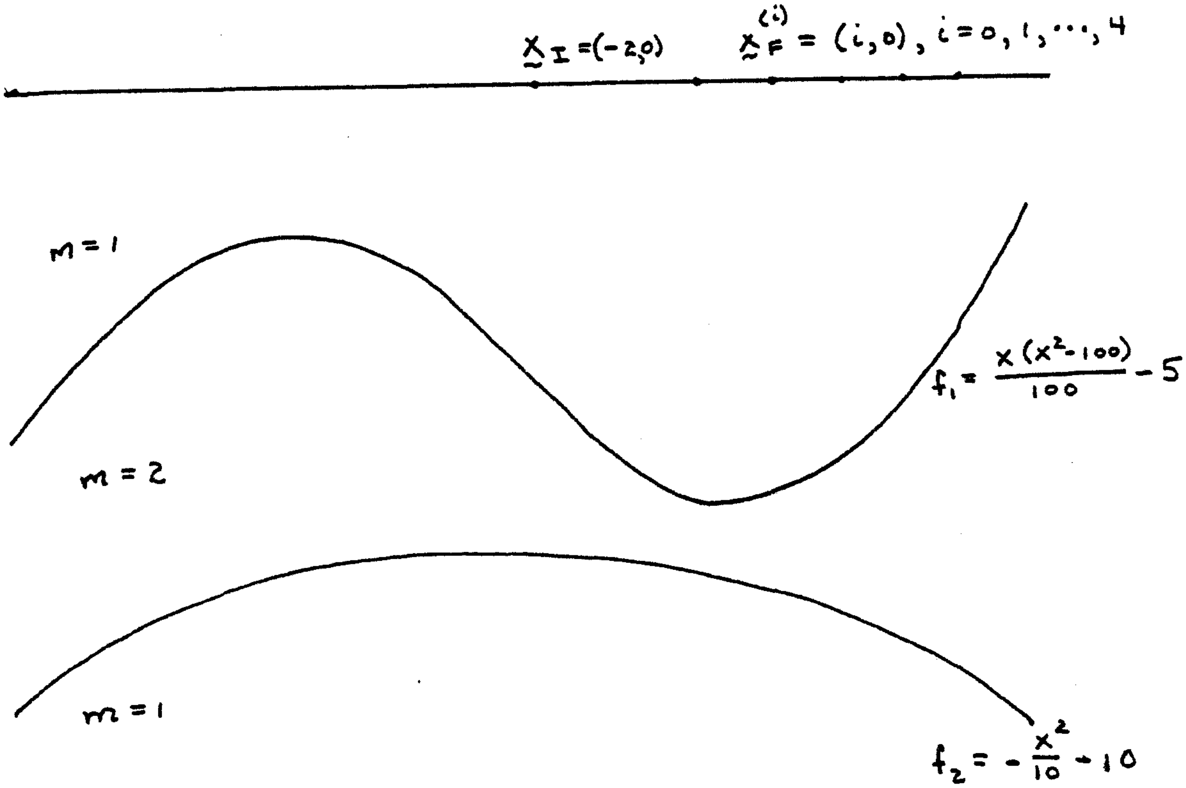
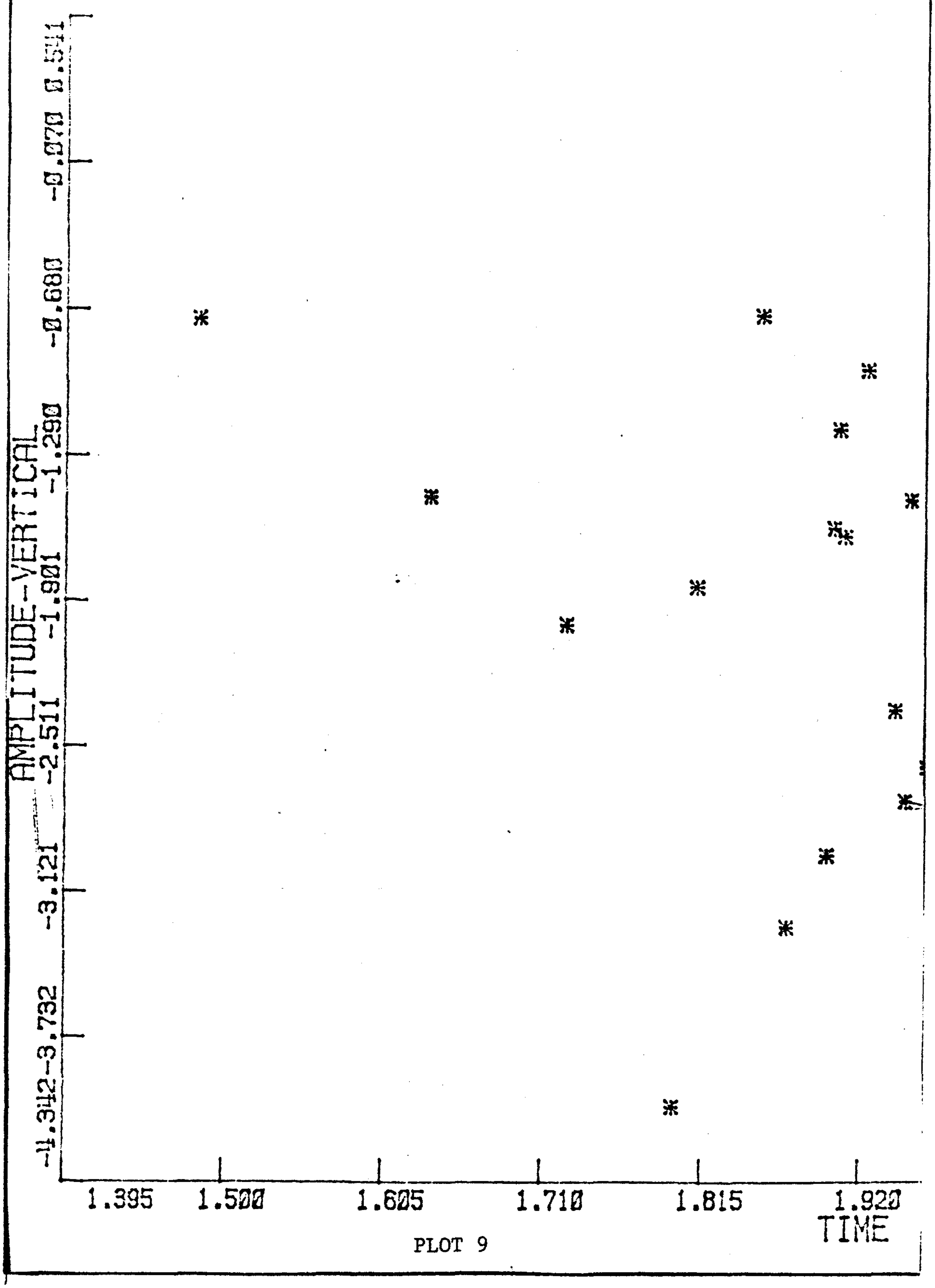

$\underset{\sim}{x}_I = (-2,0)$
$\underset{\sim}{x}_F^{(i)} = (i,0),\ i=0,1,\cdots,4$
m = 1
m = 2
$f_1 = \frac{x(x^2-100)}{100} - 5$
m = 1
$f_2 = -\frac{x^2}{10} - 10$

Figure 13. Geometry for Example 5

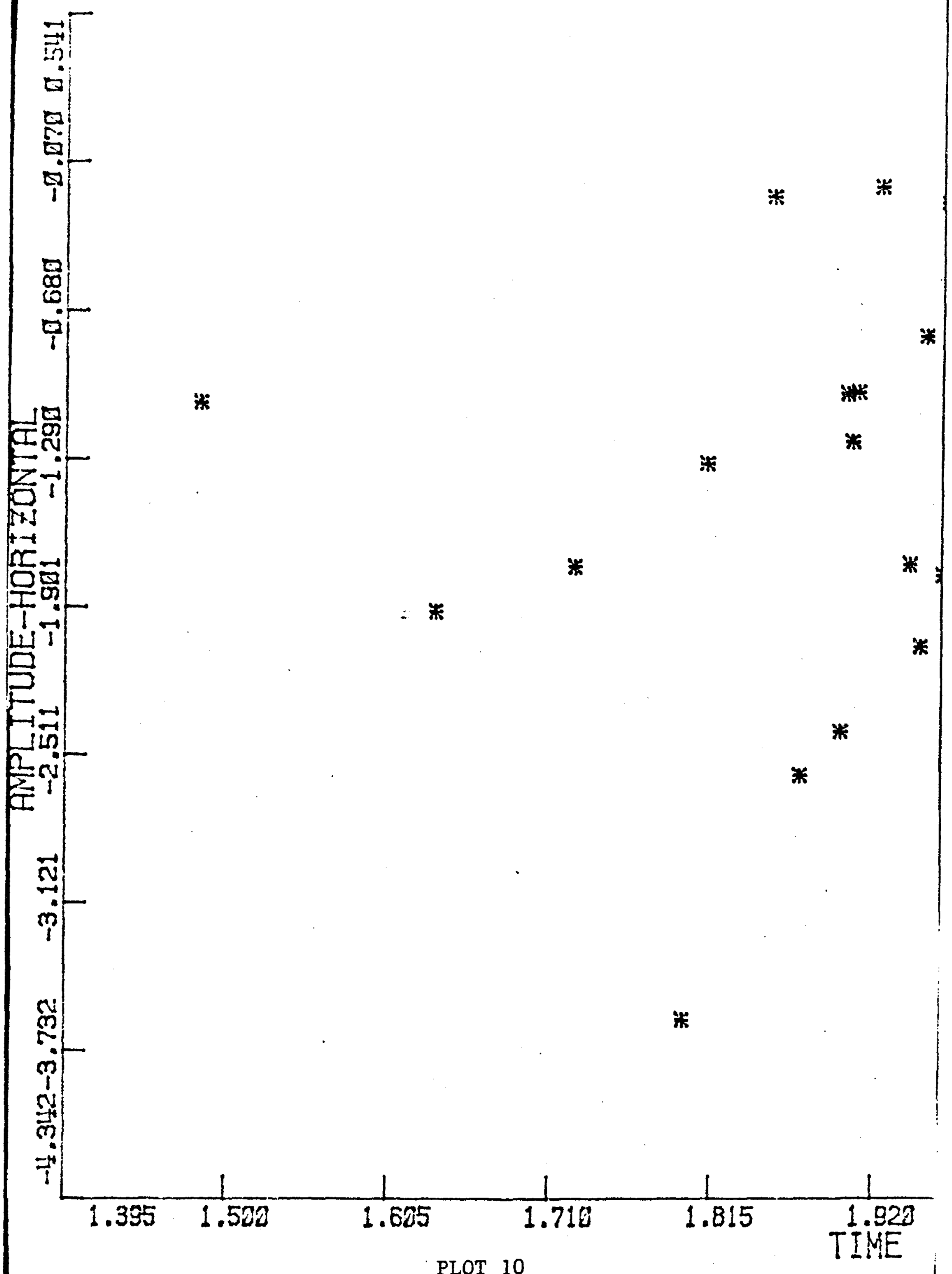
Graph #2 is a gather from the five stations indicated in Figure 12. Although two distinct structures are visible, more stations are required to refine their precise shapes. 228 rays are computed at each station. The rays included are rays of the families shown in Figure 14.

TABLE 2

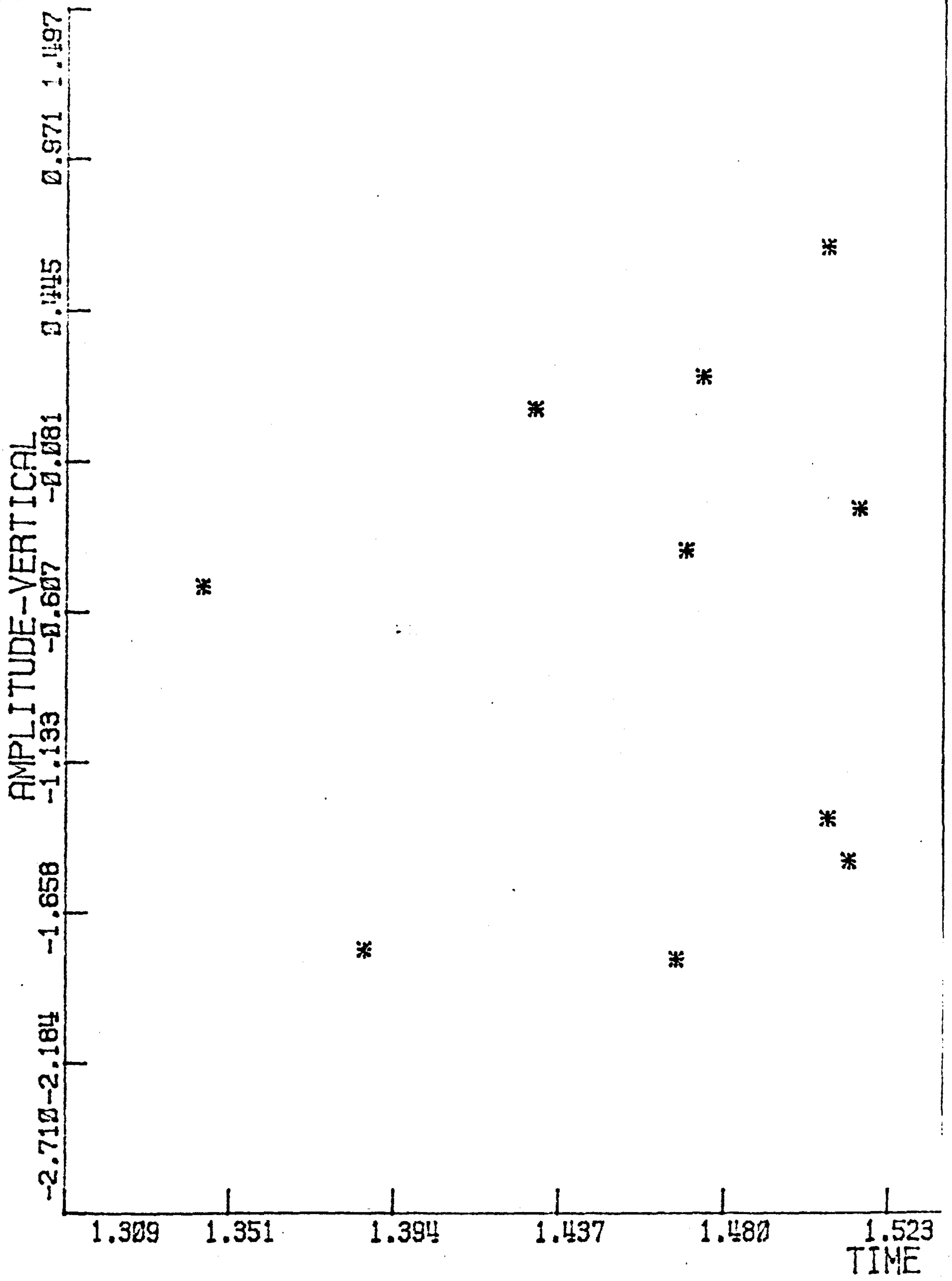
```
THE AMPLITUDE FACTOR IS  0.4067016E-01
THE ARRIVAL TIME IS  0.2052540E+02
      2
AMP=  0.4067016E-01TIME=  0.2052540E+02



THE AMPLITUDE FACTOR IS  0.8020618E+00
THE ARRIVAL TIME IS  0.1922112E+02
      3
AMP=  0.8020618E+00TIME=  0.1922112E+02


THE AMPLITUDE FACTOR IS  0.3722455E+00
THE ARRIVAL TIME IS  0.2171618E+02
      2
AMP=  0.3722455E+00TIME=  0.2171618E+02



THE AMPLITUDE FACTOR IS  0.1169690E+00
THE ARRIVAL TIME IS  0.1907521E+02
      3
AMP=  0.1169690E+00TIME=  0.1907521E+02



THE AMPLITUDE FACTOR IS  0.1298163E+00
THE ARRIVAL TIME IS  0.2119133E+02
      3
AMP=  0.1298163E+00TIME=  0.2119133E+02



THE AMPLITUDE FACTOR IS  0.1372871E+01
THE ARRIVAL TIME IS  0.1969878E+02
      3
AMP=  0.1372871E+01TIME=  0.1969878E+02


THE AMPLITUDE FACTOR IS  0.4592402E+00
THE ARRIVAL TIME IS  0.2247403E+02
      2
AMP=  0.4592402E+00TIME=  0.2247403E+02



THE AMPLITUDE FACTOR IS  0.1154697E+00
THE ARRIVAL TIME IS  0.1900575E+02
      3
AMP=  0.1154697E+00TIME=  0.1900575E+02


THE AMPLITUDE FACTOR IS  0.1217986E+00
```


AMPLITUDE-VERTICAL
-4.342
-3.732
-3.121
-2.511
-1.9Ø1
-1.29Ø
-Ø.68Ø
-Ø.Ø7Ø
Ø.541
1.395
1.5ØØ
1.6Ø5
1.71Ø
1.815
1.92Ø
TIME

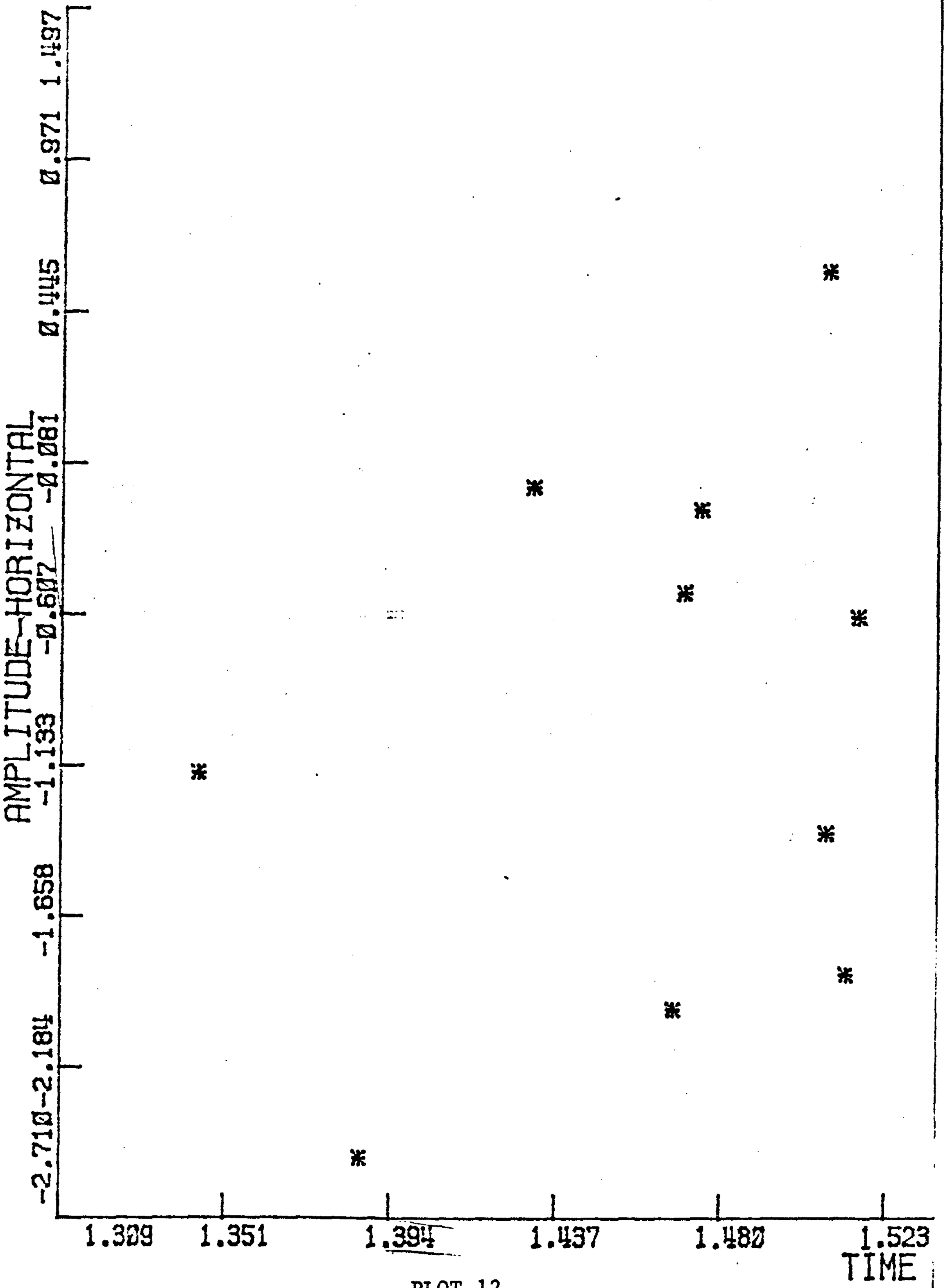
PLOT 9


AMPLITUDE-HORIZONTAL
-4.342
-3.732
-3.121
-2.511
-1.9Ø1
-1.29Ø
-Ø.68Ø
-Ø.Ø7Ø
Ø.541
1.395
1.5ØØ
1.6Ø5
1.71Ø
1.815
1.92Ø
TIME

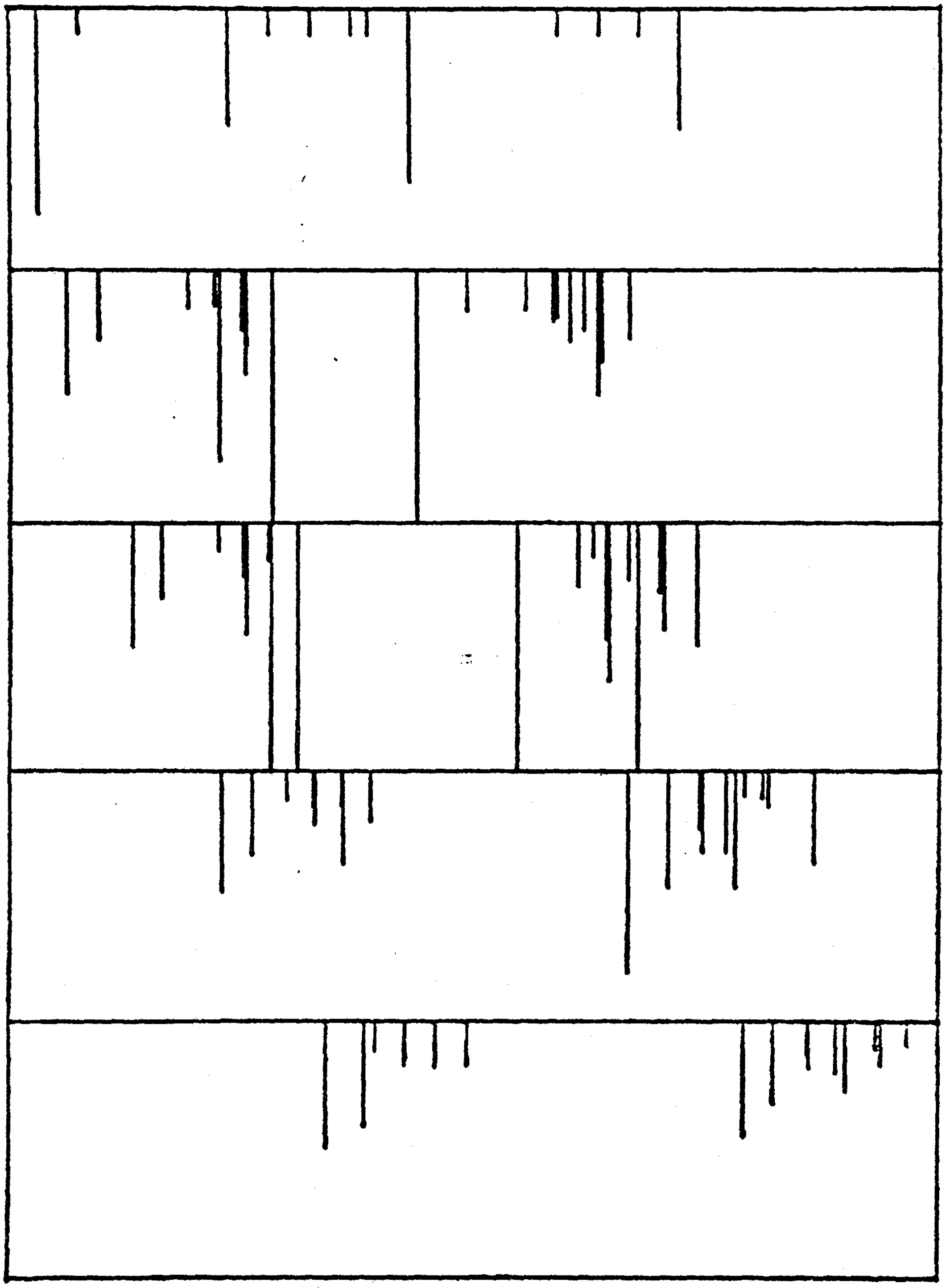
PLOT 10


AMPLITUDE-VERTICAL
-2.71Ø
-2.184
-1.658
-1.133
-Ø.6Ø7
-Ø.Ø81
Ø.445
Ø.971
1.497
1.3Ø9
1.351
1.394
1.437
1.48Ø
1.523
TIME

PLOT 11


AMPLITUDE-HORIZONTAL
1.497
0.971
0.445
-0.081
-0.607
-1.133
-1.658
-2.184
-2.710
1.309
1.351
1.394
1.437
1.480
1.523
TIME

PLOT 12



GRAPH 1

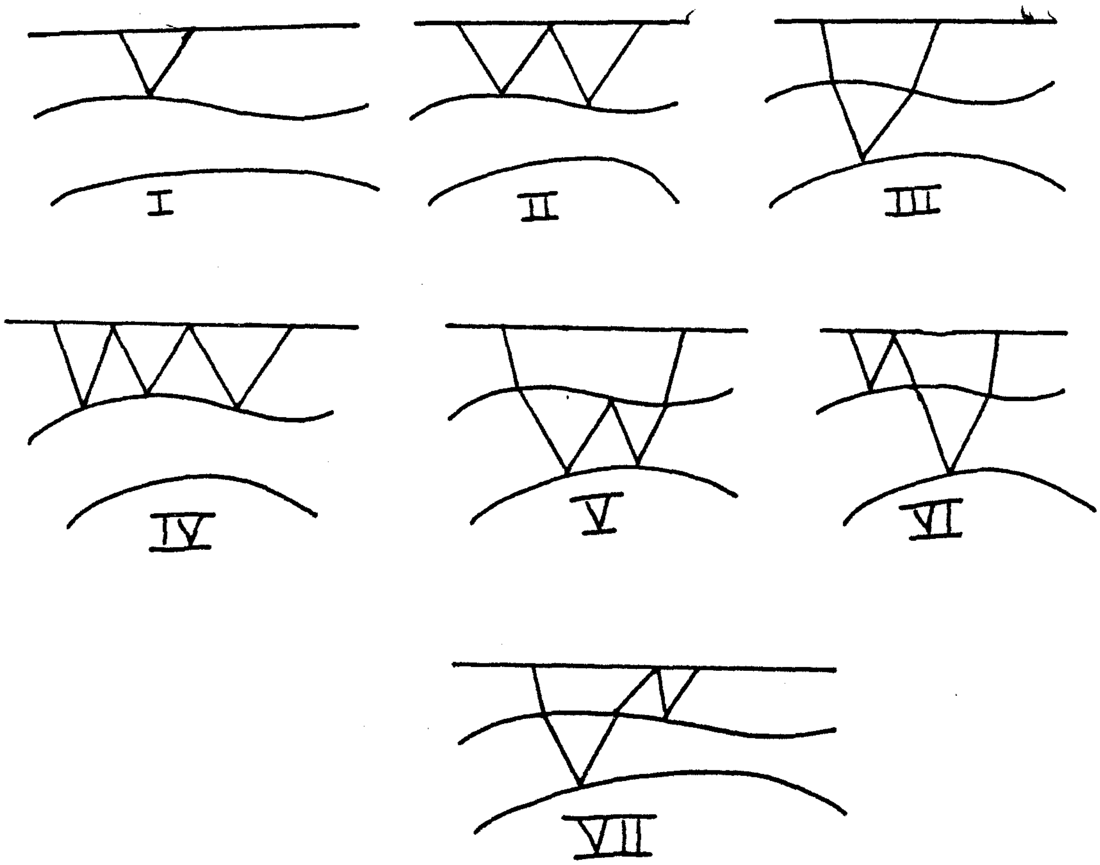


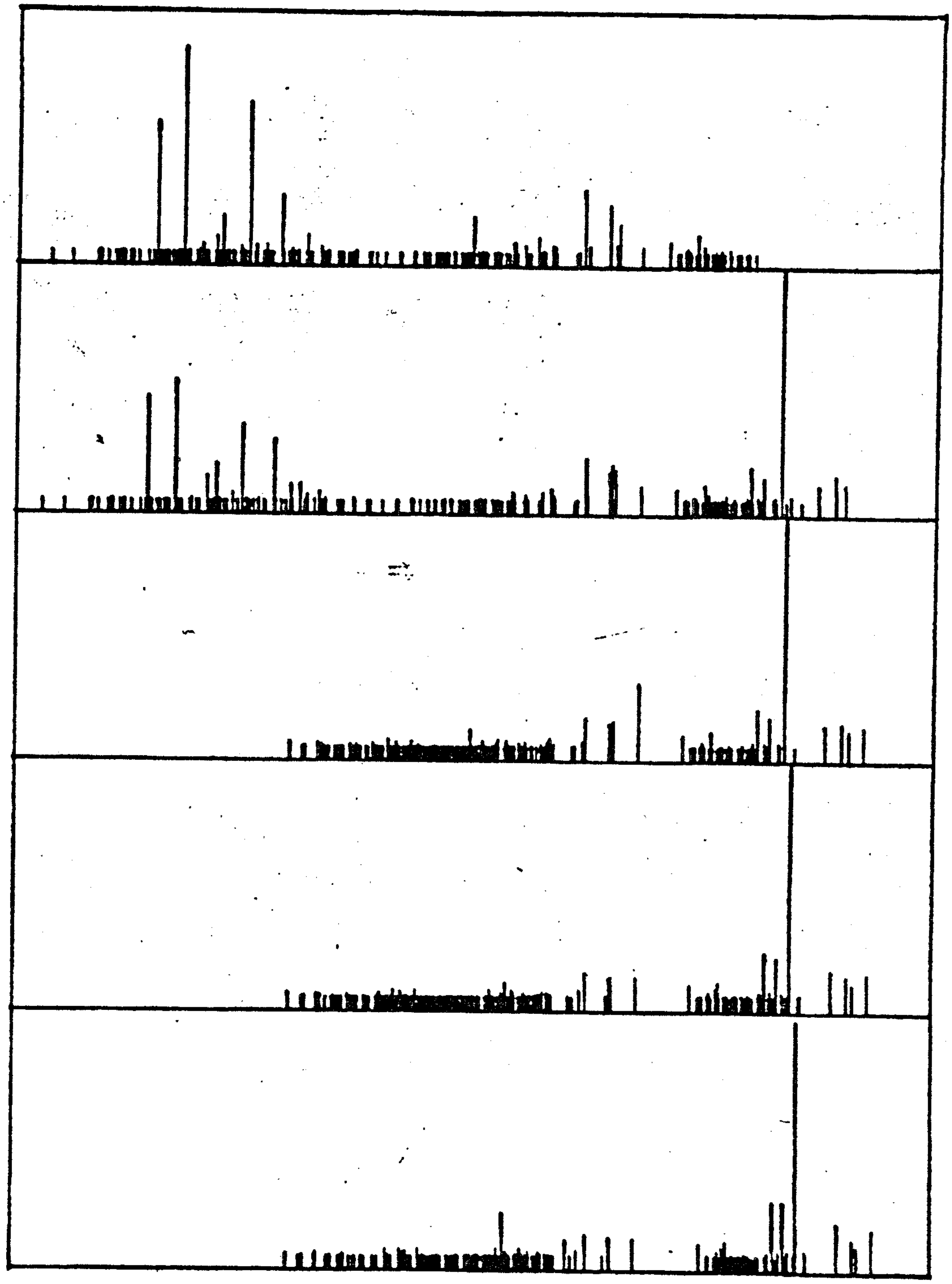
Figure 14. Ray Families for Example 5

However, as indicated by Plot #13, multiple solutions exist. The continuation method used always produces the same solutions starting from the same initial guess $\underset{\sim}{x}(0)$. No attempt is made to deflate the system to find further solutions. As illustrated by this plot, there are actually <u>three</u> solutions for the propagation type PP($\underset{\sim}{x}_I$ = (-2,0), $\underset{\sim}{x}_F$ = (4,0)).

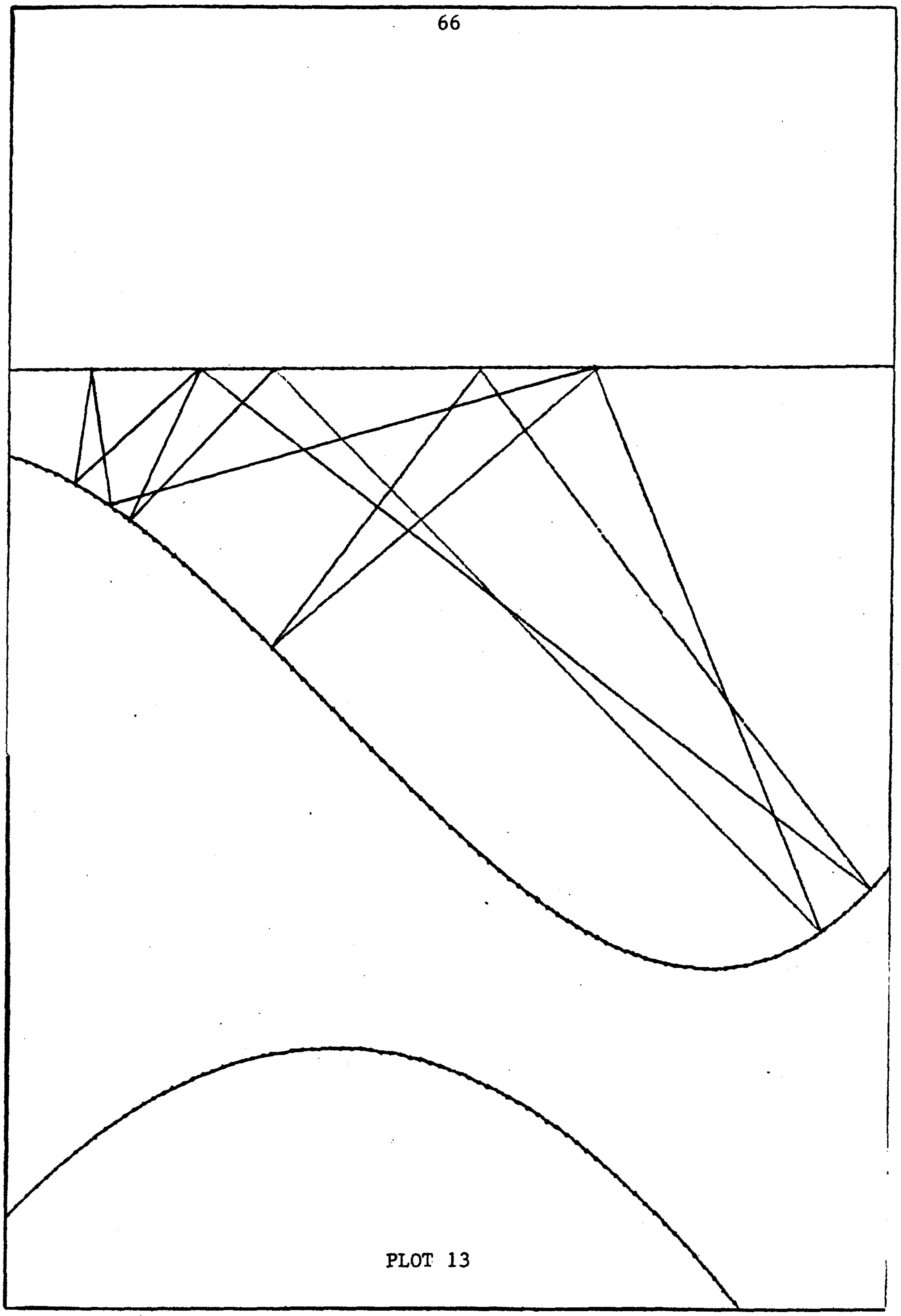
Plot 14 illustrates that some members of a family may exist while others do not. The ray of type SPP is exhibited, however there is no ray of type PPP (v_2 = 2.44, v_3 = 1.71). Since no rays of this type exist, they obviously do not appear in the gather.

One interesting aspect of this problem is that shown in Plot 15. The initial guess $\underset{\sim}{x}(0)$ always begins with the x-coordinates ordered

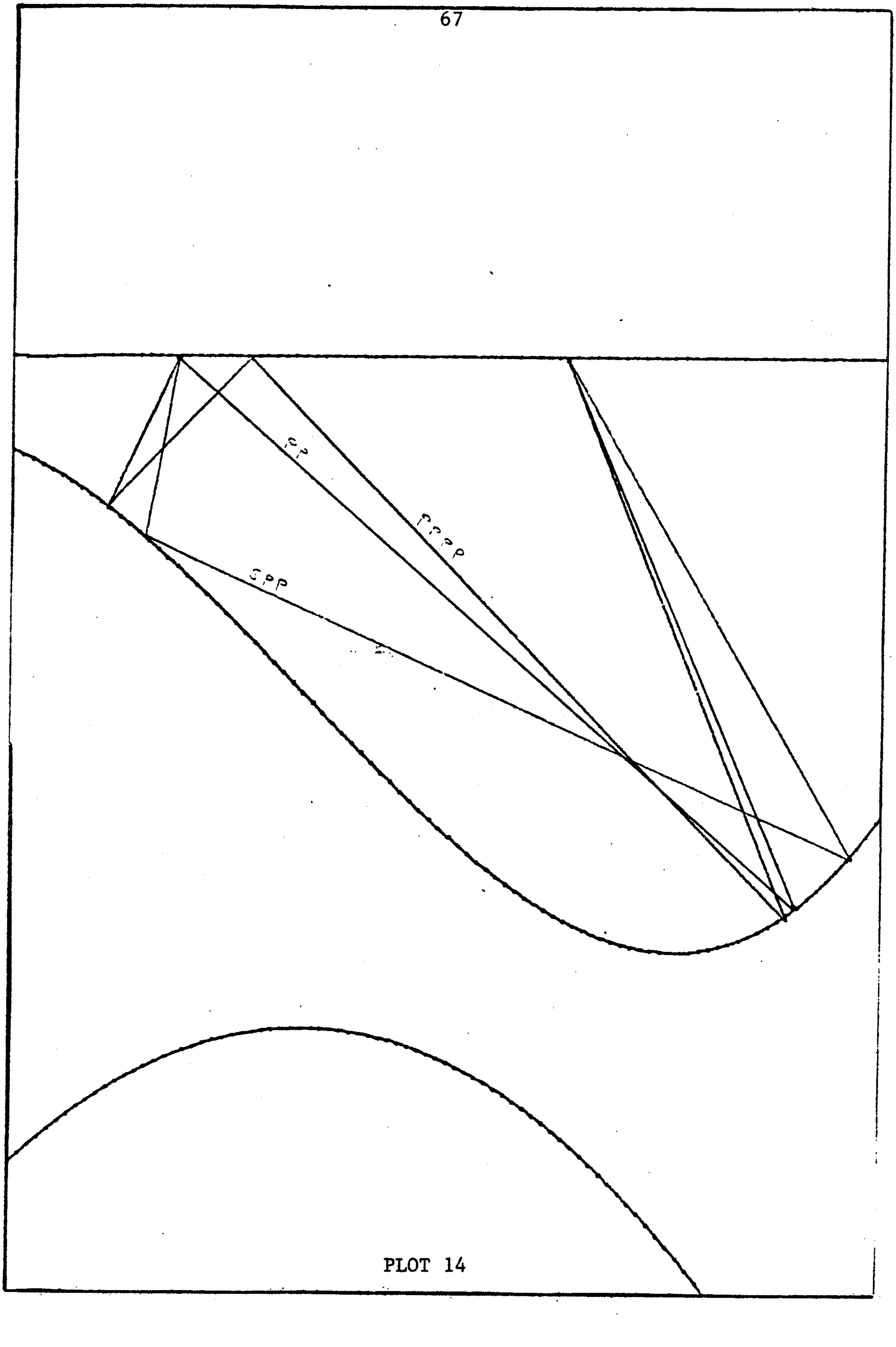
$x_1 < x_2 < x_3 < x_4 \ldots < x_N$. This means that some of the initial speeds for $\underset{\sim}{x}(0)$ must have been negative. There is no difficulty encountered due to this, and the actual solutions are exhibited in the plot.



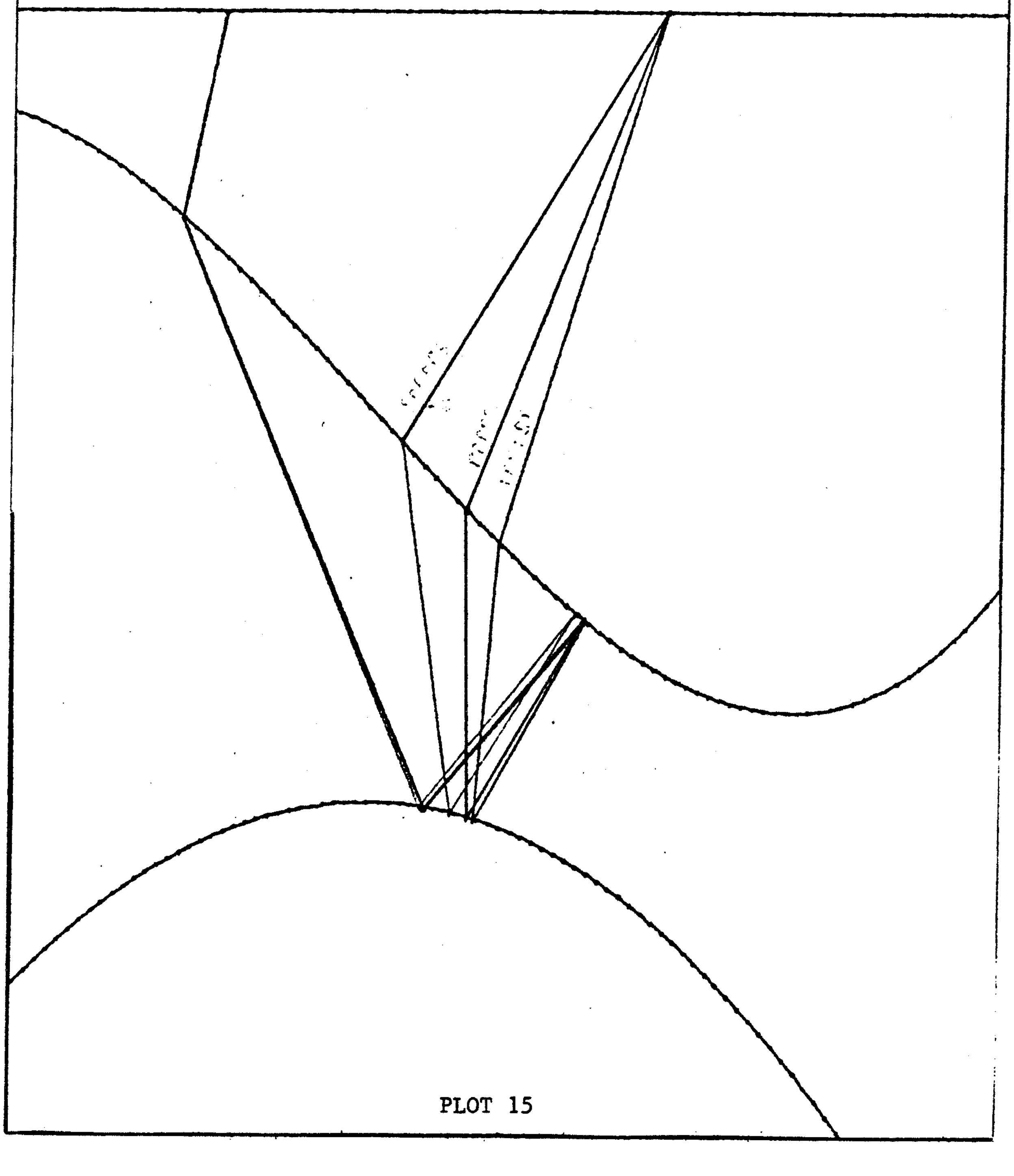
GRAPH 2



PLOT 13



PLOT 14



PLOT 15

III. THE INVERSION PROCESS

III.1. Introduction

Now that a reasonably fast and efficient method has been developed in Chapter II for solving the direct problem, we are set to attack the inversion process. Given a series of seismograms from several receivers, we seek to reconstruct the shape of the interfaces, elastic properties of each medium, and, for some problems, the location of the source, $\underset{\sim}{x}_I$. The problem, as stated, must possess non-unique solutions in general, since only the portions of the interfaces which lie on the ray paths which reach the receivers affect the data. However, given a reasonably good initial model, it will be possible to refine this by using standard nonlinear least squares techniques. The entire process should be interactive, so that a geophysicist can monitor the alteration of the model to insure a physically relevant result.

III.2. An Analytic Inversion for One Interface

One can analytically invert for one reflecting interface given a continuous distribution of first arrival times on the surface of the earth and known source location, $\underset{\sim}{x}_0$. The situation is as depicted in Figure 1.

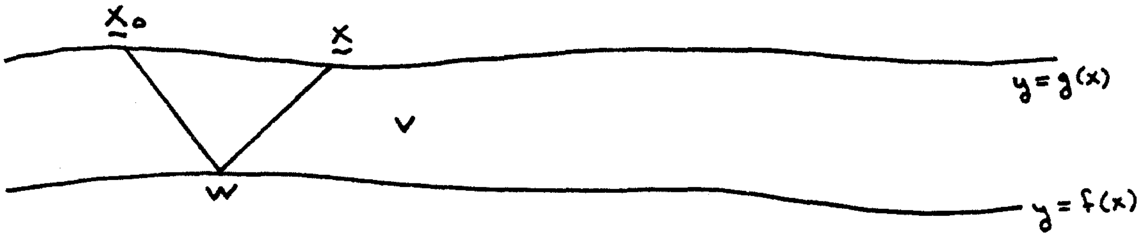


Figure 1. Inversion for One Reflecting Boundary.

We assume that $\underset{\sim}{x}_0$ is given and for all $\underset{\sim}{x}$, we know $T(\underset{\sim}{x}, \underset{\sim}{x}_0)$ where

$$T(x,x_0) = \begin{cases} \text{travel time from } \underset{\sim}{x}_0 \text{ to } \underset{\sim}{x} \text{ if PP ray} \\ \quad \text{exists for first PP arrival} \\ \\ \infty \quad \text{if PP fails to exist} \end{cases} \tag{2.1a}$$

$$y = g(x) \equiv \text{earth's surface} \tag{2.1b}$$

$$\underset{\sim}{x} = (x, g(x)) \;,\; \underset{\sim}{x}_0 = (x_0, g(x_0)) \tag{2.1c}$$

From this information, we wish to determine the shape of the reflecting interface (see Figure 1):

$$\hat{y} = f(x) \tag{2.2}$$

For ease of notation we introduce the following definitions:

$$a(z) = \tfrac{1}{2}\, v\, T\,(z,x_0) \tag{2.3a}$$

$$c(z) = \tfrac{1}{2}\,(z - x_0) \tag{2.3b}$$

Then for each z, there exists a locus of points which satisfy the travel time constraints. These correspond to all points on the ellipse, $\mathcal{M}_z$,

$$\mathcal{M}_z : \frac{\left(x - \frac{z + x_0}{2}\right)^2}{a^2(z)} + \frac{(g(x) - f(z))^2}{a^2(z) - c^2(z)} = 1 \tag{2.4}$$

To determine $f(x)$, we must eliminate the parameter z, since the solution is just the envelope of the ellipses $\mathcal{M}_z$. This is obtained by differentiating $\mathcal{M}_z$ with respect to z and setting the result equal to zero. This yields:

$$(x - \frac{z + x_0}{2})^2 \frac{v}{a^3(z)} \; T'(z, x_0) + \frac{(x - \frac{z + x_0}{2})}{a^2(z)}$$

$$+ \; \frac{(g(x))^2}{(a^2(z) - c^2(z))^2} \, (a(z)vT'(z,x_0) - c(z))$$

$$+ \; \frac{2\, f'(z) \; (g(x) - f(z))}{a^2(z) - c^2(z)} \quad = \quad 0 \qquad\qquad (2.5)$$

Thus, (2.4) and (2.5) determine f and z.

The result only applies to portions of the surface where $T(x,z) < \infty$ (see Figure 2). The region corresponding to infinite travel times cannot be recovered from the above data. In Figure 2, the portion of the boundary corresponding to AB cannot be recovered using only PP rays connecting $\underset{\sim}{x}_0$ to the surface of the earth.

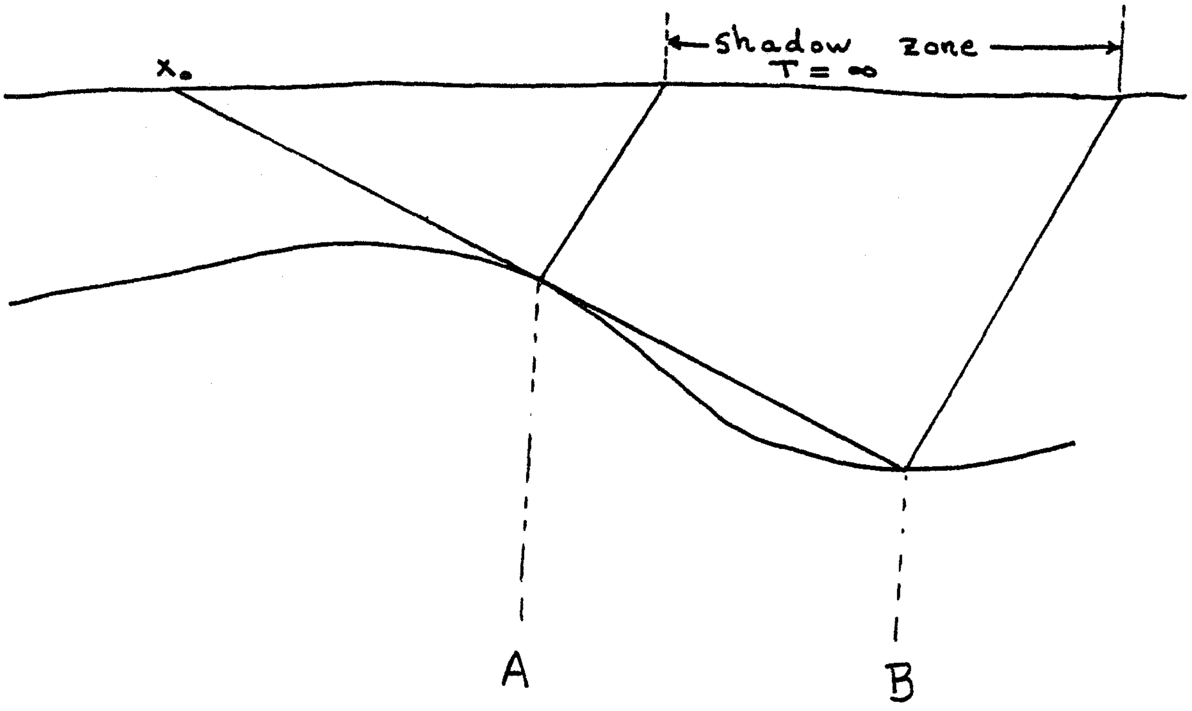


Figure 2. Non-Recoverable Portion of Interface

Also, when a region on the surface is multiply covered by PP rays, only the portion corresponding to <u>first</u> arrivals is recovered. If we wish to use this additional information (making $T(x,x_0)$ multivalued), then we can obtain the appropriate portion of the reflecting interface.

III.3. Inversion by Nonlinear Least Squares

III.3.1. Inversion Using Only Travel Times

We make the assumption that we have some algorithm to determine the type of an observed ray. (This may be included in the least squares process by allowing the redefinition of data types at some iterations, but it tends to complicate the notation. Hence we shall assume that this is automatically taken into account.) We define the observations and the computations by (j stations with k rays given at each)

$$\begin{aligned} O_{im} &= \text{observed travel time of ray } i \text{ at station } m \\ C_{im} &= \text{computed travel time of ray } i \text{ at station } m \end{aligned} \tag{3.1}$$

The function which we desire to minimize will be:

$$S = \sum_{i=1}^{k} \sum_{m=1}^{j} (O_{im} - C_{im})^2 \tag{3.2}$$

Defining $\underset{\sim}{F}$ by

$$F_{i+(m-1)k} = (O_{im} - C_{im})^2 , \quad \underset{\sim}{F} = (F_1, F_2, \ldots , F_{kj})^T \tag{3.3}$$

We may now introduce the Gauss-Newton method for the solution of the least squares problem.

$$J_n = \frac{\partial \underset{\sim}{F}}{\partial \underset{\sim}{p}} (\underset{\sim}{p}_n) \tag{3.4a}$$

$$(J_n^T J_n) [\underset{\sim}{p}_{n+1} - \underset{\sim}{p}_n] = - J_n^T \underset{\sim}{F} (\underset{\sim}{p}_n) \tag{3.4b}$$

with $\underset{\sim}{p} = (p_1, p_2, \ldots , p_N)$ the parameter vector (see Section III.3.2) and $\underset{\sim}{p}_0$ the initial guess for the true solution $\underset{\sim}{p}$.

III.3.2. Inversion Parameters

For clarity of notation, we shall confine ourselves to the case of a single receiver (i.e. $j = 1$). This inversion problem may be divided into three separate categories:

i) hypocenter locations

ii) elastic properties

iii) interface shape

Of course, any combination of i - iii may also occur.

III.3.2.i. Hypocenter Inversion

We start with the hypocenter case, since this is the simplest of the three categories. The unknowns are the two coordinates of the source location $\underset{\sim}{x}_I$. Thus, the parameter vector $\underset{\sim}{p}$ is the 2-vector representing those coordinates.

$$\underset{\sim}{p} = (x_I, y_I)^T \tag{3.5}$$

Hence, the Jacobian is simply

$$J = \begin{pmatrix} \dfrac{\partial F_i}{\partial x_I} \\ \dfrac{\partial F_i}{\partial y_I} \end{pmatrix} \tag{3.6}$$

Since the travel time can be written down explicitly from (4.1), we can obtain analytic expressions for $\dfrac{\partial c_i}{\partial x_I}$ and $\dfrac{\partial c_i}{\partial y_I}$.

$$c_i = \sum_{j=1}^{N+1} \frac{D_j^{(i)}}{v_j^{(i)}} \tag{3.7a}$$

$$\frac{\partial c_i}{\partial x_I} = \frac{1}{v_1^{(i)}} \frac{\partial D_1^{(i)}}{\partial x_I} = \frac{x_I - x_1}{D_1^{(i)}} \tag{3.7b}$$

$$\frac{\partial c_i}{\partial y_I} = \frac{1}{v_1^{(i)}} \frac{\partial D_1^{(i)}}{\partial y_I} = \frac{y_I - f_{i_1}^{(i)}(x_1)}{D_1^{(i)}} \tag{3.7c}$$

III.3.2.ii. Elastic Parameters

For the case of elastic parameters, the parameter vector $\underset{\sim}{p}$ is a 3ℓ vector, where ℓ is the number of media considered.

$$\underset{\sim}{p} = (\lambda_1, \mu_1, \rho_1, \lambda_2, \mu_2, \rho_2, \ldots, \lambda_\ell, \mu_\ell, \rho_\ell) \tag{3.8}$$

From (3.7a) and (2.9) of Chapter II we determine the appropriate derivatives:

$$\frac{\partial c_i}{\partial \lambda_n} = -\tfrac{1}{2} \sum_{k=1}^{N+1} \frac{D_k^{(i)}}{\rho_n v_n^3} \left[(n+1) \bmod 2 \quad \delta_{j_k n} \right] \tag{3.9a}$$

$$\frac{\partial c_i}{\partial \mu_n} = - \sum_{k=1}^{N+1} \frac{D_k^{(i)}}{\rho_n v_n^3} \frac{\delta_{j_k n}}{(2 - n \bmod 2)} \tag{3.9b}$$

$$\frac{\partial c_i}{\partial \rho_n} = \sum_{k=1}^{N+1} \frac{D_k^{(i)}}{2\rho_n v_n} \delta_{j_k n} \tag{3.9c}$$

where δ_{nm} is the Kronecker delta

$$\delta_{nm} = \begin{cases} 1, & n = m \\ 0, & n \neq m \end{cases} \tag{3.10}$$

III.2.iii. Interface Shapes

This third type of inversion is the most complicated. First, one must decide upon a suitable representation of the interfaces. As a first choice, we shall use simple cubic curves. It is a minor alteration to use piecewise cubic splines (or any other set of functions); however, for our simple test examples, simple cubics will suffice. One restrictive assumption which we impose is that the number of interfaces is known a priori, and that we seek only to resolve their shapes. Hence, the following structure is assumed:

$$f_{i+1} = \alpha_i x^3 + \beta_i x^2 + \gamma_i x + \varepsilon_i \qquad (3.11)$$

This leads to a $4m$ parameter vector, where m is the number of interfaces:

$$\underset{\sim}{p} = (\alpha_1, \beta_1, \gamma_1, \varepsilon_1, \ldots, \alpha_m, \beta_m, \gamma_m, \varepsilon_m) \qquad (3.12)$$

Again from (3.7a) and (2.9) of Chapter II we obtain the derivatives:

$$\frac{\partial c_i}{\partial \alpha_n} = \sum_{k=1}^{N+1} \frac{f_n(x_k) - f_{i_{k-1}}(x_{k-1})}{v_k D_k^{(i)}} x_k^3 \, \delta_{i_k n} \qquad (3.13a)$$

$$\frac{\partial c_i}{\partial \beta_n} = \sum_{k=1}^{N+1} \frac{f_n(x_k) - f_{i_{k-1}}(x_{k-1})}{v_k D_k^{(i)}} x_k^2 \, \delta_{i_k n} \qquad (3.13b)$$

$$\frac{\partial c_i}{\partial \gamma_n} = \sum_{k=1}^{N+1} \frac{f_n(x_k) - f_{i_{k-1}}(x_{k-1})}{v_k D_k^{(i)}} x_k \, \delta_{i_k n} \qquad (3.13c)$$

$$\frac{\partial c_i}{\partial \varepsilon_n} = \sum_{k=1}^{N+1} \frac{f_n(x_k) - f_{i_{k-1}}(x_{k-1})}{v_k D_k^{(i)}} \delta_{i_k n} \qquad (3.13d)$$

Note that for all cases covered in i - iii we obtain the Jacobian via:

$$\frac{\partial F_i}{\partial p_j} = -2 \left[O_i - c_i\right] \frac{\partial c_i}{\partial p_j} \qquad (3.14)$$

Thus, if we use only travel times, we can easily compute the Gauss-Newton iteration matrix explicitly. Examples of these inversion processes can be found in the Appendix.

III.3.3. Inclusion of the Amplitudes in the Inversion

If we treat the travel time and amplitude of a ray as independent quantities, the new function to be minimized is determined as follows:

$$E_{im} = \text{observed amplitude}$$

$$e_{im} = \text{computed amplitude} \qquad (3.15)$$

with the highest peak normalized to unit strength in both cases. Then we wish to minimize:

$$S = \sum_{i=1}^{k} \sum_{m=1}^{j} (E_{im} - e_{im})^2 + (O_{im} - c_{im})^2 \qquad (3.16)$$

Obtaining the derivatives of the amplitudes with respect to the parameters is a bit more complicated than the preceding travel time derivation.

This is due to the fact that a 4x4 system is solved at each interface to determine the amplitude change. The work involved in determining these derivatives exceeds that required to use a finite difference approximation (i.e. the work required to trace one additional ray per parameter). Hence we use the approximation:

$$\frac{\partial e_i}{\partial p_j} = \frac{e_i(p_j + h) - e_i(p_j)}{h} \tag{3.17}$$

The parameter h is completely at our disposal. Thus, we choose a suitably small quantity, but not so small that it is swamped by the round-off error of the computer. Typically, h is fixed somewhere between 10^{-3} and 10^{-2}. We then achieve the appropriate derivatives via:

$$\frac{\partial (E_i - e_i)^2}{\partial p_j} = 2(e_i - E_i)\frac{\partial e_i}{\partial p_j} \tag{3.18}$$

One final consideration is the appropriate weighting of the amplitude data versus the travel time data. We define

$$F_{2i - 1+2(m-1)k} = A(O_{im} - c_{im})^2$$

$$F_{2i + 2(m-1)k} = B(E_{im} - e_{im})^2 \tag{3.19}$$

and formulate the new cost function

$$\hat{S} = \sum_{i=1}^{2jk} F_i \tag{3.20}$$

One then may ask how to choose the ratio B/A. In most existing codes B/A = 0. In the current code, this is left up to the user. Presumably the user will be familiar with the reliability of the physical data, and will thus be able to make a reasonably good determination. For all examples in the Appendix, we have used either B/A = 0 or B/A = 1.

REFERENCES

1. A. Bamberger, G. Chavent, Ch. Hemon, and P. Lailly, Inversion of Normal Incidence Seismograms, 48th Annual Meeting of the Society of Exploration Geophysicists, San Francisco, California, September 13, 1978.

2. A. Bamberger, G. Chavent, and P. Lailly, About the Stability of the Inverse Problem in 1-D Wave Equations - Application to the Interpretation of Seismic Profiles, Appl. Math. Optim., 5 (1979), 1 - 47.

3. Alain Bamberger, Guy Chavent, and Patrick Lailly, Une application de la théorie contrôle a un probleme inverse de sismique, Ann. Geophys., t. 33, fasc. 1/2 (1979), 183 - 200.

4. V. Červený, I. A. Molotkov, and I. Pšenčík, Ray Method in Seismology (1977), Univerzita Karlova, Praha.

5. R. Chander, On Tracing Seismic Rays with Specified End Point, J. Geophys., 41 (1975), 173 - 177.

6. Joel N. Franklin, Matrix Theory, (1968) Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

7. David Gay, Modifying Singular Values: Existence of Solutions to Systems of Nonlinear Equations Having a Possibly Singular Jacobian Matrix, Mathematics of Computation, Volume 31, Number 140 (October 1977), 962 - 973.

8. I. M. Gelfand and S. V. Fomin, Calculus of Variations, (1963) Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

9. Philip E. Gill and Walter Murray, Algorithms for the Solution of the Nonlinear Least-Squares Problem, SIAM J. Numer. Anal., Vol. 15, No. 5 (October 1978), 977 - 992.

10. Eugene Isaacson and Herbert Bishop Keller, Analysis of Numerical Methods, (1966), John Wiley & Sons, Inc., New York.

11. Joseph B. Keller and Herbert B. Keller, Determination of Reflected and Transmitted Fields by Geometrical Optics, Journal of the Optical Society of America, Vol. 40, No. 1 (January 1950), 48 - 52.

12. Herbert B. Keller, Propagation of Stress Discontinuities in Inhomogeneous Elastic Media, SIAM Review, Vol. 6, No. 4 (October 1964), 356 - 382.

13. W. H. K. Lee and V. Pereyra, Solving Two-Point Seismic Ray-Tracing Problems in a Heterogeneous Medium. Part 2. Numerical Solutions of Two-Dimensional Velocity Models, (April 1977) (to appear).

14. V. Pereyra, H. B. Keller, and W. H. K. Lee, Computational Methods for Inverse Problems in Geophysics: Inversion of Travel Time Observations (to appear in Phys. Earth Planet. Inter., 21).

15. V. Pereyra and W. H. K. Lee, Solving Two-Point Seismic Ray-Tracing Problems in a Heterogeneous Medium. Part 1. A General Numerical Method Based on Adaptive Finite-Difference (April 1977) (to appear).

PART II

ANALYSIS OF OPTIMAL STEP SIZE SELECTION IN HOMOTOPY AND CONTINUATION METHODS

I. HISTORICAL INTRODUCTION

Perhaps the best survey of continuation methods prior to 1950 is contained in Ficken [29]. We shall begin this survey with a brief review of Ficken's exposition. He begins by dividing all continuation procedures into two basic categories: stepwise and topological. The stepwise grouping is further subdivided into set-theoretic and constructive approaches. Throughout Ficken's summary, the homotopy parameter is confined to the interval $J = [0, 1]$.

An early example of the usage of stepwise methods is due to Schwartz (1869) (detailed in Lichtenstein [49]). Given that to a point $P(0)$ in one manifold there corresponds a unique $Q(0)$ in another manifold, one wishes to show that for each $P(s)$ along a given curve there corresponds a unique $Q(s)$ for all $s \varepsilon J$. The set-theoretic method of proof consists of showing that the set $K \subset [0, 1]$ which has the desired property is both open and closed and hence must be the entire set. One drawback of this is the absence of a lower bound on the step-length. Other examples of set-theoretic arguments have been used by Schlesinger [70] (differential equations), Hadamard [34] (inversion of point-transformations in Euclidean r-space), and Lévy [45] (inversion of point-transformations in the function space L_2). Bernstein [8] applied continuation to the Dirichlet problem for the circle. Given a solution for the boundary values $\phi(\theta)$, he sought a solution with boundary values $F(\theta)$. He observed that the solution depends analytically on α for the boundary values:

$$F(\theta, \alpha) = \phi(\theta) + \alpha[F(\theta) - \phi(\theta)] \qquad (1.1)$$

which allowed him to apply continuation. He included the notion of an a priori bound and also advocated the use of a uniform steplength.

Constructive proofs have been employed by Schauder [68] (differential equations), Weinstein [83] (conformal mapping), Weyl [84], and Lewy [46], [47] (differential geometry).

The use of topological methods is best exemplified by Leray and Schauder [44] (functional equations). Topological index is employed with a priori bounds and complete continuity to obtain a pure existence theorem for a solution in [0, 1]. Additional details may be found in Leray and Schauder [44], and applications are presented in Leray [43], Dolph [25], and Cronin [17]. Theorems on uniqueness were developed by Rothe [63], [64].

The first notable use of *numerical* continuation was due to Lahaye [40], [41] (1934-35) for a single equation and [42] (1948) for systems of equations. There have been numerous authors who have employed numerical continuation. Others of note who will not be discussed below include Sidlovskaya [74] (1958), Anselone and Moore [5] (1966), and Deist and Sefor [22] (1967).

In 1953, Davidenko [18], [19] introduced the idea of examining the differential equation underlying a related homotopy. For example, suppose we wish to solve:

$$F(\underset{\sim}{x}) = \underset{\sim}{0} \tag{1.2}$$

We introduce the homotopy

$$F(\underset{\sim}{x}) - e^{-\lambda} F(\underset{\sim}{x}^{\circ}) = \underset{\sim}{0} \tag{1.3}$$

Note that for $\lambda = 0$, $\underset{\sim}{x} = \underset{\sim}{x}^{\circ}$ is a solution of (1.3). By differentiating (1.3) with respect to λ, we obtain the initial value problem

$$\frac{\partial F}{\partial \underset{\sim}{x}} \frac{d\underset{\sim}{x}}{\partial \lambda} + e^{-\lambda} F(\underset{\sim}{x}^{\circ}) = \underset{\sim}{0} , \qquad \underset{\sim}{x}(0) = \underset{\sim}{x}^{\circ} \tag{1.4}$$

Then using (1.3) in (1.4) results in

$$\frac{\partial F}{\partial \underset{\sim}{x}} \frac{d\underset{\sim}{x}}{d\lambda} + F(\underset{\sim}{x}) = \underset{\sim}{0} , \qquad \underset{\sim}{x}(0) = \underset{\sim}{x}^{\circ} \tag{1.5}$$

This is referred to as the Davidenko differential equation underlying the homotopy defined in (1.3). A similar differential equation may be obtained for any homotopy. (An exposition of Davidenko's work is contained in Rall [57].) Others who have examined this technique include Bittner [9], Kleinmichel [39], and Bosarge [10]. The extension to Banach spaces has been studied by Yakovlev [86], Meyer [52], and Avila [6].

Considerable effort has been expended on the application of continuation methods to obtain Brouwer fixed points. A good survey of this work is contained in Alexander and Yorke [3]. Scarf [66] first used the idea of "following a path" from the boundary to the fixed point. Eaves [26] uses a standard set of maps, homotopes to one of these, and follows the fixed points of the changing maps. This was jointly refined by Eaves and Scarf in [27]. A similar method was used on a 20-dimensional fixed point problem by Kellogg, Li, and Yorke [38]. Eaves' homotopy approach was made rigorous by Yorke, who then applied it to a variety of problems (see Chow, Mallet-Paret, and Yorke [16]). Although Scarf used simplicial methods, he did not introduce an extra parameter. Eaves initiated the homotopy approach, but he employed Sperner simplices. The resulting method of Kellogg et al [38], has since been superceded by a variety of superior algorithms (Li [48], Smale [75], Alexander and

Yorke [3], Chow et al. [16]). These methods were then adapted for use on two-point boundary-value problems. (Alexander [1], Alexander and Yorke [2], Chow et al. [16], Peitgen and Prüfer [55])

Before expanding on a few notable papers, let us first refer the reader to some general references on continuation methods. Extensive discussion of these techniques may be found in Ortega and Rheinboldt [53] and Wacker [80]. For theoretical details of homotopies see Eaves [28], Eaves and Scarf [27], Todd [78], or Lüthi [50]. Conference proceedings on continuation methods are presented in Karamardian [35], Wacker [80], and Peitgen [54]. Two other shorter specialized papers worthy of mention are Saigal [65] and Gould and Tolle [32].

In 1967, Roberts and Shipman [62] applied continuation methods to two-pointboundary value problems (TPBVP). The continuation parameter employed was the location of one boundary point. Given the TPBVP

$$\begin{aligned} &\text{a)} \quad \dot{x} = F(x, t) \\ &\text{b)} \quad g\big(x(a)\big) + h\big(x(b)\big) = c \end{aligned} \tag{1.6}$$

the following homotopy was used

$$\begin{aligned} &\text{a)} \quad \dot{x} = F(x, t) \\ &\text{b)} \quad g\big(x(a)\big) + h\big(x(\lambda)\big) = c \end{aligned} \tag{1.7}$$

where $\lambda \varepsilon (a, b]$. The technique was to solve initially for $\lambda - a$ small and then continue. The interval $[a, b]$ was normalized to $[0, 1]$. In regard to step size the statement is made, "No general rule is applicable at this time". The suggestion was to use a "modest" step size $(\Delta\lambda \leq .1)$. The advantage of this technique for solving TPBVP's is that the same differencing scheme may be used at each step. Two major problems are encountered. First, $\Delta\lambda$ may tend to zero before $\lambda = 1$ is reached. Second, how does one choose the initial $\Delta\lambda$ so that the solution does not blow up?

Thurston [77] in 1969 proposed a method for continuing Newton's method through limit points and bifurcation points. The method only applies when the linear "variational equations" are self-adjoint, and seems to be very heuristic. The nonlinear terms are expanded about any good approximate solution, but near critical points, quadratic terms must be retained. No theoretical results are given, but some impressive computations are displayed.

Bosarge [11] (1970) examines the Davidenko differential equation (1.5). Conditions are given on F to ensure the existence of a solution. (This is essentially a rewording of the Kantorovich criteria for the convergence of Newton's method.) Bounds are required on the first and second derivatives of F. If bounds may be obtained for $q + 1$ derivatives of F, Bosarge shows that the relaxed Newton method works and derives a bound for the relaxation constant h_q. It should be noted that

$$\lim_{q\to\infty} h_q = 1 \qquad (1.8)$$

He proves Relaxed Newton (Euler-Newton) will work for $q = 1$ with h_1 where this method is defined by

$$\begin{aligned} &a)\quad x_{n+1} = x_n - h_1 F'^{-1}(x_n)F(x_\circ) \quad , \quad n = 0,\dots, N_1 - 1 \\ &b)\quad x_{n+1} = x_n - F'^{-1}(x_n)F(x_n) \ , \quad n = N_1,\dots \qquad (1.9) \\ &c)\quad h_1 = \frac{1}{N_1} \end{aligned}$$

Also for $q = 2$, the Trapezoidal Newton scheme works with conditions on h.

$$\begin{aligned} &\text{a)}\quad y_n = x_n - \tfrac{1}{2}h_2 F'^{-1}(x_n)F(x_\circ) \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad n = 0, 1, \ldots N_2 - 1 \\ &\text{b)}\quad x_{n+1} = x_n - h_2 F'^{-1}(y_n)F(x_\circ) \qquad\qquad\qquad\qquad\qquad (1.10) \\ &\text{c)}\quad x_{n+1} = x_n - F'^{-1}(x_n)F(x_n) \quad , \quad n = N_2 \quad \ldots \\ &\text{d)}\quad h_2 = \frac{1}{N_2} \end{aligned}$$

The basic idea here is to use a Relaxed Newton iteration until the region of convergence for Newton's method is entered. One then uses regular Newton iterations. The results of this paper are applied to TPBVP's by Bosarge and Smith [12].

Wasserstrom [82] (1973) gives a nice illustration of the power of continuation methods. Several simple examples of applications are exhibited: polynomials, TPBVP's , parameter identification, eigenvalue problems. A brief historical section is included; however, no new developments are presented.

The question of the feasibility of numerical continuation was addressed by Avila [7] (1974) in an extension of his doctoral dissertation. Theorems on feasibility are developed for initial guess (4.11a) of Section II. Basically, conditions are presented so that the iterative method will converge for some positive step length $\Delta\lambda$. Only uniform steps are considered.

Deuflhard, Pesch, and Rentrop [24] applied continuation to parallel shooting (1976). The method of choosing the initial guess is that of

(4.1) a) of Section II, $x^{\circ}(\lambda_{i+1}) = x(\lambda_i)$. The TPBVP

$$y' = f(x, y, \tau) \quad , \quad X \varepsilon [a, b] \qquad (1.11)$$

$$+ \text{ Boundary Conditions}$$

is replaced by

$$\begin{aligned} & y' = f(x, y, \tau) \quad + \text{boundary conditions} \\ & \tau' = 0 \\ & h(\tau(b)\ ) = 0 \end{aligned} \qquad (1.12)$$

with the conditions on h:

$$\begin{aligned} & h(a) = 0 \\ & h'(\tau) \neq 0 \end{aligned} \qquad (1.13)$$

$$(h > 0 \text{ and } h \text{ convex}) \text{ or } (h < 0 \text{ and } h \text{ concave}).$$

Initial conditions for (1.12) are then guessed as

$$y(x_j) = s_j{}^{\circ} \qquad (1.14)$$

The following update scheme is then used

$$\begin{aligned} s^{k+1} &= s^k + \hat{\lambda}_k\ (\Delta s^k + \Delta\tau^k\ \widehat{\Delta s^k}) \quad , \quad 0 < \hat{\lambda}_k \leq 1 \\ \tau^{k+1} &= \tau^k + \hat{\lambda}_k\ \ \Delta\tau^k \\ \Delta s^k &= -J(s^k\ , \tau^k)^{-1}\ F(s^k\ , \tau^k) \\ \widehat{\Delta s}^k &= -J(s^k\ , \tau^k)^{-1}\ F_\tau(s^k\ , \tau^k) \\ \Delta\tau^k &= -\ \frac{h(\tau^k)}{h'(\tau^k)} \end{aligned} \qquad (1.15)$$

where F is the discretized version of (1.12) and $J = \frac{\partial F}{\partial \underset{\sim}{x}}$. $\hat{\lambda}_k$ is a relaxation factor for Newton's method. This may be extended if f is a function of more parameters, say $f(x, y, \tau_1, \tau_2, \tau_3, \ldots)$, by the update scheme

$$\begin{aligned} \widehat{\Delta s^k_{\tau_i}} &= -J^{-1} F_{\tau_i}(s, \tau_1, \tau_2, \ldots) \\ \tau_i^{k+1} &= \tau_i^k + \hat{\lambda}_k \Delta\tau_i^k \\ s^{k+1} &= s^k + \hat{\lambda}_k \Big(\Delta s^k + \sum_i \Delta\tau_i^k \, \widehat{\Delta s_{\tau_i}^k}\Big) \end{aligned} \tag{1.16}$$

Deuflhard [23] (1977) developed estimates for step size in terms of only local quantities. For the guesses of (4.1)a) and (4.1)b) of Section II, a ratio of the maximum feasible step size, $\Delta\beta$, is obtained. The homotopy used is

$$F(x) - (1 - \beta)F(x_\circ) = 0 \tag{1.17}$$

The bound is given by

$$\frac{\Delta\beta^b_{\max}}{\Delta\beta^a_{\max}} = \sqrt{\frac{2}{\sqrt{2} - 1}} \tag{1.18}$$

independent of the function F. For the relaxed Newton method, he also obtains the value of λ_k which minimizes $\frac{\|x^{k+1}\|}{\|\Delta x^k\|}$ where

$$\begin{aligned} x^{k+1} &= x^k + \lambda_k \Delta x_k \\ \Delta x^k &= -J(x^k)^{-1} F(x^k) \end{aligned} \tag{1.19}$$

and λ_k is the relaxation constant. (The optimal value depends on the constants in the Kantorovich theorem.) The case of a rank-deficient Jacobian is also considered. In general, Deuflhard seeks to maximize the size of the step taken, without regard to the work required.

Schmidt [72] in 1978 studied adaptive step size selection. He determined that there were four major considerations in determining the step size: (1) maximum and minimum step sizes, (2) a simple rule for determining the step size, (3) ability to recognize non-convergence, and (4) ability to recognize convergence. With $\Delta\lambda^{i+1} = \lambda^{i+1} - \lambda^{i}$, Schmidt suggests using the method:

$$\begin{aligned} &\text{a)} \quad \Delta\lambda^{i+1} = f\Delta\lambda^{i} \\ &\text{b)} \quad f = a + b\,\frac{(1-\alpha)}{\alpha} \end{aligned} \tag{1.20}$$

where a and b are fixed constants, and α is an estimate of the convergence constant for the iterative method which is given by:

$$\begin{aligned} &\text{a)} \quad \beta_k = \frac{\|\underset{\sim}{x}^{k+1} - \underset{\sim}{x}^{k}\|}{\|\underset{\sim}{x}^{k} - \underset{\sim}{x}^{k-1}\|} \\ &\text{b)} \quad \alpha = \max_{k} \beta_k \end{aligned} \tag{1.21}$$

This is applied to the problem

$$H(\underset{\sim}{x}, \lambda) \equiv K(\underset{\sim}{x})\underset{\sim}{x} - \lambda P = 0 \tag{1.22}$$

The iteration process is deemedto be non-convergent if the following three conditions all hold:

$$\begin{aligned} &\text{a)} \quad \beta_k > 1 \\ &\text{b)} \quad \|H(\underset{\sim}{x}^{k}, \lambda^{i})\| - \|H(\underset{\sim}{x}^{k-1}, \lambda^{i})\| \geq A\,\lambda^{i}\,\|P\| \\ &\text{c)} \quad \frac{\|\underset{\sim}{x}^{k} - \underset{\sim}{x}^{k-1}\|}{\max\limits_{j}\,(|x_j^{\,k-1}|)} > 1 \end{aligned} \tag{1.23}$$

where A essentially represents the maximum percentage increase allowable in the residual with respect to P. As a test for convergence the following criteria is employed:

$$\frac{\alpha}{1 - \alpha} \; ||\underset{\sim}{x}^{k} - \underset{\sim}{x}^{k-1}|| \quad < \quad \text{specified error tolerance} \qquad (1.24)$$

An attempt to minimize the computational effort in solving the numerical continuation was presented by Wacker [80] (1978). He analyzed the Newton initial guess ((4.1 a) Section II) and the Modified Newton initial guess ((4.1)b) Section II). However, he restricts himself to uniform step sizes. With this restriction, he develops theorems on a minimum step size which will guarantee convergence at every step.

Alexander and Yorke [3] (1978) developed an algebraic topological condition that guarantees that the continuation method will work. Many areas of application are cited, but the method is not applied to any in this paper.

Keller [36] (1978) introduces the idea of using pseudo-arclength as a new parameter. Given a problem

$$G(\underset{\sim}{x}, \lambda) = 0 \qquad (1.25)$$

both $\underset{\sim}{x}$ and λ are allowed to depend on the parameter s. The Davidenko differential equation is then solved with the additional condition that a normalization condition is satisfied (which mimics the condition that s be arclength, hence the notation pseudo-arclength).

$$\text{a)} \quad \frac{dG}{ds} \; (\underset{\sim}{x} \; (s), \lambda(s) \;\;) = \frac{\partial G}{\partial \underset{\sim}{x}} \; \frac{d\underset{\sim}{x}}{ds} + \frac{\partial G}{\partial \lambda} \; \frac{d\lambda}{ds} = 0$$

$$(1.26)$$

$$\text{b)} \quad N(\frac{d\underset{\sim}{x}}{ds} , \frac{d\lambda}{ds}) \;\; = 1$$

where N is the normalization. The major advantage to this is the fact that limit points in the parameter λ are regular points under the s-parametrization (i.e., s is always increasing along the solution curve). A method for estimating a Δs so that iteration will converge is presented based on estimates of the derivatives of G and the condition number of the iteration matrix.

Rheinboldt [61] (1979) examines the method of obtaining an initial guess. He suggests using a quadratic approximation:

$$\underset{\sim}{x}^{\circ}(\lambda_{\circ} + \Delta\lambda) = \underset{\sim}{x}(\lambda_{\circ}) + \Delta\lambda\frac{d\underset{\sim}{x}}{d\lambda}(\lambda_{\circ}) + \tfrac{1}{2}(\Delta\lambda)^2 \frac{d^2\underset{\sim}{x}}{d\lambda^2}(\lambda_{\circ}) \qquad (1.27)$$

The radius of convergence is estimated by extrapolating the estimates for previous values of λ. The intersection of this extrapolated curve and the quadratic in (1.27) determines the maximum allowable size of $\Delta\lambda$. No discussion is given as to how to obtain an estimate for the radius of convergence at any fixed λ value.

Schmidt [71] (1979) advocates using an approximation to the Davidenko differential equation, which results in considerable computational savings.

II. CRITERIA FOR STEP SIZE SELECTION IN THE HOMOTOPY AND CONTINUATION METHODS

II.1 Introduction

In the following sections, our purpose will be to examine the optimization of the continuation procedure. In this formulation, we desire to minimize the amount of work required over some class of methods which is assumed to be at our disposal. We shall examine several of the simplest methods in this context. The preceding chapter has given an historical overview of previous work devoted to this subject. We shall attempt to expand upon some of the procedures mentioned there, and, perhaps, unify some of these methods. We shall begin with a general formulation of the homotopy problem in Section II.2. The problem is formulated in such generality that we have no hope of obtaining a solution. It is then shown that the problem may be reformulated as an optimal control problem. Again this formulation provides no method of obtaining a solution, but does allow the introduction of the concept of feasibility. Simple theorems which parallel those of Avila [7] are presented to determine the feasibility of numerical continuation.

In Section II.3, the problem is dissected into its component parts. Also, work variables are introduced as a measure of the effort expended. The actual weighting assigned to each of these variables is dependent on both the user and the physical machinery involved.

Section II.4 addresses the topics of choice of initial guess for $\underset{\sim}{x}(\lambda + \Delta\lambda)$ and the iterative method used to converge for a fixed value of λ. It is found that high order interpolatory methods are far superior to lower order methods for small $\Delta\lambda$. It is thus suggested that these methods be used in lieu of the standard constant and linear approximations. In terms of storage requirements, the opposite is the case. High order methods require more storage than low order methods. Thus, one concludes that a high order method should be used, but not so high that the storage capacity of the given machine is surpassed.

For iterative methods, the standard Newton and chord iterations are examined along with several variants. For large systems, it is shown that one should use chord iterations until sufficiently close and only switch to Newton iterations when they are truly quadratically convergent. Two examples are presented combining the iterative methods with initial guesses, illustrating that the higher order approximations are more efficient. Also, the various iterative methods are compared for several model problems. The results are somewhat mixed for small scale problems, but the result stated above for large scale systems tends to be borne out.

The continuation method is reformulated in terms of the "power" expended; i.e., the work per unit parameter step. For this formulation with fixed initial guess and iterative method, it is possible to obtain a value for $\Delta\lambda$ which minimizes an upper bound on the power. This is given in terms of convergence constants and bounds on derivatives in Theorems 5.1, 5.2, and 5.3. Since the convergence constants usually involve a bound on the inverse of a matrix, a computational method for determining such a bound is developed via Theorem 5.4 and Corollaries 5.5 and 5.6

Section II.6 presents two computational procedures based on the theorems of Section II.5. Computational Scheme I requires that a dependence of the convergence constant on the error be specified. It is then possible to apply the theorems of Section II.5 directly to determine an optimal value of the convergence constant. One then interpolates the values of $\Delta\lambda$ to obtain the desired rate. Based on the conclusions of Section II.4, a particular chord variant iteration is proposed. Computational Scheme II employs extrapolatory techniques to obtain estimates for the radius of convergence, and extends the work of Rheinboldt [61] to higher order initial guesses.

Finally in Section II.7, the question of intermediate error tolerances is addressed. For uniform step lengths, it is shown that uniform error tolerances optimize the work.

This is merely a first attempt at examining the continuation method in its greatest generality. No tremendously startling results are obtained, although a basis for guidelines is developed. Hopefully, this may be expanded to obtain more computationally implementable results.

II.2 The Homotopy Problem

II.2.1 General Information

We shall make a distinction between a homotopy problem and a continuation problem. By the homotopy problem, we shall mean a problem for which only one value of the parameter is of physical interest [i.e., the parameter has been artificially introduced]. Continuation shall refer to a problem in which there is a naturally occurring parameter, and hence all values of the parameter are significant [e.g., Reynolds number].

Let us assume we are given a homotopy problem with parameter λ .

$$G(\underset{\sim}{x}, \lambda) = 0 \tag{2.1}$$

We shall assume that for $\lambda = \lambda_\circ$ we know a solution, say $\underset{\sim}{x}^\circ$, and desire a solution for $\lambda = \lambda_1$. Define the sequence of sets P_i as follows:

$$\begin{aligned} &P_\circ = \{[\lambda_\circ, \lambda_1]\} \\ &P_n\ (t_1, \ldots, t_{n-1}) = \{[t_\circ, t_1], [t_1, t_2], \ldots, [t_{n-1}, t_n]\} \\ &t_\circ = \lambda_\circ, t_n = \lambda_i \\ &t_i < t_{i+1} \end{aligned} \tag{2.2}$$

Note that P_n is a partition of the interval $[\lambda_\circ, \lambda_1]$ into $n+1$ pieces. We now define the set of all partitions into n subintervals by:

$$P_{n+1} = \{ P_n(t_1, \ldots, t_{n-1}) |\ \ t_i \varepsilon\ (\lambda_\circ, \lambda_1)\} \tag{2.3}$$

Finally we form the set of all partitions of $[\lambda_\circ, \lambda_1]$.

$$Q = \bigcup_{n=1}^{\infty} P_n \tag{2.4}$$

It is over this set of partitions which we seek to minimize work. We assume that we have at our disposal some set of methods available for obtaining a solution at t_{i+1}, given a solution at t_i. Thus we define

$$\eta_i \in \mathcal{N} \equiv \{\text{all available methods for obtaining } \underset{\sim}{x}^{i+1} \text{ given } \underset{\sim}{x}^{i}\}. \tag{2.5}$$

We must also introduce the concept of the work involved in producing $\underset{\sim}{x}^i$. For the purposes of this section we shall leave this fairly vague. This is actually a subjective quantity. The work may differ depending on the economics employed by the problem-solver. The following section shall pin down this concept for the class of methods which we will study. Hence, we define

$$W_{\eta_i}(z_1, z_2) = \text{work required to go from } z_1 \text{ to } z_2 \text{ in one step using method } \eta_i. \tag{2.6}$$

For any given partition of $q \varepsilon Q$, we then define

$$W_{\underset{\sim}{\eta}}(q) = \sum_{i=0}^{|q|-1} W_{\eta_i}(t_i, t_{i+1}) \tag{2.7}$$

where $\underset{\sim}{\eta} = (\eta_o, \ldots, \eta_{|q|-1})$, allowing for a change of methods from one step to the next.

With the above notation (2.2) - (2.7), we can now state the homotopy problem in the broadest generality. Assume we are given (2.1) with λ_1 fixed. Then we wish to find $q \varepsilon Q$ and $\underset{\sim}{\eta} \varepsilon \mathcal{N}^{|q|}$ such that

$$W^* = \min_{\eta} \min_{q} W_{\underset{\sim}{\eta}}(q) \tag{2.8}$$

It is abundantly clear that this problem cannot be solved for $\underset{\sim}{\eta}$ and q in this generality. Therefore, we shall examine it from various perspectives and attempt to analyze its component parts, hopefully lending insight on reasonable ways to proceed.

II.2.2 Formulation as an Optimal Control Problem for Fixed Iterative Methods

If we fix the method η_i which is used to be the same for each i, then it is possible to state the problem in the context of nonlinear optimization. Assume we seek a solution of (2.1) for $\lambda = \lambda_1$. Since we have fixed the method, say $\eta_i = \hat{\eta}$, then we can make the following definitions:

$r(t_i)$ = radius of convergence of the method $\hat{\eta}$ at $\lambda = t_i$

$\varepsilon(t_i, t_{i+1})$ = error incurred at t_{i+1} by using initial guess defined by method $\hat{\eta}$. (2.9)

We, of course, do not know either of these functions. Hence the following formulation is only of theoretic interest. However, it does allow us to make rigorous a few statements which should be fairly obvious. Since the theory does not allow for variable numbers of unknowns, we fix the number of steps to be taken as large, and assign zero work to steps which do not take us outside a predetermined error tolerance. A standard theorem of optimal control theory (see McCormick [51]) is adapted for use on this problem. First we state some preliminaries.

The nonlinear programming problem is stated as

$$\begin{aligned} &\text{minimize } f(\underset{\sim}{x}) \\ &\quad \text{subject to} \\ &\qquad g_i(\underset{\sim}{x}) \geq 0 \qquad \text{for } i = 1, \dots, m \\ &\qquad h_j(\underset{\sim}{x}) = 0 \qquad \text{for } j = 1, \dots, p \end{aligned} \qquad (2.10)$$

The generalized Lagrangian is then defined by

$$\mathcal{L}(\underset{\sim}{x}, u_\circ, u_1, \ldots, u_m, w_1, \ldots, w_p) = u_\circ f(\underset{\sim}{x}) - \sum u_i g_i(\underset{\sim}{x}) + \sum w_j h_j(\underset{\sim}{x}) \tag{2.11}$$

The theorem is then as follows:

<u>Theorem 2.1</u>. Assume f, g, and h are continuously differentiable. Necessary conditions that a point $\bar{\underset{\sim}{x}}$ be a local minimizer to Problem (2.10) are that there exist multiplier values

$$(\bar{u}_\circ, \bar{u}_1, \ldots, \bar{u}_m, \bar{w}_1, \ldots, \bar{w}_p)$$

(<u>not</u> <u>all</u> equal to zero) such that

$$\begin{aligned}
&\text{a)} \quad g_i(\bar{\underset{\sim}{x}}) \geq 0 \quad , \quad i = 1, \ldots, m \\
&\text{b)} \quad h_j(\bar{\underset{\sim}{x}}) = 0 \quad , \quad j = 1, \ldots, p \\
&\text{c)} \quad \bar{u}_i \geq 0 \quad , \quad i = 0, 1, \ldots, m \\
&\text{d)} \quad \bar{u}_i g_i(\bar{\underset{\sim}{x}}) = 0 \quad , \quad i = 1, \ldots, m \\
&\text{e)} \quad \bar{\nabla}\mathcal{L}(\bar{\underset{\sim}{x}}, \bar{u}_\circ, \ldots, \bar{u}_m, \bar{w}_1, \ldots, \bar{w}_p) = \underset{\sim}{0}
\end{aligned} \tag{2.12}$$

where ∇ is the gradient with respect to $\underset{\sim}{x}$.

If we reformulate the problem in this context, we find the constraints are

$$r(t_{i+1}) - \varepsilon(t_i, t_{i+1}) \geq 0, \quad t_{i+1} - t_i \geq 0 \tag{2.13}$$

(i.e., we are within the radius of convergence).

Hence the Lagrangian is given by

$$\mathcal{L} = u_\circ W_{\underset{\sim}{\hat{\eta}}}(\underset{\sim}{t}) - \Big(\sum u_i\big(r(t_i) - \varepsilon(t_i, t_{i+1})\big) + \sum u_{n+i}(t_{i+1} - t_i)\Big) \tag{2.14}$$

Using (2.13) and (2.14) in (2.12) we obtain a necessary condition for $\underset{\sim}{t} = (t_\circ, t_1, \ldots, t_{n+1})$ to be optimal.

$$\left.\begin{array}{ll} \text{a)} & t_{i+1} - t_i \geq 0 \\ \text{b)} & r(t_i) - \varepsilon(t_i, t_{i+1}) \geq 0 \end{array}\right\} \quad i = 0, \ldots, n$$

$$\text{c)} \quad u_j \geq 0 \quad , \quad j = 0, \ldots 2n$$

$$\left.\begin{array}{ll} \text{d)} & u_i(r(t_{i+1}) - \varepsilon(t_i, t_{i+1})\ ) = 0 \\ \text{e)} & u_{n+i}(t_{i+1} - t_i) = 0 \end{array}\right\} \quad i = 0, \ldots, n \qquad (2.15)$$

$$\text{f)} \quad t_o = \lambda_1, \quad t_{n+1} = \lambda_1$$

$$\text{g)} \quad u_0 \nabla W_{\hat{\eta}} - \sum_{i=1}^{n} u_i \nabla\big(r\ (t_i) - \varepsilon(t_i, t_{i+1})\big) - \sum_{i=1}^{n} u_{n+i} \nabla\,(t_{i+1} - t_i) = 0$$

Let us examine the significance of (2.15)g. This says if the work is increasing at some point t_i, then one would want the distance from the path to be decreasing, implying one should take smaller steps. Alternatively, if the work is decreasing, then the distance from the path should increase, implying that larger steps are required for efficiency. In the present generality, it is impossible to obtain useful qualitative results for use in actual computation. With this in mind, we proceed to study some fixed methods with more constraints applied in the following sections.

Some very trivial theorems follow directly from Theorem 2.1. Let us first introduce the idea of a feasible partition (cf. Avila [71]). If the initial guess determined by $\hat{\eta}$ at t_i produces a convergent sequence for the iterative method specified by $\hat{\eta}$, then the step $[t_{i-1}, t_i]$ is feasible. If $[t_{i-1}, t_i]$ is feasible for all $i = 1, \ldots, n$, then the partition $P_n(t_1, \ldots, t_{n-1})$ is called feasible.

Theorem 2.2. Let $I = [\lambda_\circ, \lambda_1]$ and $r\,(\tau) > 0$, for all $\tau \varepsilon I$.

If $\{t_i\}$ satisfies (2.15) b) and (2.15) f), then the partition $P_n(t_1, \ldots, t_{n-1})$ is feasible X

The preceding theorem assumes that $\underset{\sim}{x}(t_i)$ has been calculated exactly. In reality, this is impossible, hence we must take round-off errors into consideration. Consider the method of formulating an initial guess given by:

$$\underset{\sim}{x}^\circ(t_{i+1}) = \underset{\sim}{x}^F(t_i) \tag{2.16}$$

where $\underset{\sim}{x}^\circ \equiv$ initial guess, $\underset{\sim}{x}^F \equiv$ accepted value for $\bar{\underset{\sim}{x}}(t_i)$.

We shall assume a tolerance has been specified at t_i, say $\delta(t_i)$. $\underset{\sim}{x}^F$ is accepted as a solution if

$$\| \underset{\sim}{x}^F(t_i) - \bar{\underset{\sim}{x}}(t_i) \| \leq \delta(t_i) \tag{2.17}$$

where $G(\underset{\sim}{\bar{x}}, t_i) = 0$.

We may bound the quantity $\varepsilon(t_i, t_j)$ as follows:

$$\begin{aligned}
\varepsilon(t_i, t_j) = \| \underset{\sim}{x}^\circ(t_j) - \bar{\underset{\sim}{x}}(t_j) \| &= \| \underset{\sim}{x}^F(t_i) - \bar{\underset{\sim}{x}}(t_i) - \bar{\underset{\sim}{x}}(t_j) \| \\
&\leq \| \underset{\sim}{x}^F(t_i) - \bar{\underset{\sim}{x}}(t_i) \| + \bar{\underset{\sim}{x}}(t_i) - \underset{\sim}{x}(t_i) \| \\
&\leq \delta(t_i) + \| \bar{\underset{\sim}{x}}(t_i) - \bar{\underset{\sim}{x}}(t_i) \| \qquad (2.18)
\end{aligned}$$

Thus we may state:

Theorem 2.3. Let the initial guess be given by (2.16). If $\| \bar{\underset{\sim}{x}}(t_i) - \bar{\underset{\sim}{x}}(t_{i+1}) \| \leq r(t_{i+1}) - \delta(t_i)$ for $i = 0, \ldots, n-1$ then $P_n(t_1, \ldots, t_{n-1})$ is a feasible partition. X

Immediately we see that it is required that $r(t_{i+1}) > \delta(t_i)$. This is essentially the problem in Avila's example of infeasibility (and the solution curve is not differentiable). This example is cited below.

Define

$$s(x;\ \alpha,\beta.\ a) = \begin{cases} a\ , & x \leq -\beta \\ 0\ , & -\alpha \leq x \leq \alpha \\ -a\ , & x \geq \beta \end{cases} \tag{2.19}$$

and on the intervals $(-\beta, -\alpha)$ and (α, β) by straight-line segments so that s is continuous everywhere in R^1.

$G(x, t)$ is then defined for $t \varepsilon [0, 1]$ by:

$$G(x,\ t) = \begin{cases} t, & 0 \leq t \leq \frac{1}{2} \\ t + s(x-t;\ \frac{1}{2}t - \frac{1}{2},\ 3t - \frac{3}{2},\ 5t - \frac{5}{2}\ ) \end{cases} \tag{2.20}$$

Functional iteration is used for the iterative convergence process

$$x^{n+1} = G(x^n,\ t) \tag{2.21}$$

For G defined in (2.20), the radius of convergence of (2.21) may be determined explicitly as

$$r(t) = \begin{cases} 1 + t, & 0 \leq t \leq \frac{1}{2} \\ t - \frac{1}{2}. & \frac{1}{2} < t \leq 1 \end{cases} \tag{2.22}$$

Thus we see that as t approaches $\frac{1}{2}$ from above, the radius of convergence tends toward 0. Since $\delta(t)$ is set a priori as positive and bounded away from 0, then we may take

$$\begin{aligned} &\inf_{t\varepsilon(\frac{1}{2},1)} \delta(t) = \overline{\delta} > 0 \\ &\lim_{t \downarrow \frac{1}{2}} r(t) = 0 \end{aligned} \tag{2.23}$$

Thus for $\varepsilon < \overline{\delta}$ we obtain the result

$$r(\tfrac{1}{2} + \varepsilon) - \delta(\tfrac{1}{2} + \varepsilon) = \varepsilon - \delta(\tfrac{1}{2} + \varepsilon) < \varepsilon - \overline{\delta} < 0 \qquad (2.24)$$

Hence the numerical continuation process need not be feasible. (In fact, Avila has shown that by using (2.16) and (2.21), one cannot proceed past the value $t = \frac{1}{2}$, even if $\delta(t) = 0$.)

II.3 Components of the Problem and Definition of Work Variables.

II.3.1 Basic Components of the Problem

Before examining any specific methods, we should first state exactly what we mean by a "method". Each method may be decomposed into several component parts, each of which may be analyzed if we freeze all others.

Let us begin by assuming we wish to solve (2.1). Given that we have converged to a solution at λ(to within some specified error tolerance), how do we obtain an initial guess for the solution at $\lambda + \Delta\lambda$? This determination shall comprise the first component of our method, i.e., the choice of $\underset{\sim}{x}^\circ(\lambda + \Delta\lambda)$ given $\underset{\sim}{x}^F(\lambda)$ such that $\| \underset{\sim}{x}^F(\lambda) - \bar{\underset{\sim}{x}}(\lambda) \| \leq \delta(\lambda)$.

Secondly, given the guess $\underset{\sim}{x}^\circ(\lambda + \Delta\lambda)$, how do we proceed to solve $G(\underset{\sim}{x}, \lambda + \Delta\lambda) =0$? This convergence process shall be deemed the second component of the method. In the following sections we shall confine ourselves to the very simplest iterative methods.

A third consideration is the choice of intermediate error tolerances. And the final, and traditionally the most intensely studied factor, the choice of the step length $\Delta\lambda$.

At this point we shall begin to consider continuation problems rather than the homotopy problem. (Although a reparametrization of any homotopy can convert it to a continuation problem, this may do more harm than good.) If a sufficiently large number of steps are required for the homotopy problem, then all estimates are essentially the same as considering it as a continuation problem. Only when examining the intermediate error tolerances do we return to examine the homotopy problem again.

Before our examination of particular methods, a standardization of the work function shall be imposed. This will correspond to one particular choice for W_{η_i} of (2.6).

II.3.2 Work Variables

Work will be decomposed into two components: multiplicative operations (m-op's) and storage requirements. By multiplicative operations we mean either multiplication or division. The storage is the actual amount of computer memory required (exclusive of the program for continuation). Thus we make the following definitions.

a) N_1 = # of m-op's to perform on LU factorization

b) N_2 = # of m-op's to evaluate G

c) N_3 = # of m-op's to evaluate $G_{\underset{\sim}{x}}$ (3.1)

d) N_4 = # of m-op's to evaluate G_λ

e) N_5 = # of m-op's to perform one back substitution in a factorized system

f) $N(G;\ i,\ j)$ = # of m-op's to compute $\dfrac{\partial G^{i+j}}{\partial \underset{\sim}{x}^i \partial \lambda^j}$

a) S_1 = storage locations required for $G_{\underset{\sim}{x}}$

b) S_2 = storage locations required for G (3.2)

c) S_3 = # scalars stored aside from G or its derivatives

d) $S(G;\ i,\ j,)$ = storage locations required for $\dfrac{\partial G^{i+j}}{\partial \underset{\sim}{x}^i \partial \lambda^j}$

For most of the methods of the following sections, we shall not require anything higher than first derivatives. Hence our standard function to be minimized will be given by:

$$W = A \sum_{i=1}^{5} N_i y_i + B \sum_{i=1}^{3} S_i y_{i+5} \tag{3.3}$$

where this is to be minimized over the considered methods. Note that the y_i's, N_i's and S_i's are all method dependent. A and B are simply constants, the relative weights assigned to m-op's versus storage.

II.4 Initial Guesses and Iterative Methods

II.4.1 Computation of the Initial Guess for $\underset{\sim}{x}^{\circ}(\lambda + \Delta\lambda)$

First, we turn our attention to one-step methods of approximation. That is, methods which use information only at the point λ to approximate the value at $\lambda_1 + \Delta\lambda$. The two most popular of these methods are

$$\begin{aligned} &\text{a)}\quad \underset{\sim}{x}^{\circ}(\lambda + \Delta\lambda) = \bar{\underset{\sim}{x}}(\lambda) \equiv \underset{\sim}{x}_a \\ &\text{b)}\quad \underset{\sim}{x}^{\circ}(\lambda + \Delta\lambda) = \bar{\underset{\sim}{x}}(\lambda) + \Delta\lambda\, \dot{\bar{\underset{\sim}{x}}}(\lambda) \equiv \underset{\sim}{x}_b \end{aligned} \qquad (4.1)$$

where $\cdot \equiv \dfrac{\partial}{\partial\lambda}$.

Avila [4] has analyzed the feasibility of continuation methods using (4.1) a) and functional iteration or Newton's method, whereas Wacker [34] has done the same for both (4.1 a) and (4.1) b) using Newton's method; in both cases only a uniform step length is considered.

(4.1) a) and b) can, of course, be generalized to arbitrary order approximations. Thus, we shall group all other truncated Taylor series into the classification

$$\text{c)}\quad \underset{\sim}{x}^{\circ}(\lambda + \Delta\lambda) = \sum_{i=0}^{n} \frac{(\Delta\lambda)^i}{i!}\, \bar{\underset{\sim}{x}}^{(i)} \equiv \underset{\sim}{x}_{c,n} \qquad (4.1)$$

We shall also consider some general multistep approximations

Let $\lambda_\circ + \sum_{i=1}^{k} \Delta\lambda_i = \lambda_k$

$$\text{d)}\quad \underset{\sim}{x}^{\circ}(\lambda_{m+1}) = \sum_{j=0}^{m} \bar{\underset{\sim}{x}}(\lambda_j) \prod_{\substack{k=0\\k\neq j}}^{m} \frac{\lambda_{m+1} - \lambda_k}{\lambda_j - \lambda_k} \equiv \underset{\sim}{x}_{d,m} \qquad (4.1)$$

e) $\underset{\sim}{x}^{\circ}(\lambda_{m+1}) = H_{2m+1}\,(\lambda_{m+1}) \equiv \underset{\sim}{x}_{e,m}$ (4.1)

where H_{2m+1} is the Hermite polynomial of order $2m+1$ using data at the points λ_i, i=0, ..., m.

We must now weigh the error involved in using each such approximation versus the work involved per step. For this purpose, we use only the # of m-op's. TABLE 1 illustrates each case. The vector G is assumed to have length ν.

$\underset{\sim}{x}^{\circ}(\lambda + \Delta\lambda)$	error ε	work
$\underset{\sim}{x}_a$	$(\Delta\lambda)\,\lVert\dot{\underset{\sim}{x}}\,(z_1)\rVert$	0
$\underset{\sim}{x}_b$	$\frac{1}{2}\,(\Delta\lambda)^2\,\lVert\ddot{\underset{\sim}{x}}(z_2)\rVert$	$N_4 + N_5$
$\underset{\sim}{x}_{c,i}$	$\frac{(\Delta\lambda)^{i+1}}{(i+1)!}\,\lVert\underset{\sim}{x}^{(i+1)}(z_{3,i})\rVert$	$N_4 + N_5 + \sum_{j=2}^{i}\sum_{k=2}^{j} N(G;j,k)$
$\underset{\sim}{x}_{d,i}$	$\frac{\lVert\underset{\sim}{x}^{(i+1)}(z_{4,i})\rVert}{(i+1)!}\,\prod_{k=0}^{i}\left[\sum_{j=1}^{k}\Delta\lambda_{i-j}\right]$	$(2i+1)(i+1) + i\nu$
$\underset{\sim}{x}_{e,i}$	$\frac{\lVert\underset{\sim}{x}^{(2i+2)}(z_{5,i})\rVert}{(2i+2)!}\cdot\left(\prod_{k=0}^{i}\left[\sum_{j=0}^{k}\Delta\lambda_{i-j}\right]\right)^2$	$N_5 + N_4 + i(6i+4) + 2i\nu$

TABLE 1

By themselves, we may not use these figures to determine which is best to use. However, if we use them in tandem with some particular iterative method, we may be able to make a determination.

II.4.2 Work Analysis for Some Specific Iterative Methods

II.4.2.i Newton's Method; $\underset{\sim}{x}^{n+1} = \underset{\sim}{x}^{n} - J^{-1}(\underset{\sim}{x}^{n})f(\underset{\sim}{x}^{n})$

We now determine the work required for Newton's method to converge within a preset error tolerance, say ε_F. At each iteration we require 1 LU factorization, 1 back substitution, 1 evaluation of f and 1 evaluation of $f_{\underset{\sim}{x}}$. Hence from (3.1), the work per iteration is $N_1 + N_2 + N_3 + N_5$. By the Kantorovich theorem (Ortega and Rheinboldt [53]) we have the following. We wish to solve the system $f(\underset{\sim}{x}) = 0$, with $f_{\underset{\sim}{x}} = J$.

Theorem 4.1

Given

$$
\begin{aligned}
&\text{a)}\quad \| J(\underset{\sim}{x}^{\circ})^{-1} \| \leq \beta \\
&\text{b)}\quad \| J(\underset{\sim}{x}^{\circ})^{-1} f(\underset{\sim}{x}^{\circ}) \| = \| \underset{\sim}{x}^{\circ} - \underset{\sim}{x} \| \leq \alpha \\
&\text{c)}\quad \| J(\underset{\sim}{x}) - J(\underset{\sim}{y}) \| \leq \gamma \| \underset{\sim}{x} - \underset{\sim}{y} \| \quad \text{in} \quad \| \underset{\sim}{x} - \underset{\sim}{\bar{x}} \| \leq 2 \| \underset{\sim}{x}^{\circ} - \underset{\sim}{\bar{x}} \|
\end{aligned}
\tag{4.2}
$$

If $h = \alpha\beta\gamma < \frac{1}{2}$, then

$$\varepsilon^{\nu} = \| \underset{\sim}{x}^{\nu} - \underset{\sim}{\bar{x}} \| \leq \frac{(2h)^{2^{\nu}}}{\beta\gamma 2^{\nu}} \tag{4.3}$$

X

Let $\beta\gamma = \hat{C}$. Then (4.3) may be restated as

$$\hat{C}\,\varepsilon^{\nu} \leq \frac{(2\alpha\hat{C})^{2^{\nu}}}{2^{\nu}} \leq (2\alpha\hat{C})^{2^{\nu}} = (2\, \| \underset{\sim}{x}^{\circ} - \underset{\sim}{x}^{1} \| \, \hat{C})^{2^{\nu}} \tag{4.4}$$

Thus we may guarantee that the error is less than ε_F if $\hat{C}\varepsilon_F \leq (2\alpha\hat{C})^{2^{\nu}}$. This yields an upper bound on the number of steps needed for convergence, say $\hat{\nu}_N$, as

$$\hat{\nu}_N = \left\lceil \log_2 \frac{\log \hat{C}\, \varepsilon_F}{\log 2\, \alpha \hat{C}} \right\rceil \tag{4.5}$$

An upper bound on the work required using Newton's method is then obtainable using (3.1).

$$\hat{W}_N = \hat{\nu}_N(N_1 + N_2 + N_3 + N_5) \tag{4.6}$$

II.4.2.ii Chord Method

The chord method will be defined by

$$\underset{\sim}{x}^{n+1} = \underset{\sim}{x}^{n} - \hat{J}^{-1} f(\underset{\sim}{x}^{n}) \tag{4.7}$$

where we assume that the iteration matrix $\hat{J}$ is available.

Furthermore it is assumed that the convergence constant for the iterations is known, i.e.,

$$\varepsilon^{\nu} \leq K_C \varepsilon^{\nu-1} \tag{4.8}$$

An upper bound on the number of steps required to meet the tolerance is

$$\hat{\nu}_c = \left\lceil \frac{\log \varepsilon_F - \log \varepsilon_\circ}{\log K_c} \right\rceil \tag{4.9}$$

Only one evaluation of f and one back substitution are required per iteration. Hence, from (3.1) and (4.9) we obtain an upper bound on the work:

$$\hat{W}_c = \hat{\nu}_c\ (N_2 + N_5) \tag{4.10}$$

II.4.2.iii Standard Modified Newton

This method involves the use of the Newton matrix at the point $\underset{\sim}{x}^\circ$ (i.e., $J(\underset{\sim}{x}^\circ)$) as $\hat{J}$ in (4.7). Thus the iterations are defined by

$$\underset{\sim}{x}^{n+1} = \underset{\sim}{x}^{n} - J^{-1}(\underset{\sim}{x}^{\circ})f(\underset{\sim}{x}^{n}) \qquad (4.11)$$

To factor $J(\underset{\sim}{x}^{\circ})$ initially requires $N_1 + N_2$. At each iteration we shall require $N_3 + N_5$. The upper bound on the number of non-Newton steps necessary (the first step is a Newton step) may be obtained by considering the error relationship

$$\varepsilon_F \le K_m^{\nu} \varepsilon^1 \le K_m^{\nu} \hat{C}(2\alpha)^2 \qquad (4.12)$$

Thus we obtain the upper bound

$$\hat{\nu}_m = \left\lceil \frac{\log \frac{\varepsilon_F}{4\hat{c}\;\alpha^2}}{\log K_m} \right\rceil = \left\lceil \frac{\log \varepsilon_F - \log\,[\hat{C}(2\alpha)^2]}{\log K_m} \right\rceil \qquad (4.13)$$

This then yields $\hat{W}_m$ as

$$\hat{W}_m = N_1 + N_3 + (1 + \hat{\nu}_m)\;(N_2 + N_5) \qquad (4.14)$$

II.4.2.iv Special Modified Newton

This method uses a Newton step to start the iteration process and also to end it. The last Newton iteration is used when $\varepsilon^{\nu} \le \varepsilon_F^{\frac{1}{2}}$ From (4.3), (4.11), and (4.12), we have

$$\varepsilon_F \le (K_m)^{\frac{\hat{\nu}_m}{2}} \varepsilon_2 = (K_m)^{\frac{\hat{\nu}_m}{2}} \varepsilon_2^{-\frac{1}{2}}\varepsilon_2 \qquad (4.15)$$

Setting $\varepsilon_2^{-\frac{1}{2}} = K_m^{\gamma}$ yields $\gamma = -\frac{1}{2}\dfrac{\log \varepsilon_2}{\log K_m}$ and (4.15) then gives us

$$\varepsilon_F^{\frac{1}{2}} \le (K_m)^{(\hat{\nu}_m - \frac{\log \varepsilon_2}{\log K_m})} \qquad (4.16)$$

The upper bound on the number of steps is simply the ceiling of the exponent

$$\hat{\nu}_s = \left\lceil \frac{1}{2}\ \hat{\nu}_m - \frac{\log \varepsilon_2}{\log K_m} \right\rceil \qquad (4.17)$$

For purposes of comparison, we will use the approximation

$$\log \varepsilon_2 \approx \log\left[(2\alpha)^2\ \hat{C}\ \right] \qquad (4.18)$$

Using (4.18) in (4.17) results in

$$\hat{\nu}_s \approx \left\lceil \frac{1}{2}\left(\hat{\nu}_m - \frac{\log[(2\alpha)^2\hat{C}]}{\log K_m}\right) \right\rceil$$

$$= \left\lceil \frac{1}{2}\left(\frac{\log \varepsilon_F - \log\ [(2\alpha)^2\hat{C}]}{\log K_m} - \frac{\log\ [(2\alpha)^2\hat{C}]}{\log K_m}\right) \right\rceil$$

$$(4.19)$$

$$= \left\lceil \frac{\log \varepsilon_F}{2\log K_m} - \frac{\log\ [(2\alpha)^2\hat{C}]}{\log K_m} \right\rceil$$

Hence we obtain the upper bound on the work

$$\hat{W}_s = (\hat{\nu}_s + 2)\ [N_2 + N_5] + 2(N_1 + N_3) \qquad (4.20)$$

II.4.2.v Reverse Modified Newton

In this method a Newton step is used only for the last step. Some chord matrix is used for all other iterations (perhaps, using $J(\underset{\sim}{x}^F(\lambda - \Delta\lambda))$ would be advisable since it is available at no cost). Here we need to satisfy the relation

$$\varepsilon_F = K_s^{\hat{\nu}_R}\ \varepsilon^\circ \qquad (4.21)$$

Solving for $\hat{\nu}_R$ gives

$$\hat{\nu}_R = \left\lceil \frac{\log \dfrac{\varepsilon_F}{\varepsilon^\circ}}{\log K_s} \right\rceil \qquad (4.22)$$

The corresponding upper bound on the work is

$$\hat{W}_R = (1 + \hat{\nu}_R)\ (N_2 + N_5) + N_1 + N_3 \qquad (4.23)$$

II.4.2.vi Convergence Sphere Approximation

For the purpose of comparing the preceding methods, we shall assume that the initial guess lies within the intersection of the convergence domains of all the methods. Also, we shall use the approximation

$$\varepsilon^\circ \approx 2\alpha \qquad (4.24)$$

in order to compare the work required for the various methods. This is a reasonable approximation for ε° small, but might be quite bad for large ε°. However, α and ε° enter only through $\log \varepsilon^\circ$ and $\log 2\alpha$, thus extending the region of validity of the approximation.

II.4.3 Two Examples of the Usage of the Work Function

As an example of how we may employ the preceding work estimates to develop useful algorithms, we start with a comparison of one-step initial guesses to multi-step guesses. For the present, we shall assume that the length of the parameter steps has been predetermined. The vector G is assumed to have length ν. The initial guesses of (4.1) b) and (4.1) e) with $m = 1$ will be compared.

II.4.3.i Newton's Method

The work to compute $\underset{\sim}{x}_b$ ((4.1) b) is simply an evaluation of G_λ and one back substitution, i.e., $N_4 + N_5$. ν multiplications are also required. From TABLE 1 the error $\varepsilon_b = \frac{1}{2} \|\ddot{\underset{\sim}{x}}\| (\Delta\lambda)^2$.

To compute $\underset{\sim}{x}_{e,1}$, we need to do 2 G_λ-evaluations and 2 back substitutions per point. However, one of these is again required at the next step; so, an average of only one is performed per step.

Hence, the average work is also $N_4 + N_5$, but $2\nu + 10$ multiplications are required. From TABLE 1 the error is $\varepsilon_{e,1} = \frac{1}{6}\|\underset{\sim}{x}^{(4)}\|\,(\Delta\lambda)^4$. For $\Delta\lambda$ small and $\underset{\sim}{x}\varepsilon C^5$ we have $\varepsilon_{e,1} = 0(\varepsilon_b{}^2)$. This implies we will need at least one extra Newton step to converge. Therefore

$$T = W_b - W_{e,1} = N_1 + N_2 + N_3 + N_5 - (\nu + 10) \tag{4.25}$$

When is $T > 0$? Almost always. Since N_1 is in general a ν^3 operation, T is positive for all but the smallest values of ν. Thus using $\underset{\sim}{x}_{e,1}$ is more efficient from the standpoint of m-op's required.

However, if we analyze the storage requirements, we find that $\underset{\sim}{x}_{e,1}$ requires the storage of 2 extra ν-vectors. Thus

$$S = S_b - S_{e,1} = 2\nu > 0 \tag{4.26}$$

For full systems, this does not seem particularly significant, but for banded systems with a small bandwidth this could be a large percentage of the storage requirements. Here it is up to the individual to develop a weighting system for determining the relative cost of m-op's (computer time) and storage.

II.4.3.ii Chord Method

In analyzing the chord method, we need to consider the number of iterations required to bring ε_b down to $\varepsilon_{e,1}$. Using the approximate relationship $\varepsilon_{e,1} \approx \varepsilon_b^2 = K_c^n\,\varepsilon_b$, solve for n to get

$$n = \frac{\log \varepsilon_b}{\log K_c} \tag{4.27}$$

Then the difference in the work required is

$$T = W_b - W_{e,1} = n(N_2 + N_5) - (\nu + 10) \tag{4.28}$$

In general, N_5 is a ν^2 operation. Hence T is negative only if $n = 0$, i.e., $\varepsilon_b \approx \varepsilon_{e,1}$. But for small step lengths we know that $\varepsilon_{e,1} \approx \varepsilon_b^2$ is a good approximation. Thus, $\underset{\sim}{x}_{e,1}$ is again the clear cut winner in terms of m-ops. The storage considerations are exactly the same as in II.4.3.i (see 4.26).

II.4.4 A Digression Concerning Choice of Method

Three model problems shall be examined. For these models we shall assume a reasonable Newton convergence rate ($2\alpha = \varepsilon° = .56$). For a given error tolerance, we calculate the number of work-equivalent iterations for each method, and the required convergence constant to obtain the same error bound.

II.4.4.i Scalar Polynomial

We seek to find a root of the n^{th}-degree polynomial

$$y = x^n + a_1 x^{n-1} + \ldots + a_n \tag{4.29}$$

From (3.1) we determine the work variables:

$$N_1 = 0,\ N_2 = n - 1,\ N_3 = n - 1,\ N_5 = 1 \tag{4.30}$$

The upper bounds on the work required are obtained from (4.6), (4.10), (4.14), (4.20) and (4.23):

$$\begin{aligned} \hat{W}_N &= (2n - 1)\hat{\nu}_N \\ \hat{W}_M &= n - 1 + n\hat{\nu}_M \\ \hat{W}_C &= n\hat{\nu}_c \\ \hat{W}_S &= 2(2n-1) + n\hat{\nu}_S \\ \hat{W}_R &= n - 1 + n\hat{\nu}_R \end{aligned} \tag{4.31}$$

A table may now be used to illustrate the equivalent work required to bring the error within a desired error tolerance. In

TABLES 2 and 3, $\hat{K}_\alpha$ denotes the value of K_α such that the upper bound on the work using method α is the same as for Newton's method. TABLE 2 depicts the case $n = 2$. TABLE 3 is the general case for $n \geq 8$.

TABLE 2

$n = 2$

ν_N	ν_m	$\hat{K}_m$	ν_s	$\hat{K}_s$	ν_c	$\hat{K}_c$	ν_R	$\hat{K}_R$	ε_F
2	2	.56	-	-	3	.56	2	.74	10^{-1}
3	4	.42	1	.31	4	.36	4	.64	10^{-2}
4	5	.19	3	.31	6	.23	5	.44	10^{-4}
5	7	.08	4	.13	7	.07	7	.29	10^{-8}
6	8	.01	6	.05	9	.01	8	.10	10^{-16}

TABLE 3

$n \geq 8$

ν_N	ν_m	$\hat{K}_m$	ν_s	$\hat{K}_s$	ν_c	$\hat{K}_c$	ν_R	$\hat{K}_R$	ε_F
2	2	.56	-	-	3	.56	2	.74	10^{-1}
3	4	.42	1	.31	5	.44	4	.64	10^{-2}
4	6	.26	3	.31	7	.29	6	.51	10^{-4}
5	8	.11	5	.19	9	.13	8	.33	10^{-8}
6	10	.02	7	.08	11	.03	10	.16	10^{-16}

We can write a general formula for the $\hat{K}$'s. If $\varepsilon_F = 10^{-q}$ and $\varepsilon^\circ = 10^{-p}$ then

$$\begin{aligned} \hat{K}_m &= 10^{\frac{2p-q}{\nu_m}} \\ \hat{K}_s &= 10^{\frac{4p-q}{2\nu_s}} \\ \hat{K}_c &= 10^{\frac{p-q}{\nu_c}} \\ \hat{K}_R &= 10^{\frac{2p-q}{2\nu_R}} \end{aligned} \qquad (4.32)$$

This is valid for all the following tables.

TABLES 1 and 2 indicate that for scalar polynomials it is vastly superior to use Newton's method; note that the reverse Newton is the next best in terms of work.

II.4.4.ii Full Matrix System

Let us solve the system

$$f_j(\underset{\sim}{x}) \equiv \sum_{i=1}^{n} a_{ij}x_i x_j - b_j = 0 \quad , \quad j = 1, \ldots, n \qquad (4.33)$$

The work variables of (3.1) become

$$N_1 = n^3, \; N_2 = 3n^2, \; N_3 = 1, \; N_5 = n^2 \qquad (4.34)$$

Thus the corresponding work bounds are given by

$$\begin{aligned} \hat{W}_N &= (n^3 + 4n^2 + 1)\,\nu_N \\ \hat{W}_m &= n^3 + 1 + 4n^2\,\nu_m \\ \hat{W}_c &= 4n^2\nu_c \\ \hat{W}_s &= 2(n^3 + 4n^2 + 1) + 4n^2\nu_s \end{aligned} \qquad (4.35)$$

We could again formulate tables similar to TABLES 1 and 2, but it is more convenient to look at the general expression for the

$\hat{K}$'s and ν's for $n \geq 3$.

$$\nu_m = \left\lfloor \frac{1}{4} n(\nu_N - 1) + \nu_N \right\rfloor \quad , \quad \hat{K}_m = 10^{-\frac{2\nu_N}{\nu_N(n+4)-n}}$$

$$\nu_s = \left\lfloor \frac{1}{4} n(\nu_N - 2) + \nu_N \right\rfloor \quad , \quad \hat{K}_s = 10^{-\frac{2\nu_N}{\nu_N(n+4) - 2\nu_N}}$$

$$\nu_c = \left\lfloor \frac{1}{4} n\nu_N + \nu_N \right\rfloor \quad , \quad \hat{K}_c = 10^{-\frac{2\nu_N - 1}{\nu_N(n+4)}} \tag{4.36}$$

$$\nu_R = \nu_s \quad , \quad \hat{K}_R = 10^{-\frac{2\nu_N + 3}{4\nu_s}}$$

A comparison of the chord method to the Newton method now yields mixed results. For large n, the ratio of Newton steps to any modified Newton or chord scheme tends to $4/n$. This indicates we can do many "chord-type" iterations and still cost less than one Newton step. However, the size of the "trade-off" constant $\hat{K}$ decreases with ν_N, at the rate $10^{-\frac{2\nu_N}{(n+4)\nu_N}}$. Since most current codes set a limit on the number of allowable Newton steps the behavior for fixed ν_N is important. If we fix ν_N, then for any of the $\hat{K}$'s in (4.36) we have

$$\lim_{n\to\infty} \hat{K}_\alpha(n, \nu_N) = 1 \tag{4.37}$$

Thus for the large dimensional problems, this indicates the chord method is preferable to Newton's method even though

$$\lim_{\nu_N \to \infty} \hat{K}_\alpha(n, \nu_N) = 0 \tag{4.38}$$

A comparison of the two behaviors is given in TABLE 3 for $\hat{K}_c$.

TABLE 3

$\hat{K}_c$

ν_N n	2	3	4	5	6
10	.719	.645	.517	.349	.173
50	.918	.892	.843	.761	.634
100	.956	.942	.915	.867	.789
250	.982	.976	.964	.943	.907
500	.990	.987	.981	.971	.952
1000	.995	.993	.990	.985	.975
2000	.997	.996	.995	.992	.987

Since Newton becomes more efficient the closer you come to the root, this suggests beginning with chordsteps and shifting to Newton when sufficiently close. This fits in well with the generation of an initial guess in continuation methods. At the point λ we use Newton's method for the final step with iteration matrix $J(\underset{\sim}{x}^n, \lambda)$. This is also used to determine the tangent vector. This may then be used as $\hat{J}$ [see (4.7)]. This will be further discussed in succeeding sections.

II.4.4.iii A Banded System

Consider the problem

$$\nabla^2\phi - f(\phi, x, y) = 0 \tag{4.39}$$

on a rectangular domain with homogeneous boundary conditions where

$$f = Ax + Bx^2 + C + D\phi + E\phi^2 + Fy + Gy^2 \tag{4.40}$$

Let us discretize the domain using n points in each direction. Then the work variables of (3.1) are

$$N_1 = n^3, \quad N_2 = 2n^2, \quad N_3 = 9n^2 \ , \quad N_5 = n^2 \qquad (4.41)$$

and the work bounds are given by

$$\begin{aligned} \hat{W}_N &= (n^3 + 11n^2)\hat{\nu}_N \\ \hat{W}_c &= 11n^2\hat{\nu}_c \\ \hat{W}_R = \hat{W}_m &= n^3 + 11n^2\hat{\nu}_m \\ \hat{W}_s &= 2(n^3 + 11n^2) + 11n^2\hat{\nu}_s \end{aligned} \qquad (4.41)$$

The same basic results as in Section II.4.4.ii are obtained. That is both (4.37) and (4.38) are valid.

II.5 Power Formulation

II.5.1 Optimality Theorems

For the general continuation problem, we obviously do not wish to minimize work, as such, since we have no particular parameter value in mind. Instead we should seek to minimize the work per parameter step length, which we shall call the power. (By way of analogy, if the parameter is time, we may actually consider λ to be power.)

To do this, we represent both $\hat{W}$ and the step size $\Delta\lambda$ as functions of the error ε.

Thus, our desire is to minimize:

$$P(\varepsilon) = \frac{\hat{W}(\varepsilon)}{\Delta\lambda(\varepsilon)} \tag{5.1}$$

We note that this is not the actual power that is expended, but an upper bound on that power.

First we turn to Newton's method.

Theorem 5.1 Assume ε_F is given, and Newton's method (section II.4.2.i) is used with initial guess (4.1) a).

Assume

$$\begin{aligned} &\text{a)} \quad \varepsilon^{\circ} < 1 \\ &\text{b)} \quad \varepsilon^{n} \leq (\varepsilon^{n-1})^{2} \\ &\text{c)} \quad \|\dot{\underset{\sim}{x}}\| \leq M \qquad \text{for} \qquad 0 \leq \Delta\lambda \leq \Lambda \end{aligned} \tag{5.2}$$

Then a lower bound on the optimal step size (i.e., minimization of P in (5.1) using initial guess (4.1) a)) is given by

$$\overline{\Delta\lambda} = \frac{\varepsilon_F^{\frac{1}{2^{n_F}}}}{M} \tag{5.3}$$

where n_F is defined by

$$n_F \varepsilon_F^{2^{-\frac{1}{n_F}}} = \min_{n \geq 1} n\, \varepsilon_F^{\frac{1}{2^n}} \qquad (5.4)$$

Proof: The work involved in Newton's method may be derived from (4.9), with $2\alpha\hat{C} = \varepsilon_o$, as:

$$\hat{W}(\varepsilon°) = \left\lceil \log_2 \frac{\log \varepsilon_F}{\log \varepsilon_o} \right\rceil \sum_{\substack{i=1 \\ i \neq 4}}^{5} N_i = \mathcal{K} \left\lceil \log_2 \frac{\log \varepsilon_F}{\log \varepsilon_o} \right\rceil \qquad (5.5)$$

From (4.1) a) and (5.2) c) we obtain

$$\varepsilon(\Delta\lambda) = \Delta\lambda \, ||\dot{\underset{\sim}{x}}(z)|| \leq M \quad \Delta\lambda \qquad (5.6)$$

whence

$$\Delta\lambda(\varepsilon) \geq \frac{\varepsilon}{M} \qquad (5.7)$$

Using (5.5) and (5.7) in (5.1) yields

$$\frac{W}{\Delta\lambda} \leq \frac{\mathcal{K} M}{\varepsilon_o} \left\lceil \log_2 \frac{\log \varepsilon_F}{\log \varepsilon_o} \right\rceil \equiv \overline{P} \, (\varepsilon°) \qquad (5.8)$$

Note that the ceiling in (5.8) can be replaced by an integer n by redefining (5.8) as follows:

$$\overline{P}(\varepsilon°) = \frac{M\,n}{\varepsilon°}\mathcal{K} \,, \quad \varepsilon° \in I_n \equiv \left[\varepsilon_F^{\frac{1}{2^{n-1}}} \,, \quad \varepsilon_F^{\frac{1}{2^n}} \right] \qquad (5.10)$$

Since $\frac{1}{\varepsilon°}$ is monotonic over each I_n, we need only minimize over the discrete set $\left\{ \varepsilon_F^{\frac{1}{2^n}} \right\}$. Thus we must minimize

$$g(n) = n\, \varepsilon_F^{-\frac{1}{2^n}} \qquad (5.10)$$

But $n_F \varepsilon_F^{-\frac{1}{2^{n_F}}} = \min_{n\geq 1} n\varepsilon_F^{-\frac{1}{2^n}} = \min_{n\geq 1} g(n)$ from (5.4).

Hence n_F minimizes $\overline{P}$ in (5.9). The minimum is attained for $\varepsilon^\circ = \varepsilon_F^{\frac{1}{2^{n_F}}}$. But from (5.7)

$$\Delta\lambda_{\text{opt.}} \geq \frac{\varepsilon_F^{\frac{1}{2^{n_F}}}}{M} = \overline{\Delta\lambda} \qquad X$$

Obviously if $\| \dot{\underset{\sim}{x}} \|$ is small then $\overline{\Delta\lambda}$ is a good approximation to the optimal step-length. This, of course, presupposes the knowledge of $\| \dot{\underset{\sim}{x}} \|$. We can use the values of $\| \dot{\underset{\sim}{x}} \|$ at previously computed λ values and extrapolate to estimate the bound M.

The above result is easily extended to other types of initial guesses.

<u>Theorem 5.2</u> Let conditions (5.2) a) and b) of Theoremn 5.1 hold, and in addition let (see (4.1) c):

$$\| \dot{\underset{\sim}{x}}^{(N+1)} \| \leq M_N \quad \text{for} \quad 0 \leq \Delta\lambda \leq \Lambda \quad \text{along the path} \tag{5.11}$$

$$y(\Delta\lambda) = \underset{\sim}{x}_{c,n}$$

Then using Newton's method and initial guess (4.1) c), a lower bound on the optimal step length is given by

$$\Delta\lambda = \left[\frac{\varepsilon_F^{\frac{1}{2^{n_F}}}}{M_N} \right]^{\frac{1}{N+1}} \tag{5.12}$$

where n_F is defined by

$$n_F \varepsilon_F^{-\frac{1}{(N+1)2^{n_F}}} = \min_{n\geq 1} n\, \varepsilon_F^{-\frac{1}{(N+1)2^{n}}} \tag{5.13}$$

Proof: Simply use $\Delta\lambda \geq \left[(N+1)!\, M_N \varepsilon\right]^{\frac{1}{N+1}}$ from TABLE 1 in the proof of theorem (5.1). X

Similarly, this can easily be extended to other iterative methods. If we examine any of the modified chord methods discussed in Section II.4.2, we can write the work as $A + Bn$, where A and B are method dependent constants and m is the number of iterations required. However, an extra parameter, the covergence constant K, also appears. If we denote the points where $m(\varepsilon) = i$ (an integer) by m_i, then we can formulate a general theorem.

Theorem 5.3

Let (5.2) a) , (5.2) b), and (5.11) hold.

Let the iterative method used have the work function defined by

$$W = A + Bn, \quad m_{n-1}(\varepsilon_{F,}\ K) < \varepsilon \leq m_n(\varepsilon_F, K) \tag{5.14}$$

with convergence constant K. Then the lower bound on the optimal step size $\overline{\Delta\lambda}$ yields the power:

$$\overline{P} = \frac{A + Bn_F}{(m_{n_F})^{\frac{1}{(N+1)2^{n_F}}}} \cdot \frac{M_N}{\left[(N+1)!\, M_N\right]^{\frac{1}{N+1}}} \tag{5.15}$$

where n_F is defined by

$$n_F\, m_{n_F}^{\frac{1}{(N+1)2^{n_F}}} = \min_{n\geq 1} n\, m_n^{\frac{1}{(N+1)2^{n}}} \tag{5.16}$$

and the lower bound $\overline{\Delta\lambda}$ is given by

$$\overline{\Delta\lambda} = \left[\frac{m_{n_F}}{M_N}\right]^{\frac{1}{N+1}}$$

Proof: Replace $\varepsilon^{\frac{1}{2^{n_F}}}$ by m_{n_F} in Theorem 5.2. All steps follow the proof of Theorem 5.1 mutatis mutandis. X

The most restrictive aspects of these theorems is that we require $\varepsilon < 1$. However, they can all be modified by replacing ε by $2h$ (with a slight change of various constants). In that form, the theorem then replaces conditions (5.2) a) and b) by (4.2) with $h = \alpha\beta\gamma < \frac{1}{2}$.

Recalling that β is a bound on the inverse of the Jacobian, this leads us to an examination of a technique to obtain such a bound.

II.5.2 A Bound for the Inverse of a Matrix.

Theorem 5.4

Let L be lower triangular.

Define $\hat{L}$ by

$$\text{a)}\quad \hat{\ell}_{jj} = |\ell_{jj}| \qquad\qquad (5.17)$$

$$\text{b)}\quad \hat{\ell}_{ij} = -|\ell_{ij}| \qquad \text{for } i \neq j$$

Then $\|\hat{L}^{-1}\|_{\infty} \geq \|L^{-1}\|_{\infty}$

where $\|\cdot\|_{\infty}$ is the maximum absolute row sum.

Proof: We can write L and $\hat{L}$ as the difference of diagonal and nilpotent matrices.

$$\begin{aligned} &a)\quad L = D - N \\ &b)\quad \hat{L} = \hat{D} - \hat{N} \end{aligned} \tag{5.18}$$

where by (5.17) we have

$$\begin{aligned} &a)\quad \| \hat{N} \|_\infty \geq \| N \|_\infty \\ &b)\quad \| \hat{D} \|_\infty = \| D \|_\infty \end{aligned} \tag{5.19}$$

and $\hat{n}_{ij} \geq 0$ for all i,j. Thus all elements of $[\hat{N}]^K$ are non-negative for any positive K.

But given L in nxn, and N nilpotent the inverse may be written as a finite expansion:

$$\begin{aligned} &a)\quad L^{-1} = D + N + N^2 + N^3 + \ \ldots + N^{n-1} \\ &b)\quad \hat{L}^{-1} = \hat{D} + \hat{N} + \hat{N}^2 + \hat{N}^3 + \ \ldots + \hat{N}^{n-1} \end{aligned} \tag{5.20}$$

If we now take the norm of L^{-1} we obtain

$$\begin{aligned} \|L^{-1}\|_\infty &= \| D + N + N^2 + \ldots + N^{n-1} \|_\infty \\ &\leq \| D \|_\infty + \| N \|_\infty + \ldots + \| N^{n-1} \|_\infty \\ &\leq \| \hat{D} \|_\infty + \| \hat{N} \|_\infty + \ldots + \| \hat{N}^{n-1} \|_\infty \\ &= \| \hat{D} + \hat{N} + \ldots + \hat{N}^{n-1} \|_\infty = \| \hat{L}^{-1} \|_\infty \qquad X \end{aligned}$$

Corollary 5.5

Let U be upper triangular

Define $\hat{U}$ by

$$\begin{aligned} &a)\quad \hat{U}_{jj} = |U_{jj}| \\ &b)\quad \hat{U}_{ij} = - |U_{ij}|| \qquad \text{for } i \neq j \end{aligned} \tag{5.22}$$

Then $\| \hat{U}^{-1} \|_1 \geq \|U^{-1}\|_1$

where $\| \cdot \|_1$ is the maximum absolute column sum.

<u>Proof</u>: $\| U\|_1 = \| U^T\|_\infty$ and $\| \hat{U}^{-1}\|_1 = \| \hat{U}^{T^{-1}} \|_\infty$

Thus by theorem 5.4

$$\| U^{T^{-1}} \|_\infty \geq \| U^{T^{-1}} \|_\infty \tag{5.23}$$

Hence

$$\| \hat{U}^{-1}\|_1 \geq \| U^{-1}\|_1 \qquad X$$

<u>Corollary 5.6</u> Let $J = LU$ be a nonsingular nxn matrix.

Then

$$\| J^{-1}\|_\infty \leq c(n) \| \hat{L}^{-1}\|_\infty \| \hat{U}^{-1}\|_1 \tag{5.24}$$

<u>Proof</u>: By the properties of matrix norms:

$$\| J^{-1}\|_\infty \leq \| L^{-1}\|_\infty \| U^{-1}\|_\infty \tag{5.25}$$

Also for $\| \cdot\|_\infty$ and $\| \cdot\|_1$, there exist constants $c(n)$ dependent only on n, such that for an nxn matrix A,

$$\| A\|_\infty \leq c(n)\| A\|_1 \tag{5.26}$$

Using (5.21), (5.23) and (5.26) in (5.25) yields

$$\begin{aligned} \| J^{-1}\|_\infty &\leq \| L^{-1}\|_\infty \| U^{-1}\|_1 \; c(n) \\ &\leq \| \hat{L}^{-1}\|_\infty \| \hat{U}^{-1}\|_1 \; c(n) \qquad X \end{aligned} \tag{5.27}$$

The usefulness of these theorems is evidenced by the small operations count required to obtain this bound. Computationally, changing the signs of ℓ_{ij} and u_{ij} can be done by simply altering the sign bit. Then one only need solve the system.

$$\hat{L}\hat{\underset{\sim}{y}} = \hat{\underset{\sim}{x}} \tag{5.28}$$

with $\underset{\sim}{x} = (1, 1, \ldots 1)^T$. It is easy to show that with $\| \underset{\sim}{x} \|_\infty = 1$ and $\hat{L}\underset{\sim}{y} = \underset{\sim}{x}$ that

$$\| \hat{\underset{\sim}{y}}\|_\infty \geq \| \underset{\sim}{y}\|_\infty \tag{5.29}$$

But this is just the definition of the induced norm.

$$\| \hat{\underset{\sim}{y}} \|_\infty = \max_{\| \underset{\sim}{x} \| = 1} \| \hat{L}^{-1} \underset{\sim}{x} \|_\infty = \| \hat{L}^{-1} \|_\infty \tag{5.30}$$

Similarly one can show that $\| \hat{\underset{\sim}{z}} \|_\infty = \| \hat{U}^{-1} \|_\infty$ where

$$\hat{U} \hat{\underset{\sim}{z}} = \hat{\underset{\sim}{x}} \tag{5.31}$$

Thus only 2 back substitutions are required to obtain a bound on $\| J^{-1} \|$, given that J has been LU-factored. The maximum amount of work required is $2n^2 + 1$ operations. To actually compute $\| J^{-1} \|$ is an order n^3 operation.

II.6 COMPUTATIONAL PROCEDURES

II.6.1 Computational Scheme I.

In reality we really cannot determine ε° or h (see (4.2) and (4.3)) very well a priori. Thus we need to develop a criterion based on computable quantities. If we return to the method suggested in Section II.4.4.ii, using $J(\underset{\sim}{x}^F, \lambda)$ as the iteration matrix at the point $\lambda + \Delta\lambda$, then Theorems 5.1 - 5.3 can be reformulated in terms of the convergence constant K_c (see (4.8)).

It is necessary to make an assumption on how K_c depends on $\Delta\lambda$ or ε°. The simplest assumption is that it is a linear function, say $K_c = \hat{A}(\Delta\lambda)$. Since we have the solution within our desired accuracy at λ, we set $K_c = 0$ at $(\underset{\sim}{x}^F, \lambda)$. By now choosing some increment $\Delta\lambda$ and doing one iteration at $\lambda + \Delta\lambda$, we obtain an estimate for K_c at $\lambda + \Delta\lambda$(which is of course dependent on the method of initial guess):

$$K_c \approx \frac{\|f(\underset{\sim}{x}^1, \lambda + \Delta\lambda)\|}{\|f(\underset{\sim}{x}^\circ, \lambda + \Delta\lambda)\|} \qquad (6.1)$$

Since $\Delta\lambda$ is known, we can now determine an estimate of $\hat{A}$. How does this help? Let us reformulate Theorem 5.3 as an example. The work is now expressed as a function of K_c, as is the step length.

$$\text{a)} \quad \hat{W} = A + Bn(K_c) \;, \qquad m_{n-1} < K_c \leq m_n$$

$$\text{b)} \quad \Delta\lambda \approx \frac{K_c}{\hat{A}} \qquad (6.2)$$

$$\text{c)} \quad P = \hat{A}M\,\frac{A + Bn}{K_c}$$

Since $n(K_c)$ is known as a function of K_c, let us consider this as a continuous function of K_c, i.e., define:

$$\hat{n}(K_c) = n(K_c) \ , \qquad 0 < K_c < 1 \tag{6.3}$$

Then by minimizing P, we obtain a value for $K_c = K_{opt}$ given by

$$\frac{\partial P}{\partial K_c} = 0 \qquad \frac{\partial}{\partial K_c} \qquad \frac{A + B\hat{n}(K_c)}{K_c} = 0 \tag{6.4}$$

Once we have obtained K_{opt} by solving (6.4) (which is very easy for all the methods of Section II.4), we use a linear interpolation to determine a better choice for $\Delta\lambda$. We have $K_i(0) = 0$, $K_c(\Delta\lambda) = K$. (Note K may be greater than 1, indicating the method may be diverging at $\lambda + \Delta\lambda$.) Thus we determine $\Delta\lambda_2$ such that $K_c(\lambda + \Delta\lambda_2) = K_{opt}$.

If the iterations fail to converge at $\Delta\lambda_2$, the process can be repeated, or a higher order interpolation scheme could be used since we now have values of K_c for three $\Delta\lambda$ points. Since we do not compute a new Jacobian during this process, the cost of obtaining K_c is order n^2. Thus several failures are not significant in comparison with the cost of factoring the system (for n large).

II.6.2 Computational Scheme II

The following scheme is quite similar to a method suggested recently by Rheinboldt [61], but more attention is paid here to obtaining an estimate for the radius of convergence. Also, higher order extrapolation methods are incorporated.

Define:

$$\begin{aligned} &\text{a)} \quad S_N(\underset{\sim}{x}^\nu) = \underset{\sim}{x}^\nu - J^{-1}(\underset{\sim}{x}^\nu)f(\underset{\sim}{x}^\nu) \\ &\text{b)} \quad S_c(\underset{\sim}{x}^\nu) = \underset{\sim}{x}^\nu - J^{-1}(\underset{\sim}{x}^\circ)f(\underset{\sim}{x}^\nu) \end{aligned} \tag{6.5}$$

Then Newton's method is

$$\underset{\sim}{x}^{\nu+1} = S_N(\underset{\sim}{x}^{\nu}) \tag{6.6}$$

and the modified chord method is

$$\underset{\sim}{x}^{\nu+1} = S_c(\underset{\sim}{x}^{\nu}) \tag{6.7}$$

The function S in the following may refer to either S_N or S_c. For convergence of (6.6) or (6.7) we need

$$\| S(\underset{\sim}{x}) - S(\underset{\sim}{y}) \| \leq \rho \quad \| \underset{\sim}{x} - \underset{\sim}{y} \| \tag{6.8}$$

$$\text{for } \underset{\sim}{x}, \underset{\sim}{y} \in \mathcal{S}_\sigma(\underset{\sim}{x}^\circ)$$

where $0 \leq \rho < 1$, $\sigma > 1$ and $\mathcal{S}_\sigma = \{\underset{\sim}{x} | \; \|\underset{\sim}{x} - \underset{\sim}{x}^\circ\| \leq \sigma\}$. We shall use the local approximations to ρ, say ρ_ν, given by

$$\| S(\underset{\sim}{x}^{\nu}) - S(\underset{\sim}{x}^{\nu-1}) \| = \| \underset{\sim}{x}^{\nu+1} - \underset{\sim}{x}^{\nu} \| = \rho_\nu \; \| \underset{\sim}{x}^{\nu} - \underset{\sim}{x}^{\nu-1} \| \tag{6.9}$$

If the method is converging, we obtain

$$\frac{\| \Delta \underset{\sim}{x}^{\nu+1} \|}{\| \Delta \underset{\sim}{x}^{\nu} \|} = \frac{\| \underset{\sim}{x}^{\nu+1} - \underset{\sim}{x}^{\nu} \|}{\| \underset{\sim}{x}^{\nu} - \underset{\sim}{x}^{\nu-1} \|} = \rho_\nu \tag{6.10}$$

Or, expressing ρ_ν as a function of $\varepsilon^\nu = \| \underset{\sim}{x}^\nu - \bar{\underset{\sim}{x}} \|$, we have

$$\frac{\| \varepsilon^{\nu+1} - \varepsilon^{\nu} \|}{\| \varepsilon^{\nu} - \varepsilon^{\nu-1} \|} = \rho_\nu \tag{6.11}$$

Thus we have the two sets of data $\{\| \varepsilon^\nu \|\}$ and $\{\rho_\nu\}$. We may extrapolate to find an approximation to the radius of convergence at λ, say $r_\lambda = \| \bar{\varepsilon} \|$, where $\| \bar{\varepsilon} \|$ is the extrapolated value for which $\rho_\nu = 1$.

If we have taken some finite number of steps i, then we have the values $\left\{ r_{\lambda_j} \right\}_{j=1}^{i}$. We may now extrapolate r_{λ_j} versus λ_j to obtain an estimate of the radius of convergence for $\lambda > \lambda_i$. We shall denote this curve by $\bar{r}(\lambda)$.

Assuming that we now have $\bar{x}(\lambda_j)$ for $j \leq i$, where $\bar{x}$ and λ_j satisfy (2.1), then we attempt to approximate $\bar{x}(\lambda_i + \Delta\lambda)$ by

$$\bar{\underset{\sim}{x}}(\lambda_i + \mu) = \bar{\underset{\sim}{x}}(\lambda_i) + \mu\dot{\bar{\underset{\sim}{x}}}(\lambda_i) + \frac{\mu^2}{2} \ddot{\underset{\sim}{x}} (z) \qquad (6.12)$$

where $\lambda_i \leq z \leq \lambda_i + \mu$

Then using (4.1) a) as initial guess, we should choose μ so that

$$\| \underset{\sim}{x}_a(\lambda_i + \mu) - \bar{\underset{\sim}{x}}(\lambda_i) \| < \bar{r} (\lambda_i + \mu) \qquad (6.13)$$

Thus to first order we obtain

$$\mu \| \dot{\bar{\underset{\sim}{x}}}(\lambda_i) \| < \bar{r}(\lambda_i + \mu) \qquad (6.14)$$

If $\bar{r}$ is linear, quadratic, or cubic, we can solve explicitly for μ. It is obvious that for higher order approximations, one obtains

$$\mu^n \| \bar{\underset{\sim}{x}}^{(n)} (\lambda_i) \| < n! \, \bar{r} (\lambda_i + \mu). \qquad (6.15)$$

II.7 INTERMEDIATE ERROR TOLERANCES

For the homotopy problem, we do not desire the solution at intermediate values of λ. However, it is necessary to compute the solution accurately enough so that we may proceed. Thus we should examine the choice of intermediate error tolerances ε_F. Let us assume that we have a priori decided on the steps to be taken, i.e., we have $\{\lambda_i\}_{i=1}^N$. For Newton's method we may write the work in terms of ε_{F_i} as:

$$\hat{W} = \mathcal{K}\left[\log \frac{\log \varepsilon_{F_1}}{\log c} + \sum_{i=2}^{N-1} \log \frac{\log \varepsilon_{F_i}}{\log(c+\varepsilon_{F_{i-1}})} + \log \frac{\log \varepsilon_{F_N}}{\log(c+\varepsilon_{N-1})}\right] \qquad (7.1)$$

where ε_{F_N} is fixed, c is a bound on $\frac{(\Delta\lambda_i)^n \,\|\underset{\sim}{x}^{(n)}\|}{n!}$ throughout the interval, and $\mathcal{K}$ is a constant.

To minimize with respect to ε_{F_i} we take the derivatives with respect to these variables and set them equal to 0.

$$\frac{\partial W}{\partial \varepsilon_{F_i}} = 0 \Longrightarrow (c+\varepsilon_{F_i}) \log (c+\varepsilon_{F_i}) - \varepsilon_{F_i} \log \varepsilon_{F_i} = 0 \qquad)7.2)$$

for i = 1, ..., N-1.

This implies that we should choose all the ε_{F_i} to be the same, i.e., a uniform choice determined by (7.2). (This is a consequence of the fact that a uniform bound on c was imposed.)

Allowing the c's to vary, say c_i, will yield different values for ε_{F_i} at each step determined by (7.2) with c replaced by c_i. In either case, we cannot determine the optimal value for ε_{F_i} using calculable quantities.

REFERENCES

1. J.C. Alexander, "Bifurcation of Zeroes of Parameterized Functions", J. Functional Analysis, (to appear).

2. J.C. Alexander and James A. Yorke, "Calculating Bifurcation Invariants as Elements in the Homotopy of the General Linear Group", Journal of Pure and Applied Algebra, 13, (1978), 1-8.

3. J.C. Alexander and James A. Yorke, "The Homotopy Continuation Method: Numerically Implementable Topological Procedures", Transactions of the American Mathematical Society, Volume 242, (August 1978), 271-283.

4. Eugene Allgower and Kurt Georg, "Simplicial and Continuation Methods for Approximating Fixed Points and Solutions to Systems of Equations" SIAM Review, Volume 22, No. 1 (January 1980), 28-85.

5. P. Anselone and R. Moore, "An Extension of the Newton Kantorovich Method for Solving Non-linear Equations with an Application to Elasticity", Journal Math. Anal. Appl., Volume 13 (1966), 476-501.

6. J. Avila, "Continuation Methods for Nonlinear Equations", Ph.D. Thesis, University of Maryland, College Park (1971).

7. J. Avila, Jr., "The Feasibility of Continuation Methods for Nonlinear Equations", SIAM J. Numer. Anal., Volume II., No. 1 (March 1974), 102-122.

8. S. Bernstein, "Sur la Généralisation du Problème de Dirichlet (I)", Mathematische Annalen, Volume 62 (1906), 253-271; (II), Mathematische Annalen, Volume 69 (1910), 82-136.

9. L. Bittner, "Einige Kontinuierlich Analogien von Iterationsverfahren , Funktionalanalysis, Approximationstheorie," Numerische Mathematik, ISNM7 (1976), Birkhahser-Verlag, Basel, 114-135.

10. W. Bosarge,"Infinite Dimensional Iterative Methods and Applications", IBM Houston Sci. Center Rep. 320.2347 (1968), Houston, Texas.

11. W.E. Bosarge, Jr., "Iterative Continuation and the Solution of Nonlinear Two-Point Boundary Value Problems", Numer. Math., Volume 17 (1971), 268-283.

12. W.E. Bosarge, Jr., and C.L. Smith, "Some Numerical Results for Iterative Continuation in Nonlinear Boundary-value Problems", IBM J. Res. Develop. (July 1971), 323-327.

13. D.C. Brabston and H.B. Keller, "A Numerical Method for Singular Two-Point Boundary Value Problems", SIAM J. Numer. Anal., Volume 14, No. 5 (October 1977), 779-791.

14. B. Bunow and J.P. Kernevez, "Numerical Exploration of Bifurcating Branches of Solutions to Reaction-Diffusion Equations Describing the Kinetics of Immobilized Enzymes", (to appear).

15. James Callahan, "Singularities and Plane Maps", American Mathematical Monthly (March 1974), 211-240.

16. S.N. Chow, J. Mallet-Paret and J.A. Yorke, "Finding Zeroes of Maps: Homotopy Methods That are Constructive with Probability One", (to appear).

17. J. Cronin, "The Existence of Multiple Solutions of Elliptic Differential Equations", American Mathematical Society Transactions, Volume 66 (1949), 289-307.

18. D. Davidenko, "On a New Method of Numerically Integrating a System of Nonlinear Equations", Dokl. Akad. Nauk SSSR, Volume 88 (1953), 601-604 (In Russian).

19. D. Davidenko, "On the Approximate Solution of a System of Nonlinear Equations", Ukrain. Mat. Z., Volume 5 (1953), 196-206 (In Russian).

20. D.W. Decker and C.T. Kelley, "Newton's Method at Singular Points I", (to appear).

21. D.W. Decker and C.T. Kelley, "Newton's Method at Singular Points II", (to appear).

22. F. Deist and L. Sefor, "Solution of Systems of Nonlinear Equations by Parameter Variation", Comput. J., Volume 10 (1967), 78-82.

23. P. Deuflhard, "A Modified Newton Method for the Solution of Ill-Conditioned Systems of Nonlinear Equations with Application to Multiple Shooting", Numer. Math., Volume 22 (1974), 289-315.

24. P. Deuflhard, H.-J. Pesch, and P. Rentrop, "A Modified Continuation Method for the Numerical Solution of Nonlinear Two-Point Boundary Value Problems by Shooting Techniques", Numer. Math. 26 (1976), 327-343.

25. C.L. Dolph, "Nonlinear Integral Equations of the Hammerstein Type", American Mathematical Society Transactions, Volume 66 (1949), 289-307.

26. B.C. Eaves, "Homotopies for the Continuation of Fixed Points", Math. Programming, Volume 3 (1972), 1-22.

27. B.C. Eaves and H. Scarf, "The Solution of Systems of Piecewise Linear Equations", Math. Operations Research, Volume 1 (1976), 1-27.

28. B. Curtis Eaves, "A Short Course in Solving Equations with PL Homotopies", SIAM-AMS Proceedings, Volume 9 (1976), 73-144.

29. F.A. Ficken, "The Continuation Method for Functional Equations", Comm. Pure Appl. Math., Volume 4 (1951), 435-456.

30. C.W. Gear, "Initial Value Problems: Practical Theoretical Developments", (1978) (to appear).

31. Kurt Georg, "On Tracing an Implicitly Defined Curve by Quasi-Newton Steps and Calculating Bifurcation by Local Perturbations", (to appear in SIAM J. on Sci. and Stat. Comp.).

32. F.J. Gould and J.W. Tolle, " A Unified Approach to Complementarity in Optimization", Discrete Math. Volume 7 (1974), 225-271.

33. Andreas Griewank and M.R. Osborne, "On Newton's Method for Singular Problems", (1978) (to appear).

34. J. Hadamard, "Sur les transformations ponctuelles", Bulletin de la Société Mathématique de France, Volume 34 (1906), 71-84.

35. S. Karamardian, ed., Fixed Points: Algorithms and Applications, Proc. Conf. on Computing Fixed Points with Applications (1977), Clemson University, Academic Press, New York.

36. Herbert B. Keller, "Global Homotopies and Newton Methods", Recent Advances in Numerical Analysis (ed., C. de Boor and G.H. Golub) (1978), Academic Press, New York.

37. Herbert B. Keller, "Numerical Solution of Bifurcation and Nonlinear Eigenvalue Problems", in Applications of Bifurcation Theory, (1977) Academic Press, Inc., New York.

38. R.B. Kellogg, T.Y. Li, and J.A. Yorke, "A Method of Continuation for Calculating a Brouwer Fixed Point", Fixed Points, Algorithms, and Applications (S. Karamardian, ed.) (1977), Academic Press, New York.

39. H. Kleinmichel, "Stetige Analoga und Iterationsverfahren für Nichtlineare Gleichungen in Banachraumen", Math. Nachr., Volume 37 (1968), 313-344.

40. E. Lahaye, "Solution of Systems of Transcendental Equations", Acad. Roy. Belg. Bull. Cl. Sci., Volume 5 (1948), 805-522.

41. E. Lahaye, "Sur la Representation des Racines Systems d'Équations Transcendantes", Deuxieme Congres National Des Sciences, Volume 1, (1935), 141-146.

42. E. Lahaye, "Solution of Systems of Transcendental Equations", Acad. Roy. Belg. Bull. Cl. Sci., Volume 5 (1948), 805-822.

43. J. Leray, "Les Problèmes de Representation Conforme d'Helmholtz; Theorie des Sillages et Des Prones", Commentarii Mathematici Helvetici, Volume 8 (1935-6), 149-180, 250-263.

44. J. Leray and J. Schauder, "Topologie et Équations Fonctionelles", Paris Ecole Normale Superieure, Annales Scientifiques (1934), 45-78.

45. P. Lévy, "Sur les Fonctions de Lignes Implicites", Bulletin de la Sociêtê Mathêmatique de France, Volume 48 (1920), 13-27.

46. H. Lewy, "On the Existence of a Closed Convex Surface Realizing a Given Riemannian Metric", National Academy of Sciences, Proceedings, Volume 24 (1938), 104-106.

47. H. Lewy, "On Differential Geometry in the Large, I (Minkowski's Problem)", American Mathematical Society Transactions, Volume 43 (1938), 258-270.

48. T.Y. Li, "A Rigorous Algorithm for Fixed Point Computation", (to appear).

49. L. Lichtenstein, Kontinuitätsmethode im Gebiete der Konformen Abbildung, Encyklopädie der Mathematischen Wissenschaften, IIC3, (Potentialtheorie, Konforme Abbildung)§46 (1918), 346-352.

50. H.J. Lüthi, "Komplementaritäts - und Fixpunktalgorithmen in der Mathematischen Programmierung, Spieltheorie und Ökonomie", Lecture Notes in Economic and Mathematical Systems, No. 129 (1976), Springer-Verlag, Heidelberg-New York.

51. G.P. McCormick, "Optimality Criteria in Nonlinear Programming", in Nonlinear Programming, SIAM-AMS Proceedings, Volume IX (1976), AMS, Providence.

52. G. Meyer, "On Solving Nonlinear Equations with a One-Parameter Operator Embedding", SIAM J. Numer. Anal., Volume 5 (1968), 739-752.

53. J.M. Ortega and W.C. Rheinboldt, Iterative Solutions of Nonlinear Equations in Several Variables , Academic Press, New York-London.

54. H.O. Peitgen, Ed., Approximation of Fixed Points and Functional Differential Equations , Springer Lecture Notes (1979), Springer-Verlag, New York.

55. H.O. Peitgen and M. Prüfer, "A Constructive Approach to Global Bifurcation, the Schauder Continuation Method and Coincidence Problems", Approximation of Fixed Points and Functional Differential Equations ,H.O. Peitgen, ed., Springer Lecture Notes (1979) Springer-Verlag, New York.

56. L.B. Rall, "Convergence of the Newton Process to Multiple Solutions", Numerische Mathematik, Volume 9 (1966), 23-37.

57. L.B. Rall, "Davidenko's Method for the Solution of Nonlinear Operator Equations", MRC Tech. Summary Rep. 948 (1968), University of Wisconsin, Madison.

58. G.W. Reddien, "On Newton's Method for Singular Problems", SIAM J. Numer. Anal., Volume 15, No. 5 (October 1978), 993-996.

59. G.W. Reddien, "Newton's Method and High Order Singularities", (to appear).

60. Werner C. Rheinboldt, "A Unified Convergence Theory for a Class of Iterative Processes", SIAM J. Numer. Anal., Volume 5, No. 1 (March 1968), 42-63.

61 Werner C. Rheinboldt, "Solution Field of Nonlinear Equations and Continuation Methods", Technical Report ICMA-79-04, Institute for Computational Mathematics and Applications, Department of Mathematical Statistics, University of Pittsburgh (March 1979).

62. S.M. Roberts and J.S. Shipman, "Continuation in Shooting Methods For Two-Point Boundary Value Problems", Journal of Mathematical Analysis and Applications, Volume 18 (1967), 45.58.

63. E. Rothe, "Zur Theorie der Topologischen Ordnung und der Vektorfelder in Banachschen Räumen", Compositio Mathematica, Volume 5, (1937-8), 177-197.

64. E. Rothe, "Topological Proofs of Uniqueness Theorems in the Theory of Differential and Integral Equations", American Mathematical Society Bulletin, Volume 45 (1939), 606-613.

65. R. Saigal,"Fixed Point Computing Methods", Encyclopedia of Computer Science and Technology, Volume 8 (1977), Marcel-Dekher Inc., New York.

66. Herbert Scarf, "The Approximation of Fixed Points of a Continuous Mapping", SIAM J. Appl. Math., Volume 15, No. 5 (September 1967), 1328-1343.

67. J. Schauder, "Einige Anwendungen der Topologie der Funktionalräume", Recueil Mathématique (Sbornik), Volume 43 (1936), 747-753.

68. J. Schauder, "Über Lineare Elliptische Differential Gleichungen Zweiter Ordnung", Mathematische Zeitschrift, Volume 38 (1934), 257-282; in particular, 278.

69. Karl Scherer, "On the Best Approximation of Continuous Functions by Splines", SIAM Journal on Numerical Analysis, Volume 7 (1970), 418-423.

70. L. Schlesinger, "Zur Theorie der Linearen Differentialgleichungen im Anschluss an das Riemannsche Problem (Dritte Abhandlung)", Journal fur die Reine und Angewandte Mathematik, Volume 130 (1905), 26-46.

71. C. Schmidt, "Approximating Differential Equations that Describe Homotopy Paths", Report 7931, Center for Mathematical Studies in Business and Economics, Department of Economics and Graduate School of Business, University of Chicago (1979).

72. William F. Schmidt, "Adaptive Step Size Selection for Use with the Continuation Method", International Journal for Numerical Methods in Engineering, Volume 12 (1978), 677-694.

73. Rüdiger Seydel, "Numerical Computation of Branch Points in Nonlinear Equations", TUM-MATH 7907, Technische Universität München, Institut Für Mathematik (March 1979).

74. N. Sidlovskaya, "Application of the Method of Differentiation with Respect to a Parameter to the Solution of Nonlinear Equations in Banach Spaces", Leningrad Gos. Univ. Ucen. Zap. Ser. Mat. Nauk, 33 (1958), 3-17 (In Russian).

75. S. Smale, "A Convergent Process of Price Adjustment and Global Newton Methods", J. Math. Econ., 3 (1976), 1-14.

76. R. Szeto, "Corrected Newton Process - Stutter Theorem", (to appear).

77. G.A. Thurston, "Continuation of Newton's Method Through Bifurcation Points", Journal of Applied Mechanics (September 1969), 425-430.

78. M.J. Todd, "The Computation of Fixed Points and Applications", Springer Lecture Notes in Economics and Mathematical Systems, 124 (1976), Springer-Verlag, Heidelberg-New York.

79. M. van Veldhuizen, "A Note on Partial Pivoting and Gaussian Elimination", Numer. Math., 29 (1977), 1-10.

80. H. Wacker, ed., Continuation Methods (1978), Academic Press, New York.

81. Hansjörg Wacker, E. Zarzer, and W. Zulehner, "Optimal Stepsize Control for the Globalized Newton Method", in Continuation Methods (1978), Academic Press, Inc.

82. E. Wasserstrom, "Numerical Solutions by the Continuation Method", SIAM Review, Volume 15, No. 1 (January 1973), 89-119.

83. A. Weinstein, "Der Kontinuitätsbeweis des Abbildungssatzes für Polygone", Mathematische Zeitschrift, Volume 21 (1924), 72-84.

84. H. Weyl, "Über die Bestimmung einer Geschlossen Konvexen Fläche durch ihr Linienelement", Vierteljahrschrift der Natur forschenden Gesellschaft in Zürich, Volume 61 (1916), 40-72.

85. Olof Widlund, "On Best Bounds for Approximation by Piecewise Polynomial Functions", Numer. Math. Volume 27 (1977), 327-338.

86. M.N. Yakovlev, "On Some Methods of Solving Nonlinear Equations", Trudy Mat. Inst. Steklov., Volume 84 (1965), 8-40; English translation, TR68-75, University of Maryland, College Park (1968).

APPENDIX

One example of each type of inversion possible is illustrated in the series of following plots. In all cases the horizontal scale is the iteration number and the vertical scale is the maximum norm of the residual. The initial guesses used are off on the order of 100%. The exact solution is calculated using the data from Example II.7.5. Five receivers are utilized and the 18 most direct rays are traced for each receiving station. Thus 90 rays are traced at each iteration. The criterion used for convergence is that all components of the residual have an absolute relative error of less than 1%.

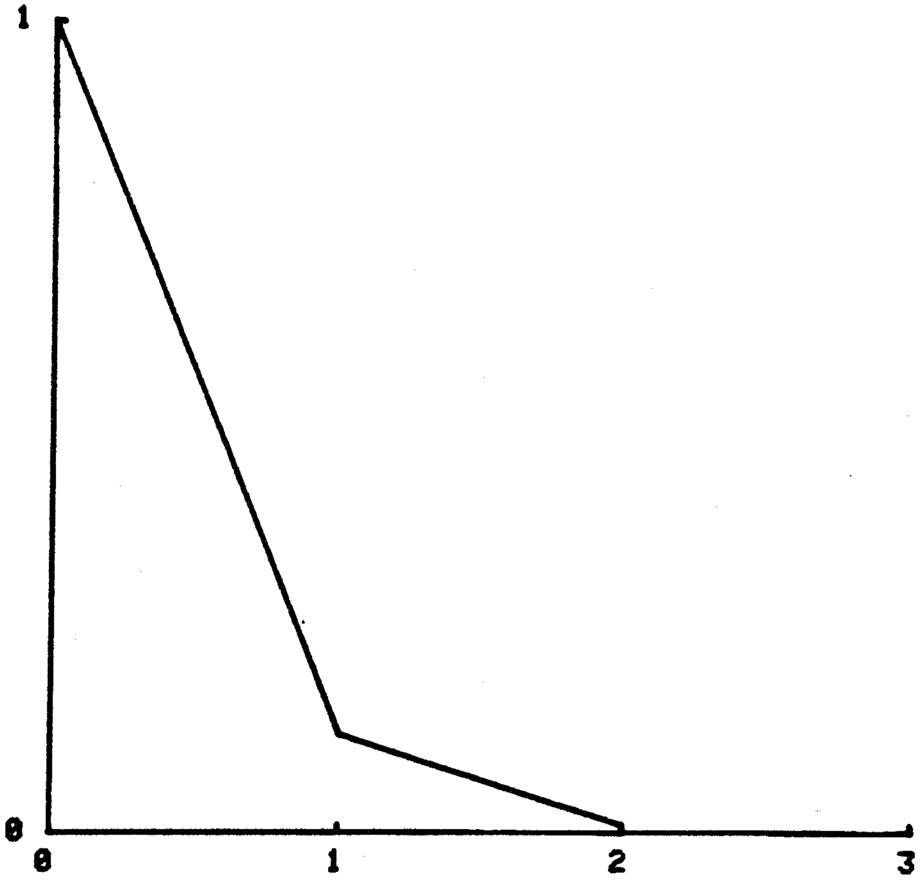
Plot 1 illustrates inversion for source location using only travel times. The process converges in 2 iterations to within the specified tolerance.

Inversion for the elastic parameters is shown in Plots 2 - 3. Plot 2 uses travel times only and Plot 3 both travel time and amplitude. This is a 3-parameter inversion as the parameters μ, λ, and ρ are also assumed unknown. Ten iterations were required for Plot 2 and 12 iterations for Plot 3.

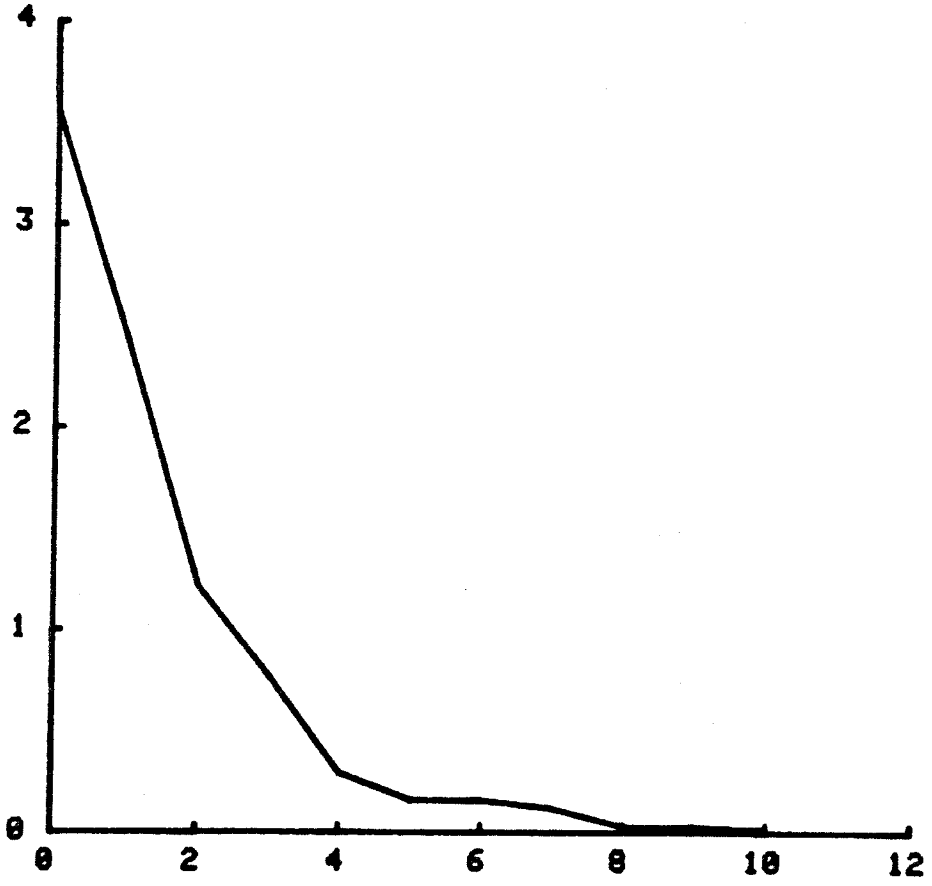
Plot 4 (travel time) and Plot 5 (travel time and amplitude) represent the inversion for interface shapes. Both required 4 iterations for convergence. There are 4 unknown parameters, the constants in the cubic representing the interface.

The following table explains the inversions depicted in Plots 6 - 13 (TT = travel time, A = amplitude).

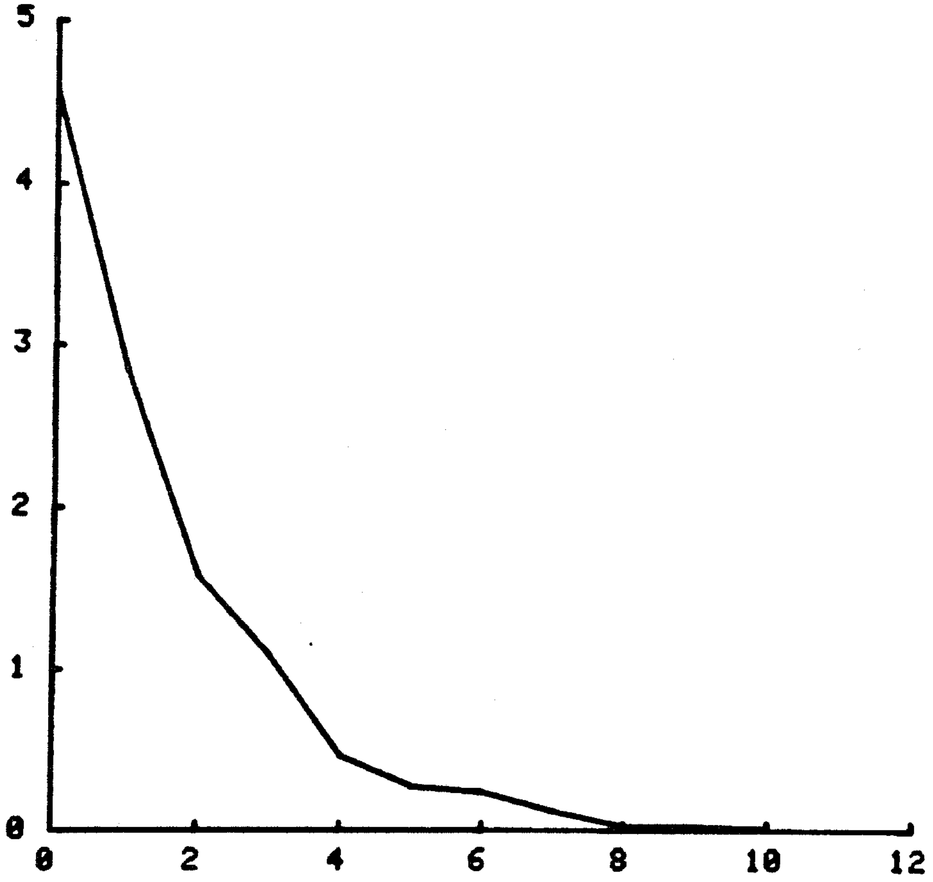
Plot #	Inversion Parameters	Data	Iteration Required to Converge
6	hypocenter, elastic parameters	TT only	6
7	hypocenter, elastic parameters	TT and A	8
8	hypocenter, interfaces	TT only	5
9	hypocenter, interfaces	TT and A	10
10	elastic parameters, interfaces	TT only	14
11	elastic parameters, interfaces	TT and A	9
12	hypocenter, elastic parameters, interfaces	TT only	17
13	hypocenter, elastic parameters, interfaces	TT and A	14



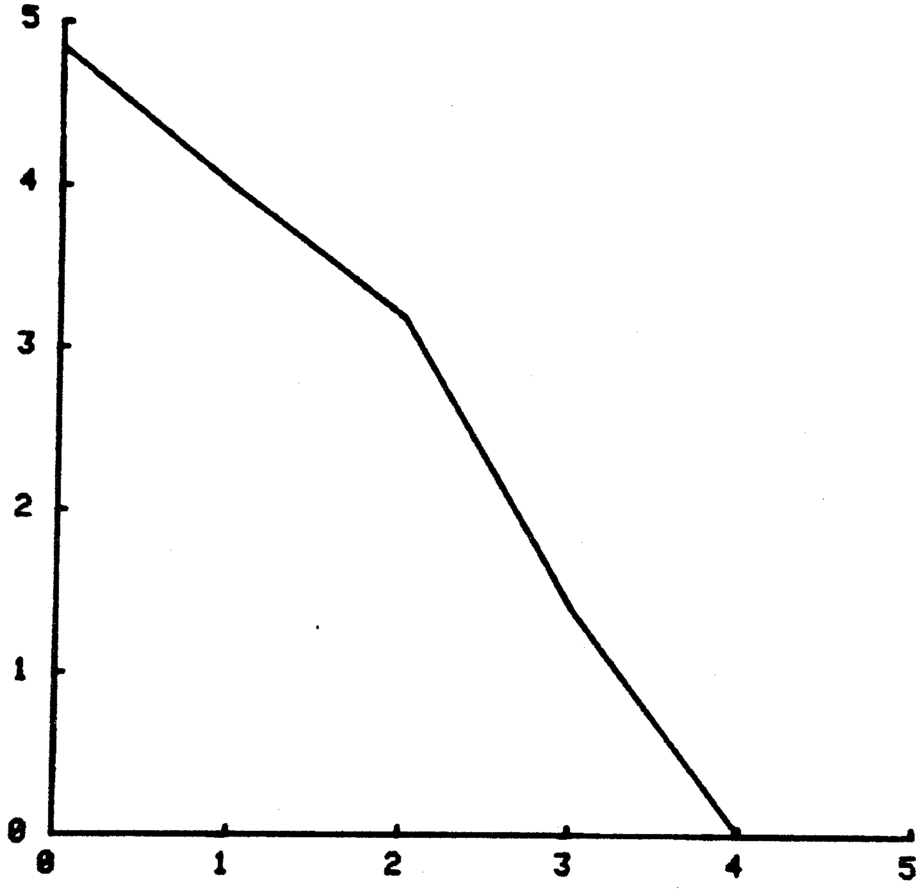
PLOT 1



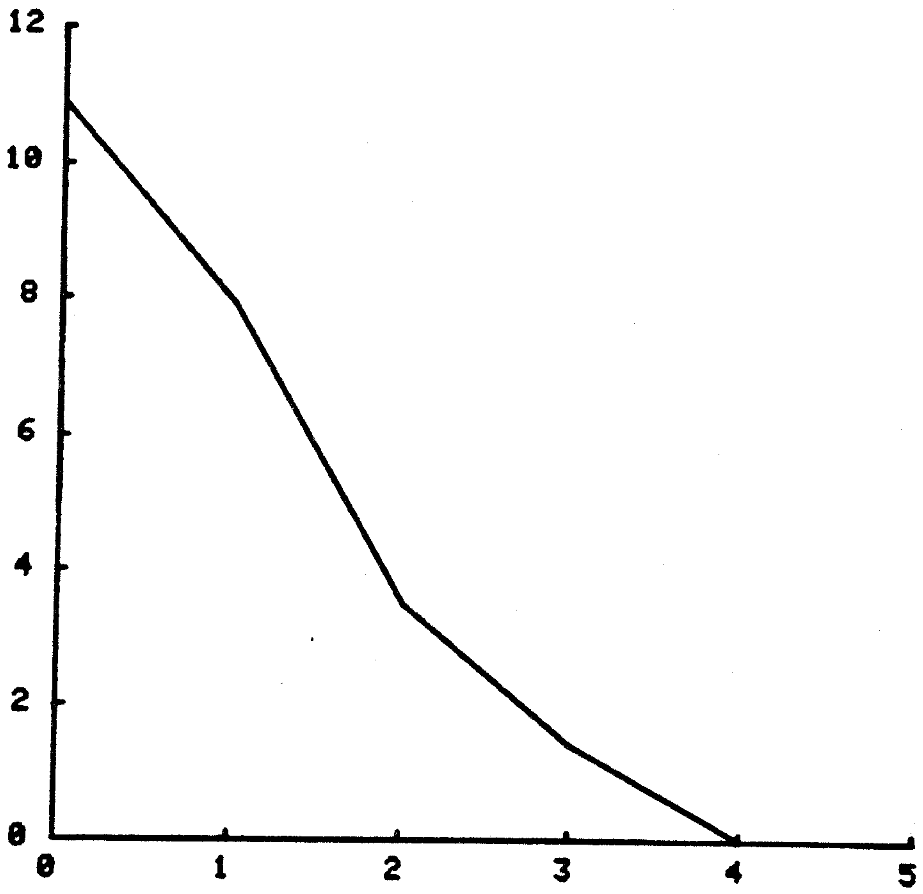
PLOT 2



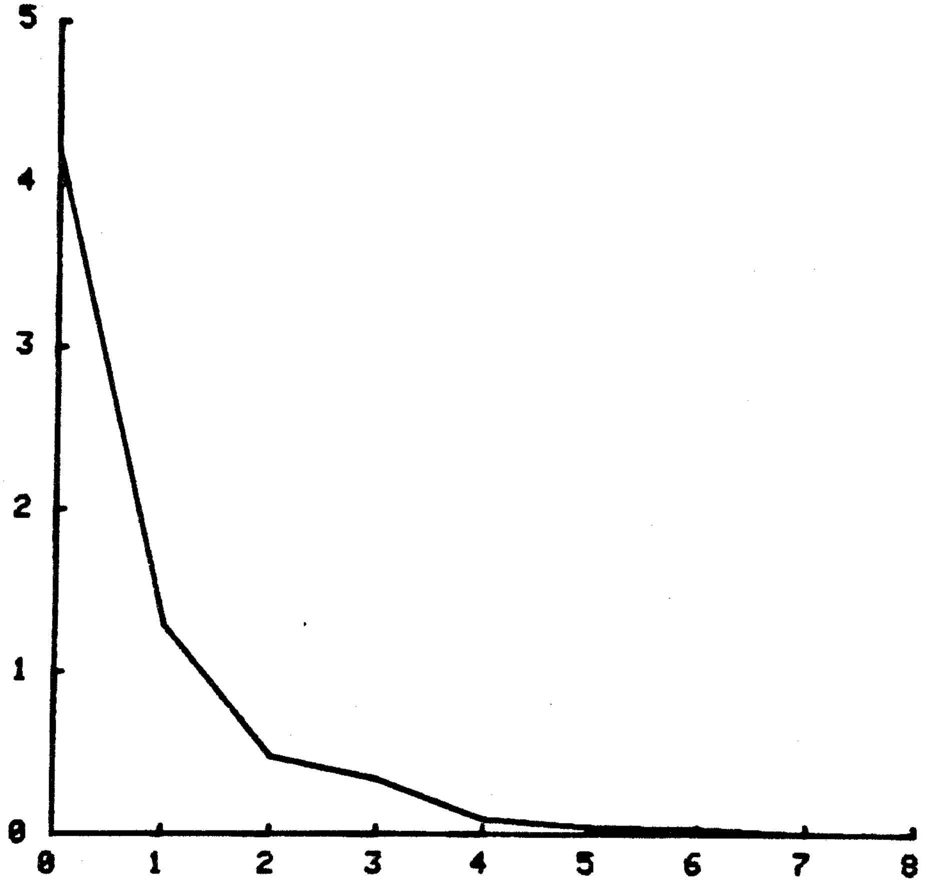
PLOT 3



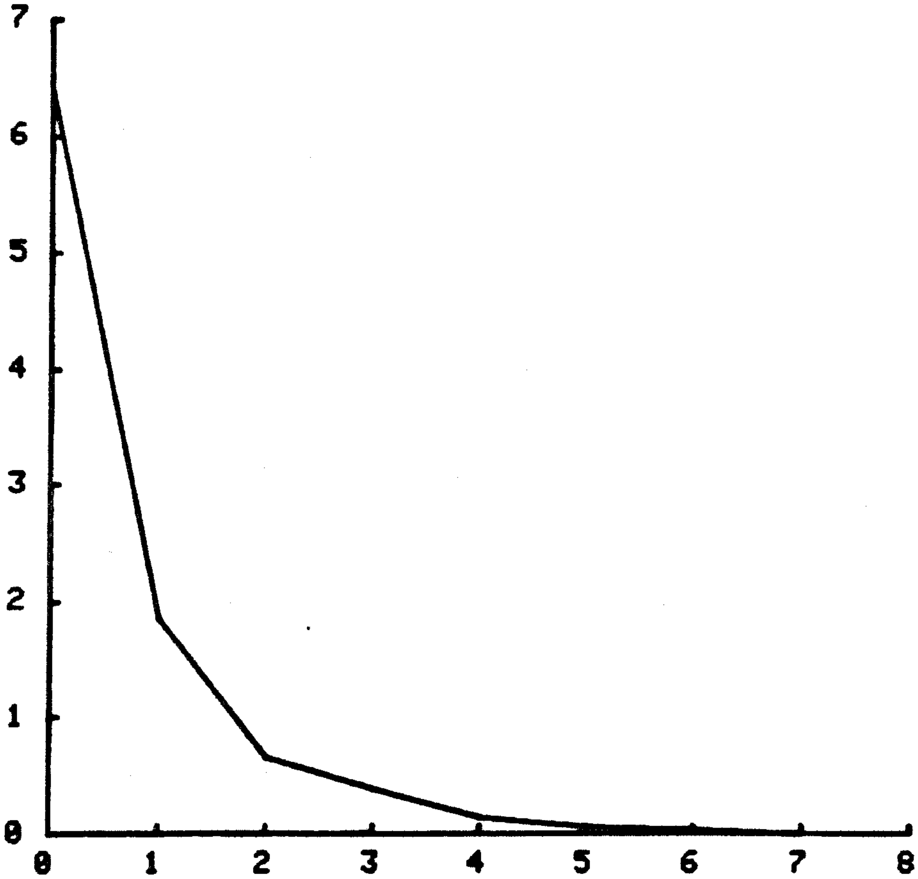
PLOT 4



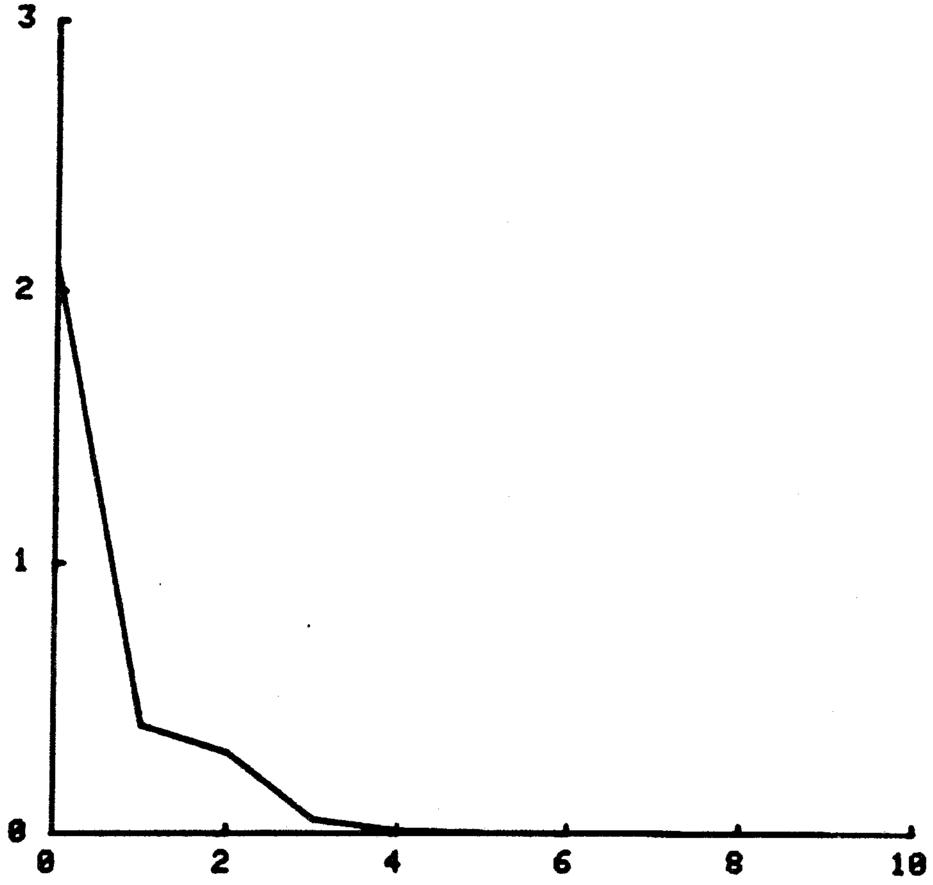
PLOT 5



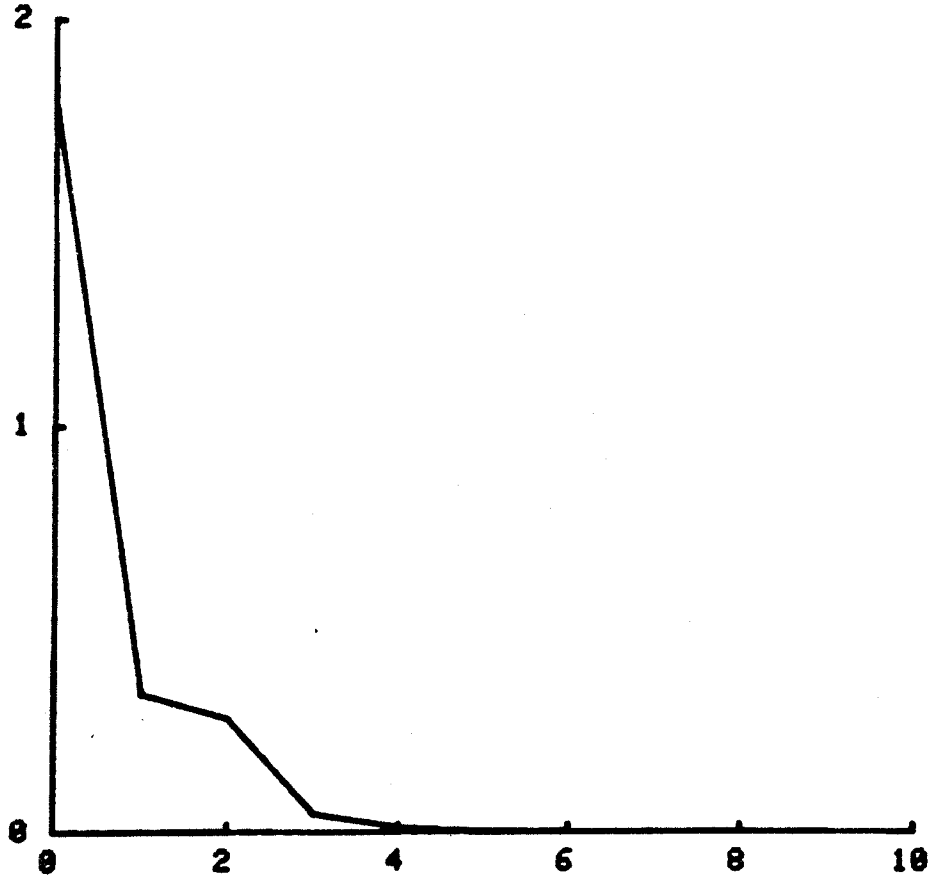
PLOT 6



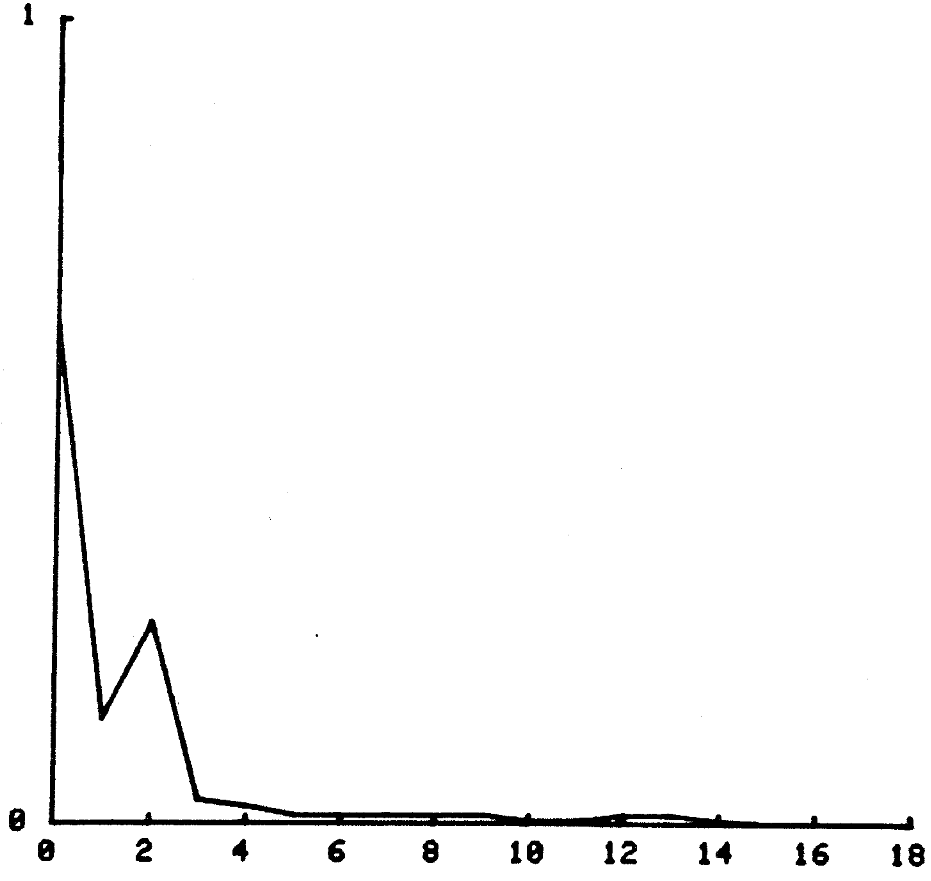
PLOT 7



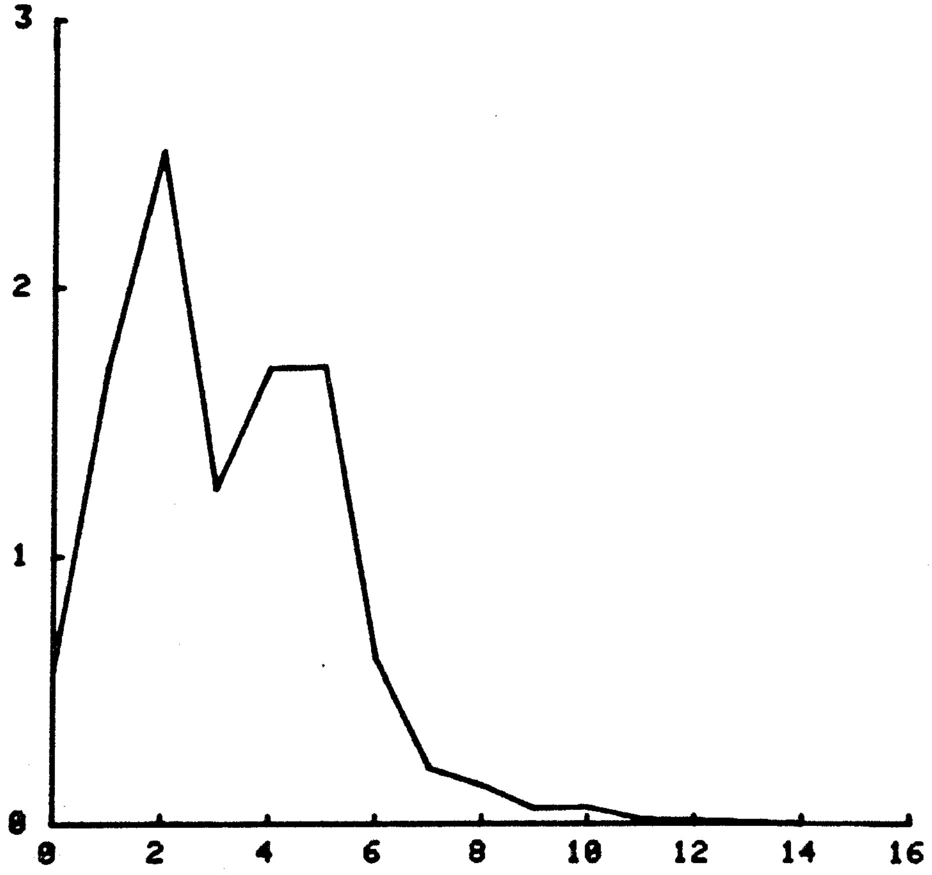
PLOT 8



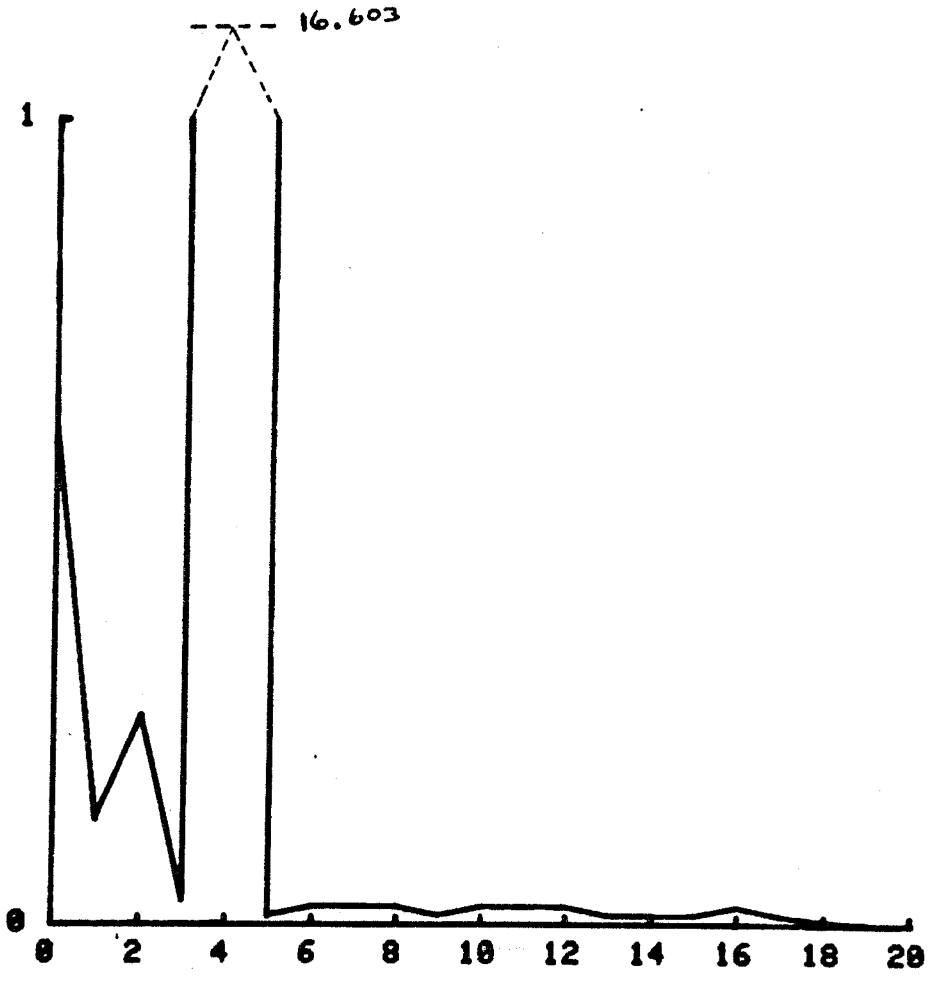
PLOT 9



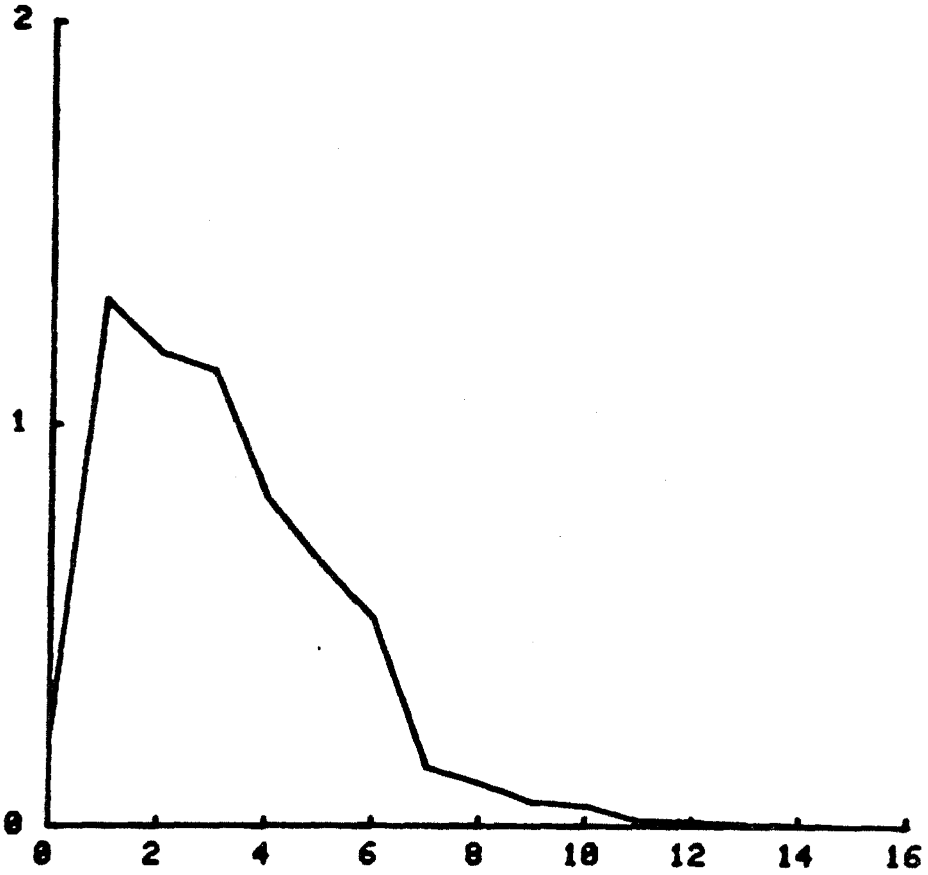
PLOT 10



PLOT 11



PLOT 12



PLOT 13