THE APPLICATION OF THE FINITE ELEMENT METHOD TO THE NEUTRON TRANSPORT EQUATION

by

William Russell Martin

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Nuclear Engineering)
in The University of Michigan
1976

Doctoral Committee:

Professor James J. Duderstadt, Chairman
Professor Ziya A. Akcasu
Assistant Professor John C. Lee
Assistant Professor Dianne O'Leary
Professor George J. Fix (Carnegie-Mellon University)

ABSTRACT

THE APPLICATION OF THE FINITE ELEMENT METHOD

TO THE NEUTRON TRANSPORT EQUATION

by

William Russell Martin

Chairman: James J. Duderstadt

This paper examines the theoretical and practical application of the finite element method to the neutron transport equation.

The theoretical examination which is applicable to the general transport equation in arbitrary geometry includes a derivation of the equivalent integral law (or weak form) of the first order neutron transport equation, to which the finite element method (Galerkin approach) is applied, resulting in a system of algebraic equations. We show that in principle the system of equations can be solved with certain physical restrictions concerning the criticality of the medium. The convergence of this approximate solution to the exact solution with mesh refinement is examined, and a non-optimal estimate of the convergence rate is obtained analytically. It is noted that the numerical results indicate a faster convergence rate and several approaches to obtain this result analytically are outlined.

The practical application of the finite element method involved the development of a computer code capable of solving the neutron transport equation in 1-D plane geometry. Vacuum, reflecting, or specified incoming boundary conditions may be analyzed, and all are treated as natural boundary conditions. The incorporation of the reflecting boundary conditions is seen to result in an ambiguity, which must be resolved by

consideration of the direction in which neutrons travel. Discontinuous phase space finite elements are introduced, and it is seen that discontinuous angular elements effectively match the analytical discontinuities in the angular flux at $\mu = 0$ for plane geometry. In addition, the use of discontinuous spatial elements is shown to result in treating continuity of the angular flux at an interface as a natural interface condition in the direction of neutron travel.

The time-dependent transport equation is also examined and it is shown that the application of the finite element method in conjunction with the Crank-Nicholson time discretization method results in a system of algebraic equations which is readily solved.

Numerical results are given for several critical slab eigenvalue problems, including anisotropic scattering, and the results compare extremely well with benchmark results. It is seen that the finite element code is more efficient than a standard discrete ordinates code for certain problems. Precise numerical tests are made on the convergence rate of the approximate solution (L_2 norm) with mesh refinement and also with the eigenvalue error. These results indicate $O(h^{k+1})$ solution error in the L_2 norm and $O(h^{2k+1})$ error in the eigenvalue, where h is the mesh spacing and k the degree of the finite element. A problem with severe heterogeneities is considered and it is shown that the use of discontinuous spatial and angular elements results in a marked improvement in the results. Finally, time-dependent problems are examined and it is seen that the phenomenon of angular mode separation makes the numerical treatment of the transport equation in slab geometry a considerable challenge, with the result that the angular mesh has a dominant effect on obtaining acceptable solutions to the time-dependent transport equation.

ACKNOWLEDGMENT

The author wishes to thank his wife, Patricia Ann, for her constant support and encouragement throughout this investigation. Her patience was clearly needed and deeply appreciated. The author is also grateful to his daughter, Amy Leigh, who has been quite a joy to the author during his research.

The author wishes to express his appreciation for the aid and encouragement of his thesis committee chairman, Dr. James Duderstadt, during the course of this investigation. His ready availability to field questions, along with his physical and mathematical insight that resulted in helpful answers, were instrumental to the author in his research.

The author also wishes to express his gratitude to Dr. George Fix, whose help during the early stages of this investigation was invaluable. His numerical analysis course, Math 571-572, was undoubtedly the highlight of the author's long association with mathematics courses at The University of Michigan, and was a constant reference throughout this investigation.

The author also gives thanks to Dr. Ziya Akcasu, Dr. John Lee, and Dr. Dianne O'Leary, who participated as members of the author's thesis committee.

Financial aid during this research, which was provided by the National Science Foundation (Grant ENG-7303868) and the Horace H. Rackham School of Graduate Studies (Post-Doctoral Fellowship), was deeply appreciated.

Finally, Madelyn Hudkins deserves a word of thanks for her rapid and accurate typing of the thesis.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (continued)

LIST OF NOMENCLATURE

x,y,z	usual orthogonal axes
$\underline{r}$	spatial coordinate, specified by coordinates (x,y,z)
θ	polar angle with respect to x-axis
ϕ	azimuthal angle about x-axis
μ	cosine of polar angle θ, $\mu = \cos\theta$
$\hat{\Omega}$	direction of neutron motion, specified by angular coordinates (θ,ϕ)
R	spatial domain
∂R	boundary of R
4π	angular domain
V	$R \times 4\pi$, phase space. A position in phase space is specified by $(\underline{r},\hat{\Omega})$
$\Gamma = \Gamma^+ \cup \Gamma^-$	$\partial R \times 4\pi$, incoming and outgoing boundaries of V
t	time
$\hat{n}$	unit normal to ∂R
$Q(\underline{r},\hat{\Omega},t)$	angular neutron flux at $(\underline{r},\hat{\Omega})$. The number of neutrons per unit area moving in direction $\hat{\Omega}$ about $d\hat{\Omega}$ at $\underline{r}$ at time t
$S(\underline{r},\hat{\Omega},t)$	volumetric source of neutrons at $(\underline{r},\hat{\Omega})$. The number of neutrons produced at $\underline{r}$ in $d\underline{r}$ at $\hat{\Omega}$ in $d\hat{\Omega}$ at time t
$\Sigma_t(\underline{r})$	total macroscopic cross-section for collision at $\underline{r}$, cm^{-1}
$\Sigma_f(\underline{r})$	total macroscopic cross-section for fission at $\underline{r}$, cm^{-1}

$\Sigma_a(\underline{r})$	total macroscopic cross-section for absorption collision at $\underline{r}$, cm^{-1}
$\Sigma_s(\underline{r}, \hat{\Omega}' \to \hat{\Omega})$	differential macroscopic scattering cross-section at $\underline{r}$ to scatter neutrons from $\hat{\Omega}'$ to $\hat{\Omega}$
$\Sigma_s(\underline{r})$	total macroscopic scattering cross-section at $\underline{r}$, cm^{-1} $\Sigma_s(\underline{r}) = \int_{4\pi} d\hat{\Omega}' \, \Sigma_s(\underline{r}, \hat{\Omega}' \to \hat{\Omega})$
$\mathcal{H}^0$	space of trial functions which are square integrable over V. A zero order Sobolev space
$\mathcal{H}^1$	space of trial functions which have square integrable gradients over V. A first order Sobolev space
(Q, Ψ)	inner product, $\iint_V d\underline{r} \, d\hat{\Omega} \, Q(\underline{r}, \hat{\Omega}) \Psi(\underline{r}, \hat{\Omega})$
$\|Q\|_0$	L_2 norm, $(Q, Q)^{1/2}$
$\|Q\|_1$	one norm, $\left[(Q, Q) + (\nabla Q, \nabla Q)\right]^{1/2}$
$\langle Q, \Psi \rangle_\pm$	incoming(-) and outgoing (+) boundary inner products, $\int_{\Gamma^\pm} d\Omega \, d\hat{\Omega} \, \lvert\hat{\Omega} \cdot \hat{m}\rvert \, Q(\underline{r}, \hat{\Omega}) \Psi(\underline{r}, \hat{\Omega})$
$\langle Q \rangle_\pm$	boundary norms, $\left[\langle Q, Q \rangle_\pm\right]^{1/2}$ (actually semi-norm, because $\langle Q \rangle_\pm = 0$ does not $\Rightarrow Q = 0$)
$\Psi(\underline{r}, \hat{\Omega})$	arbitrary test function in $\mathcal{H}^1$
K	collision operator, defined $K\Psi \equiv \Sigma_t(\underline{r}) \Psi(\underline{r}, \hat{\Omega}) - \int_{4\pi} d\hat{\Omega}' \Sigma_s(\underline{r}, \hat{\Omega}' \to \hat{\Omega}) \Psi(\underline{r}, \hat{\Omega}')$
$a(Q, \Psi)$	bilinear form
$Q_h(\underline{r}, \hat{\Omega})$	finite element approximation to actual solution

S^h	finite element subspace of H^1
h	parameter dependent on mesh spacing (as mesh is refined, $h \to 0$)
$\underline{\underline{A}}$	matrix of coefficients for steady state problem
$\underline{\Phi}$	column vector of nodal fluxes
$\underline{s}$	column vector due to sources and incoming boundary conditions
$\underline{\underline{K}}$	matrix due to collision terms
$\underline{\underline{T}}$	matrix due to transport term
$\underline{\underline{B}}$	matrix due to boundary terms
$\underline{\underline{I}}$	matrix due to interface terms
	(note $\underline{\underline{A}} = \underline{\underline{T}} + \underline{\underline{D}} + \underline{\underline{B}} + \underline{\underline{I}}$)
c	eigenvalue for criticality problem
$\underline{\underline{M}}$	RHS matrix for generalized matrix eigenvalue problem, $\underline{\underline{A}}\,\underline{\Phi} = c\,\underline{\underline{M}}\,\underline{\Phi}$
$\|\underline{v}\|_\infty$	maximum norm, $\max\{|v_1|, |v_2|, \ldots, |v_N|\}$
$\|\underline{\underline{A}}\|_\infty$	subordinate matrix norm
v	magnitude of velocity of neutron, $v = \|\underline{v}\|$ and $\underline{v} = v\,\hat{\Omega}$
$\underline{\underline{P}}$	mass matrix in time-dependent finite element approximation, $\frac{1}{v}\,\underline{\underline{P}}\,\dot{\underline{\Phi}} + \underline{\underline{A}}\,\underline{\Phi}(t) = \underline{S}(t)$
$\underline{\Phi}(t)$	time-dependent vector of nodal fluxes
$\dot{\underline{\Phi}}(t)$	time derivative of $\underline{\Phi}(t)$
t^n	time at step n ($t^n = n\ \Delta t$ for uniform timesteps)
$\underline{\Phi}(n)$	$\underline{\Phi}$ at time t^n, $\underline{\Phi}(t^n)$
$\underline{\underline{L}}$	lower diagonal matrix with unit diagonal
$\underline{\underline{U}}$	upper diagonal matrix

CHAPTER I

INTRODUCTION

I.A. Opening Remarks

The finite element method is one of the most powerful and versatile methods available for solving partial differential equations encountered in engineering and physics. Originally created in the mid 1950's by civil engineers for application to the large and complicated problems that are commonly analyzed in structural mechanics, the finite element method in the time since has developed into a commonly used tool in such diverse areas as fluid mechanics, electrostatics, magnetohydrodynamics, neutron diffusion, and heat transfer. As noted by Zienkiewicz[1,2], one of the early investigators, the finite element method has in this relatively short period become a standard tool in engineering science.

The popularity of the finite element method has been in part due to the remarkable success most users have had with the method, although in many cases where complicated, irregular domains must be treated, the finite element method is the only recourse. This success with the finite element method is due to the firm mathematical foundations that the method is based upon, which guarantee success for a large class of problems. Although early investigators who employed the finite element method were certainly not aware of the mathematical basis for the method's success, they were guided by physical intuition and understanding that were later to be confirmed mathematically. As noted by Strang and Fix[3], once it was discovered that the finite element method was in fact a Ritz approximation wherein the potential energy was minimized over a class of suit-

able trial functions, the method obtained instantly a sound theoretical basis.

The finite element method is in actuality a special case of a very general method of obtaining approximate solutions to partial differential equations[4]--the expansion of the solutions in terms of a set (possibly infinite) of trial functions, where the expansion coefficients are determined by some criterion. For example, one could expand the solution in terms of polynomials of a given degree and then ask that the least squares error over the domain of interest be minimized by the approximate solution. Or, as in the classical Ritz method, one might require that the approximate solution minimize a functional that has the original partial differential equation as its Euler equation. In any case, the criterion (least squares error, minimization of a functional, etc.) determines how the expansion coefficients are to be determined and the approximate solution (generally) will be close in some sense to the actual solution being sought. The major departure the finite element method makes from the traditional trial function expansion methods is that the domain of interest is first subdivided into small regions, or finite elements. The solution is then expanded in terms of a specially constructed finite dimensional subspace (the finite element subspace). This space has as its basis a set of polynomials (or other convenient functions) which are individually local over a specific collection of finite elements. That is, the basis functions are non-zero only over a localized part of the physical domain. This local structure leads to computational advantages in that the resultant matrix of coefficients is sparse (i.e., contains a large number of zero entries). But perhaps more importantly, this

choice of local basis functions results in approximation of the solution by piece-wise polynomials, over the domain of interest, rather than an expansion in terms of functions defined (and non-zero) over the entire domain. As mathematicians have discovered in recent years, piece-wise approximation is generally preferable to region-wise approximation, and precise error bounds can be found which guarantee convergence of the approximate solution for quite general problems. Thus the choice of local basis functions defined over a few adjacent finite elements was originally motivated by convenience because interface conditions and boundary conditions are much more easily imposed on the trial functions over a small regular domain (i.e., a finite element) than over a large irregular domain that might characterize a typical problem. This choice has since proven to be an optimal choice in terms of approximation theory.

The finite element method then consists of expanding the solution in terms of trial functions which are piece-wise polynomials (or other functions) over the domain of interest. For most applications, the expansion coefficients are determined by a Ritz (variational) or Galerkin (residual) technique. In most applications it can be demonstrated theoretically that the approximate solution is as "close" (in an appropriate norm) to the actual solution as any member of the approximating subspace. But if the approximating subspace is a space of piece-wise polynomials of degree k, which are typical finite element subspaces, and the mesh spacing is h, then a priori one knows from approximation theory that the approximate solution will agree within $O(h^{k+1})$ of the exact

solution, if the exact solution is sufficiently well-behaved[3]. Thus the convergence of the finite element method is well-established for a large class of problems and one typically does not need to worry about convergence to the actual solution.

I.B. Application of the Finite Element Method in Nuclear Reactor Analysis

Although the development of the finite element method has paralleled the development of the nuclear industry and the resultant need for numerical solution of neutron diffusion and transport equations, it has largely been ignored by nuclear reactor analysts until relatively recently. In hindsight, this is somewhat surprising in view of the superior results obtained when Kaper, Leaf, and Lindeman[5] applied the finite element method to the 2-D neutron diffusion equation, compared with a conventional finite difference code. Hansen and Kang[6], in a recent review of the finite element method in reactor analysis, also note that numerical results have generally shown the overall superiority of the finite element method, at least when applied to the diffusion equation. These results included the investigations made by Kang and Hansen[7] and Semenza, Lewis, and Rossow[8]. But the situation with the neutron diffusion equation is quite different than with the neutron transport equation, which is the subject of this work. The neutron diffusion equation is similar to the partial differential equations for which the finite element method has been successful--elliptic, self-adjoint systems which result in symmetric, positive-definite matrices which are known a priori to have positive solutions which are guaranteed to converge to the actual solution as the mesh is refined. However, the neutron transport equation, which is not

self-adjoint and not elliptic (at least in the conventional sense), does not possess these desirable properties and is somewhat of an unknown quantity as far as the finite element method is concerned.

However, the need for a reliable, efficient means of solving the neutron transport equation is strong. In a review[9] of the limitations of existing transport methods, Lathrop notes that the finite element method is a promising method for solving the neutron transport equation in multi-dimensional form. In addition, Froehlich[10] notes that current applications of the finite element method to the 1-D and 2-D transport equations appear promising. The finite element method is regarded by some experts as having such promise for a number of reasons:

(1) Non-orthogonal meshes are readily treated since triangular and/or rectangular elements are typically used. In addition, triangular or rectangular isoparametric elements (having one or more curved sides) have been widely used in structural analysis to analyze curved boundaries. (The restriction to orthogonal meshes is inherent in conventional discrete ordinate codes and also in finite difference methods in general). It should be noted that this freedom to represent quite general configurations is of particular significance for applications to hexagonal fuel assemblies and core geometries (i.e. LMFBR, GCFR, HTGR) which are very crudely represented by orthogonal meshes.

(2) The ray effect[11] and related phenomena of flux oscillations and instabilities (well-known problems in discrete ordinates codes) have been shown to be suppressed or eliminated in finite element codes.

(3) The system of algebraic equations resulting from a finite element treatment of the transport equation is characterized by a sparse structure which facilitates storage and solution. (This is in contrast to the full matrices which arise in integral transport equation solutions.)

(4) Boundary conditions are conveniently and rigorously included (formulation of boundary conditions is an ongoing problem in spherical harmonic (P_N) methods).

(5) Treatment of anisotropic scattering and sources appears to be readily feasible within the finite element method (anisotropic scattering is not conveniently handled by integral transport methods) when applied to the first order transport equation.

(6) The order of convergence can be varied depending on the choice of approximating polynomials. Thus, the use of higher order finite element methods (e.g., cubics, quartics, quintics) may drastically reduce computing times to achieve a given accuracy (generally finite difference techniques result in a fixed order of convergence equivalent to linear finite elements).

On the basis of the above considerations, it appears evident that the finite element method has several advantages over conventional transport methods. However, to date the finite element method has not been extensively examined for application to the neutron transport equation. Such an investigation is the subject of this dissertation. The following section summarizes the previous work that has been done in this area and this chapter is then concluded with an outline of the remainder of this dissertation.

I.C. Previous Applications to Neutron Transport

The application of the finite element method to transport calculations was initially proposed in 1971 by Kaper, Leaf, Lindeman[12] and Ohnishi[13,14]. Kaper, et al., formulated a finite element solution of the variational principle associated with the second-order even-parity form of the one-dimensional transport equation in slab geometry (monoenergetic, isotropic sources and scattering). Although no actual computations were presented, a specific algorithm for setting up and solving the resulting set of algebraic equations was included. Ohnishi discussed possible variations in application of the finite element scheme to the transport equation. For example, Ohnishi noted one could apply finite elements either in space or angle alone, retaining discrete ordinates methods and conventional finite difference methods for the angle and space variables, respectively.

In 1972, Ukai[15] examined the theoretical bases for the application of the finite element method to the multi-dimensional transport equation with energy dependence, and anisotropic scattering. Ukai employed the more general integral law (Galerkin or residual) formulation of the finite element method rather than the variational (Ritz) approach taken by others. The advantage of the Galerkin method is that the transport equation and boundary conditions are used directly, whereas the Ritz method cannot be applied because the transport equation is not self-adjoint. (In order to employ the Ritz procedure, one is forced to use the even-parity form of the transport equation, which is self-adjoint and hence has an associated variational principle). Ukai derives theoretical error bounds and orders of convergence for the finite element

method in addition to proving that a unique solution exists to the finite element formulation of the transport equation.

Miller, Lewis, and Rossow[16-17] have investigated the use of the finite element method in phase space for both the one-dimensional and two-dimensional transport equations (monoenergetic, isotropic scattering and sources). In addition, Lewis, Miller, and Henry[18] examined the application of the finite element method to the integral transport equation. Numerical studies indicated that the finite element method compares quite favorably with traditional methods. Specifically, Briggs, Miller, and Lewis[19] found that the finite element approach eliminated the ray effect and that computing times were comparable to that of conventional discrete ordinates codes for a given solution accuracy. Moreover, one achieves considerably more flexibility to examine problems with non-orthogonal geometry. However, the attendant problem with finite element methods, the need to store the coefficient matrix and invert it directly, is magnified with the extremely large problems encountered in multi-dimensional neutron transport. This problem is considered by Kaper, Leaf, and Lindeman[20] to be a major obstacle in the application of the method to transport problems. In this study, Kaper, et al have utilized the second order approach similar to Miller, Lewis, and Rossow[16-17] to incorporate the finite element method, although a direct LU decomposition solver was used rather than the conjugate gradient technique employed by Miller, et al.

In an attempt to neutralize the concerns with the direct inversion of the finite element coefficient matrix, Yuan, Lewis, and Miller[21] investigated the use of block iterative methods to solve the resultant

system of algebraic equations. However, these results were not conclusive and it would appear at this time that direct inversion of finite element coefficient matrices is still the most favorable solution technique.

Reed, Hill, Brinkley, and Lathrop[22] successfully applied the finite element method to the spatial variables of the 2-D transport equation, retaining the discrete ordinates method for the angular variables. Since the discrete ordinates method involves solving for mesh fluxes along the lines of neutron flight, the mesh was restricted to triangles which lay on horizontal bands. This method has been incorporated into the production level 2-D discrete ordinates code, TRIPLET[22]. In addition, TRIPLET allows the use of discontinuous spatial trial functions at the triangle boundaries. A 1-D analog of TRIPLET, ONETRAN[23], has recently been released.

Pitkaranta and Silvennoinen[24,25] examined both the one-group and multi-group forms of the 1-D transport equation in spherical geometry. The Ritz procedure was used (applied to the even-parity form of the transport equation), and the numerical results demonstrated the applicability of the finite element method to realistic multigroup transport calculations. In addition, the finite element solutions they obtained converged considerably faster than the solutions obtained from discrete ordinates methods.

Pitkaranta[26] has also applied the finite element method to the second-order transport equation with a non-self-adjoint variational principle that effectively solves both the even and odd parity equations simultaneously in selected sub-regions of the physical domain. For

certain problems with strong heterogeneities Pitkaranta found that this method offered advantages compared to a finite element formulation involving only the even parity equation. In addition, the solution of both even and odd parity equations allows the calculation of the angular flux which is generally not computable when only the even parity equation is solved, although the scalar flux can be obtained without solving the odd parity equation.

In summary, the major applications of the finite element method to the neutron transport equation have included a variational (Ritz) formulation of the second order (even parity) form of the transport equation, and a residual (Galerkin) formulation to treat the spatial variables of the transport equation combined with a conventional discrete ordinates treatment in the angular variables. Both of these applications have been extended to 2-D geometries. However, the application of phase space finite elements to the first order transport equation has only been examined theoretically--no numerical results have been reported.

I.D. Outline of Remaining Chapters

This section outlines the remainder of this dissertation which contains the results of the investigation into the application of the finite element method to the first order neutron transport equation.

Chapter II begins with a derivation of the integral law formulation (weak form) of the general neutron transport equation. The necessary mathematical concepts and definitions are introduced and then the finite element approximation is applied to the equivalent integral law. The resulting system of algebraic equations is examined to ensure unique-

ness of the solution and a discussion is presented concerning the convergence of the approximate solution to the exact solution.

Chapter III considers the specific application of the finite element method to the transport equation in 1-D plane geometry. After the desired form of the equation is presented, the equivalent integral law is derived and the finite element approximation is applied. A discussion of the treatment of boundary conditions as natural or essential is included, along with specific calculation of the matrix elements for the various boundary conditions. The choice of the finite element subspace is explained, indicating the construction of the basis functions. Observations are made concerning the mathematical properties of the resulting system of algebraic equations and a discussion of the 1-D error analysis is made. Some of the difficulties in the analysis are mentioned and then an explicit demonstration of the convergence rate is made.

Chapter IV extends the formulation of the finite element method presented in Chapter III to include use of discontinuous angular and spatial finite elements. The additional matrix elements needed to incorporate discontinuous spatial elements are explicitly calculated.

Chapter V examines the application of the finite element method to the time-dependent neutron transport equation. It is seen that once the time-independent transport equation is solved with the finite element method, this extension to include time dependence is quite straightforward.

Chapter VI discusses the structure and properties of the coefficient matrix and the method which is used to invert it.

Chapter VII discusses the implementation of the method on the computer, including the methods used to minimize execution time and computer storage.

Chapter VIII presents the numerical results obtained with a computer code (FTRAN) written to apply the finite element method as formulated in the earlier chapters. These results include: verification of the code; precise numerical tests to examine the convergence of the method with mesh refinement (including eigenvalue convergence rates); application to some classic problems of neutron transport (critical slab problem and the Milne problem); application to a problem with strong heterogeneities; and two typical problems in time-dependent neutron transport (equilibrium decay and wave propagation).

Chapter IX presents the conclusions based on the results of this investigation and also recommendations for future effort in this and related areas of transport problems.

CHAPTER II

ANALYTICAL FORMULATION OF THE FINITE ELEMENT METHOD AS APPLIED TO THE NEUTRON TRANSPORT EQUATION

This chapter contains results which are applicable to the general time-independent first order neutron transport equation in arbitrary geometry.

II.A. Integral Law Formulation (Weak Form) of the Neutron Transport Equation

Consider the steady-state neutron transport equation[27,28] in arbitrary geometry with a specified angular flux on the incoming boundary:

$$\hat{\Omega} \cdot \nabla Q(\underline{r}, \hat{\Omega}) + \Sigma_t(\underline{r})\, Q(\underline{r}, \hat{\Omega}) = \int_{4\pi} d\hat{\Omega}'\, \Sigma_s(\underline{r}, \hat{\Omega}' \to \hat{\Omega})\, Q(\underline{r}, \hat{\Omega}') + S(\underline{r}, \hat{\Omega}) \tag{1}$$

subject to

$$Q(\underline{r}, \hat{\Omega}) = Q_0(\underline{r}, \hat{\Omega}) \quad \text{on} \quad \Gamma^- \tag{2}$$

where

$Q(\underline{r}, \hat{\Omega})$ - angular flux

$S(\underline{r}, \hat{\Omega})$ - volumetric source

$\Sigma_t(\underline{r})$ - total cross-section

$\Sigma_s(\underline{r}, \hat{\Omega}' \to \hat{\Omega})$ - differential scattering cross-section

and the terms relating to the geometry of the phase space are defined

$\mathcal{R}$ - spatial domain

4π - angular domain

$\mathcal{V}$ - phase space, $\mathcal{R} \times 4\pi$

$\partial\mathcal{R}$ - boundary of $\mathcal{R}$

$\hat{n}$ - unit outward normal to $\partial\mathcal{R}$

Γ - $\partial\mathcal{R} \times 4\pi$

Γ^+ - outgoing boundary, all $(\underline{r},\hat{\Omega}) \in \Gamma$ such that $\hat{\Omega}\cdot\hat{n} \geq 0$

Γ^- - incoming boundary, all $(\underline{r},\hat{\Omega}) \in \Gamma$ such that $\hat{\Omega}\cdot n \leq 0$

Although specified incoming flux boundary conditions are assumed, the analysis is generalizable to reflecting boundary conditions or vacuum boundary conditions (e.g., set $Q_0(\underline{r},\hat{\Omega}) = 0$).

The goal of this section is to derive the integral law (or weak form) corresponding to Eqs. (1) and (2) and demonstrate the equivalence of the two formulations of the neutron transport equation.

First we must introduce the concept of a Sobolev space[3,29,30], which is crucial for the analysis of the finite element method. Using standard notation and terminology[30], a zero order Sobolev space is the space of all square integrable (in the Lebesque sense) real functions defined in the phase space $\mathcal{V}$. This is also seen to be the definition of the usual real Hilbert space, and is denoted H^0. The first order Sobolev space, H^1, is that space of real functions whose generalized first spatial derivatives are in H^0. The concept of a generalized derivative is not to be considered in this work, but suffice it to say that if derivatives defined in only the ordinary sense are considered, the space H^1 would not be complete[29]. Including functions whose generalized derivatives are

square integrable completes the space, similar to the completion of $\mathcal{H}^0$ by extending ordinary integration with the concept of Lebesque integration. These Sobolev spaces can then be concisely defined:

$$\mathcal{H}^0 = \left\{ \Psi(\underline{x},\hat{\Omega}) \;\middle|\; \iint_V d\underline{x}\, d\hat{\Omega}\, |\Psi(\underline{x},\hat{\Omega})|^2 < \infty \right\} \tag{3}$$

$$\mathcal{H}^1 = \left\{ \Psi(\underline{x},\hat{\Omega}) \;\middle|\; \iint_V \left[|\Psi(\underline{x},\hat{\Omega})|^2 + |\nabla\Psi(\underline{x},\hat{\Omega})|^2 \right] < \infty \right\} \tag{4}$$

The following inner products and norms are defined:

$$\text{inner product} = (Q,\Psi) \equiv \iint_V d\underline{x}\, d\hat{\Omega}\, Q(\underline{x},\hat{\Omega})\Psi(\underline{x},\hat{\Omega}) \tag{5}$$

$$L_2 \text{ norm} = \|Q\|_0 \equiv (Q,Q)^{1/2} \tag{6}$$

$$\text{one-norm} = \|Q\|_1 \equiv \left[(Q,Q) + (\nabla Q, \nabla Q) \right]^{1/2} \tag{7}$$

$$\text{boundary inner product} = \langle Q,\Psi \rangle_{\pm} \equiv \int_{\Gamma^{\pm}} ds\, d\hat{\Omega} \left| \hat{\Omega}\cdot\hat{n} \right| Q(\underline{x},\hat{\Omega})\Psi(\underline{x},\hat{\Omega}) \tag{8}$$

$$\text{boundary norm (semi-norm)} = \langle Q \rangle_{\pm} \equiv \langle Q,Q \rangle_{\pm}^{1/2} \tag{9}$$

To develop the integral law, multiply Eq. (1) by an arbitrary $\Psi(\underline{x},\hat{\Omega}) \in \mathcal{H}^1$, and integrate over the phase space V:

$$\iint_V d\underline{r}\, d\hat{\Omega}\, \hat{\Omega}\cdot\nabla Q(\underline{r},\hat{\Omega})\,\Psi(\underline{r},\hat{\Omega}) + \iint_V d\underline{r}\, d\hat{\Omega}\, \Sigma_t(\underline{r})\, Q(\underline{r},\hat{\Omega})\Psi(\underline{r},\hat{\Omega})$$
$$= \iint_V d\underline{r}\, d\hat{\Omega}\, \Psi(\underline{r},\hat{\Omega}) \int_{4\pi} d\hat{\Omega}'\, \Sigma_s(\underline{r},\hat{\Omega}'\to\hat{\Omega})\, Q(\underline{r},\hat{\Omega}')$$
$$+ \iint_V d\underline{r}\, d\hat{\Omega}\; S(\underline{r},\hat{\Omega})\,\Psi(\underline{r},\hat{\Omega}) \qquad (10)$$

Now, using the definition of the "collision operator" K,

$$K\Psi(\underline{r},\hat{\Omega}) \equiv \Sigma_t(\underline{r})\Psi(\underline{r},\hat{\Omega}) - \int_{4\pi} d\hat{\Omega}'\, \Sigma_s(\underline{r},\hat{\Omega}'\to\hat{\Omega})\Psi(\underline{r},\hat{\Omega}') \qquad (11)$$

and the inner product notation above, Eq. (10) can be concisely stated:

$$(\hat{\Omega}\cdot\nabla Q, \Psi) + (KQ,\Psi) = (S,\Psi) \qquad (12)$$

Now integrate the first term of Eq. (12) by parts to find

$$-(Q, \hat{\Omega}\cdot\nabla\Psi) + \langle Q,\Psi\rangle_+ - \langle Q,\Psi\rangle_- + (KQ,\Psi) = (S,\Psi) \qquad (13)$$

Substituting the known boundary conditions, Eq. (2), into Eq. (13), and rearranging, we arrive at the final integral law, or weak form, of Eqs. (1) and (2):

$$-(Q, \hat{\Omega}\cdot\nabla\Psi) + \langle Q,\Psi\rangle_+ + (KQ,\Psi)$$
$$= (S,\Psi) + \langle Q_0,\Psi\rangle_-\,, \quad \text{All } \Psi \in H^1 \qquad (14)$$

A solution $Q(\underline{r},\hat{\Omega}) \in H^1$ of Eq. (14) which is valid for all $\Psi(\underline{r},\hat{\Omega}) \in H^1$ is then termed a <u>weak</u> solution[15,31] to the original PDE + b.c., Eqs. (1) and (2). The modifier "weak" expresses the fact that

Eqs. (1) and (2) are being satisfied in an integral sense, rather than a pointwise sense, hence a weaker requirement is being fulfilled[31]. However, if the solution $Q(\underline{r},\hat{\Omega})$ to Eq. (14) possesses a derivative in the usual sense, then in fact $Q(\underline{r},\hat{\Omega})$ is also a solution to Eqs. (1) and (2). To demonstrate this, assume $\tilde{Q}(\underline{r},\hat{\Omega})$ satisfies Eq. (14),

$$-(\tilde{Q}, \hat{\Omega}\cdot\nabla\psi) + \langle\tilde{Q},\psi\rangle_+ + (K\tilde{Q},\psi) = (S,\psi) + \langle Q_0,\psi\rangle_- \quad , \quad \text{All } \psi \in H^1 . \tag{15}$$

Integrate the first term of Eq. (15) by parts,

$$(\hat{\Omega}\cdot\nabla\tilde{Q},\psi) + \langle\tilde{Q},\psi\rangle_- - \langle\tilde{Q},\psi\rangle_+ + \langle\tilde{Q},\psi\rangle_+ + (K\tilde{Q},\psi) = (S,\psi) + \langle Q_0,\psi\rangle_- \quad , \text{All } \psi \in H^1 . \tag{16}$$

Now since ψ is arbitrary, choose a subspace $H_s^1 \subset H^1$ such that $\psi = 0$ on Γ^- if $\psi \in H_s^1$ (actually ψ vanishes in a generalized sense[20]). But then we have

$$(\hat{\Omega}\cdot\nabla\tilde{Q},\psi) + (K\tilde{Q},\psi) = (S,\psi) \quad , \text{All } \psi \in H_s^1 . \tag{17}$$

But for this to be valid in general,

$$\hat{\Omega}\cdot\nabla\tilde{Q} + K\tilde{Q} = S \quad \text{in } V . \tag{18}$$

Now to retrieve the boundary conditions, substitute Eq. (18) into Eq. (16)

$$\langle\tilde{Q},\psi\rangle_- = \langle Q_0,\psi\rangle_- \quad , \text{All } \psi \in H^1 .$$

But again, for this to be valid in general, it must hold that

$$\tilde{Q} = Q_0 \quad \text{on } \Gamma^- \quad . \tag{19}$$

But Eqs. (18) and (19) are identical to Eqs. (1) and (2), and by the uniqueness of the solution to the transport equation[32], Q and $\tilde{Q}$ must be identical. Hence the solution to Eq. (14), if it possesses a derivative in the usual sense, is also a solution to Eqs. (1) and (2).

II.B. Remarks on the Integral Law Formulation

Using the standard notation for a bilinear form, $a(\Phi,\Psi)$, the transport operator contained in Eq. (14) can be written

$$a(Q,\Psi) = -(Q, \hat{\Omega}\cdot\nabla\Psi) + \langle Q,\Psi\rangle_+ + (KQ,\Psi) \tag{20}$$

and the integral law, Eq. (14), is then conveniently expressed

$$a(Q,\Psi) = (S,\Psi) + \langle Q_0,\Psi\rangle_- \quad , \text{All } \Psi \in H^1 \quad . \tag{21}$$

First, it should be noted that the boundary condition, Eq. (2), is embodied in the integral law and is not imposed as an additional restriction on the space of admissible trial functions H^1. Thus the boundary conditions are natural boundary conditions[3].

An alternate integral law formulation of Eqs. (1) and (2) could have been derived by using Eq. (12)

$$(\hat{\Omega}\cdot\nabla Q,\Psi) + (KQ,\Psi) = (S,\Psi) \tag{12}$$

and requiring the space of admissible trial functions to satisfy the boundary conditions (in a generalized sense[29]), i.e.,

$$H_B^1 = \left\{ \psi(\underline{r}, \hat{\Omega}) \,\middle|\, \psi \in H^1 \text{ AND } \psi = Q_0 \text{ ON } \Gamma^- \right\} \tag{22}$$

The integral law would then be

$$(\hat{\Omega} \cdot \nabla Q, \psi) + (KQ, \psi) = (S, \psi), \text{ ALL } \psi \in H_B^1 . \tag{23}$$

Now the boundary conditions would be <u>essential boundary conditions</u> because they are imposed directly on the space of trial functions[3].

Ukai[15] notes that the essential boundary condition formulation may be more difficult to use from a computational standpoint, and this formulation tends to make the mathematical analysis somewhat more difficult as well. As the numerical results to be presented later will indicate, the overall accuracy of the method is also decreased somewhat with the use of the essential boundary condition formulation.

II.C. <u>Properties of the Bilinear Form $a(Q,\psi)$</u>

The properties of the bilinear form $a(Q,\psi)$, which is a representation of the transport operator, will now be discussed.

First consider the collision operator K:

$$KQ \equiv \Sigma_t(\underline{r}) Q(\underline{r}, \hat{\Omega}) - \int_{4\pi} d\hat{\Omega}' \Sigma_s(\underline{r}, \hat{\Omega}' \to \hat{\Omega}) Q(\underline{r}, \hat{\Omega}')$$

Assuming that the scattering cross-section only depends on the angle between the incoming and outgoing neutrons, $\hat{\Omega}' \cdot \hat{\Omega}$,

$$\Sigma_s(\underline{r}, \hat{\Omega}' \to \hat{\Omega}) = \Sigma_s(\underline{r}, \hat{\Omega}' \cdot \hat{\Omega}) \tag{24}$$

we can expand $\Sigma_s(\underline{r}, \hat{\Omega}' \cdot \hat{\Omega})$ in terms of Legendre polynomials in

$\hat{\Omega}' \cdot \hat{\Omega}$,

$$\Sigma_s(\underline{r}, \hat{\Omega}' \cdot \hat{\Omega}) = \sum_{\ell=0}^{\infty} \left(\frac{2\ell+1}{4\pi}\right) b_\ell(\underline{r})\, P_\ell(\hat{\Omega}' \cdot \hat{\Omega}) \tag{25}$$

But by the addition theorem[27], this can be expressed in terms of spherical harmonics

$$P_\ell(\hat{\Omega}' \cdot \hat{\Omega}) = \sum_{m=-\ell}^{+\ell} \left(\frac{4\pi}{2\ell+1}\right) \overline{Y}_{\ell m}(\hat{\Omega}')\, Y_{\ell m}(\hat{\Omega})$$

Hence

$$KQ = \Sigma_t(\underline{r})\, Q(\underline{r}, \hat{\Omega}) - \int_{4\pi} d\hat{\Omega}' \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{+\ell} b_\ell(\underline{r})\, \overline{Y}_{\ell m}(\hat{\Omega}')\, Y_{\ell m}(\hat{\Omega})$$

Now form the inner product (KQ, Q) and consider the scattering term (MQ, Q), where $(KQ, Q) = (\Sigma_t Q, Q) - (MQ, Q)$,

$$(MQ, Q) = \int_R d\underline{r} \int_{4\pi} d\hat{\Omega}\, Q(\underline{r}, \hat{\Omega}) \int_{4\pi} d\hat{\Omega}' \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{+\ell} b_\ell(\underline{r})\, \overline{Y}_{\ell m}(\hat{\Omega}')\, Y_{\ell m}(\hat{\Omega})\, Q(\underline{r}, \hat{\Omega}')$$

Interchanging summation and integration, which is valid because the expansion in Eq. (25) is a uniformly convergent series

$$(MQ, Q) = \int_R d\underline{r} \sum_{\ell=0}^{\infty} b_\ell(\underline{r}) \sum_{m=-\ell}^{+\ell} \int_{4\pi} d\hat{\Omega}\, Q(\underline{r}, \hat{\Omega})\, Y_{\ell m}(\hat{\Omega}) \; (x) \int_{4\pi} d\hat{\Omega}'\, Q(\underline{r}, \hat{\Omega}')\, Y_{\ell m}(\hat{\Omega}')$$

But either angular integral can be bounded from above using the Schwartz inequality[33]:

$$\int_{4\pi} d\hat{\Omega}'\, Q(\underline{r},\hat{\Omega}')\, Y_{\ell m}(\hat{\Omega}') \leq \left[\int_{4\pi} d\hat{\Omega}' |Q(\underline{r},\hat{\Omega}')|^2 \int_{4\pi} d\hat{\Omega}' |Y_{\ell m}(\hat{\Omega}')|^2\right]^{1/2}$$

hence

$$(MQ,Q) \leq \int_{\mathcal{R}} d\underline{r} \sum_{\ell=0}^{\infty} b_\ell(\underline{r}) \sum_{m=-\ell}^{+\ell} \int d\hat{\Omega} |Q(\underline{r},\hat{\Omega})|^2 \int_{4\pi} d\hat{\Omega}' |Y_{\ell m}(\hat{\Omega}')|^2$$

But by normalization of the $Y_{\ell m}(\hat{\Omega})$,

$$\int d\hat{\Omega} |Y_{\ell m}(\hat{\Omega})|^2 = 1$$

Hence

$$(MQ,Q) \leq \int_{\mathcal{R}} d\underline{r} \int d\hat{\Omega} |Q(\underline{r},\hat{\Omega})|^2 \sum_{\ell=0}^{\infty} (2\ell+1)\, b_\ell(\underline{r}) \qquad (26)$$

But the total cross-section term is readily obtained

$$(\Sigma_t Q,Q) = \int_{\mathcal{R}} d\underline{r} \int_{4\pi} d\hat{\Omega}\, \Sigma_t(\underline{r}) |Q(\underline{r},\hat{\Omega})|^2 \qquad (27)$$

therefore combining Eqs. (26) and (27),

$$(KQ,Q) \geq \int_{\mathcal{R}} d\underline{r} \int_{4\pi} d\hat{\Omega} |Q(\underline{r},\hat{\Omega})|^2 \left\{ \Sigma_t(\underline{r}) - \sum_{\ell=0}^{\infty} (2\ell+1)\, b_\ell(\underline{r}) \right\}$$

Defining

$$\gamma_0 \equiv \min_{\underline{r} \in \mathcal{R}} \left\{ \Sigma_t(\underline{r}) - \sum_{\ell=0}^{\infty} (2\ell+1)\, b_\ell(\underline{r}) \right\},$$

$$(KQ,Q) \geq \gamma_0 \|Q\|_0^2 . \qquad (28)$$

Therefore, if $\gamma_0 > 0$, the collision operator is positive bounded below[4]. Physically this can be related to the criticality of the region $\mathcal{R}$, as follows.

Define $C(\underline{r}) = \dfrac{\sum_{\ell=0}^{\infty} (2\ell+1)\, b_\ell(\underline{r})}{\Sigma_t(\underline{r})}$. Then $\gamma_0 = \min_{\underline{r} \in \mathcal{R}} \left[(1 - C(\underline{r}))\, \Sigma_t(\underline{r}) \right]$

and clearly $\gamma_0 > 0$ is a sufficient condition for $\mathcal{R}$ to be sub-critical, since the net number of secondary neutrons per collision is < 1, and leakage from $\mathcal{R}$ will cause the system to be more sub-critical. For a single region problem with isotropic scattering,

$$\gamma_0 = \Sigma_t (1-C) = \Sigma_a$$

and $\gamma_0 = 0$ corresponds to pure scattering, which will result in criticality only in an ∞- medium.

Now consider $a(\mathcal{Q},\mathcal{Q})$ in the form of Eq. (20) and after an integration by parts:

$$a(\mathcal{Q},\mathcal{Q}) = -(\mathcal{Q}, \hat{\Omega}\cdot\nabla\mathcal{Q}) + \langle \mathcal{Q},\mathcal{Q}\rangle_+ + (K\mathcal{Q},\mathcal{Q}) \qquad (20)$$

$$a(\mathcal{Q},\mathcal{Q}) = (\hat{\Omega}\cdot\nabla\mathcal{Q},\mathcal{Q}) + \langle \mathcal{Q},\mathcal{Q}\rangle_- + (K\mathcal{Q},\mathcal{Q}) \qquad (29)$$

Noting that the inner product is real and $\mathcal{Q}$ is real, Eqs. (20) and (29) can be combined to obtain:

$$a(\mathcal{Q},\mathcal{Q}) = (K\mathcal{Q},\mathcal{Q}) + \frac{1}{2}\left[\langle \mathcal{Q}\rangle_+^2 + \langle \mathcal{Q}\rangle_-^2\right] \qquad (30)$$

and clearly if Eq. (28) is valid, then

$$a(\mathcal{Q},\mathcal{Q}) \geq \gamma_0 \|\mathcal{Q}\|_0^2 \qquad (31)$$

because of the positivity of the boundary norms. Hence for $\gamma_0 > 0$, $a(\mathcal{Q},\mathcal{Q})$ is positive bounded below[4]. It is interesting to note that if $\mathcal{Q}$ were complex and a complex inner product is used, then Eq. (30) cannot be obtained. This can also be shown in a negative sense by showing

that if Eq. (31) is valid for complex Φ then $a(\Phi,\Psi)$ is symmetric, which is a contradiction:

Assume Eq. (31) is valid for all complex Φ. Express Φ in terms of its real and imaginary parts,

$$\Phi = u + iv$$

where u, v are real, and substitute into $a(\Phi,\Phi)$:

$$\begin{aligned} a(\Phi,\Phi) &= a(u+iv,\, u+iv) \\ &= a(u,u) + a(u,iv) + a(iv,u) + a(iv,iv) \\ &= a(u,u) - i\,a(u,v) + i\,a(v,u) + a(v,v) \\ &= a(u,u) + a(v,v) + i\,[a(v,u) - a(u,v)] \end{aligned}$$

Now Eq. (31) implies $a(\Phi,\Phi)$ is real; therefore,

$$a(v,u) = a(u,v)$$

and $a(\Phi,\Psi)$ is symmetric for real functions. But clearly $a(\Phi,\Psi)$ is not symmetric due to the transport term and therefore we are restricted to real function spaces when Eq. (31) is used.

A physical explanation may suffice to show that $a(\Phi,\Phi)$ is positive definite for $\gamma_0 = 0$. For $\gamma_0 = 0$, Eq. (30) becomes

$$a(\Phi,\Phi) \geq \frac{1}{2}\left[\langle \Phi \rangle_+^2 + \langle \Phi \rangle_-^2\right] \qquad (32)$$

Now assume $a(\Phi,\Phi) = 0$. But Eq. (32) implies $\Phi = 0$ on Γ^+ and Γ^-. Hence the physical situation is a finite region with no net production or loss of neutrons within the interior with identically zero incoming and outgoing fluxes. Clearly this can be satisfied only when the flux is identically zero, or,

$$a(\Omega,\Omega)=0 \Longrightarrow \Omega=0 \ . \tag{33}$$

This relation expresses the fact that $a(\Omega,\Omega)$ is positive definite[4] for a system with pure scattering (c = 1), although the stronger condition of boundedness from below cannot be shown. Mathematically this difference is significant because Eq. (31) ensures a unique solution exists, while Eq. (33) ensures that if a solution exists, it is unique[4]. Since on physical grounds we expect a solution to exist for reasonable source configurations, this difference is minimal.

Thus the problem of finding the (weak) solution to Eq. (14) is well-posed in that a unique solution exists which depends continuously on the data. The next section will discuss the method by which Eq. (14) is to be solved.

II.D. The Finite Element Approximation

Equation (14) is as formidable to solve as the original partial differential equation and boundary conditions, Eqs. (1) and (2). But Eq. (14) is in a convenient form for obtaining an approximate solution. This is done by choosing a specially constructed finite-dimensional subspace $S^h \subset \mathcal{H}^1$. S^h is a finite element subspace with the superscript h being a parameter dependent on the mesh spacing. Mathematically the approximation consists of a sequence of approximate solutions in the subspaces S^h as the mesh is refined ($h \to 0$). This sequence $\{\Omega_h\}$ will be guaranteed to converge to the actual solution (in the energy norm) because:

(i) Q is in $\mathcal{H}^1$ which is complete by construction

(ii) The sequence of subspaces $S^h \subset \mathcal{H}^1$ are dense in $\mathcal{H}^1$

(iii) By (ii) there exists a sequence $\{\psi^h\}$ that converges to Q or $\{\psi_h\} \to Q$ as $h \to 0$

(iv) (to be shown) The finite element method yields a solution that is as close to the actual solution as any member of the subspace (sequence of subspaces) S^h, hence $\{Q_h\} \to Q$ as $h \to 0$.

In practice, of course, one chooses a given mesh refinement and computes the approximate solution, but it is reassuring to know that the process of refining the mesh leads to the actual solution.

The details of the construction of a finite element subspace will be deferred to the section concerning the application in plane geometry. For this general discussion, assume that a subspace $S^h \subset \mathcal{H}^1$ has been constructed, and S^h is N-dimensional with basis:

$$\psi_i^h(\underline{x}, \hat{\Omega}) \quad , \quad i = 1, 2, \cdots, N$$

where again the superscript h corresponds to the mesh spacing. The approximation then consists of finding a $Q_h(\underline{x}, \hat{\Omega}) \in S^h$ such that

$$a(Q_h, \psi^h) = (S, \psi^h) + \langle Q_0, \psi^h \rangle \quad , \text{All } \psi^h \in S^h . \tag{34}$$

This is now the approximate integral law and can be seen to be equivalent to a Galerkin approximation because since $S^h \subset \mathcal{H}^1$, the actual solution also satisfies Eq. (34), which implies

$$a(Q - Q_h, \psi^h) = 0$$

which is the usual Galerkin[25] requirement that the residual (error) be orthogonal to the space of trial functions with respect to the energy inner product.

Since S^h is finite dimensional, and $Q_h \in S^h$, $Q_h(\underline{r},\hat{\Omega})$ can be expanded

$$Q_h(\underline{r},\hat{\Omega}) = \sum_{j=1}^{N} \phi_j \psi_j^h(\underline{r},\hat{\Omega})$$

and Eq. (34) only has to hold for the basis functions of S^h:

$$\sum_{j=1}^{N} \phi_j a(\psi_j^h,\psi_i^h) = (S,\psi_i^h) + \langle Q_0,\psi_i^h\rangle_- \quad , \qquad i=1,2,\cdots,N \tag{35}$$

Or, defining the matrix $\underline{\underline{A}}$,

$$\underline{\underline{A}} = \{A_{ij}\} \quad , \quad A_{ij} \equiv a(\psi_j^h,\psi_i^h)$$

and the vectors

$$\underline{\phi} = \text{col}(\phi_1,\phi_2,\cdots,\phi_N) \quad , \quad \underline{S} = \text{col}(S_1,S_2,\cdots,S_N)$$

where $S_i = (S,\psi_i^h) + \langle Q_0,\psi_i^h\rangle_-$, the following system of algebraic equations is obtained:

$$\underline{\underline{A}}\ \underline{\phi} = \underline{S} \tag{36}$$

Writing out the terms explicitly, one obtains

$A_{ij} = T_{ij} + B_{ij} + K_{ij}$ where

$$T_{ij} = -\iint_V d\underline{r}\,d\hat{\Omega}\ \psi_j^h(\underline{r},\hat{\Omega})\,\hat{\Omega}\cdot\nabla\psi_i^h(\underline{r},\hat{\Omega}) \tag{37}$$

$$B_{ij} = \int_{\Gamma^+} d\underline{s}\, d\hat{\Omega}\, \hat{\Omega}\cdot\hat{n}\, \Psi_i^h(\underline{r},\hat{\Omega})\, \Psi_j^h(\underline{r},\hat{\Omega}) \tag{38}$$

$$K_{ij} = \iint_V d\underline{r}\, d\hat{\Omega}\, \Sigma_t(\underline{r})\, \Psi_i^h(\underline{r},\hat{\Omega})\Psi_j^h(\underline{r},\hat{\Omega}) - \iint_V d\underline{r}\, d\hat{\Omega}\, \Psi_i^h(\underline{r},\hat{\Omega}) \int_{4\pi} d\hat{\Omega}'\, \Sigma_s(\underline{r},\hat{\Omega}'\to\hat{\Omega})\Psi_j^h(\underline{r},\hat{\Omega}') \tag{39}$$

and

$$S_i = \iint_V d\underline{r}\, d\hat{\Omega}\, S(\underline{r},\hat{\Omega})\Psi_i^h(\underline{r},\hat{\Omega}) - \int_{\Gamma^-} d\underline{s}\, d\hat{\Omega}\, \hat{\Omega}\cdot n\, Q_0(\underline{r},\hat{\Omega})\Psi_i^h(\underline{r},\hat{\Omega}) \tag{40}$$

II.E. Properties of the Matrix of Coefficients $\underline{\underline{A}}$

The following properties of the matrix $\underline{\underline{A}}$ will be demonstrated:

1. $\underline{\underline{A}}$ is non-singular (i.e., det $\underline{\underline{A}} \neq 0$)
2. $\underline{\underline{A}}$ is non-symmetric
3. $\underline{\underline{A}}$ is off-diagonal dominant for decreasing mesh ($h \to 0$)

To show that det $\underline{\underline{A}} \neq 0$, it is sufficient to show

$$\underline{\underline{A}}\,\underline{C} = \underline{0} \Rightarrow \underline{C} = \underline{0} \tag{41}$$

i.e., the only solution to the homogeneous problem is the trivial solution, hence $\underline{\underline{A}}$ can be inverted and det $\underline{\underline{A}} \neq 0$. To prove Eq. (41), assume

$$\underline{\underline{A}}\,\underline{C} = \underline{0} \tag{42}$$

for some $\underline{C} \neq \underline{0}$. Pre-multiply Eq. (42) by $\underline{C}^*$ to obtain

$$\underline{C}^*\underline{\underline{A}}\,\underline{C} = 0 \tag{43}$$

Now choose the element $\xi^h \in S^h$ that has $\underline{C}$ as its expansion vector and form

$$a(\xi^h, \xi^h) = a\left(\sum_{k=1}^{N} C_k \psi_k^h , \sum_{j=1}^{N} C_j \psi_j^h\right)$$

$$= \sum_{k=1}^{N} \sum_{j=1}^{N} C_k C_j \, a(\psi_k^h, \psi_j^h)$$

$$= \sum_{k=1}^{N} \sum_{j=1}^{N} C_k C_j A_{jk} = \underline{C}^* \underline{\underline{A}} \, \underline{C}$$

Thus by Eq. (43) $a(\xi^h, \xi^h) = 0$. But by Eq. (32), $\xi^h = 0$. Since the ψ_i^h, $i = 1, \ldots, N$ are linearly independent basis functions, we conclude $\underline{C} = \underline{0}$ and Eq. (41) is proven. $\underline{\underline{A}}$ is therefore non-singular and the system of Eqs. (36) has a unique solution.

An alternate proof[15] of the non-singular nature of $\underline{\underline{A}}$ is given because it yields information concerning the stability of the solution with its data.

Define Q_h to be the solution to Eq. (34) and combine the data into one term:

$$a(Q_h, \psi^h) = (f, \psi^h) \quad , \text{ All } \psi^h \in S^h .$$

In particular, this holds for $\psi^h = Q_h$, hence

$$a(Q_h, Q_h) = (f, Q_h)$$

But by Eq. (30),

$$(KQ_h, Q_h) + \tfrac{1}{2}\langle Q_h \rangle^2 = (f, Q_h)$$

By the Schwartz inequality,

$$(f, Q_h) \leq \|f\|_0 \, \|Q_h\|_0$$

and by Eq. (28),

$$(KQ_h, Q_h) \geq \gamma_0 \|Q_h\|_0^2$$

thus since $\langle Q_h \rangle^2 \geq 0$,

$$\gamma_0 \|Q_h\|_0^2 \leq \|f\|_0 \|Q_h\|_0$$

or

$$\|Q_h\|_0 \leq \frac{1}{\gamma_0} \|f\|_0 \tag{44}$$

Since we also have

$$\gamma_0 \|Q_h\|_0^2 + \tfrac{1}{2} \langle Q_h \rangle^2 \leq \|f\|_0 \|Q_h\|_0 \tag{45}$$

then Eq. (44) and (45) yield

$$\gamma_0 \|Q_h\|_0^2 + \tfrac{1}{2} \langle Q_h \rangle^2 \leq \frac{1}{\gamma_0} \|f\|_0^2 \tag{46}$$

which is the desired inequality. Note that an immediate consequence of Eq. (46) is the non-singular nature of $\underline{\underline{A}}$: (assuming $\gamma_0 \geq 0$)

If f = 0, then the matrix equation is

$$\underline{\underline{A}} \, \underline{\xi}^h = \underline{0}$$

But by Eq. (46), $Q_h = 0$, thus its expansion vector $\underline{\xi}^h = \underline{0}$ because the basis functions are linearly independent. Hence $\underline{\underline{A}} \, \underline{\xi}^h = \underline{0} \Rightarrow \underline{\xi}^h = \underline{0}$ which implies $\underline{\underline{A}}$ is non-singular.

Another consequence of this is that Eq. (44) expresses the continuous behavior of the solution on the data[15]:

If Q_1, Q_2 are solutions for data f_1 and f_2 respectively, then

$$\| Q_1 - Q_2 \|_0 \leq \frac{1}{\gamma_0} \| f_1 - f_2 \|_0 \tag{47}$$

and this is independent of the mesh spacing h. Thus the system of equations, Eq. (34), should have a stable solution as the mesh is refined ($h \rightarrow 0$), assuming $\gamma_0 \geq 0$.

In most applications of the finite element method, the matrix $\underline{\underline{A}}$ is symmetric. However, in this application the transport term destroys the symmetry of $\underline{\underline{A}}$:

Defining $$T_{ij} = -(\Psi_j^h, \hat{\Omega} \cdot \nabla \Psi_i^h) \quad \text{(transport term)} \tag{48}$$

$$K_{ij} = (K\Psi_j^h, \Psi_i^h) \quad \text{(collision term)} \tag{49}$$

$$B_{ij} = \langle \Psi_j^h, \Psi_i^h \rangle_+ \quad \text{(boundary term)} \tag{50}$$

we have $$A_{ij} = T_{ij} + B_{ij} + K_{ij}$$

Now integrate T_{ij} by parts,

$$\begin{aligned} T_{ij} &= -(\Psi_j^h, \hat{\Omega} \cdot \nabla \Psi_i^h) \\ &= (\hat{\Omega} \cdot \nabla \Psi_j^h, \Psi_i^h) + \langle \Psi_j^h, \Psi_i^h \rangle_- - \langle \Psi_j^h, \Psi_i^h \rangle_+ \end{aligned}$$

Since the basis functions Ψ_i^h are all real,

$$T_{ij} = T_{ji} + \langle \Psi_j^h, \Psi_i^h \rangle - \langle \Psi_j^h, \Psi_i^h \rangle_+$$

For a typical finite element subspace S^h, only a few ψ_j^h will be non-zero on the boundary $\Gamma^{\pm}$, since the basis functions are local. Therefore, $T_{ij} = -T_{ji}$ for most i, j, and $\underline{\underline{T}}$ is nearly anti-symmetric. In addition, $T_{ii} = 0$ for most i and $\underline{\underline{T}}$ is off-diagonal dominant. Also, since T_{ij} involves a derivative, its "size" is larger than the remainder of A_{ij} for small enough mesh. Thus $\underline{\underline{A}}$ can be off-diagonal dominant.

The boundary term B_{ij} is clearly symmetric and the collision term K_{ij} will be symmetric if the scattering cross-section depends only upon the angle between the incident and outgoing neutrons:

$$K_{ij} = \left(K\psi_j^h, \psi_i^h\right) = \iint_V d\underline{r}\, d\hat{\Omega}\, \Sigma_t(\underline{r})\, \psi_i^h(\underline{r},\hat{\Omega})\, \psi_j^h(\underline{r},\hat{\Omega})$$

$$- \iint_V d\underline{r}\, d\hat{\Omega}\, \psi_i^h(\underline{r},\hat{\Omega}) \int_{4\pi} d\hat{\Omega}'\, \Sigma_s(\underline{r},\hat{\Omega}'\to\hat{\Omega})\, \psi_j^h(\underline{r},\hat{\Omega}')$$

and interchange of $\hat{\Omega}, \hat{\Omega}'$ implies

$$K_{ij} = K_{ji} \quad \text{if} \quad \Sigma_s(\underline{r},\hat{\Omega}'\to\hat{\Omega}) = \Sigma_s(\underline{r},\hat{\Omega}'\cdot\hat{\Omega})$$

Therefore $\underline{\underline{A}}$ consists of a symmetric portion and a nearly anti-symmetric portion due to the transport term.

II.F Convergence and Error Analysis for the Diffusion Equation

The key reason for the success of the finite element method is that the approximate solution, whether it be obtained from a variational (Ritz) finite element formulation or the residual (Galerkin) finite element formulation used in this investigation, generally is as close (in a suitable norm) to the actual solution as any member of the approximating subspace S^h. Thus the error analysis may rely on standard results of ap-

proximation theory, using S^h (i.e., piece-wise approximation), independent of the particular physical problem being considered. Therefore the convergence of the solution is generally guaranteed and the order of convergence with mesh refinement may be predicted in most cases. These remarks are certainly valid for second-order, elliptic, self-adjoint systems, such as the one-speed or within-group neutron diffusion equations, the heat conduction equation, Laplace's equation, etc. However, when one attempts to derive theoretical error bounds and convergence rates for the neutron transport equation, which is first-order and non-self-adjoint, certain problems are encountered. To appreciate these problems and provide some insight as to where progress may be achieved, the error analysis for a typical second-order, elliptic, self-adjoint system will be briefly reviewed. It is also worthwhile to note that Fix[34] has examined the eigenvalue problem for the multi-group diffusion equations, which are not self-adjoint, but this will not be discussed.

Consider the mono-energetic neutron diffusion equation with zero flux (Dirichlet) boundary conditions on Γ_1 and zero current (Neumann) boundary conditions on Γ_2 :

$$-D(\underline{r})\nabla^2 Q(\underline{r}) + \Sigma_a(\underline{r})Q(\underline{r}) = S \quad \text{in } R \tag{51}$$

subject to

$$Q(\underline{r}) = 0 \quad , \quad \underline{r} \in \Gamma_1 \tag{52}$$

$$-D\nabla Q(\underline{r}) = 0 \quad , \quad \underline{r} \in \Gamma_2 \tag{53}$$

where $\Gamma = \Gamma_1 + \Gamma_2$ is the boundary of R.

To develop the integral law, we can proceed as we did with the transport equation, and multiply Eq. (51) by an arbitrary $\psi \in H_B^1$ where

$$H_B^1 = \left\{ \psi(\underline{r}) \,\middle|\, \int_R d\underline{r} \left[|\psi(\underline{r})|^2 + |\nabla\psi(\underline{r})|^2 \right] < \infty \right.$$

and

$$\left. \psi(\underline{r}) = 0 \quad \text{on } \Gamma_1 \right\}$$

(note the Dirichlet boundary conditions are imposed on the space H_B^1) and integrate over R:

$$\int_R d\underline{r} \left(-D\nabla^2 Q\,\psi + \Sigma_a Q \psi \right) = \int_R d\underline{r}\, S\psi$$

Now integrate by parts, using the boundary conditions to eliminate the surface term:

$$\int_R d\underline{r} \left[D\nabla Q \nabla\psi + \Sigma_a Q\psi - S \right] = 0$$

or

$$(D\nabla Q, \nabla\psi) + (\Sigma_a Q, \psi) = (S, \psi)$$

It is easily shown that the Neumann boundary condition is a natural boundary condition for this problem while the Dirichlet boundary condition is an essential boundary condition that must be imposed on the space of trial functions[3]. In addition, if there are internal boundaries, say between two dissimilar materials with diffusion coefficients D_A and D_B, then it is also seen that the current continuity $-D_A \nabla Q_A = -D_B \nabla Q_B$ at the interface is a natural boundary condition and therefore can be ignored.

In any event, we have the bilinear form

$$a(\Phi,\Psi) = (S,\Psi) \quad , \text{ All } \Psi \in H_B^1 \tag{54}$$

where $$a(\Phi,\Psi) = (D\nabla\Phi, \nabla\Psi) + (\Sigma_a \Phi, \Psi)$$

to which the finite element approximation is made by choosing a subspace $S^h \subset H_B^1$ and finding the solution $\Phi_h \in S^h$ for which Eq. (51) is valid for all $\Psi^h \in S^h$.

It is a standard calculation[3] to show there exist positive constants C_1 and C_2 independent of Φ and Ψ such that

$$a(\Phi,\Phi) \geq C_1 \|\Phi\|_1^2 \tag{55}$$

$$a(\Phi,\Psi) \leq C_2 \|\Phi\|_1 \|\Psi\|_1 \tag{56}$$

for any $\Phi, \Psi \in H_B^1$. This is a statement to the effect that the energy norm,

$$\|\Phi\|_E \equiv \sqrt{a(\Phi,\Phi)}$$

and the one-norm, $\|\Phi\|_1$, are equivalent[3]. The error estimates are then obtained with Eqs. (55) and (56) by starting with the energy norm of the error in the approximate solution and then adding and subtracting an arbitrary element $\Psi^h \in S^h$,

$$a(\Phi-\Phi_h, \Phi-\Phi_h) = a(\Phi-\Phi_h, \Phi-\Psi^h+\Psi^h-\Phi_h)$$

Since the bilinear form is linear in either function,

$$a(Q-Q_h, Q-Q_h) = a(Q-Q_h, Q-\psi^h) + a(Q-Q_h, \psi^h-Q_h) \quad (57)$$

But the residual $Q-Q_h$ is orthogonal to the space S^h in the energy norm, i.e.,

$$a(Q, \psi^h) = (f, \psi^h)$$
$$a(Q_h, \psi^h) = (f, \psi^h)$$

hence $$a(Q-Q_h, \psi^h) = 0 \quad , \quad \text{All } \psi^h \in S^h .$$

Therefore, the second term on the RHS of Eq. (57) is zero, and we have

$$a(Q-Q_h, Q-Q_h) = a(Q-Q_h, Q-\psi^h) , \quad \text{All } \psi^h \in S^h . \quad (58)$$

Now using Eq. (55), the LHS of Eq. (58) is bounded from below

$$C_1 \|Q-Q_h\|_1^2 \le a(Q-Q_h, Q-Q_h) \quad (59)$$

and the RHS of Eq. (58) is bounded from above using Eq. (56)

$$a(Q-Q_h, Q-\psi^h) \le C_2 \|Q-Q_h\|_1 \|Q-\psi^h\|_1 \quad (60)$$

Combining Eqs. (59) and (60),

$$\|Q-Q_h\|_1 \le C \|Q-\psi^h\|_1 , \quad \text{All } \psi^h \in S^h \quad (61)$$

which is the crucial estimate, because this states that the approximate solution Q_h is as close to the actual solution Q (in the one-norm), as any arbitrary $\psi^h \in S^h$. But from the approximation theory of piecewise polynomials, we know

$$\|Q - \psi^h\|_1 \leq K h^k |Q|_{k+1} \tag{62}$$

that is, S^h can approximate the solution Q to within $O(h^k)$ in the one-norm, where k is the degree of the piecewise polynomials, K is a constant independent of Q and $|Q|_1$ is the semi-norm (for k = 1)

$$|Q|_2 = \left[\int d\underline{x} \, |\nabla^2 Q|\right]^{1/2}$$

which is constant for a given Q.

It is a standard procedure (the Nitsche "trick") in numerical analysis[3,35] to obtain the error estimates in the L_2-norm, which adds $O(h)$ to the error estimate:

$$\|Q - Q_h\|_0 \leq C h^{k+1} |Q|_{k+1}. \tag{63}$$

This step is not a natural result of the finite element approximation but must be obtained with some additional analysis. The reason for this is that the approximate solution is optimal in the "energy" norm $a(Q,Q)$ which is equivalent to the one-norm, and not the L_2-norm.

In addition, the eigenvalue convergence rates may be computed[3]:

$$|\lambda_h - \lambda| \leq C \|Q - Q_h\|_1^2 \tag{64}$$

i.e., the error in the eigenvalue is equivalent to the error in the eigenfunction, measured in the one-norm. Therefore, using Eq. (62)

$$|\lambda_h - \lambda| \leq C h^{2k} \tag{65}$$

or $O(h^2)$ convergence for linear elements, $O(h^4)$ for quadratic elements, etc.

II.G Convergence and Error Analysis for the Transport Equation

To the extent possible, the analysis above will be repeated for the transport equation. As we will find, however, the rigorous analysis yields a convergence which is not optimal in the exponent of h, i.e., the L_2-error for the approximation solution decreases as $O(h^k)$ rather than $O(h^{k+1})$ as with the diffusion equation.

To obtain the rigorous error bounds, begin with Eq. (30) combined with Eq. (28),

$$a(\Phi,\Phi) \geq \gamma_0 \|\Phi\|_0^2 + \frac{1}{2}\left[\langle \Phi \rangle_+^2 + \langle \Phi \rangle_-^2 \right] \tag{66}$$

Thus $a(\Phi,\Phi)$ is bounded from below by the L_2-norm and the boundary norms. Comparing this with the equivalent bounds Eq. (55), for the diffusion equation, where $a(\Phi,\Phi)$ is bounded from below by the one-norm, we see that Eq. (66) is a considerably weaker condition.

To bound $a(\Phi,\Psi)$ from above, start with Eq. (20)

$$a(\Phi,\Psi) = -(\Phi, \hat{\Omega}\cdot\nabla\Psi) + (K\Phi,\Psi) + \langle \Phi,\Psi \rangle_+ \tag{20}$$

Noting that the collision operator is a bounded operator:

$$(K\Phi,\Phi) = (\Sigma_t \Phi,\Phi) - (M\Phi,\Phi)$$

where $(M\Omega,\Omega)$ is the scattering term defined in obtaining Eq. (26). Since $(M\Omega,\Omega) \geq 0$,

$$(K\Omega,\Omega) \leq (\Sigma_t\Omega,\Omega) = \Sigma_0 \|\Omega\|_0^2$$

where Σ_0 is the maximum value of the total cross-section in the domain $\mathcal{R}$. Now use the Schwartz inequality and the symmetry of K to obtain

$$|(K\Omega,\Psi)| \leq (K\Omega,K\Omega)^{1/2} (\Psi,\Psi)^{1/2}$$

or

$$|(K\Omega,\Psi)| \leq \Sigma_0 \|\Omega\|_0 \|\Psi\|_0 \tag{67}$$

using Eq. (66). Similarly, using the Schwartz inequality,

$$|(\Omega, \hat{\Omega}\cdot\nabla\Psi)| \leq \|\Omega\|_0 \|\Psi\|_1 \tag{68}$$

and

$$|\langle\Omega,\Psi\rangle_+| \leq \langle\Omega\rangle_+ \langle\Psi\rangle_+ \tag{69}$$

Now using the following identity[15]

$$|a|\,|b| \leq \epsilon |a|^2 + \frac{1}{4\epsilon}|b|^2 \quad , \epsilon > 0$$

Eqs. (67) and (68) become (with $\epsilon = \frac{\gamma_0}{4}$)

$$|(K\Omega,\Psi)| \leq \frac{\gamma_0}{4}\|\Omega\|_0^2 + \frac{\Sigma_0^2}{\gamma_0}\|\Psi\|_0^2 \tag{70}$$

$$|(\Omega,\hat{\Omega}\cdot\nabla\Psi)| \leq \frac{\gamma_0}{4}\|\Omega\|_0^2 + \frac{1}{\gamma_0}\|\Psi\|_1^2 \tag{71}$$

and Eq. (69) becomes (with $\epsilon = \frac{1}{4}$)

$$\left| \langle Q, \psi \rangle_+ \right| \leq \frac{1}{4} \langle Q \rangle_+^2 + \langle \psi \rangle_+^2 \tag{72}$$

Therefore, using Eqs. (70)-(72) to bound the terms on the RHS of Eq. (20), we obtain the desired upper bound (using $\|\psi\|_0 \leq \|\psi\|_1$, $\langle Q \rangle_-^2 \geq 0$)

$$\begin{aligned} |a(Q,\psi)| \leq{}& \frac{\gamma_0}{2} \|Q\|_0^2 + \frac{1}{\gamma_0}\left(1+\Sigma_0^2\right)\|\psi\|_1^2 \\ &+ \frac{1}{4}\left[\langle Q \rangle_+^2 + \langle Q \rangle_-^2\right] + \langle \psi \rangle_+^2 \end{aligned} \tag{73}$$

Now let us consider the error in the approximate solution measured in the energy norm:

"energy" error = $a(Q-Q_h, Q-Q_h)$

and noting that the residual $Q-Q_h$ is orthogonal to S^h,

$$a(Q-Q_h, Q-Q_h) = a(Q-Q_h, Q-\psi^h+\psi^h-Q_h) \tag{74}$$

or

$$a(Q-Q_h, Q-Q_h) = a(Q-Q_h, Q-\psi^h) + a(Q-Q_h, \psi^h-Q_h)$$

the second term on the RHS of Eq. (74) vanishes. Therefore,

$$a(Q-Q_h, Q-Q_h) = a(Q-Q_h, Q-\psi^h) \tag{75}$$

for arbitrary $\psi^h \in S^h$. Now apply Eq. (66) to the LHS of Eq. (75) and Eq. (73) to the RHS of Eq. (75) to obtain the inequality:

$$\gamma_0 \|\Theta-\Theta_h\|_0^2 + \frac{1}{2}\left[\langle \Theta-\Theta_h\rangle_+^2 + \langle \Theta-\Theta_h\rangle_-^2\right]$$
$$\leq \frac{\gamma_0}{2}\|\Theta-\Theta_h\|_0^2 + \frac{1}{\gamma_0}(1+\Sigma_t^2)\|\Theta-\psi^h\|_1^2$$
$$+ \frac{1}{4}\langle \Theta-\Theta_h\rangle_+^2 + \frac{1}{4}\langle \Theta-\Theta_h\rangle_-^2 + \langle \Theta-\psi^h\rangle_+^2$$

Now using the following inequality[15]

$$\langle \psi\rangle_\pm \leq \|\psi\|_1^2$$

and rearranging, we obtain

$$\frac{\gamma_0}{2}\|\Theta-\Theta_h\|_0^2 + \frac{1}{4}\left[\langle \Theta-\Theta_h\rangle_+^2 + \langle \Theta-\Theta_h\rangle_-^2\right]$$
$$\leq \left[\frac{1}{\gamma_0}(1+\Sigma_t^2)+1\right]\|\Theta-\psi_h\|_1^2$$

Now defining $C = 2\left[\frac{1}{\gamma_0}(1+\Sigma_t^2)+1\right]$ we obtain the final inequality ("energy" inequality)

$$\gamma_0\|\Theta-\Theta_h\|_0^2 + \frac{1}{2}\langle \Theta-\Theta_h\rangle^2$$
$$\leq C\|\Theta-\psi^h\|_1^2, \quad \text{All } \psi^h \in S^h \tag{76}$$

This energy inequality states that the finite element approximation, when measured in the L_2-norm, is as close to the actual solution as an arbitrary element of S^h when measured in the one-norm. Therefore, the inequality is not as sharp as for the diffusion equation, where both sides were measured in the same norm (see Eq. (61)).

This results in a loss of O(h) for the convergence rate because from approximation theory, S^h can approximate a sufficiently smooth function Q to within $O(h^k)$ in the one-norm:

$$\|Q-\psi^h\|_1^2 \leq Ch^k |Q|_{k+1} \tag{77}$$

where $|Q|_{k+1}$ is defined as before with the diffusion equation. Equation (77) and the energy inequality, Eq. (76), immediately yield for the error in the approximate solution in the L2-norm:

$$\|Q-Q_h\|_0 \leq C_1 h^k |Q|_{k+1}$$

where C_1 is a constant independent of h.

Thus the error estimate for the transport equation is decreased by O(h) compared to the estimate for the diffusion equation, Eq. (63).

However, the numerical results, which have been reported previously as a part of this investigation[36] and which are discussed in detail in a later section, would indicate that the convergence rate is actually

$$\|Q-Q_h\|_0 \leq Ch^{k+1} |Q|_{k+1} \tag{78}$$

i.e., $O(h^{k+1})$ convergence for finite elements of degree k. Therefore, one might expect that the analytical proof may be sharpened to obtain this more rapid convergence.

One approach to demonstrating this is to consider the truncation error associated with the finite element approximation and obtain explicit bounds on the resulting error in the approximate solution. This has been done in Chapter III for linear elements and the results, which rely on numerical results for the norm of the inverse of the coefficient matrix, support the $O(h^{k+1})$ convergence.

However, this approach is not entirely satisfying due to the need for obtaining numerical estimates for the norm of the inverse of the matrix of coefficients, $\underline{\underline{A}}$. It would be desirable to obtain an improved estimate of the convergence rate in a more rigorous manner, similar to the preceding proof for the non-optimal estimate. However, a rigorous demonstration of this optimal convergence rate was not obtained during the course of this investigation and the following discussion is therefore devoted to several possible approaches for obtaining the optimal result.

One approach might be to split up the transport bilinear form,

$$a(\Phi,\Psi) = -(\Phi, \hat{\Omega}\cdot\nabla\Psi) + \langle \Phi,\Psi\rangle_t + (K\Phi,\Psi) \tag{79}$$

into its symmetric and anti-symmetric portions, which is possible because any bilinear form can be reduced to a sum of a symmetric bilinear form and an anti-symmetric bilinear form[37],

$$a(\Phi,\Psi) = a_s(\Phi,\Psi) + a_a(\Phi,\Psi) \tag{80}$$

where

$$a_s(\Phi,\Psi) = \quad \text{symmetric bilinear form} \tag{81}$$

$$a_a(\Phi,\Psi) = \quad \text{anti-symmetric bilinear form} \tag{82}$$

By the definition of symmetric and anti-symmetric bilinear forms,

$$a_s(\Phi,\Psi) = a_s(\Psi,\Phi) \tag{83}$$

$$a_a(\Phi,\Psi) = -a_a(\Psi,\Phi) \tag{84}$$

Now, using the following identity, which is merely an integration by parts of the first term in Eq. (79),

$$\begin{aligned}(-\Phi, \hat{\Omega}\cdot\nabla\Psi) = (\hat{\Omega}\cdot\nabla\Phi, \Psi) \\ - \langle \Phi,\Psi\rangle_+ + \langle \Phi,\Psi\rangle_- \quad ,\end{aligned} \tag{85}$$

the symmetric and anti-symmetric parts in Eq. (80) are readily obtained

$$a_s(\Phi,\Psi) = (K\Phi,\Psi) + \frac{1}{2}\left[\langle \Phi,\Psi\rangle_+ + \langle \Phi,\Psi\rangle_-\right] \tag{86}$$

$$a_a(\Phi,\Psi) = -(\Phi, \hat{\Omega}\cdot\nabla\Psi) + \frac{1}{2}\left[\langle \Phi,\Psi\rangle_+ - \langle \Phi,\Psi\rangle_-\right] \tag{87}$$

It is easily verified that

$$a_s(\phi,\psi) = a_s(\psi,\phi) \tag{88}$$

$$a_a(\phi,\psi) = -a_a(\psi,\phi) \tag{89}$$

and

$$a(\phi,\psi) = a_s(\phi,\psi) + a_a(\phi,\psi) \tag{90}$$

Now if one assumes that the anti-symmetric portion of a (ϕ,ψ) is bounded by the symmetric portion of a (ϕ,ψ), i.e.,

$$|a_a(\phi,\psi)| \leq |a_s(\phi,\psi)| \qquad \text{All } \phi,\psi \in H^1 \tag{91}$$

then it is easy to show that the approximate solution minimizes the error in the energy over the subspace $S^h \subset H^1$. That is, if ϕ_h is the approximate solution and ψ^h is an arbitrary element of S^h, then

$$a(\phi-\phi_h,\phi-\phi_h) \leq a(\phi-\psi^h,\phi-\psi^h) \qquad \text{All } \psi^h \in S^h. \tag{92}$$

However, the validity of Eq. (91) is questionable because it involves bounding the derivative terms by surface terms and volume terms, which at first glance does not appear valid. Therefore this approach will not

be pursued any further, except to remark that the result--minimization of the energy over the space S^h, would appear to be valid.

A more promising approach may be to employ a method similar to that used by Lasaint and Raviart[38] to analyze the rate of convergence of the solution method employed in TRIPLET[22], which is the finite element transport code which utilizes discontinuous elements in space and conventional discrete ordinates methods in angle. Although the analysis is concerned with a somewhat different set of equations, involving only spatial terms, some of the results should certainly be applicable to the current analysis. For example, the general results obtained by Lasaint and Raviart were also not optimal in the convergence rate, in that $O(h^k)$ rates were predicted. However, when they restricted the spatial domain to be partitioned into an orthogonal mesh, then they were able to prove $O(h^{k+1})$ convergence for finite elements of degree k. Thus this approach may be a fruitful means of obtaining the optimal convergence rate for this application.

The final attempt to reconcile the predicted and observed convergence rates involves an examination of the eigenvalue convergence rate and its relationship to the solution convergence rate. As will be seen, an eigenvalue convergence rate which is consistent with the $O(h^{k+1})$ solution convergence rate is observed numerically, which adds additional credence to the argument for the improved error estimate, which is $O(h^{k+1})$ convergence for finite elements of order k on a mesh with spacing h. To do this, the discussion will consist of the following. First, we will assume that the finite element solution is the best approximation to the actual solution within S^h, with the L_2 norm being the

measure of error. Using this assumption, the error in the energy will be derived. Then the eigenvalue problem is formulated and the error in the eigenvalue is derived, using the error in the energy. It is then seen that this predicted eigenvalue convergence rate, which depended of course on the optimal convergence in the solution, agrees with the numerical results for the error in the eigenvalue. Thus the numerical results for the eigenvalue problem are consistent with the improved estimate in the error in the solution and of course they are also consistent with the numerical results for the error in the solution.

The eigenvalue convergence rate will depend on the predicted error in the energy, or

$$\|Q-Q_h\|_E = \left[a(Q-Q_h, Q-Q_h)\right]^{1/2} \tag{93}$$

As was noted above, it appears reasonable that in fact the finite element solution Q_h minimizes the error in the energy over the space S^h; however, the proof of this has not been performed. To obtain the eigenvalue prediction, it will be assumed that indeed the finite element method yields the optimal solution in the space S^h. That is, if Q_h is the approximate solution and Q is the actual solution, then

$$\|Q-Q_h\|_0 \leq Ch^{k+1}|Q|_{k+1} \tag{94}$$

where $|Q|_k$ is the usual semi-norm as used previously, h is the mesh spacing, and k is the degree of the finite element. Using the results of Lasaint and Raviart[38], the approximate solution loses $O(h^{1/2})$ in the error on the boundaries,

$$\langle Q-Q_h\rangle_{\pm} \leq C h^{k+\frac{1}{2}} |Q|_{k+1} \quad . \tag{95}$$

Again, it should be noted that we are assuming Q_h is the optimal solution, which is consistent with the observed numerical results. We will now derive the eigenvalue convergence rate using Eqs. (94) and (95), and compare with the numerical results, which are independent of the numerical results obtained for the error (L_2 norm) in the solution.

To obtain the eigenvalue error, first consider the error in the energy,

$$\|Q-Q_h\|_E^2 = a(Q-Q_h, Q-Q_h) \tag{96}$$

Using Eq. (30), Eq. (96) becomes

$$\|Q-Q_h\|_E^2 = \left(K(Q-Q_h), Q-Q_h\right) + \frac{1}{2}\left[\langle Q-Q_h\rangle_+^2 + \langle Q-Q_h\rangle_-^2\right] \tag{97}$$

But the first term of Eq. (97) may be bounded from above using Eq. (67),

$$\left(K(Q-Q_h), Q-Q_h\right) \leq \Sigma_0 \|Q-Q_h\|_0^2 \tag{98}$$

and therefore with Eq. (98), Eq. (97) becomes

$$\|Q-Q_h\|_E^2 \leq \Sigma_0 \|Q-Q_h\|_0^2 + \frac{1}{2}\left[\langle Q-Q_h\rangle_+^2 + \langle Q-Q_h\rangle_-^2\right] \tag{99}$$

But now Eqs. (94) and (95) can be applied immediately to Eq. (99), yielding

$$\|Q-Q_h\|_E^2 \leq C_1 h^{2k+2} |Q|_{k+1}^2 + C_2 h^{2k+1} |Q|_{k+1}^2 \qquad (100)$$

which expresses the error in the energy, assuming the approximate solution Q_h is indeed the optimal choice over the space S^h. For decreasing mesh (i.e., $h \rightarrow 0$), one then obtains $O(h^{2k+1})$ convergence in the energy, but we will carry along for now the additional $O(h^{2k+2})$ term, since the numerical results appear to substantiate this term as well.

Now we will use the error in the energy, Eq. (100), to derive the error in the eigenvalue.

The eigenvalue problem is formulated in the following manner. The physical situation consists of a physical domain R which for simplicity is assumed to be a homogeneous medium with isotropic scattering. In order for this region to be critical, the number of neutrons being produced in R must be equal to the number of neutrons being lost via collisions or leakage. If there is a net loss of neutrons, the region R is said to be sub-critical, and if there is a net gain of neutrons, the region R is said to be super-critical. The eigenvalue problem then consists of a balance relationship, with the eigenvalue C being the factor by which the production term is scaled to achieve criticality. This relationship is the following

$$\hat{\Omega}\cdot\nabla Q(\underline{r},\hat{\Omega}) + Q(\underline{r},\hat{\Omega}) = c\int_{4\pi} d\hat{\Omega}'\, Q(\underline{r},\hat{\Omega}') \tag{101}$$

where the unit of length is taken to be mean free paths, $\frac{1}{\Sigma_t}$. The eigenvalue C is defined

$$C = \frac{\nu\Sigma_f + \Sigma_s}{\Sigma_t} \tag{102}$$

where Σ_f is the fission cross section and ν is the average number of neutrons produced per fission event. Of course, as defined C would appear to be a fixed number; however, in practice one would adjust the fission concentration, thus effectively changing the value of C. Therefore, for the ensuing discussion, it will be assumed that C is the term which is scaled to achieve criticality in the given region R.

To derive the equivalent integral law formulation of the eigenvalue problem, multiply Eq. (101) by an arbitrary $\Psi(\underline{r},\hat{\Omega}) \in H^1$ and integrate over the phase space V. After an integration by parts as with the earlier formulation with the source problem, the equivalent integral law may be phrased:

Find $Q(\underline{r},\hat{\Omega}) \in H^1$ and a real number C such that

$$a(Q,\Psi) = c\, m(Q,\Psi), \quad \text{All } \Psi \in H^1 \tag{103}$$

where

$$a(Q,\Psi) = -(Q,\hat{\Omega}\cdot\nabla\Psi) + (Q,\Psi) + \langle Q,\Psi\rangle_+ \tag{104}$$

$$m(Q,\Psi) = \left(\int_{4\pi} d\hat{\Omega}' Q(\underline{r},\hat{\Omega}'), \Psi \right) \tag{104}$$

and it has been assumed for convenience that vacuum boundary conditions are imposed on Γ^-.

Now consider the scattering bilinear form $m(Q,\Psi)$, which is defined in Eq. (104) for the case of isotropic scattering. Although the present analysis is restricted to isotropic scattering, the analyses can be generalized to include a general anisotropic scattering kernel of the form:

$$\Sigma_s(\underline{r},\hat{\Omega}'\cdot\hat{\Omega}) = \sum_{\ell=0}^{L} \left(\frac{2\ell+1}{4\pi}\right) b_\ell P_\ell(\hat{\Omega}'\cdot\hat{\Omega})$$

For the present analysis, though, we shall assume the Legendre coefficients are zero, except of course for b_0, which is the total scattering cross-section Σ_s (plus $\nu\Sigma_f$ when applicable). However, for the eigenvalue problem, the b_0 term is included within the eigenvalue C, as noted in Eq. (102).

Writing out the scattering bilinear form,

$$m(Q,\Psi) = \int_R d\underline{r} \int_{4\pi} d\hat{\Omega}\, \Psi(\underline{r},\hat{\Omega}) \int_{4\pi} d\hat{\Omega}'\, Q(\underline{r},\hat{\Omega}')$$

and letting $\Psi = Q$, we have

$$m(Q,Q) = \int_R d\underline{r} \int_{4\pi} d\hat{\Omega}\, Q(\underline{r},\hat{\Omega}) \int_{4\pi} d\hat{\Omega}'\, Q(\underline{r},\hat{\Omega}') \tag{105}$$

Applying the Schwartz inequality to both angular integrals of Eq. (105),

$$m(Q,Q) \le 4\pi \int_R d\underline{r} \left[\int_{4\pi} d\hat{\Omega} |Q(\underline{r},\hat{\Omega})|^2 \right]^{1/2} \left[\int_{4\pi} d\hat{\Omega}' |Q(\underline{r},\hat{\Omega}')|^2 \right]^{1/2}$$

$$= \int_R d\underline{r} \int_{4\pi} d\hat{\Omega} \, |Q(\underline{r},\hat{\Omega})|^2 \, 4\pi$$

$$= 4\pi \, \|Q\|_0^2$$

Therefore, the scattering bilinear form is bounded from above,

$$m(Q,Q) \le 4\pi \|Q\|_0^2 \tag{106}$$

The significance of Eq. (106) is that one obtains $O(h^{k+1})$ error in the solution when the scattering "norm" is used:

$$m(Q-Q_h, Q-Q_h) \le 4\pi \|Q-Q_h\|_0^2$$

or

$$m(Q-Q_h, Q-Q_h) = O(h^{2k+2})$$

Thus the error in the scattering bilinear form is

$$m(Q-Q_h, Q-Q_h) = O(h^{2k+2}) \tag{108}$$

while the error in the energy is

$$a(Q-Q_h, Q-Q_h) = O(h^{2k+1}) \tag{109}$$

$$c^h = \frac{a(\phi_h, \phi_h)}{m(\phi_h, \phi_h)} \tag{112}$$

Then the error in the eigenvalue could be calculated as

$$\epsilon = |c - c^h| = \left| \frac{a(\phi, \phi)}{m(\phi, \phi)} - \frac{a(\phi_h, \phi_h)}{m(\phi_h, \phi_h)} \right| \tag{113}$$

and then if one could show

$$a(\phi_h, \phi_h) = a(\phi, \phi) + O(h^{2k+1}) \tag{114}$$

$$m(\phi_h, \phi_h) = m(\phi, \phi) + O(h^{2k+2}) \tag{115}$$

the resultant error ϵ would be easily calculated to be

$$\epsilon = |c - c^h| \sim O(h^{2k+1}) \tag{116}$$

This estimate of the error is consistent with eigenvalue results reported in other applications, in that the error in the eigenvalue is equivalent to the error in the energy, i.e.,

$$\Delta\lambda \sim a(\phi - \phi_h, \phi - \phi_h) \tag{117}$$

and since

$$a(\phi - \phi_h, \phi - \phi_h) \sim O(h^{2k+1})$$

the result seems plausible. Also, as will be reported in Sec. VIII.D., the observed eigenvalue convergence was $O(h^{2k+1})$ (at least for k = 1) and thus there is additional numerical evidence for this estimate of the eigenvalue error.

These estimates will now be used to substantiate an estimate of the error in the approximate eigenvalue which is obtained by the finite element approximation. Of course, Eq. (103) is not the eigenvalue problem which is actually solved by the finite element method. The finite element approximation is to choose a finite element subspace $S^h \subset \Phi^1$ and seek a solution $Q_h \in S^h$, as stated below:

Find a real number C^h and a solution $Q_h \in S^h$ such that for all $\psi^h \in S^h$,

$$a(Q_h, \psi^h) = C^h m(Q_h, \psi^h) \tag{108}$$

where

$$\begin{aligned} a(Q_h, \psi^h) = &- (Q_h, \hat{\Omega} \cdot \nabla \psi^h) \\ &+ \langle Q_h, \psi^h \rangle_+ + (Q_h, \psi^h) \end{aligned} \tag{109}$$

and

$$m(Q_h, \psi^h) = \left(\int_{4\pi} d\hat{\Omega}' Q_h(r, \hat{\Omega}'), \psi^h \right) \tag{110}$$

If the actual eigenfunction Q and the approximate eigenfunction Q_h were known, then the eigenvalues could be computed

$$C = \frac{a(Q, 0)}{m(Q, 0)} \tag{111}$$

However, the proof of Eqs. (116) and (117) could not be completed, although similar results could be obtained by a simple application of the triangle inequality

$$\left[a(\phi_h,\phi_h)\right]^{1/2} = \left[a(\phi,\phi)\right]^{1/2} + O\left(h^{k+\frac{1}{2}}\right) \tag{118}$$

and

$$\left[m(\phi_h,\phi_h)\right]^{1/2} = \left[m(\phi,\phi)\right]^{1/2} + O\left(h^{k+1}\right) \tag{119}$$

but these expressions were not helpful in obtaining the eigenvalue estimate.

To summarize the results of this section, a non-optimal estimate of the convergence rate is obtained rigorously. However, numerical results indicate that the optimal convergence is achieved in practice and thus there is incentive to obtain the optimal rate analytically. Several approaches are outlined to obtain this optimal estimate but none were successful in predicting the numerical evidence. In addition, it is also noted that the eigenvalue rates are consistent with the observed L_2 solution convergence rates, thus providing additional assurance that indeed one should be able to demonstrate the optimal convergence rate analytically.

CHAPTER III

SPECIFIC APPLICATION TO 1-D PLANE GEOMETRY

III.A. Derivation of the 1-D Neutron Transport Equation

Consider the plane geometry illustrated in Figure 1, where the independent variables are the spatial coordinate x and the angular coordinate μ, where $\mu = \cos\theta$ and θ is the angle that the neutron velocity subtends with the positive x-axis.

The general geometry neutron transport equation, Eq. (1),

$$\hat{\Omega}\cdot\nabla Q(\underline{r},\hat{\Omega}) + \Sigma_t(\underline{r})\, Q(\underline{r},\hat{\Omega}) = \int_{4\pi} d\hat{\Omega}'\, \Sigma_s(\underline{r};\hat{\Omega}'\to\hat{\Omega})\, Q(\underline{r},\hat{\Omega}') + S(\underline{r},\hat{\Omega}) \tag{1}$$

is transformed to plane geometry as follows. Assume the scattering kernel may be expanded in Legendre polynomials:

$$\Sigma_s(\underline{r},\hat{\Omega}'\to\hat{\Omega}) = \Sigma_s(\underline{r},\hat{\Omega}'\cdot\hat{\Omega})$$

or

$$\Sigma_s(\underline{r},\hat{\Omega}'\cdot\hat{\Omega}) = \sum_{\ell=0}^{L} \frac{2\ell+1}{4\pi}\, b_\ell(\underline{r})\, P_\ell(\hat{\Omega}'\cdot\hat{\Omega})$$

Now the directions $\hat{\Omega}$ and $\hat{\Omega}'$ correspond to the angular coordinates (θ,φ), $(\theta';\varphi')$ respectively, where θ is as defined in Figure 1 and φ is the azimuthal angle around the x-axis. Now since there is aximuthal symmetry, Q does not depend on φ. The volume element $d\hat{\Omega}'$ can be expressed

$$d\hat{\Omega}' = \sin\theta'\, d\theta'\, d\varphi'$$

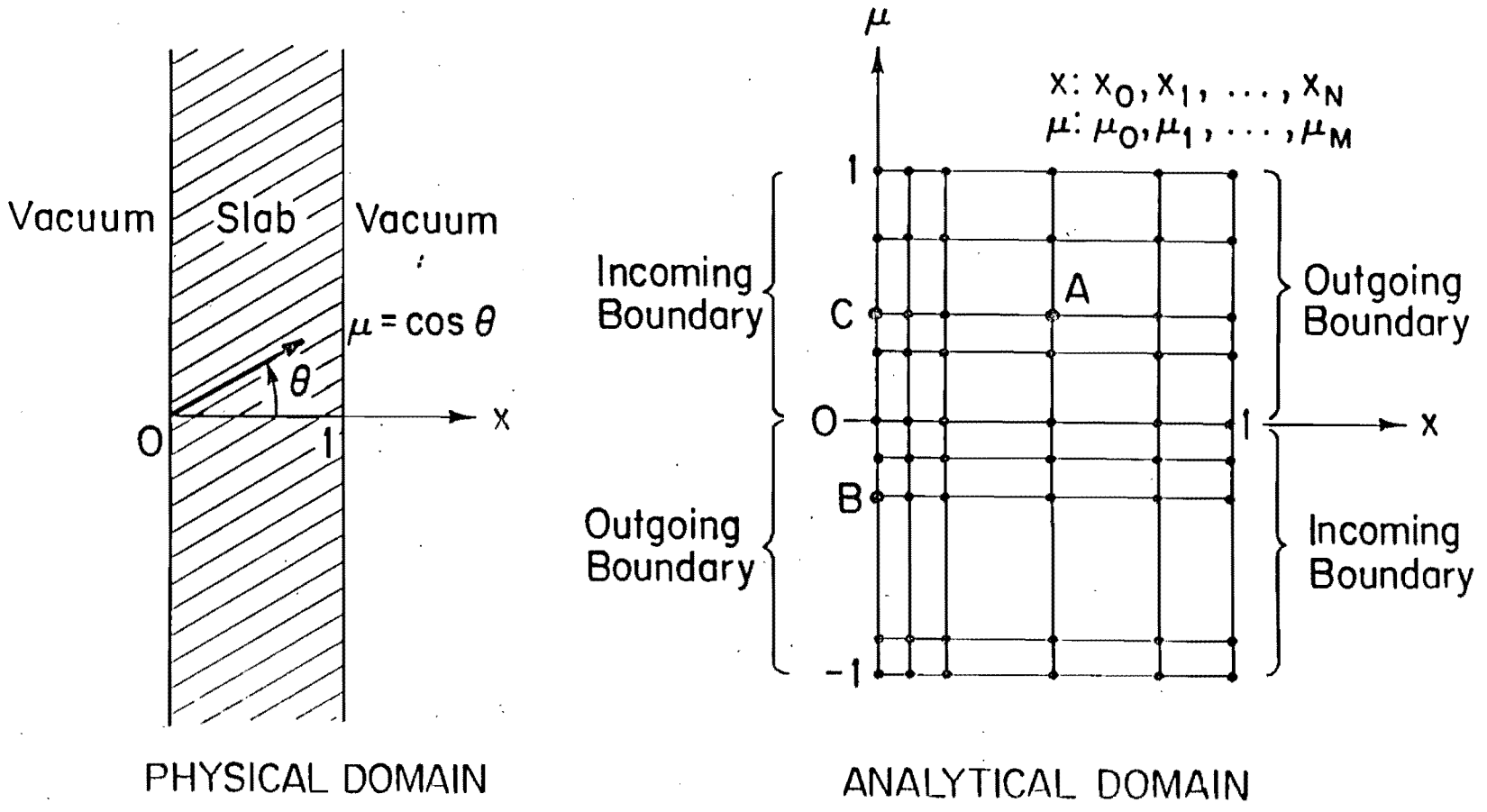


FIGURE 1. 1-D Plane Geometry

and the scattering integral may be integrated over $d\varphi'$, after substituting for $P_\ell(\hat{\Omega}'\cdot\hat{\Omega})$:

$$\int d\hat{\Omega}'\,\Sigma_s(\underline{r},\hat{\Omega}'\cdot\hat{\Omega})\,Q(\underline{r},\hat{\Omega}')$$

$$= \int_0^{\pi} d\theta' \int_0^{2\pi} d\varphi' \sum_{\ell=0}^{L} \frac{2\ell+1}{4\pi} b_\ell(x)\, P_\ell(\hat{\Omega}'\cdot\hat{\Omega})$$

$$= \int_{-1}^{1} d\mu' \int_0^{2\pi} d\varphi' \sum_{\ell=0}^{L} \frac{2\ell+1}{4\pi} b_\ell(x) \Bigg\{ P_\ell(\mu) P_\ell(\mu')$$

$$+ 2\sum_{m=1}^{\ell} \frac{(\ell-m)!}{(\ell+m)!} P_\ell^m(\mu) P_\ell^m(\mu') \cos[m(\varphi-\varphi')] \Bigg\} Q(x,\mu)$$

But $$\int_0^{2\pi} d\varphi' \cos m(\varphi-\varphi') = 0 \quad \text{if } m \neq 0,$$

(note that m is an integer)

therefore,

$$\int_{4\pi} d\hat{\Omega}'\,\Sigma_s(\underline{r},\hat{\Omega}'\cdot\hat{\Omega})\,Q(\underline{r},\hat{\Omega}')$$

$$= 2\pi \int_{-1}^{1} d\mu' \sum_{\ell=0}^{L} \frac{2\ell+1}{4\pi} b_\ell(x)\, P_\ell(\mu) P_\ell(\mu')\, Q(x,\mu')$$

$$= \sum_{\ell=0}^{L} \frac{2\ell+1}{2} b_\ell(x)\, P_\ell(\mu) \int_{-1}^{1} d\mu'\, P_\ell(\mu')\, Q(x,\mu')$$

Also

$$\hat{\Omega}\cdot\nabla Q \longrightarrow \mu \frac{\partial \psi}{\partial x}$$

$$\Sigma_t(\underline{r})\,Q(\underline{r},\hat{\Omega}) \longrightarrow \Sigma_t(x)\,Q(x,\mu)$$

$$S(\underline{r},\hat{\Omega}) \longrightarrow S(x,\mu)$$

and Eq. (1) becomes

$$\mu \frac{\partial Q}{\partial x} + \Sigma_t(x)\, Q(x,\mu) = \sum_{\ell=0}^{L} \frac{2\ell+1}{2} b_\ell(x)\, P_\ell(\mu) \int_{-1}^{1} d\mu'\, P_\ell(\mu')\, Q(x,\mu') + S(x,\mu) \qquad (120)$$

Assuming for the purposes of the present analysis that x varies from 0 to 1, then the allowable boundary conditions at x = 0 and x = 1 are as follows:

(i) Vacuum boundary conditions at x = 0,1:

$$Q(0,\mu) = 0, \quad \mu > 0$$
$$Q(1,\mu) = 0, \quad \mu < 0$$

(ii) Specified incoming flux at x = 0,1:

$$Q(0,\mu) = Q_0(0,\mu), \quad \mu > 0$$
$$Q(1,\mu) = Q_0(1,\mu), \quad \mu < 0$$

(iii) Reflecting boundary conditions at x = 0 and (i) or (ii) at x = 1:

$$Q(0,\mu) = Q(0,-\mu)$$

plus $Q(1,\mu) = 0$

or $Q(1,\mu) = Q_0(1,\mu)$

III.B Integral Law for Plane Geometry

Define the space of allowable trial functions

$$H^1 = \left\{ \Psi(x,\mu) \,\middle|\, \int_0^1 dx \int_{-1}^{1} d\mu \left[|\Psi(x,\mu)|^2 + \left| \frac{\partial \Psi}{\partial x} \right|^2 \right] < \infty \right\}$$

and proceed in an analogous manner with the general formulation, multiplying Eq. (120) by $\Psi(x,\mu)$ where Ψ is arbitrary and integrate over the phase space:

$$\int_0^1 dx \int_{-1}^1 d\mu\, \mu \frac{\partial Q}{\partial x} \psi(x,\mu) + \int_0^1 dx \int_{-1}^1 d\mu\, \Sigma_t(x)\, Q(x,\mu)\psi(x,\mu)$$
$$- \int_0^1 dx \int_{-1}^1 d\mu\, \psi(x,\mu) \sum_{\ell=0}^{L} \frac{2\ell+1}{2} b_\ell(x) P_\ell(\mu) \int_{-1}^1 d\mu' P_\ell(\mu') Q(x,\mu') \qquad (121)$$
$$= \int_0^1 dx \int_{-1}^1 d\mu\, \psi(x,\mu)\, S(x,\mu)$$

Now integrate the first term of Eq. (121) by parts

$$\int_0^1 dx \int_{-1}^1 d\mu\, \mu \frac{\partial Q}{\partial x} \psi(x,\mu) = - \int_0^1 dx \int_{-1}^1 d\mu\, Q(x,\mu)\, \mu \frac{\partial \psi}{\partial x}$$
$$+ \int_{-1}^1 d\mu\, \mu \left[Q(1,\mu)\psi(1,\mu) - Q(0,\mu)\psi(0,\mu) \right] \qquad (122)$$

For simplicity, assume specified incoming boundary conditions at x = 0 and x = 1 and substitute them explicitly into Eq. (122):

$$\int_0^1 dx \int_{-1}^1 d\mu\, \mu \frac{\partial Q}{\partial x} \psi(x,\mu) = - \int_0^1 dx \int_{-1}^1 d\mu\, Q(x,\mu) \frac{\partial \psi}{\partial x}\, \mu$$
$$+ \int_0^1 d\mu\, \mu\, Q(1,\mu)\psi(1,\mu) - \int_{-1}^0 d\mu\, \mu\, Q(0,\mu)\psi(0,\mu) \qquad (123)$$
$$+ \int_{-1}^0 d\mu\, \mu\, Q_2(1,\mu)\psi(1,\mu) - \int_0^1 d\mu\, \mu\, Q_0(0,\mu)\psi(0,\mu)$$

and substitute Eq. (123) into Eq. (121) to yield the final form of the integral law

$$
\begin{aligned}
&-\int_0^1 dx \int_{-1}^1 d\mu\, \mu\, Q(x,\mu)\, \frac{\partial \psi}{\partial x} + \int_0^1 dx \int_{-1}^1 d\mu\, \Sigma_t(x)\, Q(x,\mu)\psi(x,\mu) \\
&\quad + \int_0^1 d\mu\, \mu\, Q(1,\mu)\psi(1,\mu) - \int_{-1}^0 d\mu\, \mu\, Q(0,\mu)\,\psi(0,\mu) \\
&\quad - \int_0^1 dx \int_{-1}^1 d\mu\, \psi(x,\mu) \sum_{\ell=0}^{L} \frac{2\ell+1}{2}\, b_\ell(x)\, P_\ell(\mu) \int_{-1}^1 d\mu'\, P_\ell(\mu')\, Q(x,\mu') \\
&= \int_0^1 dx \int_{-1}^1 d\mu\, S(x,\mu)\,\psi(x,\mu) + \int_0^1 d\mu\, \mu\, Q_0(0,\mu)\,\psi(0,\mu) \\
&\qquad - \int_{-1}^0 d\mu\, \mu\, Q_d(1,\mu)\,\psi(1,\mu) \quad , \text{All } \psi \in H^1 \qquad (124)
\end{aligned}
$$

Thus the task is to find the solution $Q(x,\mu) \in H^1$ such that Eq. (124) is satisfied. Although Eq. (124) is valid only for specified incoming boundary conditions, only changes in the boundary terms are needed to account for vacuum or reflecting boundary conditions. These specific changes will be included in the next section which treats the finite element approximation to solve Eq. (124).

The remarks made earlier concerning the equivalence of the integral law and the partial differential equation + boundary conditions hold for the 1-D case, as would be expected, and no additional comments will be made.

III.C Finite Element Approximation

Proceeding in an analogous manner with Chapter II, choose a finite element subspace $S^h \subset H^1$ with basis

$$\psi_i^h(x,\mu) \quad , i = 1, 2, \cdots, N$$

and seek a solution

$$\Theta_h(x,\mu) = \sum_{j=1}^{N} \phi_j \psi_j^N(x,\mu)$$

such that Eq. (124) is valid for the ψ_i^N, i = 1,2, .., N. This generates a system of equations

$$\underline{\underline{A}}\,\underline{\phi} = \underline{S} \tag{125}$$

where

$$A_{ij} = T_{ij} + K_{ij} + B_{ij}$$

$$S_i = \int_0^1 dx \int_{-1}^1 d\mu\, S(x,\mu)\,\psi_i^N(x,\mu) + R_i$$

and

$$T_{ij} = -\int_0^1 dx \int_{-1}^1 d\mu\, \mu\, \psi_j^N(x,\mu)\,\frac{\partial \psi_i^N}{\partial x}$$

$$K_{ij} = \int_0^1 dx \int_{-1}^1 d\mu\, \Sigma_t(x)\,\psi_i^N(x,\mu)\,\psi_j^N(x,\mu)$$

$$- \int_0^1 dx \sum_{\ell=0}^{L} \frac{2\ell+1}{2} b_\ell(x) \int_{-1}^1 d\mu\, P_\ell(\mu)\,\psi_i^N(x,\mu)$$

$$(x) \int_{-1}^1 d\mu'\, P_\ell(\mu')\,\psi_j^N(x,\mu')$$

and the boundary matrix elements B_{ij} and boundary source term R_i depend on the boundary conditions as explained below.

(i) Vacuum boundary conditions

$$B_{ij} = \int_0^1 d\mu\, \mu\, \psi_i^h(1,\mu)\psi_j^h(1,\mu) - \int_{-1}^0 d\mu\, \mu\, \psi_i^h(0,\mu)\psi_j^h(0,\mu)$$

$$R_i = 0$$

(ii) Specified incoming flux

$$B_{ij} = \int_0^1 d\mu\, \mu\, \psi_i^h(1,\mu)\psi_j^h(1,\mu) - \int_{-1}^0 d\mu\, \mu\, \psi_i^h(0,\mu)\psi_j^h(0,\mu)$$

(same B_{ij} as for vacuum b.c.)

$$R_i = \int_0^1 d\mu\, \mu\, \psi_i^h(0,\mu)\, Q_0(0,\mu) + \int_{-1}^0 d\mu\, \mu\, \psi_i^h(1,\mu) Q_0(1,\mu)$$

(iii) Reflecting boundary conditions

In this case there is an ambiguity that must be resolved on the basis of physical reasoning. Note that the boundary integral in Eq. (122) prior to incorporation of a boundary condition is

$$\text{boundary term} = \int_{-1}^1 d\mu\, \mu \left[Q(1,\mu)\psi(1,\mu) - Q(0,\mu)\psi(0,\mu) \right]$$

The term at $x = 1$ is evaluated in the same manner as above, depending on whether vacuum or specified boundary conditions are imposed at $x = 1$ and will not be further considered. However, the term at $x = 0$

$$B_0 = -\int_{-1}^1 d\mu\, \mu\, Q(0,\mu)\, \psi(0,\mu) \tag{126}$$

can incorporate the reflecting boundary conditions

$$Q(0,\mu) = Q(0,-\mu)$$

in two ways (both clearly allowed mathematically):

(1) Set $\quad Q(0,\mu) = Q(0,-\mu)\,, \quad \mu > 0$

and substitute into Eq. (126)

$$B_0 = -\int_{-1}^{0} d\mu\, \mu\, Q(0,\mu)\Psi(0,\mu) - \int_{0}^{1} d\mu\, \mu\, Q(0,-\mu)\Psi(0,\mu)$$

$$= -\int_{-1}^{0} d\mu\, \mu\, Q(0,\mu)\Psi(0,\mu) + \int_{-1}^{0} d\mu\, \mu\, Q(0,\mu)\Psi(0,-\mu)$$

or

$$B_0 = -\int_{-1}^{0} d\mu\, \mu\, Q(0,\mu)\left[\Psi(0,\mu) - \Psi(0,-\mu)\right] \qquad (127)$$

which generates matrix elements (assuming vacuum b.c. at x = 1)

$$B_{ij} = -\int_{-1}^{0} d\mu\, \mu\, \Psi_j^N(0,\mu)\left[\Psi_i^N(0,\mu) - \Psi_i^N(0,-\mu)\right] \qquad (128)$$
$$+ \int_{0}^{1} d\mu\, \mu\, \Psi_i^N(1,\mu)\, \Psi_j^N(1,\mu)$$

$$R_i = 0$$

The alternative approach to incorporate the reflecting boundary conditions at x=0 is

(2) Set $Q(0,\mu) = Q(0,-\mu)\ ,\ \mu < 0$

and obtain

$$B_0 = -\int_{-1}^{1} d\mu\, \mu\, Q(0,\mu)\Psi(0,\mu)$$

$$= -\int_{-1}^{0} d\mu\, \mu\, Q(0,-\mu)\Psi(0,\mu) - \int_{0}^{1} d\mu\, \mu\, Q(0,\mu)\Psi(0,\mu)$$

$$= \int_{0}^{1} d\mu\, \mu\, Q(0,\mu)\Psi(0,-\mu) - \int_{0}^{1} d\mu\, \mu\, Q(0,\mu)\Psi(0,\mu)$$

or

$$B_0 = \int_{0}^{1} d\mu\, \mu\, Q(0,\mu)\left[\Psi(0,-\mu) - \Psi(0,\mu)\right]$$

(or equivalently, $B_0 = -\int_{-1}^{0} d\mu\, \mu\, Q(0,-\mu)\left[\Psi(0,\mu) - \Psi(0,-\mu)\right]$ for comparison with Eq. 127).

This will generate matrix elements

$$B_{ij} = \int_{0}^{1} d\mu\, \mu\, \Psi_j{}^{h}(0,\mu)\left[\Psi_i{}^{h}(0,-\mu) - \Psi_i{}^{h}(0,\mu)\right] + \int_{0}^{1} d\mu\, \mu\, \Psi_i{}^{h}(1,\mu)\Psi_j{}^{h}(1,\mu) \qquad (129)$$

$$R_i = 0$$

Clearly the matrix elements B_{ij} are different for the two possible methods of incorporating the reflecting boundary conditions. However, one can differentiate between the two on the basis of physical arguments and arrive at the decision as to the proper choice.

The first choice,

$$Q(0,\mu) = Q(0,-\mu) \quad , \mu > 0$$

sets the incoming flux at x = 0 equal to the outgoing flux at x = 0. Thus, if a neutron leaves the region at x = 0, it returns at x = 0 in the reflected plane. Clearly this represents the physical phenomenon of specular reflection. However, the second choice

$$Q(0,\mu) = Q(0,-\mu) \quad , \mu < 0$$

sets the outgoing flux at x = 0 equal to the incoming flux. This does not represent the physical situation, since one would not expect the outgoing flux to be determined entirely by the incoming flux. In addition, mathematically this leads to an ill-posed problem because now the incoming boundary is the "free" boundary whereas the flux on the outgoing boundary is known. Physically this does not make sense and it is interesting to note that the numerical results support this observation in that correct results are obtained with the first choice whereas erroneous results are obtained with the second choice. Thus, it is important to incorporate the reflecting boundary condition in the direction of neutron travel. As will be discussed in a later chapter, this distinction between incorporating a condition in the direction of neutron travel versus against the neutron travel must be kept in mind when allowing discontinuities in the spatial mesh.

III.D. Construction of Finite Element Subspace

No specific mention has been made concerning the specific construction of the subspace S^h, or, equivalently, the construction of the

$\psi_i^N(x,\mu)$, i = 1,2,...., N. The purpose of this section is to examine the possible choices and indicate the reasons for the final selection.

In general the finite element method is employed in a nodal fashion, that is, the expansion coefficients for the solution are nodal parameters which are typically the value of the solution at the node or the value of one of its derivatives. Typically one speaks of a Lagrangian finite element scheme as a scheme that employs Lagrange interpolating polynomials over each finite element and the value of the solution at the nodes are the expansion coefficients and hence are the unknowns in the system of equations. One could also employ Hermite interpolation polynomials on a given finite element mesh; however, in this case each node would include additional parameters representing the value of the solution derivative(s) as well as the solution. Lagrange finite element schemes typically result in continuity of the solution everywhere in the mesh, although derivatives may be discontinuous across interelement boundaries. Hermitian schemes typically result in continuity of one or more partial derivatives everywhere within the mesh.

Since the transport equation is only first order, one can expect at most continuity of the angular flux and even this may be too restrictive for the angular variable in slab geometry. Therefore, the use of Lagrangian elements, which preserve continuity but no derivatives, would appear to be a proper choice for the transport equation.

One may also choose between a tensor product formulation versus a general formulation dependent on the geometry of the finite elements. For example, multi-dimensional finite element basis functions may be easily constructed from a one-dimensional basis as follows.

Assume the x-axis is partitioned into a mesh with nodes x_1, x_2, ... x_L; the y-axis is partitioned with nodes y_1, y_2, ..., y_N; and the z-axis with nodes z_1, z_2....,Z_M. Now the total number of nodes is LMN and with Lagrangian elements one will need LMN basis functions, one for each node. A convenient representation is to choose a one-dimensional basis

$$\psi_1(\xi), \psi_2(\xi), \ldots$$

and form a basis function for the (1,1,1) node as a tensor product of three 1-D basis functions

$$\psi_{1,1,1}(x,y,z) = \psi_1(x)\,\psi_1(y)\,\psi_1(z)$$

or for the (4,9,2) node

$$\psi_{4,9,2}(x,y,z) = \psi_4(x)\,\psi_9(y)\,\psi_2(z) \quad .$$ Thus

the solution will be expanded as

$$Q(x,y,z) = \sum_{j=1}^{LMN} \phi_j\,\psi_j(x,y,z)$$

where each

$$\psi_j(x,y,z) = \psi_{jx}(x)\,\psi_{jy}(y)\,\psi_{jz}(z)$$

where j_x is the x-node corresponding to the global node j, etc. If the one-dimensional basis consisted of linear finite elements, then the 3-D basis would be trilinear. An obvious restriction here is that the mesh must be orthogonal, similar to a finite difference mesh, although a variation on this has been reported[39] which allows local mesh refinement of small rectangles (or cubes) which lie entirely within a larger

element, hence bypassing the restriction that all nodes lie on unbroken, orthogonal coordinate lines.

The more general method is to shape the basis functions to the particular element. For example a triangular element, since it has three nodes, can uniquely represent a linear polynomial over a 2-D surface, with each node contributing a piece of data. That is, a general linear polynomial in two variables

$$f(x,y) = a_0 + a_1 x + a_2(y)$$

needs the value of three parameters to be uniquely determined. If the values of the solution at the nodes of a triangle are taken to be the three parameters, then the linear polynomial is uniquely determined within the triangle. With Lagrangian elements, where the nodal parameters are the values of the unknown solution, a convenient basis for a triangle consists of three linear polynomials, each of which is unity at one node and zero at the other nodes. This will result in the expansion coefficients being identical to the solution values at the nodes. This concept of choosing a basis consisting of functions which are unity at a particular node and zero at all other nodes is typical of Lagrangian elements, and is used for the tensor product basis functions discussed earlier or for quite general multi-dimensional schemes.

For this investigation, tensor product Lagrangian basis functions are used in the spatial (x) and angular (μ) directions. The reasons for the choice of Lagrangian type elements was discussed earlier and the reasons for the choice of tensor product basis functions are simplicity of application and the computational ease with which the matrix elements are calculated.

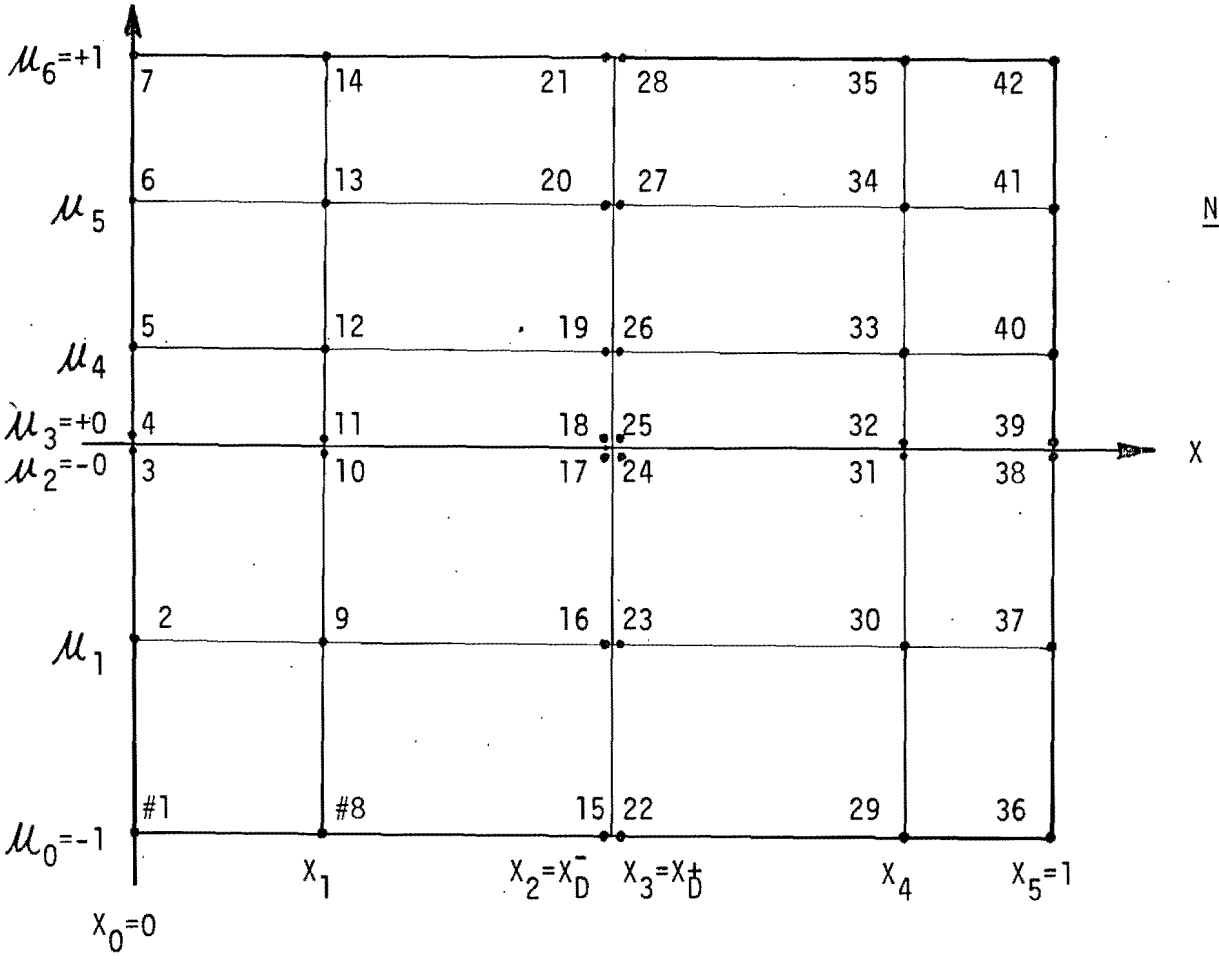
The explicit construction of the basis functions is based on the nodal numbering scheme illustrated in Figure 2. Assuming N + 1 spatial nodes $x_0, x_1, \ldots, x_N$ ranging from x = 0 to x = 1 and M + 1 angular nodes $\mu_0, \mu_1, \ldots, \mu_M$ from $\mu_0 = -1$ to $\mu_M = +1$, a basis function at a globally numbered node j is entirely determined by the corresponding x-node jx and μ-node $j\mu$. Thus j = 1,2,..., (M+1)(N+1) and to each node j corresponds the nodal pair (jx, $j\mu$) which uniquely determines the basis function corresponding to the node j:

$$\Psi_j^h(x,\mu) = \Psi_{jx}(x)\,\Psi_{j\mu}(\mu)$$

where the superscript h is suppressed for the individual 1-D basis functions $\Psi_k(z)$.

The one-dimensional basis functions are constructed as follows. For linear elements, a basis function is unity at the node it is identified with, and zero at the adjacent nodes, varying linearly between the nodes. For a mesh $z_0, z_1 \ldots$ the basis function corresponding to the node z_i has the following functional dependence:

$$\Psi_i(z) = \begin{cases} 0, & z \le z_{i-1} \\ \dfrac{z - z_{i-1}}{z_i - z_{i-1}}, & z_{i-1} \le z \le z_i \\ \dfrac{z_{i+1} - z}{z_{i+1} - z_i}, & z_i \le z \le z_{i+1} \\ 0, & z \ge z_{i+1} \end{cases}$$



NOTE :

-Discontinuous Angular Elements at μ=0

-Discontinuous Spatial Elements at X=X_D

FIGURE 2. Nodal Numbering

Note that $\psi_i(z_j)=\delta_{ij}$ as desired and the $\psi_i(z)$ is continuous but does not possess a derivative at z_i . Figure 3 illustrates the basis functions on a typical mesh, which indicates why these functions are called "tent" functions or "hill" functions in the finite element literature.

For quadratic basis functions, the two nodes of either end of an interval are not sufficient to uniquely determine the polynomial. Therefore, an additional node, termed an interior node, is placed midway between the given nodes, which are referred to as principal nodes. Again the prescription is that the basis function is unity at the given node (whether interior or principal) and zero at the others. The functional dependence for the quadratic basis functions is easily determined because the zeroes are known, and the normalization of unity yields the constant:

For a principal node, (z_{i-1}, z_{i+1} are then interior nodes)

$$\psi_i(z) = \begin{cases} 0, & z \le z_{i-2} \text{ OR } z \ge z_{i+2} \\ \dfrac{(z-z_{i-2})(z-z_{i-1})}{(z_i-z_{i-2})(z_i-z_{i-1})}, & z_{i-2} \le z \le z_i \\ \dfrac{(z-z_{i+1})(z-z_{i+2})}{(z_i-z_{i+1})(z_i-z_{i+2})}, & z_i \le z \le z_{i+2} \end{cases}$$

and for an interior node (z_{i+1}, z_{i-1} are principal nodes)

$$\psi_i(z) = \begin{cases} 0, & z \le z_{i-1} \text{ OR } z \ge z_{i+1} \\ \dfrac{(z-z_{i-1})(z-z_{i+1})}{(z_i-z_{i-1})(z_i-z_{i+1})}, & z_{i-1} \le z \le z_{i+1} \end{cases}$$

Figure 4 illustrates the quadratic basis functions. Note that the basis functions are continuous and have derivatives except at the principal

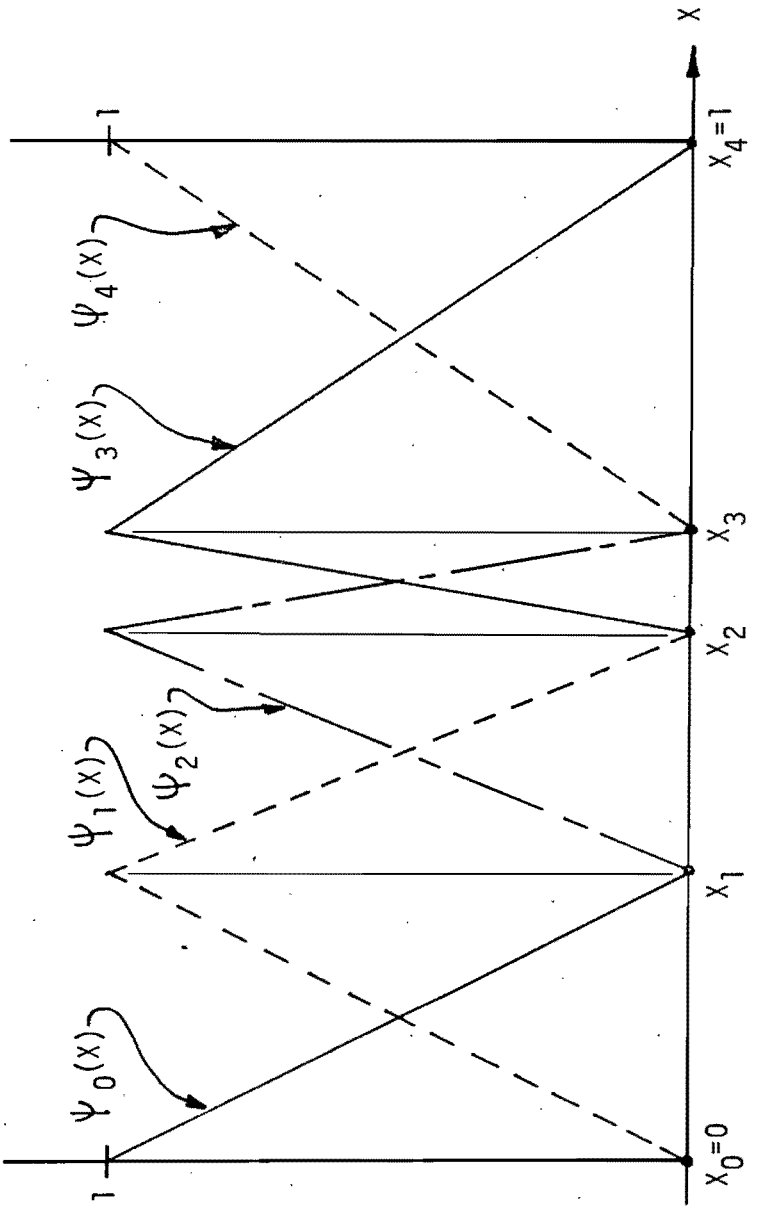


FIGURE 3. Linear Basis Functions

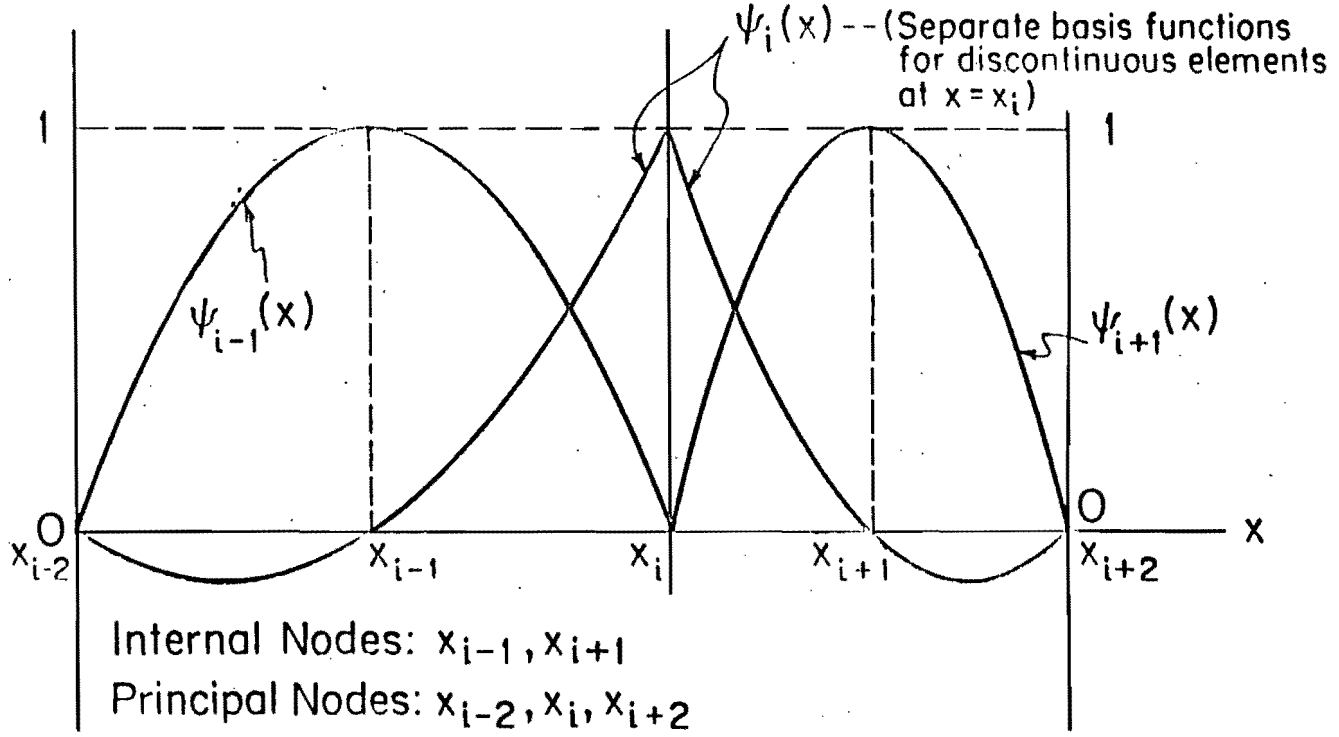


FIGURE 4. Quadratic Basis Functions

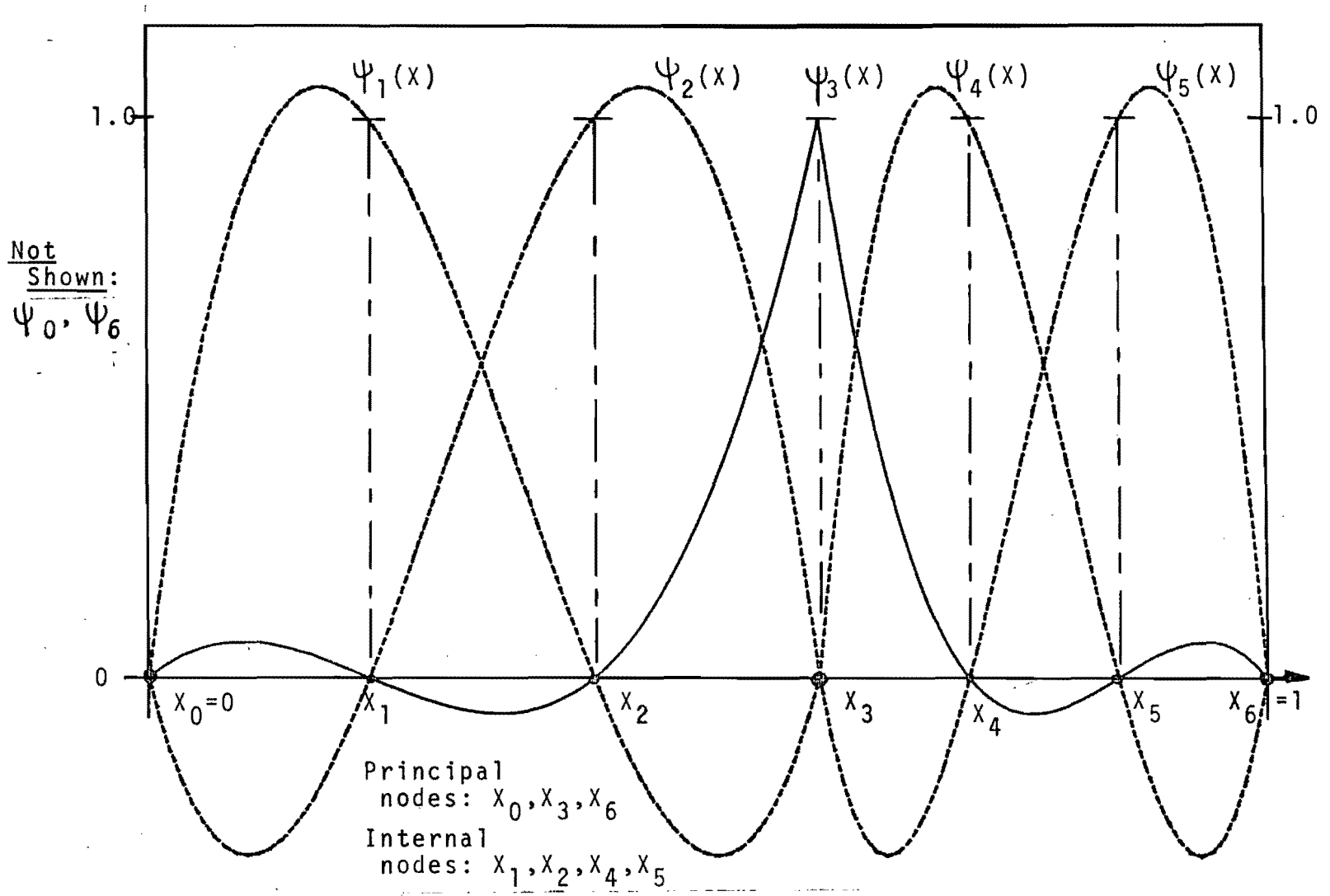


Figure 5. Cubic Basis Functions

nodes, although the basis function corresponding to the interior node has derivatives of all orders within (z_{i-1}, z_{i+1}) .

The cubic basis functions are the obvious generalization of the quadratic functions, with two evenly spaced interior nodes. The process of obtaining their functional dependence is similar to the quadratic case and will not be repeated, although Figure 5 illustrates the functional form. Also, one never needs this explicit functional form anyway because all basis functions are only considered on the standard interval $[-1,1]$, as will be discussed shortly.

III.E. Calculation of Matrix Elements

This section discusses the explicit computation of the matrix elements A_{ij}, including methods for minimizing the calculations required of the computer.

With a tensor product formulation, the matrix elements become:

$$T_{ij} = -\left[\int_0^1 dx\, \psi_{jx}(x) \frac{\partial \psi_{ix}}{\partial x}\right]\left[\int_{-1}^{1} d\mu\, \mu\, \psi_{i\mu}(\mu)\psi_{j\mu}(\mu)\right]$$

$$K_{ij} = \left[\int_0^1 dx\, \Sigma_t(x)\psi_{ix}(x)\psi_{jx}(x)\right]\left[\int_{-1}^{1} d\mu\, \psi_{i\mu}(\mu)\psi_{j\mu}(\mu)\right]$$

$$- \sum_{\ell=0}^{L}\left(\frac{2\ell+1}{2}\right)\left\{\left[\int_0^1 dx\, \psi_{ix}(x)\psi_{jx}(x) b_\ell(x)\right]\left[\int_{-1}^{1} d\mu\, P_\ell(\mu)\psi_{i\mu}(\mu)\right]\right.$$

$$\left.(x)\left[\int_{-1}^{1} d\mu'\, P_\ell(\mu')\psi_{j\mu}(\mu')\right]\right\}$$

$$B_{ij} = \int_0^1 d\mu\, \mu\, \psi_{i\mu}(\mu)\, \psi_{j\mu}(\mu) \left[\psi_{ix}(1)\psi_{jx}(1)\right] - \int_{-1}^{0} d\mu\, \mu\, \psi_{i\mu}(\mu)\, \psi_{j\mu}(\mu) \left[\psi_{ix}(0)\psi_{jx}(0)\right] .$$

Note that the uncoupling of the spatial and angular integrals significantly reduces the number of unique integrals to be computed.

Although the range of integration for the integrals is shown to be over the full angular or spatial interval, in actuality the basis functions are local, extending only over a few mesh intervals as discussed above and most of these matrix elements are zero. There is a notable exception though with the scattering term, where the coupling between two angular nodes is (suppressing the $P_\ell(\mu)$)

$$\int_{-1}^{1} d\mu\, \psi_{i\mu}(\mu) \int_{-1}^{1} d\mu'\, \psi_{j\mu}(\mu')$$

which is a product of integrals rather than an integral of the product of basis functions which appears in all of the other matrix elements. Therefore, the presence of scattering couples all of the angular nodes, although the spatial coupling is not affected. Computationally this results in an increase in the number of non-zero matrix elements.

Although at first glance it might appear that the computation of these matrix elements is still a formidable task, in practice the task is relatively easy. The reason for this is that all integrals are done on the standard interval [-1, 1] by Gaussian quadrature and then mapped into the particular mesh integral by a simple linear transformation.

For example, if an integral (I) is being performed from x_i to x_{i+1} (or μ_i to μ_{i+1}) then one has the transformation:

$$X = \frac{X_{i+1} - X_i}{2} \xi + \frac{X_{i+1} + X_i}{2}$$

which results in

$$\int_{X_i}^{X_{i+1}} dX \, \Psi_i(X) \Psi_{i+1}(X) = \frac{X_{i+1} - X_i}{2} \int_{-1}^{1} d\xi \, \Psi_a(\xi) \Psi_b(\xi)$$

or for an angular integral

$$\int_{\mu_i}^{\mu_{i+1}} d\mu \, \mu \, \Psi_i(\mu) \Psi_{i+1}(\mu) = \frac{\mu_{i+1} - \mu_i}{2} \int_{-1}^{1} d\xi \left[\frac{\mu_{i+1} - \mu_i}{2} \xi + \frac{\mu_{i+1} + \mu_i}{2} \right] \Psi_a(\xi) \Psi_b(\xi)$$

where $\Psi_a(\xi)$, $\Psi_b(\xi)$ are generic basis functions of which there are only k+1 for the entire system, where k = 1,2, or 3 for linear, quadratic, or cubic elements, respectively. Figure 6 illustrates the generic basis functions on the standard interval [-1,1] for quadratic elements.

Thus the construction of the (N+1)(M+1) dimensional subspace S^h of trial functions corresponding to a mesh of N+1 spatial and M+1 angular nodes, has been reduced to specifying a few unique basis functions (e.g. 2 unique basis functions for linear elements in space and angle). More important though, the computation of the matrix elements A_{ij} is immensely simplified. For example, a linear finite element mesh with 100 spatial nodes and 10 angular nodes will result in 30,000 non-zero matrix elements A_{ij}, each of which may be composed of up to 7 different integrals, or perhaps 150,000 integrals to be computed. However, use of the standard

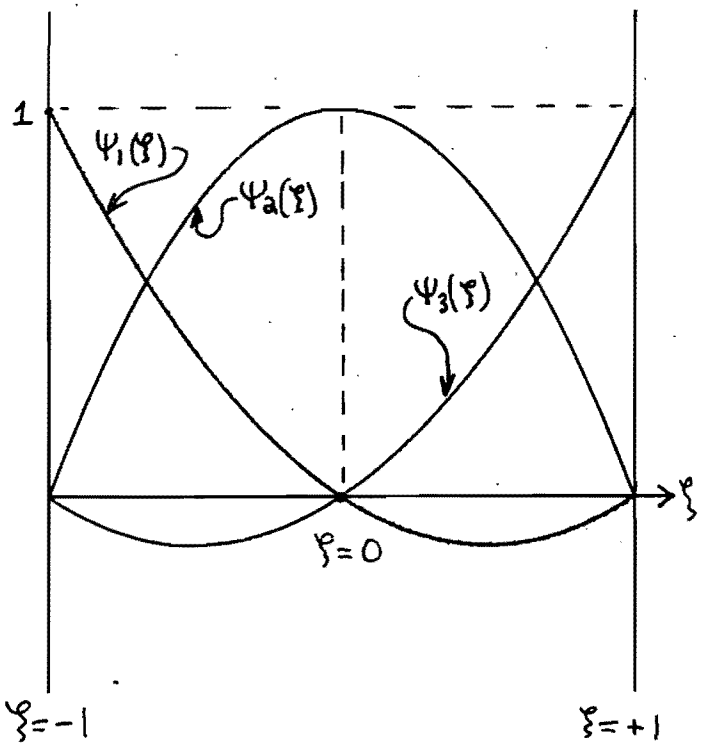


FIGURE 6

Quadratic Basis Functions On Standard Interval [-1,1]

interval results in only 20 unique integrals to be evaluated, clearly a considerable savings in effort.

III.F. Equivalent Finite Difference Relations

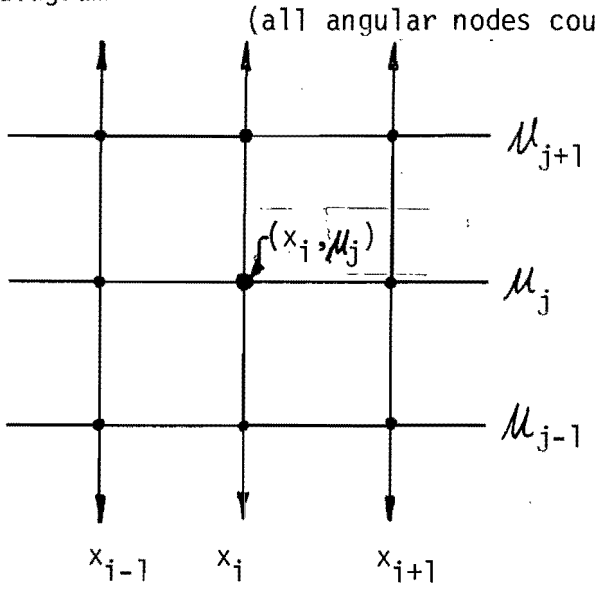
This section derives the equivalent finite difference relations which correspond to the use of linear finite elements on a uniform mesh. These relations are obtained for three typical nodes--one in the interior, one on the incoming boundary, and one on the outgoing boundary of a homogeneous slab in a vacuum. Figure 1 indicates a typical mesh with the three nodes indicated. The reason for obtaining these equivalent relations is twofold. First the truncation error associated with the approximation can be determined immediately and can be used for estimates of the pointwise error, as will be done later. Secondly, the relations offer insight as to the manner in which the solution is propagated from one node to another.

To determine the equivalent finite difference relations, one only needs to evaluate those matrix elements A_{IJ} which are non-zero, assuming node I (with coordinates x_i, μ_j) is the node in question. That is, the equation in which ϕ_I appears is

$$\sum_{J=1}^{(N+1)(M+1)} A_{IJ}\phi_J = S_I$$

and the calculation of the A_{IJ} will immediately yield the finite difference relations, which relate the nodal fluxes to one another.

Interior Node (A) The nodes which interact with node I are shown in the following diagram:



Thus the 8 surrounding nodes and all of the angular nodes within the interacting spatial distance will interact with node (x_i, μ_j). Note that the global indexing I, J is completely different from the local nodal numbering i,j which is used in this section for convenience (i.e., $\phi_I = \phi_{i,j}$). The summation $\sum A_{IJ}\phi_J$ is then calculated

$$\sum_{J=1}^{(N+1)(M+1)} A_{IJ}\phi_J = \left(\frac{\mu_{j+1}+\mu_j}{2}\right)\left(\frac{\phi_{i+1,j+1}-\phi_{i-1,j+1}}{2\Delta x}\right)\left(\frac{\Delta x \Delta \mu}{6}\right)$$

$$+ \quad 4\mu_j\left(\frac{\phi_{i+1,j}-\phi_{i-1,j}}{2\Delta x}\right)\left(\frac{\Delta x \Delta \mu}{6}\right)$$

$$+\left(\frac{\mu_j+\mu_{j-1}}{2}\right)\left(\frac{\phi_{i+1,j-1}-\phi_{i-1,j-1}}{2\Delta x}\right)\left(\frac{\Delta x \Delta \mu}{6}\right)$$

$$+ \sum_t \left[\phi_{i-1,j+1} + 4\phi_{i,j+1} + \phi_{i+1,j+1}\right. \tag{130}$$

$$+ \quad 4\phi_{i-1,j} + 16\phi_{i,j} + 4\phi_{i+1,j}$$

$$+ \quad \phi_{i-1,j-1} + 4\phi_{i,j-1} + \phi_{i+1,j-1}\Big] \frac{\Delta x \Delta\mu}{36}$$

$$+ \frac{\Sigma_s}{2}\Bigg\{\Bigg[\phi_{i-1,0}\frac{\Delta\mu}{2} + \phi_{i-1,M}\frac{\Delta\mu}{2} + \sum_{j=1}^{M-1}\phi_{i-1,j}\Delta\mu\Bigg]$$

$$+ 4\Bigg[\phi_{i,0}\frac{\Delta\mu}{2} + \phi_{i,M}\frac{\Delta\mu}{2} + \sum_{j=1}^{M-1}\phi_{i,j}\Delta\mu\Bigg]$$

$$+ \quad \Bigg[\phi_{i+1,0}\frac{\Delta\mu}{2} + \phi_{i+1,M}\frac{\Delta\mu}{2} + \sum_{j=1}^{M-1}\phi_{i+1,j}\Delta\mu\Bigg]\Bigg\}\frac{\Delta x\Delta\mu}{6}$$

Since the original partial differential equation is

$$\mu\frac{\partial\psi}{\partial x} + \Sigma_t\psi(x,\mu) - \frac{\Sigma_s}{2}\int_{-1}^{1}d\mu'\psi(x,\mu') = S(x,\mu)$$

the approximation to each term can be obtained by inspection. In particular, $\mu\frac{\partial\psi}{\partial x}$ is replaced by a weighted (in angle) centered space difference scheme, or "leap-frog" scheme[40]:

$$\mu\frac{\partial\psi}{\partial x} \approx \frac{1}{6}\left(\frac{\mu_{j+1}+\mu_j}{2}\right)\left(\frac{\psi_{i+1,j+1}-\psi_{i-1,j+1}}{2\Delta x}\right)$$

$$+ \frac{2}{3}\mu_j\left(\frac{\psi_{i+1,j}-\psi_{i-1,j}}{2\Delta x}\right) \qquad (131)$$

$$+ \frac{1}{6}\left(\frac{\mu_j+\mu_{j-1}}{2}\right)\left(\frac{\psi_{i+1,j-1}-\psi_{i-1,j-1}}{2\Delta x}\right)$$

The collision term $\Sigma_t Q(x,\mu)$ becomes

$$\Sigma_t Q(x,\mu) \approx$$

$$\Sigma_t \Big[\frac{1}{36} Q_{i-1,j+1} + \frac{1}{9} Q_{i,j+1} + \frac{1}{36} Q_{i+1,j+1} + \frac{1}{9} Q_{i-1,j} + \frac{4}{9} Q_{i,j} + \frac{1}{9} Q_{i+1,j} + \frac{1}{36} Q_{i-1,j-1} + \frac{1}{9} Q_{i,j-1} + \frac{1}{36} Q_{i+1,j-1} \Big] \tag{132}$$

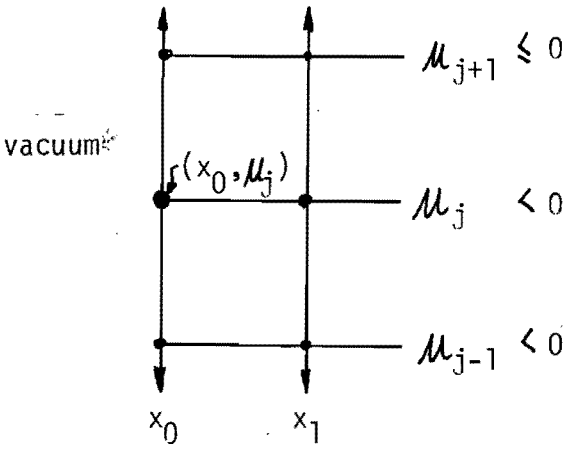
which is a cell average of the total collision term. The scattering term is approximated as

$$\frac{\Sigma_s}{2} \int_{-1}^{1} d\mu' Q(x,\mu')$$

$$\approx \frac{\Sigma_s}{2} \Bigg\{ \frac{1}{6} \Big[Q_{i-1,0} \frac{\Delta\mu}{2} + Q_{i-1,M} \frac{\Delta\mu}{2} + \sum_{j=1}^{M-1} Q_{i-1,j} \Delta\mu \Big] + \frac{2}{3} \Big[Q_{i,0} \frac{\Delta\mu}{2} + Q_{i,M} \frac{\Delta\mu}{2} + \sum_{j=1}^{M-1} Q_{i,j} \Delta\mu \Big] + \frac{1}{6} \Big[Q_{i+1,0} \frac{\Delta\mu}{2} + Q_{i+1,M} \frac{\Delta\mu}{2} + \sum_{j=1}^{M-1} Q_{i+1,j} \Delta\mu \Big] \tag{133}$$

which is exactly a mid-point rule for the scattering integral, weighted by the spatial node location.

Outgoing boundary node (B) The nodes which interact with B are shown in the following diagram:



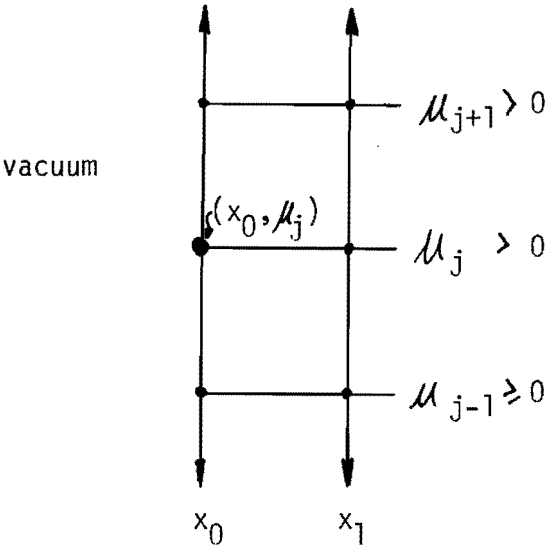
and an evaluation of the Ith equation yields the following approximations for the various terms:

$$\mu \frac{\partial Q}{\partial X} \approx \frac{1}{6}\left(\frac{\mu_{j+1}+\mu_j}{2}\right)\left(\frac{Q_{1,j+1}-Q_{0,j+1}}{\Delta X}\right) + \frac{2}{3}\mu_j\left(\frac{Q_{1,j}-Q_{0,j}}{\Delta X}\right) + \frac{1}{6}\left(\frac{\mu_j+\mu_{j-1}}{2}\right)\left(\frac{Q_{1,j-1}-Q_{0,j-1}}{\Delta X}\right) \tag{134}$$

which is identical to the approximation for the interior node except the derivative is taken over one cell.

The approximate relations for collision term $\Sigma_t Q(X,\mu)$ and the scattering term $\frac{\Sigma_s}{2}\int_{-1}^{1} d\mu' Q(X,\mu')$ are similar to the relations for the interior node (A) and will not be repeated. The interesting point is that the transport term is treated similarly for the interior node and the outgoing boundary node.

Incoming boundary node (C) The nodes which interact with node C are shown below:



and the approximations to the various terms are as follows:

$$\mu \frac{\partial Q}{\partial x} \approx \frac{1}{6}\left(\frac{\mu_{j+1}+\mu_j}{2}\right)\left(\frac{Q_{1,j+1}+Q_{0,j+1}}{\Delta x}\right) + \frac{2}{3}\mu_j\left(\frac{Q_{1,j}+Q_{0,j}}{\Delta x}\right) + \frac{1}{6}\left(\frac{\mu_j+\mu_{j-1}}{2}\right)\left(\frac{Q_{1,j-1}+Q_{0,j-1}}{\Delta x}\right) \tag{135}$$

which is significantly different than the corresponding term for nodes A or B. Note that the derivative term is no longer replaced by a difference, but a sum, which accounts for the fact that a transportive term is not desired at the incoming boundary. The reason this difference relation results in a natural vacuum boundary condition may be motivated as follows:

Requiring $\frac{Q_1 + Q_0}{\Delta X}$ to approximate $\frac{\partial Q}{\partial X}$ over the interval $[x_0, x_1]$ is tantamount to requiring

$$\frac{Q_1 + Q_0}{\Delta X} = \frac{Q_1 - Q_0}{\Delta X}$$

which can only be true in general if $Q_0 = 0$. Thus this relation attempts to force Q to be zero on the incoming boundary, which is desired.

The other terms, $\Sigma_t Q(x,\mu)$ and $\frac{\Sigma_s}{2}\int_{-1}^{1} d\mu' Q(x,\mu')$ are similar to the expressions for the boundary node (B).

To calculate the truncation error associated with these approximations, consider the $\mu \frac{\partial Q}{\partial x}$ term for the interior node A.

Expand the terms in a Taylor series about (x_i, μ_j), e.g.,

$$Q_{i+1,j-1} \approx Q_{i,j} + \left.\frac{\partial Q}{\partial x}\right|_{i,j} \Delta x - \left.\frac{\partial Q}{\partial \mu}\right|_{i,j} \Delta\mu + \frac{1}{2}\left.\frac{\partial^2 Q}{\partial x^2}\right|_{i,j} \Delta x^2$$

$$+ \frac{1}{2}\left.\frac{\partial^2 Q}{\partial \mu^2}\right| \Delta\mu^2 - \left.\frac{\partial^2 Q}{\partial x \partial \mu}\right| \Delta x \Delta\mu + O(h^3)$$

(where $h \sim \Delta X$ and $h \sim \Delta\mu$)

and substitute into the approximate relation, Eq. (131), for $\mu \frac{\partial Q}{\partial x}$

$$\mu \frac{\partial Q}{\partial x} \approx \frac{1}{6}\left(\mu_j + \frac{\Delta\mu}{2}\right)\left[\frac{2\left.\frac{\partial Q}{\partial x}\right| \Delta x + 2\left.\frac{\partial^2 Q}{\partial x \partial \mu}\right| \Delta x \Delta\mu + O(h^3)}{2\Delta X}\right]$$

$$+ \frac{2}{3}\mu_j\left[\frac{2\left.\frac{\partial Q}{\partial x}\right| \Delta x + O(h^3)}{2\Delta X}\right]$$

$$+ \frac{1}{6}\left(\mu_j - \frac{\Delta\mu}{2}\right)\left[\frac{2\left.\frac{\partial Q}{\partial x}\right| \Delta x - 2\left.\frac{\partial^2 Q}{\partial x \partial \mu}\right| \Delta x \Delta\mu + O(h^3)}{2\Delta X}\right]$$

or $$\mu \frac{\partial Q}{\partial x} \approx \mu_j \frac{\partial Q}{\partial x}(x_i, \mu_j) + O(h^a)$$

thus the truncation error for the transport term is $O(h^2)$.

Similarly, the collision term, Eq. (132), is

$$\Sigma_t Q(x, \mu) \approx \Sigma_t Q(x_i, \mu_j) + O(h^a)$$

and the scattering integral, Eq. (133), becomes

$$\frac{\Sigma_s}{2} \int_{-1}^{1} d\mu' Q(x, \mu') \approx \frac{\Sigma_s}{2} \int_{-1}^{1} d\mu' Q(x_i, \mu') + O(h^2)$$

The boundary transport terms are not quite as accurate as the interior terms, because the derivative term is essentially a one-sided difference which has $O(h)$ error. However, for the mesh cell (on the boundary), the overall accuracy is still $O(h^2)$. In addition, this apparent loss of accuracy is only within $O(h)$ of the boundary, thus the overall truncation error will remain $O(h^2)$ for linear elements.

III.G. Explicit Demonstration of $O(h^2)$ Convergence with Linear Finite Elements

As the numerical results presented in Chapter VIII indicate, the observed convergence rate to the actual solution for finite elements of degree k is $O(h^{k+1})$ when the error is measured in the L_2-norm. However, as discussed in Chapter II this estimate is difficult to show theoretically. This section is intended to provide more evidence that indeed $O(h^{k+1})$ convergence should be obtained, by explicitly demonstrating that $O(h^2)$ convergence is obtained with linear finite elements (k=1). This is done by calculating the truncation error in the finite element ap-

proximation and bounding the error in the solution with the truncation error result and the observed norm of the inverse of the matrix of coefficients. A similar procedure has been suggested by Fried[41] and by Kang and Hansen[42] in applications in other areas.

Define the approximate solution vector as $\underline{Q}_h$, which is the vector of the nodal angular fluxes which are determined by a linear finite element scheme on a uniform mesh with mesh spacing h. Define the vector of the exact nodal fluxes as $\underline{Q}$. Now the results of the previous section indicate that the linear finite element scheme on a uniform mesh is $O(h^2)$ accurate, hence

$$\underline{\underline{A}}\left(\underline{Q} - \underline{Q}_h\right) = \underline{C}\, h^4 \tag{136}$$

where the additional h^2 is due to the scaling of the elements of $\underline{\underline{A}}$ (all integrals are $\iint_{\Delta x \Delta \mu} dx d\mu = O(h^2)$) and $\underline{C}$ is a constant vector independent of h.

Define the "sup" norm as the maximum absolute component of a vector, i.e.,

$$\|\underline{C}\|_\infty \equiv \text{MAX}\left(|C_1|, |C_2|, |C_3|, \cdots, |C_N|\right)$$

then the matrix norm consistent with this norm is defined as follows[43]:

$$\|\underline{\underline{A}}\|_\infty \equiv \max_i \left[\sum_j |A_{ij}|\right]$$

i.e., the maximum absolute row sums of the matrix $\underline{\underline{A}}$.

Using these definitions, formally solve Eq. (136) for the error

$$\underline{Q} - \underline{Q}_h = \underline{\underline{A}}^{-1}\, \underline{C}\, h^4 \tag{137}$$

which is possible because $\underline{\underline{A}}$ is non-singular. Now take the sup-norm of each side of Eq. (137):

$$\|\underline{Q} - \underline{Q}_h\|_\infty \leq \|\underline{\underline{A}}^{-1}\|_\infty \|\underline{C}\|_\infty h^4 \tag{138}$$

The left hand side of Eq. (138) is the maximum nodal error in the approximate solution. At this point, since little is available in the literature concerning the norms of asymmetric matrices (let alone their inverses), numerical results are used to estimate $\|\underline{\underline{A}}^{-1}\|_\infty$. The matrix $\underline{\underline{A}}$ was assembled for several different meshes and the inverse was explicitly calculated using the LU method (to be described later) and back-solving for N right hand sides, each of which being a unit vector. Then the norm of $\underline{\underline{A}}^{-1}$ was calculated in accordance with the definition above. The results are illustrated in Figure 7 , which also has an insert tabulating the results. Note that a good fit is obtained with all of the data points, and this fit indicates

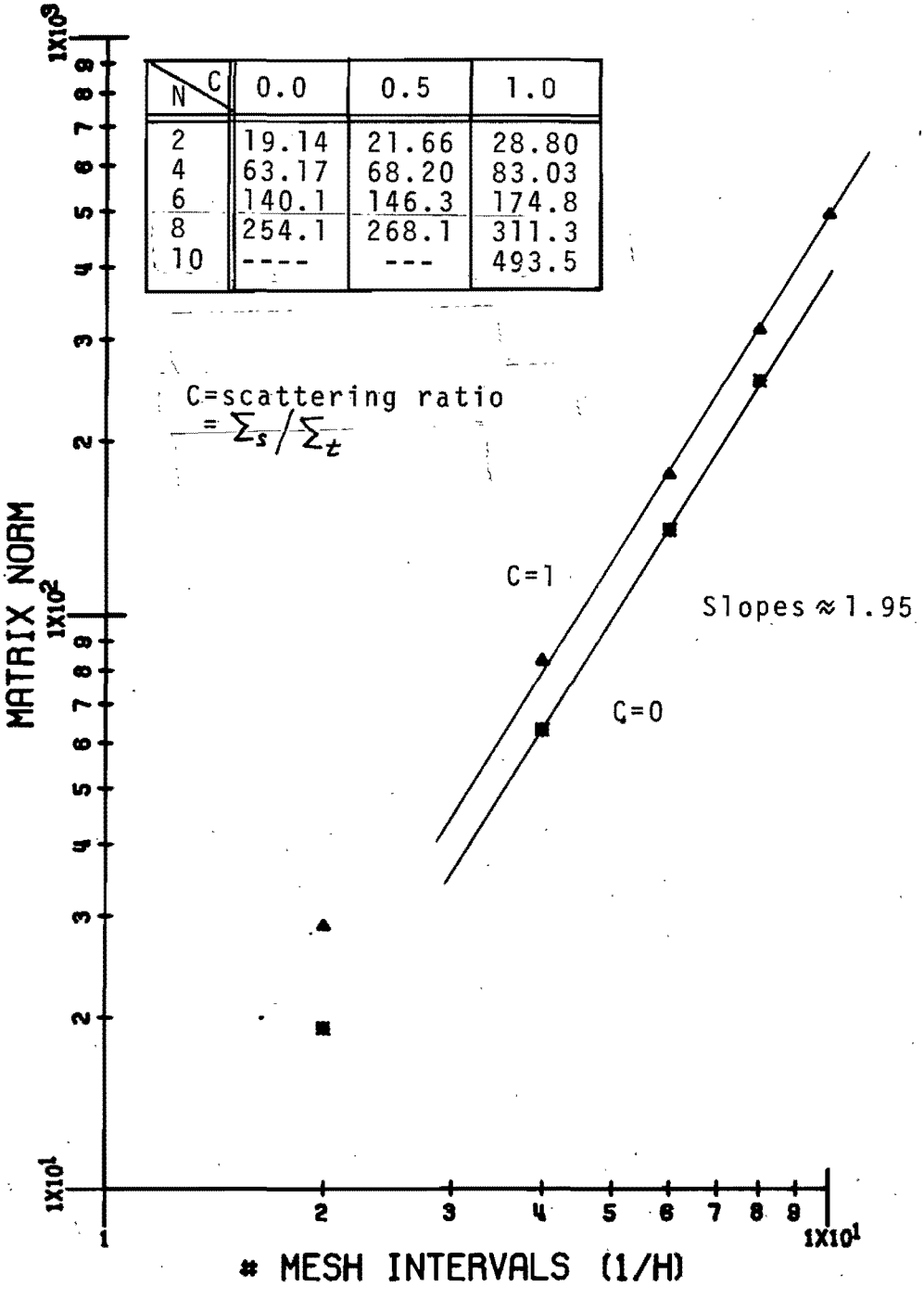
$$\|\underline{\underline{A}}^{-1}\|_\infty = O\left(\frac{1}{h^2}\right)$$

which when substituted into Eq. (138) yields

$$\|\underline{Q} - \underline{Q}_h\|_\infty \leq C_\infty h^2 \tag{139}$$

where C_∞ is a constant. Thus the pointwise (nodal) error is $O(h^2)$ for linear elements.

This error estimate can be extended to the L_2-norm as follows. Define $\tilde{Q}(x,\mu)$ as the interpolant of the exact flux $Q(x,\mu)$ in S^h. Thus, $\tilde{Q}(x,\mu) = Q(x,\mu)$ at the nodes. From approximation theory one obtains



N \ C	0.0	0.5	1.0
2	19.14	21.66	28.80
4	63.17	68.20	83.03
6	140.1	146.3	174.8
8	254.1	268.1	311.3
10	----	---	493.5

Figure 7. Matrix Norm $\|\underline{\underline{A}}^{-1}\|_\infty$ vs. Mesh Spacing

$$\| Q - \tilde{Q} \|_0 \leq C_0 h^2 \tag{140}$$

But the vector of $\tilde{Q}$ is exactly $\underline{Q}$ as used above, since $\tilde{Q}$ and Q coincide at the nodes. But $Q_h(x,\mu)$ and $\tilde{Q}(x,\mu)$ are each piecewise linear functions which differ at most by $C_\infty h^2$ at the nodes. Therefore,

$$\| \tilde{Q} - Q_h \|_0 \equiv \left[\int_0^1 dx \int_{-1}^{1} d\mu \, | \tilde{Q}(x,\mu) - Q_h(x,\mu) |^2 \right]^{1/2}$$

$$\leq \sqrt{2}\, C_\infty h^2$$

But from the triangle inequality

$$\| Q - Q_h \|_0 = \| Q - \tilde{Q} + \tilde{Q} - Q_h \|_0$$

$$\leq \| Q - \tilde{Q} \|_0 + \| \tilde{Q} - Q_h \|_0$$

$$\leq C_0 h^2 + \sqrt{2}\, C_\infty h^2$$

OR $$\| Q - Q_h \| \leq C h^2$$

which is the desired error estimate in the L_2 norm. Thus this explicit demonstration of the convergence, which depended on a numerical estimate for the inverse matrix norm, supports the numerical results.

These results would seem to indicate $O(h^{k+1})$ for general finite elements of degree k because one would not expect $\| \underline{A}^{-1} \|_\infty$ to alter its dependence on the mesh spacing. This coupled with the decreased truncation error in the approximation, which should be $O(h^{k+1})$, will result in the expected $O(h^{k+1})$ pointwise error estimate, and therefore the $O(h^{k+1})$ L_2 estimate.

CHAPTER IV

DISCONTINUOUS PHASE SPACE FINITE ELEMENTS

IV.A. Discontinuous Angular Finite Elements

In plane geometry it is well-known that the transport equation may have discontinuities in the angular flux at $\mu = 0$, such as at an interface or boundary[27]. An extreme example of such a discontinuity would be streaming in an absorbing region, where the flux may change abruptly upon a small change in direction. This will also be true in general for 2-D or 3-D orthogonal geometries because the geometry allows situations where a continuous change in angle (e.g., from incoming to outgoing direction) can result in a discontinuous change in the angular flux. Thus angular discontinuities in the angular flux are a frequent occurrence in transport in orthogonal geometries.

The finite element scheme we have developed thus far assumes continuity of the angular flux in both space and angle. For the angular variable, continuity is not required during the derivation of the integral law. However, use of Lagrange basis functions (see Sec. III.D) will result in an approximate solution which is continuous throughout the angular domain, even though the actual solution is discontinuous. Thus the actual discontinuous solution is being approximated by a continuous function, and poor results will probably occur, especially near interfaces or strong absorbers. This will be demonstrated later with actual computations.

As noted above, the presence of an angular flux discontinuity in the angular domain does not affect the validity of the integral law formulation of the transport equation in 1-D plane geometry. Analyt-

ically this is expected because the transport equation in plane geometry (or orthogonal geometries in general) has no angular derivatives, hence continuity of the solution in the angular domain is not a requirement. Secondly, the derivation of the integral law only involves (for the angular variable) simple integrations, which are valid in the presence of discontinuities if care is taken to evaluate the integrals properly.

Therefore, we can rather easily extend our finite element treatment by allowing the approximate solution to be discontinuous in angle (e.g., at $\mu = 0$). The most straightforward approach, since there are no angular derivatives to consider, is to add an additional node at $\mu = 0$, splitting the basis function into two parts, one part for $\mu \leq 0$ and one part for $\mu \geq 0$. Figure 4 illustrates the procedure assuming a quadratic angular finite element. Note that $\mu = 0$ must correspond to a principal node for the higher order elements since otherwise there will be basis functions which span the assumed discontinuity. Thus $\mu = 0$ is a double node with two unknowns, the angular fluxes for $\mu = 0+$ and $\mu = 0-$. The obvious consequence of this is the complete decoupling of the forward and backward directions, which were strongly coupled with the original scheme because of the forced continuity at $\mu = 0$. Now only scattering can couple nodes corresponding to $\mu < 0$ and $\mu > 0$ (a physically desirable situation).

It should be kept in mind that it is acceptable to separate the discussions of discontinuous angular and discontinuous spatial elements because of the tensor product formulation of the basis functions. Thus, changes in the angular basis functions to allow discontinuities will have no effect on the spatial basis functions, and vice versa.

IV.B. Discontinuous Spatial Finite Elements

Although analytically the solution to the transport equation must be everywhere continuous in the spatial domain, there may be points at which the solution exhibits a near-discontinuity. For example, the simple problem of a strong source of neutrons in a strong absorber surrounded by a vacuum will result in an angular flux with nearly discontinuous spatial dependence at the vacuum boundaries. This is due to the fact that the analytic solution is a constant everywhere in the interior of the slab, but near an incoming boundary the spatial dependence of the flux must drop to zero within a few mfp in order to meet the vacuum boundary condition. However, it is difficult for the approximate solution to follow this discontinuous behavior because it is constrained to be continuous by the choice of the approximating subspace (Lagrange basis functions).

Initially one might try to proceed as with the angular variable and simply place double nodes at the desired spatial positions, thus allowing the approximate solution to be discontinuous. This naive approach is, in fact, incorrect for several reasons. Analytically, the procedure used to derive the corresponding integral law implicitly made use of the continuity of the solution and the arbitrary member of the subspace because an integration by parts was performed over the entire domain. In fact, if these functions are allowed to be discontinuous, then the integration by parts must be done in a piecemeal fashion over the subdomains in which continuity is obeyed. This will then result in surface terms or interface terms at each point of discontinuity in the spatial mesh, which are ignored in the naive approach taken above.

Secondly, the addition of the double node decouples the solutions to either side of the discontinuity, and the interface becomes a free boundary for both sides. But physically, the regions are coupled by transport, which is embodied in the term $\partial Q / \partial x$ and hence the decoupling of the regions by adding a double node eliminates the physical process of transport across the node, unless the surface terms at each discontinuity are properly accounted for.

These interface terms can be properly accounted for in the following manner, which allows the use of discontinuous basis functions, as for the angular variable, but in such a way that the regions are coupled neutronically.

Beginning with the 1-D transport equation in plane geometry, (see Figure 1)

$$\mu \frac{\partial Q}{\partial x} + \Sigma_t(x) Q(x,\mu) = \frac{1}{2} \int_{-1}^{1} d\mu' \Sigma_s(x) Q(x,\mu') + S(x,\mu)$$

with arbitrary boundary conditions, consider only the transport term $\mu \frac{\partial Q}{\partial x}$, since the other terms will be treated in an identical manner as before.

Multiply $\mu \frac{\partial Q}{\partial x}$ by an arbitrary $\Psi \in H^1$, and integrate over the phase space. Calling this quantity T, we have

$$T = \int_0^1 dx \int_{-1}^{1} d\mu \, \mu \frac{\partial Q}{\partial x} \Psi(x,\mu) \tag{141}$$

Now assume there is a discontinuity (near-discontinuity in Q) at $x = x_D$ and integrate Eq. (141) by parts:

$$T = \int_0^{x_D^-} dx \int_{-1}^{1} d\mu \, \mu \frac{\partial Q}{\partial x} \Psi(x,\mu) + \int_{x_D^+}^{1} dx \int_{-1}^{1} d\mu \, \mu \frac{\partial Q}{\partial x} \Psi(x,\mu)$$

$$T = -\int_0^{x_D^-} dx \int_{-1}^{1} d\mu\, \mu\, Q(x,\mu)\frac{\partial \psi}{\partial x} - \int_{x_D^+}^{1} dx \int_{-1}^{1} d\mu\, \mu\, Q(x,\mu)\frac{\partial \psi}{\partial x}$$

$$+ \int_{-1}^{1} d\mu\, \mu\, Q(x,\mu)\psi(x,\mu)\Big|_{x=0}^{x=x_D^-}$$

$$+ \int_{-1}^{1} d\mu\, \mu\, Q(x,\mu)\psi(x,\mu)\Big|_{x=x_D^+}^{x=1}$$

or, combining terms,

$$T = -\int_0^1 dx \int_{-1}^{1} d\mu\, \mu\, Q(x,\mu)\frac{\partial \psi}{\partial x} + \int_{-1}^{1} d\mu\, \mu\left[Q(1,\mu)\psi(1,\mu) - Q(0,\mu)\psi(0,\mu)\right]$$

$$- \int_{-1}^{1} d\mu\, \mu\left[Q(x_D^+,\mu)\psi(x_D^+,\mu) - Q(x_D^-,\mu)\psi(x_D^-,\mu)\right] \qquad (142)$$

Note that Eq. (142) is identical to the transport term + boundary terms derived earlier in Eq. (122), except for the additional term due to the interface discontinuity, denoted by

$$I \equiv -\int_{-1}^{1} d\mu\, \mu\left[Q(x_D^+,\mu)\psi(x_D^+,\mu) - Q(x_D^-,\mu)\psi(x_D^-,\mu)\right] \qquad (143)$$

It is this interface term which must be carefully treated to allow use of discontinuous spatial elements.

The object now is to use known information to reduce the interface term, in a manner similar to the reduction of the boundary terms by explicitly substituting in the known boundary conditions. Since the only known condition at an interface is that the solution $Q(x, \mu)$ is continuous (albeit strongly varying) let us use this condition:

$$Q(x_D^+, \mu) = Q(x_D^-, \mu) \quad , \quad -1 \leq \mu \leq 1$$

But this is a similar situation to incorporating reflecting boundary conditions into the boundary terms (see Sec. III.E). Analytically there is an ambiguity because either angular flux could be eliminated in terms of the other. However, we now appeal to the physics of the situation and impose continuity in the direction of neutron travel, as was done with reflecting boundary conditions:

$$Q(x_D^+, \mu) = Q(x_D^-, \mu) \quad , \quad \mu \geq 0$$

$$Q(x_D^-, \mu) = Q(x_D^+, \mu) \quad , \quad \mu \leq 0$$

These substitutions are made in Eq. (143) for I, yielding

$$I = -\int_{-1}^{0} d\mu\, \mu\, Q(x_D^+, \mu)\left[\psi(x_D^+, \mu) - \psi(x_D^-, \mu)\right] - \int_{0}^{1} d\mu\, \mu\, Q(x_D^-, \mu)\left[\psi(x_D^+, \mu) - \psi(x_D^-, \mu)\right] \tag{144}$$

This term is additive to the 1-D integral law in Eq. (124) which was derived earlier. Now applying the finite element approximation by choosing a subspace $S^h \subset H^1$, expanding $Q_h(x, \mu)$

$$Q_h(x, \mu) = \sum_{j=1}^{N} \phi_j \psi_j^h(x, \mu)$$

and requiring the revised integral law to hold for all $\psi_j^h(x, \mu)$, $j = 1, 2, \ldots, N$, the additional term in the integral law results in the following matrix element

$$I_{ij} = -\int_{-1}^{0} d\mu\, \mu\, \psi_j^N(x_D^+, \mu)\left[\psi_i^N(x_D^+, \mu) - \psi_i^N(x_D^-, \mu)\right]$$

$$-\int_{0}^{1} d\mu\, \mu\, \psi_j^N(x_D^-, \mu)\left[\psi_i^N(x_D^+, \mu) - \psi_i^N(x_D^-, \mu)\right] \quad (145)$$

which is additive to the earlier matrix element A_{ij} defined in Eq. (125).

Note that the matrix element I_{ij} results in the following addition to the i^{th} equation:

$$E_i = \sum_{j=1}^{N} I_{ij}\, \phi_j$$

Using this observation and Figure 8, which illustrates the incoming and outgoing boundaries for the interface, the following remarks may be made:

(1) If node i is on the outgoing boundary of the LHS region ($\mu \geq 0$, $x = x_D^-$) then only the second term of the second integral in Eq. (145) is non-zero and I_{ij} is

$$I_{ij} = \int_{0}^{1} d\mu\, \mu\, \psi_j^N(x_D^-, \mu)\, \psi_i^N(x_D^-, \mu)$$

which is identical to B_{ij} in Eq. (125) for the outgoing boundary contribution to the matrix element. Thus the outgoing boundary of the interface is treated in the same way as the outgoing boundary of the entire region.

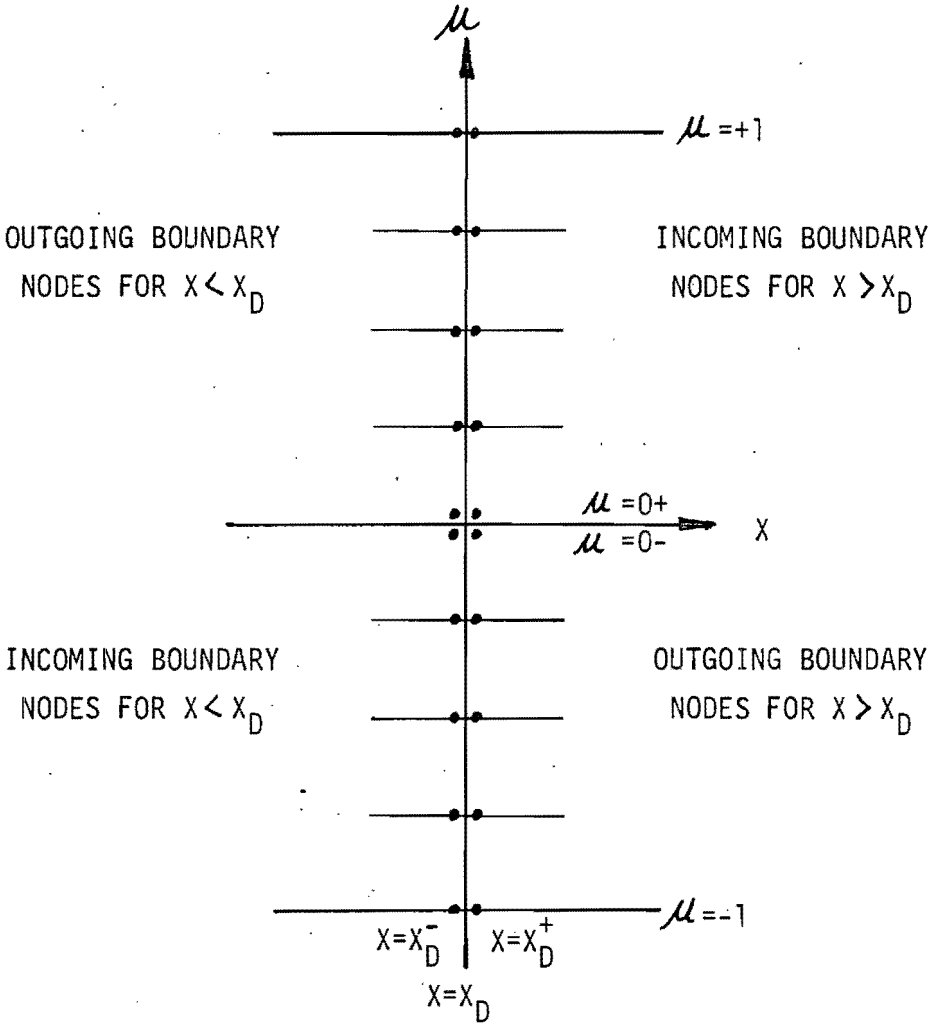


FIGURE 8. Interface Boundaries

If node i is on the outgoing boundary of the RHS region ($x = x_D^+$, $\mu \leq 0$) similar results are obtained for the other portion of the outgoing boundary.

(2) If node i is on the incoming boundary of the LHS region ($x=x_D^-$, $\mu \leq 0$), then the non-zero portion of I_{ij} is due to the second part of the first integral of Eq. (145)

$$I_{ij} = \int_{-1}^{0} d\mu\, \mu\, \psi_j^N(x_D^+, \mu)\, \psi_i^N(x_D^-, \mu)$$

This will contribute the following terms to the i^{th} equation

$$E_i = \sum_{j \in \Gamma^+ \text{ of RHS}} I_{ij}\, \phi_j$$

However, note that the summation

$$\sum_{j \in \Gamma^+ \text{ of RHS}} \phi_j\, \psi_j^N(x_D^+, \mu)$$

is exactly the angular flux expansion for the outgoing boundary of the RHS region. Defining this term as

$$Q_B(x_D^+, \mu) \equiv \sum_{j \in \Gamma^+ \text{ of RHS}} \phi_j\, \psi_i^N(x_D^+, \mu)$$

then the term E_i is

$$E_i = \int_{-1}^{0} d\mu\, \mu\, Q_B(x_D^+, \mu)\, \psi_i^N(x_D^-, \mu)$$

Comparing this with Eq. (125), it is seen that E_i is identical to the incoming boundary contribution to the source vector obtained earlier.

Similar results are obtained when node i is on the incoming boundary of the RHS region ($x = x_D^+$, $\mu \geq 0$).

Thus the effect of the discontinuous spatial finite element scheme is to solve a set of coupled source problems, each of which has an incoming angular flux which is treated as a natural boundary condition. However, in actuality the incoming fluxes are not known, and the entire system of equations must be solved simultaneously.

Since the space of trial functions S^h is no longer continuous at the specified spatial discontinuity x_D, but continuity in the direction of neutron travel is imposed within the integral law, the net result is that continuity of the angular flux in the direction of neutron motion is a natural boundary (or interface)condition. Although there was no need to mention it earlier, the previous formulation treats continuity of the angular flux as an essential condition because continuity was imposed on the space S^h. This discontinuous scheme allows more flexibility for the approximate solution to match the actual solution, as we will demonstrate later when numerical results are presented.

Implicit in the above discussion has been the fact that discontinuous angular elements were also used. Although their use is not necessary for the scheme to work, the discussion concerning the specific contributions for the incoming and outgoing boundaries becomes less precise because of the interaction of the basis function at $\mu = 0$ in both halves of the angular domain. As will be seen in the section on numerical results, though, the use of discontinuous angular elements at $\mu = 0$ should be a routine matter because of the significant improvement in the results with the small increase in effort. Therefore,

in practice one would generally use both discontinuous spatial elements and discontinuous angular elements, or the latter alone, and the above discussion is valid in its entirety.

There is a close relationship between the method used in this investigation to incorporate spatial discontinuities and the methods used to incorporate spatial discontinuities in the discrete ordinates codes TRIPLET[22] and ONETRAN[23]. These codes both employ discrete ordinates methods for the angular variables and finite element methods for the spatial variables. Although the finite element method is not employed in the manner used in this investigation, it does have the hallmarks of a finite element method in that the spatial domain is partitioned into subdomains and the solution is expanded in terms of polynomials over the subdomain. However, the algebraic equations resulting from the method used in ONETRAN or TRIPLET are somewhat different due to the fact that they are derived by different weight and integrate procedures.

As far as the equivalence of the ONETRAN/TRIPLET approach and the approach used in this investigation to treat spatial discontinuities is concerned, each method allows the incoming flux on a boundary of an element to be different than the corresponding outgoing flux of the adjacent element, but conservation of neutrons is ensured by use of surface terms. In TRIPLET this is done by explicitly expressing the jump in the angular flux at the incoming boundaries and when this term is differentiated (due to the transport term) Dirac delta functions result. Then when this resulting equation is integrated with the weighting function, the Dirac delta functions kick out the appropriate surface terms, and these terms are similar to the terms obtained above

in this investigation. In ONETRAN the method is slightly different but the end result is the same--allowing the angular flux on the incoming boundary to contribute to the effective source of neutrons within the element contributes a surface term, if the angular flux is allowed to be discontinuous on the incoming boundary.

CHAPTER V

TIME-DEPENDENT TRANSPORT

V.A. Introduction

The previous chapters have been concerned with the finite element solution of the time-independent transport equation. This restriction to steady state problems is understandable since most areas of reactor analysis are not concerned with time-dependent transport. Also, of course, the addition of the time variable complicates the equation to be solved and generally results in solution times which become excessive for even small problems. However, the time-dependent transport equation is an interesting equation and the resulting solutions can be enlightening because of the physical phenomena involved in wave propagation, reflection, scattering, etc. It is also of great significance in other applications of radiation transport (e.g., pulsed neutron experiments, laser fusion). Therefore, this brief chapter will formulate the method by which the finite element method may be applied to solve the time-dependent transport equation. The next chapter will discuss the methods by which the actual system of equations is to be solved, and numerical results are presented in Chapter VIII.

V.B. Formulation of Time-Dependent Integral Law

Begin with the general neutron transport equation[27]

$$\frac{1}{v}\frac{\partial Q}{\partial t} + \hat{\Omega}\cdot\nabla Q(\underline{r},\hat{\Omega},t) + \Sigma_t(\underline{r})\,Q(\underline{r},\hat{\Omega},t) = \int_{4\pi} d\hat{\Omega}'\,\Sigma_s(\underline{r},\hat{\Omega}'\rightarrow\hat{\Omega})\,Q(\underline{r},\hat{\Omega}',t) + S(\underline{r},\hat{\Omega},t) \tag{146}$$

subject to the initial condition

$$Q(\underline{r},\hat{\Omega},0) = Q_0(\underline{r},\hat{\Omega}) \tag{147}$$

and boundary conditions

$$Q(\underline{r},\hat{\Omega},t) = Q_s(\underline{r},\hat{\Omega},t) \text{ on } \Gamma^- \tag{148}$$

Specified incoming boundary conditions have been assumed with no loss in generality.

Defining the space H^1 as in Chapter II, choose an arbitrary $\Psi \in H^1$ and multiply Eq. (146) and integrate by parts over the phase space V:

$$\left(\frac{1}{v}\frac{\partial Q}{\partial t},\Psi\right) - \left(Q,\hat{\Omega}\cdot\nabla\Psi\right) + \left(KQ,\Psi\right) + \langle Q,\Psi\rangle_+ = \left(S,\Psi\right) + \langle Q_s,\Psi\rangle_- \tag{149}$$

where the inner product (Q,Ψ) is as defined earlier.

For the time-independent case, we proceeded by choosing a finite dimensional subspace $S^h \subset H^1$ and expanded $Q_h(\underline{r},\hat{\Omega})$ in terms of the basis functions for S^h:

$$Q_h(\underline{r},\hat{\Omega}) = \sum_{j=1}^{N} \phi_j \Psi_j^h(\underline{r},\hat{\Omega})$$

A similar procedure will be followed for the time-dependent case, except we now allow the expansion coefficients to be time-dependent:

$$Q_h(\underline{r},\hat{\Omega},t) = \sum_{j=1}^{N} \phi_j(t) \Psi_j^h(\underline{r},\hat{\Omega}) \tag{150}$$

Substitution of this expansion into Eq. (149), and requiring Eq. (149) to hold for all of the $\Psi_i^h(\underline{r},\hat{\Omega})$, i = 1,2, ..., N, one obtains the matrix differential equation

$$\frac{1}{v}\underline{\underline{P}}\,\dot{\underline{\phi}} + \underline{\underline{A}}\,\underline{\phi} = \underline{S} \tag{151}$$

where $\underline{\phi} = \begin{pmatrix} \phi_1(t) \\ \phi_2(t) \\ \vdots \\ \phi_N(t) \end{pmatrix}$ $\dot{\underline{\phi}} = \begin{pmatrix} \dot{\phi}_1(t) \\ \dot{\phi}_2(t) \\ \vdots \\ \dot{\phi}_N(t) \end{pmatrix}$ $\underline{S} = \begin{pmatrix} S_1(t) \\ S_2(t) \\ \vdots \\ S_N(t) \end{pmatrix}$

$$\underline{\underline{P}} = \{P_{ij}\}, \quad P_{ij} = (\psi_j^v, \psi_i^v),$$

and $\underline{\underline{A}}$ is identical to the $\underline{\underline{A}}$ for the steady-state equation.

Rather than a system of algebraic equations, the time-dependent equation results in a system of ordinary differential equations to be solved. An efficient method to solve Eq. (151) is the Crank-Nicholson scheme[3], which approximates the time-derivative with a forward difference, and other time-dependent terms are averaged over the present time and the incremented time:

$$\dot{\underline{\phi}} \approx \frac{\underline{\phi}(t^{m+1}) - \underline{\phi}(t^m)}{t^{m+1} - t^m}$$

$$\underline{S} \approx \frac{\underline{S}(t^{m+1}) + \underline{S}(t^m)}{2}$$

Using obvious notation, Eq. (151) becomes

$$\frac{1}{v}\underline{\underline{P}}\,\frac{\underline{\phi}^{(m+1)} - \underline{\phi}^{(m)}}{\Delta t} + \underline{\underline{A}}\,\frac{\underline{\phi}^{(m+1)} + \underline{\phi}^{(m)}}{2} = \frac{\underline{S}^{(m+1)} + \underline{S}^{(m)}}{2} \tag{152}$$

Since Eq. (147) gives $\underline{\phi}^{(0)}$, then clearly we solve for $\underline{\phi}^{(n+1)}$ in terms of $\underline{\phi}^{(n)}$, or

$$\left(\underline{\underline{P}} + \frac{v\Delta t}{2}\underline{\underline{A}}\right)\underline{\phi}^{(n+1)} = \left(\underline{\underline{P}} - \frac{v\Delta t}{2}\underline{\underline{A}}\right)\underline{\phi}^{(n)} + \frac{\underline{S}^{(n+1)} + \underline{S}^{(n)}}{2} v\Delta t \tag{153}$$

from which $\underline{\phi}^{(n+1)}$ can be solved for in terms of $\underline{\phi}^{(n)}$ and the known source term $\underline{S}$.

Thus with $\underline{\phi}^{(0)}$ known, $\underline{\phi}^{(1)}$, $\underline{\phi}^{(2)}$, ... etc., can be calculated using Eq. (153). This is especially efficient if the LU decomposition method is used to solve Eq. (153), as will be discussed in more detail in the next chapter. It should be noted that it has been assumed that $\underline{\underline{A}}$ is time-independent. This will generally be the case for neutron transport in reactors because for time scales of interest for the propagation of disturbances in the neutron flux, cross-sections or other macroscopic constants should not change appreciably. If $\underline{\underline{A}}$ is time-dependent, then the efficiency of the LU scheme is decreased somewhat. This is also discussed in the next chapter.

Note that the time-dependent incoming boundary conditions are included within the source vector $\underline{S}$, along with the time-dependent volumetric source.

V.C. Time-Dependent Transport in Plane Geometry

In plane geometry the general time-dependent transport equation, Eq. (146), becomes

$$\frac{1}{v}\frac{\partial Q}{\partial t} + \mu\frac{\partial Q}{\partial x} + \Sigma_t(x)\,Q(x,\mu,t)$$
$$= \sum_{\ell=0}^{L}\frac{2\ell+1}{2} b_\ell(x) P_\ell(\mu)\int_{-1}^{1} d\mu'\, P_\ell(\mu')\,Q(x,\mu',t) \qquad (154)$$
$$+ \; S(x,\mu,t)$$

subject to initial conditions

$$Q(x,\mu,0) = Q_0(x,\mu) \qquad (155)$$

and boundary conditions

$$\begin{aligned} Q(0,\mu,t) &= Q_0(0,\mu,t) \\ Q(1,\mu,t) &= Q_0(1,\mu,t) \end{aligned} \qquad (156)$$

Applying the finite element approximation to the equivalent integral law results in the following system of ordinary differential equations:

$$\frac{1}{v}\underline{\underline{P}}\,\dot{\underline{\Phi}} + \underline{\underline{A}}\,\underline{\Phi}(t) = \underline{S}(t) \qquad (157)$$

The specific matrix elements are

$$A_{ij} = T_{ij} + K_{ij} + B_{ij} \qquad (158)$$

$$P_{ij} = \int_0^1 dx \int_{-1}^{1} d\mu\, \psi_i^{N}(x,\mu)\,\psi_j^{N}(x,\mu) \qquad (159)$$

and the source vector is

$$S_i(t) = \int_0^1 dx \int_{-1}^1 d\mu\, S(x,\mu,t)\, \psi_i^{\sim}(x,\mu) + \int_0^1 d\mu\, \mu\, Q_0(0,\mu,t)\, \psi_i^{\sim}(0,\mu) - \int_{-1}^0 d\mu\, \mu\, Q_0(1,\mu,t)\, \psi_i^{\sim}(1,\mu) \tag{160}$$

These terms have been explicitly written out because the finite element code FTRAN solves the equations. Chapter VIII contains numerical results for a few time-dependent problems, indicating the potential of the method.

CHAPTER VI

NUMERICAL SOLUTION OF SYSTEM OF EQUATIONS

VI.A. Possible Solution Techniques

There exist a myriad of techniques for solving the system of algebraic systems represented by the matrix equation

$$\underline{\underline{A}}\,\underline{\Phi} = \underline{S} \tag{161}$$

which results from the finite element approximation to solve the integral law form of the transport equation. For example, one might choose an iterative method, such as the Gauss-Seidel or Successive Over-Relaxation (SOR) methods, or a direct method such as Gaussian eliminiation or one of its many variants (e.g., LU decomposition , Crout decomposition, etc.). The initial attempt in this investigation was to employ the SOR iterative method since it is relatively easy to implement[28]. However, it was observed fairly quickly that iterative methods would not be satisfactory for application to the matrices obtained when solving the transport equation. For example, the SOR method worked quite well with a relatively coarse spatial mesh but would not converge if the mesh were refined to any extent. Noting that the results of Sec. III.E. indicate that $\underline{\underline{A}}$ becomes increasingly off-diagonal dominant as the spatial mesh is refined, this non-convergence of the iterative methods is not surprising. It is well-known that iterative methods work very well with matrices obtained with second-order self-adjoint applications (e.g., the diffusion equation) because the matrices are diagonally dominant. However, convergence of an iterative technique for asymmetric, off-diagonal

dominant matrices characteristic of finite element transport methods cannot be shown and will probably not occur on the strength of the numerical evidence observed during this investigation.

Therefore, since the convergence of iterative methods could not be guaranteed in general for the type of matrices encountered with the transport equation, it was decided that a direct method should be used. The method chosen was the LU decomposition method, which is a well-known variant of Gaussian elimination and offers a number of advantages for the transport application. The details of the LU method and its implementation will be discussed in the next section.

VI.B. LU Decomposition Method[35]

The principle behind the LU decomposition method is to factor the matrix $\underline{\underline{A}}$ into a product of a lower triangular matrix $\underline{\underline{L}}$ and an upper triangular matrix $\underline{\underline{U}}$:

$$\underline{\underline{A}} = \underline{\underline{L}}\,\underline{\underline{U}} \tag{162}$$

where $\underline{\underline{L}}$ is zero above the main diagonal (which is unity), and $\underline{\underline{U}}$ is zero below the main diagonal. Once $\underline{\underline{A}}$ has been factored, the solution of Eq. (141) is performed with two back-substitutions as follows.

The solution of

$$\underline{\underline{L}}\,\underline{\underline{U}}\,\underline{\Phi} = \underline{S} \tag{163}$$

can be obtained by first solving

$$\underline{\underline{L}}\,\underline{X} = \underline{S} \qquad \left(\underline{X} = \underline{\underline{L}}^{-1}\,\underline{S}\right) \tag{164}$$

and then solving

$$\underline{\underline{U}}\,\underline{\Phi} = \underline{X} \qquad \left(\underline{\Phi} = \underline{\underline{U}}^{-1}\,\underline{X}\right) \tag{165}$$

for $\underline{\Phi}$. Note that Eqs. (164) and (165) simply express

$$\underline{\Phi} = \underline{\underline{U}}^{-1}\,\underline{\underline{L}}^{-1}\,\underline{S} \tag{166}$$

which is the formal solution to Eq. (163).

The key to the method is that the solution of Eqs. (164) and (165) is simple because $\underline{\underline{L}}$ and $\underline{\underline{U}}$ are triangular matrices and hence are readily inverted. In particular, if the order of $\underline{\underline{A}}$ is N (N rows and N columns) then the matrices $\underline{\underline{L}}$ and $\underline{\underline{U}}$ are of order N and the solutions to Eqs. (164) and (165) are:

$$\begin{aligned}
X_1 &= S_1 \qquad (\text{All } L_{ii} = 1) \\
X_2 &= S_2 - L_{21} X_1 \\
X_3 &= S_3 - L_{31} X_1 - L_{32} X_2 \\
&\vdots \\
X_N &= S_N - \sum_{j=1}^{N-1} L_{Nj} X_j
\end{aligned} \tag{167}$$

and

$$\begin{aligned}
\Phi_N &= X_N / U_{NN} \\
\Phi_{N-1} &= \left(X_{N-1} - U_{N-1,N} X_N\right) / U_{N-1,N-1} \\
&\vdots \\
\Phi_i &= \left(X_i - \sum_{j=i+1}^{N} U_{ij} X_j\right) / U_{ii} \\
&\vdots \\
\Phi_1 &= \left(X_1 - \sum_{j=2}^{N} U_{1j} X_j\right) / U_{11}
\end{aligned} \tag{168}$$

One obvious consequence of the LU method is that once $\underline{\underline{A}}$ has been factored into $\underline{\underline{L}}\,\underline{\underline{U}}$, the solution of problems with different source vectors is quite convenient because only back-substitutions are needed. Since the bulk of the effort is the factorization of $\underline{\underline{A}}$, not the back-substitutions, this can result in significant savings in computational time for applications involving solution of $\underline{\underline{A}}\,\underline{\Phi} = \underline{S}$ for many different $\underline{S}$. In particular, this is seen to be true for eigenvalue problems and for time-dependent problems, both of which are considered in Chapter VIII.

Let us now consider the factorization process for decomposing $\underline{\underline{A}}$ into $\underline{\underline{L}}\,\underline{\underline{U}}$. It is easily seen[35] that if $\underline{\underline{A}}$ is non-singular then $\underline{\underline{L}}$ and $\underline{\underline{U}}$ will also be non-singular. In addition, $\underline{\underline{L}}$ and $\underline{\underline{U}}$ are unique. In the discussion that follows, it will be assumed that pivoting (interchange of rows and/or columns) is not needed. In general pivoting is not required for positive definite, symmetric (hence diagonally dominant) matrices. However, we have also found that pivoting was generally not required for our matrices, although pivoting would have been beneficial for some of the precise numerical experiments, as will be discussed later.

To begin, write the original matrix $\underline{\underline{A}}$ as

$$\underline{\underline{A}}^{(1)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1m}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & \cdots & a_{2m}^{(1)} \\ \vdots & \vdots & & \vdots \\ a_{m1}^{(1)} & a_{m2}^{(1)} & \cdots & a_{mm}^{(1)} \end{bmatrix}$$

where the superscript will refer to the current step in the factorization process. The object of the first step is to replace the boxed-in elements of $\underline{\underline{A}}^{(1)}$ by zeroes. To do this, multiply row 1 (pivotal row) by the multiplier $-\ell_{j1}$, where

$$\ell_{j1} = \frac{a_{j1}^{(1)}}{a_{11}^{(1)}}$$

and add to row j, j = 2,3,...,n. This will then eliminate the first column under the diagonal. The matrix $\underline{\underline{A}}$ will then be

$$\underline{\underline{A}}^{(2)} = \begin{bmatrix} a_{11}^{(2)} & a_{12}^{(2)} & \cdots & a_{1n}^{(2)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ 0 & a_{32}^{(2)} & \cdots & \\ \vdots & \vdots & & \\ 0 & a_{m2}^{(2)} & a_{m3}^{(2)} \cdots & a_{mn}^{(2)} \end{bmatrix}$$

Now we proceed in the same fashion to replace the boxed-in column of $\underline{\underline{A}}^{(2)}$ with zeroes. Proceeding in this manner for successive columns, the inductive step is to replace the boxed-in column of $\underline{\underline{A}}^{(k)}$ by zeroes, for $k = 1, 2, \cdots, m-1$.

$$\underline{\underline{A}}^{(k)} = \begin{bmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \cdots & & a_{1m}^{(k)} \\ 0 & a_{22}^{(k)} & \cdots & & a_{2m}^{(k)} \\ 0 & 0 & a_{33}^{(k)} & \cdots & \vdots \\ \vdots & \vdots & 0 & a_{kk}^{(k)} \cdots & \\ & & & a_{k+1,k}^{(k)} & \\ & & & \vdots & \\ 0 & 0 & 0 & a_{mk}^{(k)} \cdots & a_{mm}^{(k)} \end{bmatrix}$$

Clearly the matrix will be upper triangular at the end of (n-1) steps, or

$$\underline{\underline{A}}^{(n)} = \begin{bmatrix} a_{11}^{(n)} & a_{12}^{(n)} & a_{13}^{(n)} & \cdots & a_{1m}^{(n)} \\ 0 & a_{22}^{(n)} & a_{23}^{(n)} & \cdots & a_{2m}^{(n)} \\ 0 & 0 & a_{33}^{(n)} & \cdots & \\ & & & \ddots & \\ & & & a_{ic}^{(n)} & \\ & & & & \ddots \\ & & & & a_{mm}^{(n)} \end{bmatrix}$$

This matrix $\underline{\underline{A}}^{(n)}$ is the upper triangular matrix $\underline{\underline{U}}$. The lower triangular matrix $\underline{\underline{L}}$ is the array of factors ℓ_{ij} which multiplied the pivotal row to eliminate a_{ij}. Or,

$$\underline{\underline{L}} = \begin{bmatrix} 1 & 0 & & & 0 \\ \ell_{21} & 1 & & & \\ \ell_{31} & \ell_{32} & & 1 & \\ \vdots & \vdots & & & \\ \ell_{m1} & \ell_{m2} & \cdots\cdots & & 1 \end{bmatrix}$$

It is a straightforward exercise[35] to show that indeed

$$\underline{\underline{A}} = \underline{\underline{L}}\,\underline{\underline{U}}$$

where $\underline{\underline{L}}$ and $\underline{\underline{U}}$ are as defined above.

VI.C. Eigenvalue Calculations

As will be derived in Sec. VIII.C, the critical slab eigenvalue problem results in a generalized matrix eigenvalue problem

$$\underline{\underline{A}}\,\underline{\Phi} = c\,\underline{\underline{M}}\,\underline{\Phi} \tag{169}$$

where C is the smallest number of secondary neutrons required per collision to achieve criticality in the given slab.

Since the LU method has been used to factor $\underline{\underline{A}}$, the solution of Eq. (169) is considerably simplified if the inverse power iteration[35] method is used.

To implement the inverse power iteration method[45], first guess an initial angular flux $\underline{\Phi}^{(0)}$ and eigenvalue $C^{(0)}$. Calculate an effective source vector $\underline{S}^{(0)}$,

$$\underline{S}^{(0)} = C^{(0)}\,\underline{\underline{M}}\,\underline{\Phi}^{(0)} \tag{170}$$

and solve for the next flux iterate $\underline{\phi}^{(1)}$

$$\underline{\underline{A}}\,\underline{\phi}^{(1)} = \underline{S}^{(0)} \tag{171}$$

Now calculate the new estimate for the eigenvalue C,

$$C^{(1)} = C^{(0)} \frac{\left(\underline{\underline{M}}\,\underline{\phi}^{(1)}, \underline{\underline{M}}\,\underline{\phi}^{(0)}\right)}{\left(\underline{\underline{M}}\,\underline{\phi}^{(1)}, \underline{\underline{M}}\,\underline{\phi}^{(1)}\right)} \tag{172}$$

where the vector inner product $(\underline{a}, \underline{b})$ is defined

$$(\underline{a}, \underline{b}) \equiv \sum_{i=1}^{n} a_i b_i$$

The general procedure may then be summarized:

(1) Given $\underline{\phi}^{(0)}$, $C^{(0)}$

(2) For n = 1, 2, ..., ITMAX

solve $\underline{\underline{A}}\,\underline{\phi}^{(n)} = C^{(n-1)}\,\underline{\underline{M}}\,\underline{\phi}^{(n-1)}$

where $C^{(k)} = \dfrac{\left(\underline{\underline{M}}\,\underline{\phi}^{(k)}, \underline{\underline{M}}\,\underline{\phi}^{(k-1)}\right)}{\left(\underline{\underline{M}}\,\underline{\phi}^{(k)}, \underline{\underline{M}}\,\underline{\phi}^{(k)}\right)} C^{(k-1)}$ (173)

(3) Proceed until n is greater than some predetermined limit on the eigenvalue iterations (ITMAX) or until successive values of $C^{(k)}$ and/or $\underline{\phi}^{(k)}$ agree to within a specified error ϵ (may be different).

The motivation for the updated value of the eigenvalue in Eq. (173) is that combining Eqs. (169), (170), and (171) is equivalent to solving

$$C^{(n)}\,\underline{\underline{M}}\,\underline{\phi}^{(n)} = C^{(n-1)}\,\underline{\underline{M}}\,\underline{\phi}^{(n-1)} \tag{174}$$

and taking the vector inner product of Eq. (174) with $\underline{\underline{M}}\,\underline{\phi}^{(n)}$ results in Eq. (173) when $C^{(n)}$ is solved for.

The advantage of this method is that once $\underline{\underline{A}}$ is factored into $\underline{\underline{L}}\,\underline{\underline{U}}$ for the first iteration (solving for $\underline{\phi}^{(1)}$) then each succeeding iteration only consists of back-substitutions, assuming $\underline{\underline{L}}$ and $\underline{\underline{U}}$ are stored. Thus the eigenvalue iterations are very efficient, once $\underline{\underline{A}}$ has been factored.

CHAPTER VII

COMPUTER IMPLEMENTATION

The purpose of this chapter is to discuss briefly the basic organization of FTRAN and then to discuss details of the two most significant features of FTRAN--the use of dynamic allocation of storage to allow tailoring of arrays to fit the current problem size and the use of sparse matrix techniques to take advantage of the fact that the matrix $\underline{\underline{A}}$ may be predominately zero.

In addition to a discussion of the organization of FTRAN, this chapter includes a discussion of the capabilities of FTRAN and timing results from some typical FTRAN runs.

VII.A. Basic Organization

The computer code FTRAN (Finite Element Transport) which implements the solution of the transport equation with the finite element approximation, consists of several modules which are linked together by a driver module which controls the overall flow of the code. The basic organization of FTRAN is illustrated in Figure 9, which also indicates the flow of information to achieve the final answer.

FTRAN is capable of treating the following neutron transport problems in 1-D plane geometry:

(1) Multiple regions (≤ 10)

(2) Anisotropic scattering (up to P_4) with different Legendre coefficients in each region

(3) Non-uniform spatial and/or angular meshes

(4) Vacuum, reflecting, or specified incoming boundary conditions

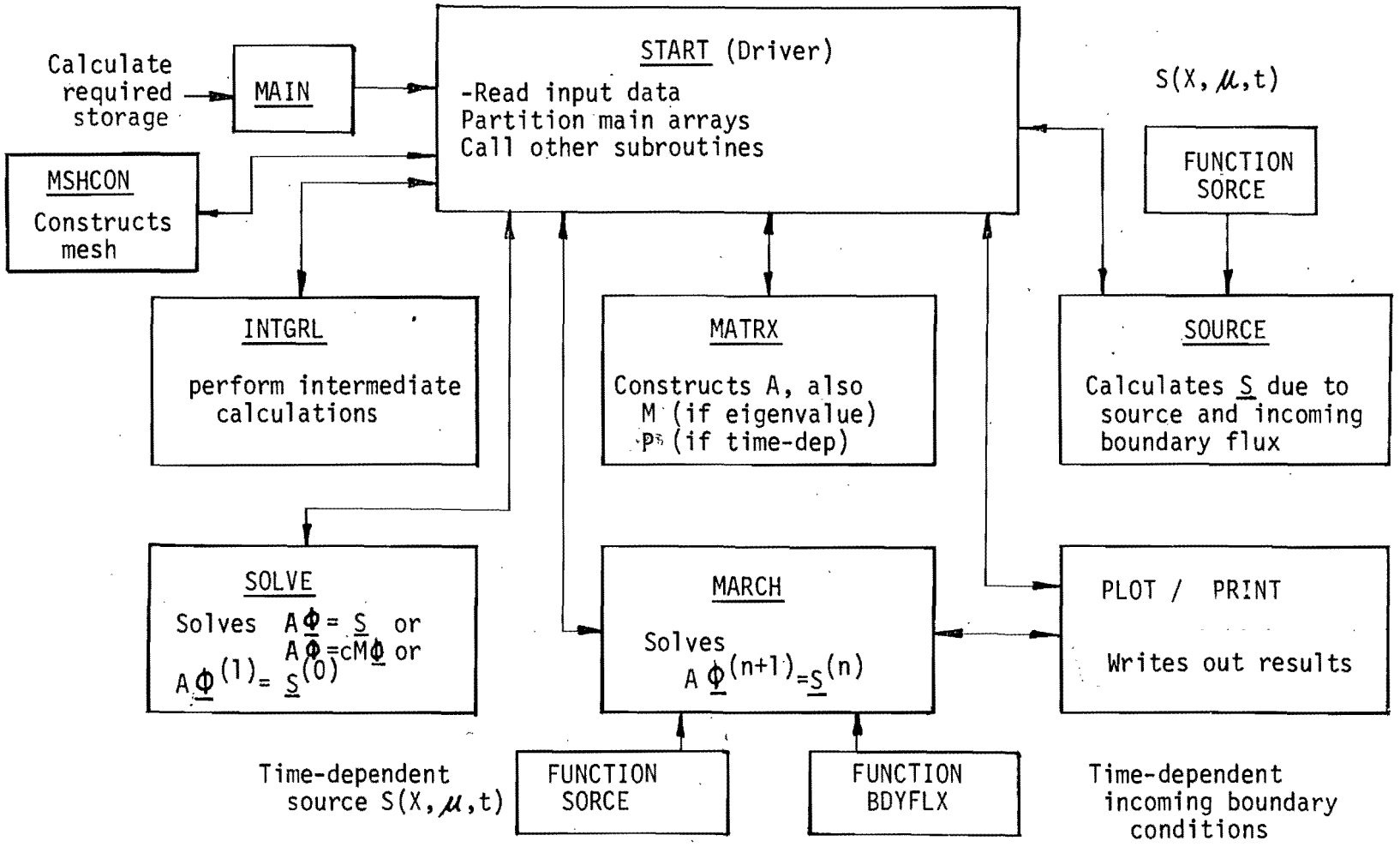


FIGURE 9. FTRAN Flow Diagram

(5) General distributed anisotropic sources

(6) Discontinuous angular finite elements at $\mu = 0$

(7) Discontinuous spatial finite elements at arbitrary spatial locations (discussed in Chapter IV).

(8) Time-dependent problems with general distributed anisotropic time-dependent sources and/or time dependent incoming boundary conditions

(9) Critical slab eigenvalue problem with multiple regions, anisotropic scattering, and vacuum or reflecting boundary conditions.

In addition, FTRAN has been coded to allow precise timing studies by numerous calls on the MTS clock at various points during the execution of the code.

As is discussed in detail in Appendices A and B, the input is handled by use of NAMELIST statements, which allows convenient interactive use of the code from remote terminals.

Using Figure 9 as a guide, we will briefly discuss the basis organization of FTRAN. The MAIN program is acutally a very short program (5 lines) and only serves to dimension two vectors (one real and one integer) of sufficient length to treat the specific problem being solved. Generally the MAIN program is compiled at run time with reasonable estimates of the required lengths of the two arrays included in the DIMENSION and CALL statements. MAIN then calls START, which is the driver subroutine for FTRAN. START reads in the input data and writes out a summary of the input data for the convenience of the user. START also partitions the two main arrays dimensioned in the MAIN program

into a number of smaller vectors corresponding to the matrix of coefficients ($\underline{\underline{A}}$), the scattering matrix ($\underline{\underline{M}}$) (if eigenvalue problem), various vectors for intermediate calculation of matrix elements, and other vectors to contain the solution, source, etc. By use of dynamic dimensioning, as is to be discussed in the next section, these vectors are made equivalent to vectors or arrays (of various orders) in other subroutines.

START then calls sequentially on MSHCON (construction of mesh), INTGRL (calculation of various integrals for matrix elements), MATRX (actual computation of matrix elements), SOURCE (computation of source vector $\underline{S}$, including contributions due to incoming fluxes), SOLVE (solution of system of equations) and MARCH (advance solution through time for time-dependent problem). Other subroutines are also used for special purpose calculations such as INTERP (interpolate function between nodes with linear, quadratic, or cubic interpolation in either variable), NORM (calculation of the max-norm of the inverse of $\underline{\underline{A}}$, $\| \underline{\underline{A}}^{-1} \|_{\infty}$), and ERROR (calculates L_2 error of the computed solution with respect to the exact solution, which is input via FUNCTION TEST.

For convenience, a distributed anisotropic source (also may be time-dependent) may be input with FUNCTION SORCE, and a time-dependent incoming flux may be input with FUNCTION BDYFLX. These functions are convenient for the user because they allow quite complicated source distributions to be input without inputting any numbers. Of course, if used, these functions are compiled at run time.

The next two sections concern the most significant features of the computer implementation--the use of dynamic dimensioning and the use of

sparse matrix techniques to allow only a fraction of the matrix $\underline{\underline{A}}$ (the non-zero part) to be stored.

VII.B. Execution Time/Storage Considerations

The solution of the 1-D transport equation, which is analytically a two-dimensional problem (x,μ), can involve an extremely large matrix $\underline{\underline{A}}$. For example, a problem with 100 spatial nodes and 10 angular nodes, which is a small problem, involves 1000 unknowns and hence the matrix $\underline{\underline{A}}$ is 1000 x 1000, or 10^6 storage locations. Clearly the storage requirements for realistic problems can easily exceed the available storage and fast memory of most computer installations, let alone the CPU time required to factor a matrix of that size. Thus there is ample incentive to develop algorithms and computer procedures to minimize the actual storage required to solve these large problems.

The obvious first step would be to store and manipulate only the non-zero elements of $\underline{\underline{A}}$, since they may be only a fraction of the total number of elements of $\underline{\underline{A}}$. To see this, note that the coupling of nodes to one another can be separated into two parts, angular coupling and spatial coupling, when tensor product basis functions are used. In general, a node will not be globally coupled to another node unless it is coupled both in space and angle. However, as noted earlier, the presence of scattering couples all of the angular nodes since there is a finite probability of being scattered into any direction for any initial direction. Therefore, the global coupling is determined entirely by the spatial coupling if there is any scattering present.*

* If there is no scattering (i.e. $\sum_s(\underline{\Lambda},\hat{\Omega}'\rightarrow\hat{\Omega})=0$), then the global coupling is determined also by the angular coupling. However, this special case is not accounted for in this investigation.

Precise calculations of the number of non-zero elements of $\underline{\underline{A}}$ will now be made. The spatial coupling of nodes can be determined by inspection of Figures 3-5, which illustrate linear, quadratic, and cubic basis functions, respectively. It is seen that at most $2I_x + 1$ nodes are coupled to a given node, where I_x is the degree of the spatial finite element. Hence, the use of linear spatial elements implies that a node is coupled to its immediate neighbors, while the use of cubic spatial elements may result in the node being coupled to its three nearest neighbors on either side. Since the angular coupling is full, the total number of nodes that may interact with a given node (counting itself) is

$$B = (2I_x + 1) N_\mu \tag{175}$$

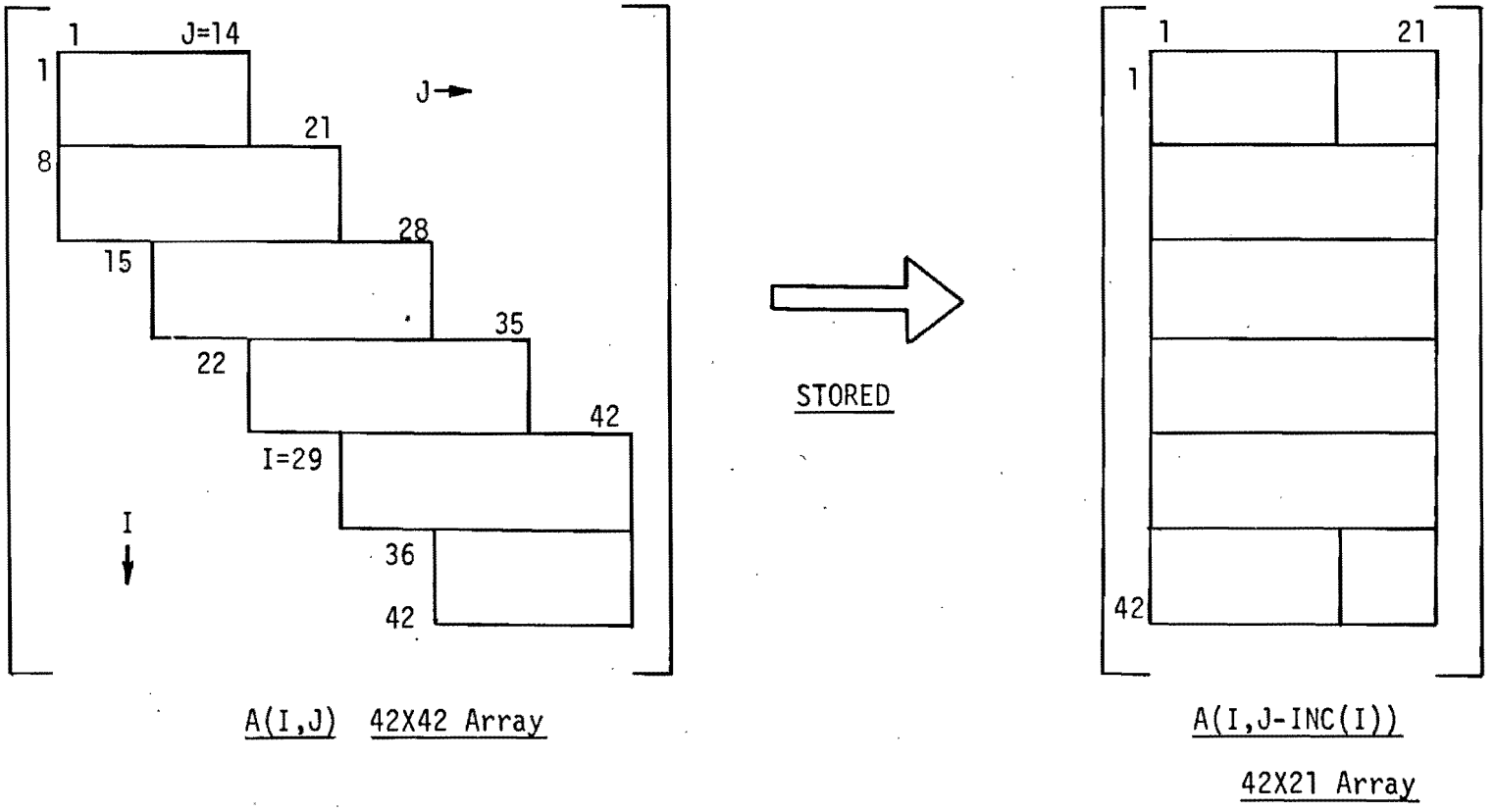
where I_x = degree of spatial element

N_μ = no. of nodes in angular mesh.

Note that B will then be related to the bandwidth of the matrix, which is here defined as the maximum row length of $\underline{\underline{A}}$. In fact, with this definition,

$$B = \text{bandwidth} = (2I_x + 1)\, N_\mu .$$

Because of the orderly coupling of nodes with the tensor product basis functions, the matrix $\underline{\underline{A}}$ will have a very precise structure which consists of blocks of length B and height N_μ distributed along the main diagonal, except near the corners of $\underline{\underline{A}}$. Figure 10 illustrates a typical matrix $\underline{\underline{A}}$, and this structure is clearly seen. With scattering present, there will be in general no zero elements within these blocks.



NOTE: Matrix corresponds to Figure 2.

FIGURE 10. Schematic of Matrix A

Although this cannot be considered an advantage, it does allow one to use standard banded matrix techniques for elimination without concern for fill-in of zeroes.

Using the well-known estimate of the number of operations (multiplies or divides) to factor a full N x N matrix[35]

$$\#\text{ops} \sim N^3 \tag{176}$$

the number of operations for the banded matrix discussed here is considerably smaller. That is, the number of multipliers to eliminate each column below the diagonal is initially $B(N_\mu - 1)$ because there are B elements in the first row which are multiplied to eliminate one element in the column, and there are $(N_\mu - 1)$ elements in each column to be eliminated. Since there are N columns, this results in (using Eq. (175))

$$\#\text{ops} \sim N N_\mu B$$
$$\text{OR} \quad \#\text{ops} \sim N N_\mu^2 (2I_x + 1) \tag{177}$$

Thus the number of operations is linear in the number of spatial unknowns and cubic in the number of angular unknowns, versus cubic in each if $\underline{\underline{A}}$ were full. This dependence is illustrated in Figure 11 and typical timing data are tabulated in Table I. Thus given an angular mesh, the execution time is linear with problem size.

However, the key point is that the zero elements outside the bands must be ignored in the elimination process, because multiplication by zero takes as much computer time as multiplication by a non-zero. Therefore the coding must take into account the block structure of $\underline{\underline{A}}$ to ensure that only non-zero elements are considered. This has been done in subroutine SOLVE for both the factorization and the back substitution algorithms.

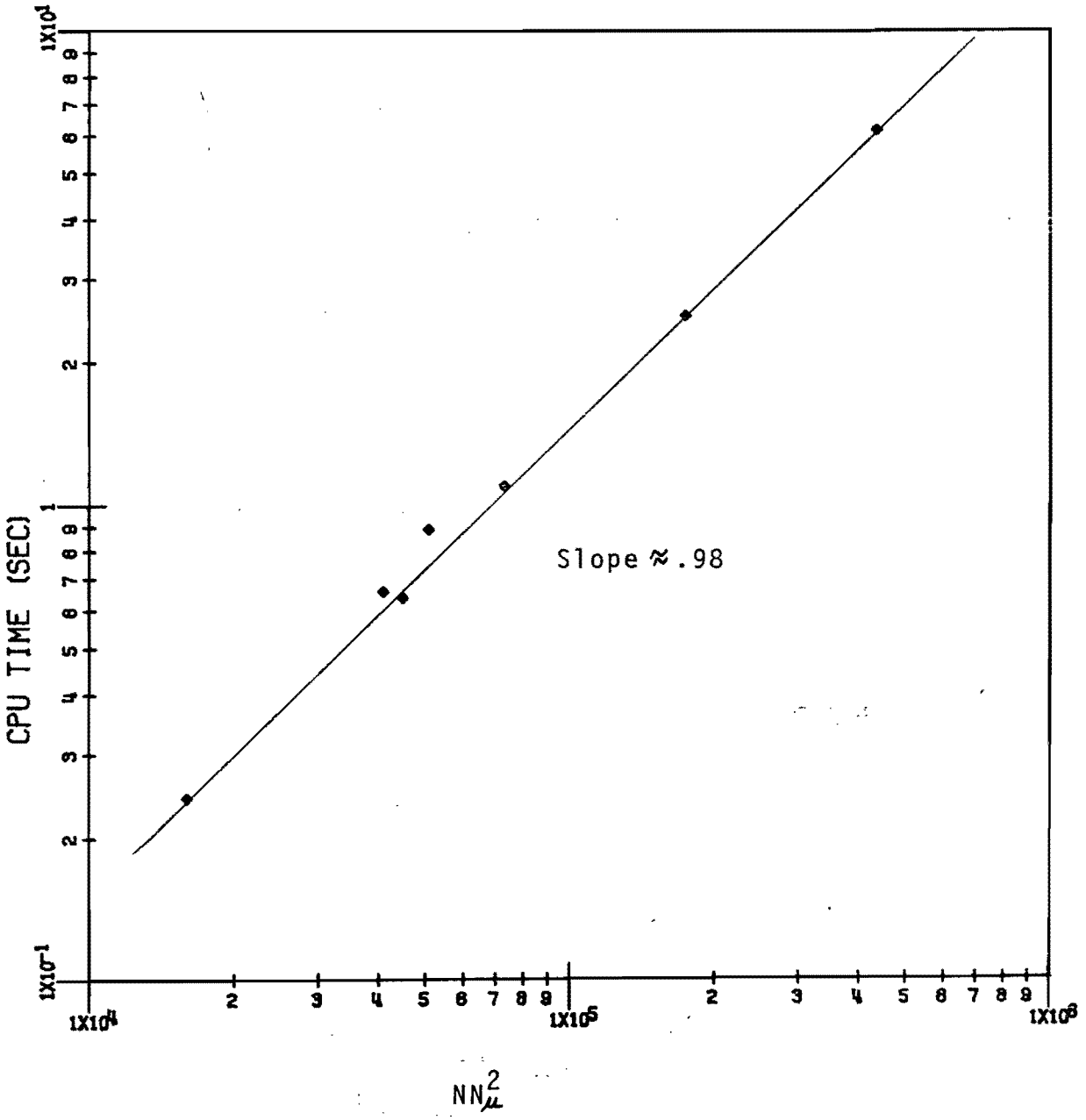


Figure 11. Execution Time vs. Problem Size

DESCRIPTION OF RUN (SS-steady state, TD-time dep)	CPU[a] Time (sec) / subroutine			
	INTGRL	MATRX	SOLVE	MARCH
49 X-nodes (linear) 4 μ-nodes (linear) 196 unknowns, SS	.004	.05	.09	---
58 X-nodes (linear) 10 μ-nodes (linear) 580 unknowns, SS	.005	.26	.90	---
101 X-nodes (linear) 12 μ-nodes (linear) 1212 unknowns, SS	.008	.53	2.50	---
101 X-nodes (quadratic) 6 μ-nodes (quadratic) 606 unknowns, SS	.007	.29	.94	---
6 X-nodes (linear) 5 μ-nodes (linear) 30 unknowns, SS	.001	.008	.017	---
21 X-nodes (linear) 4 μ-nodes (linear) 84 unknowns, TD (50 steps)	.002	.02	.02	1.4
21 X-nodes (linear) 13 μ-nodes (linear) 273 unknowns, TD (50 steps)	.004	.19	.63	7.5
23 X-nodes (linear) 10 μ-nodes (linear) 230 unknowns, TD (100 steps)	.004	.14	.32	7.2

[a] AMDAHL 470V/6

TABLE I. Timing Data

Thus the above analysis indicates that considerable savings in execution time can be realized by taking advantage of the sparse structure of $\underline{\underline{A}}$ during the factorization and back substitution steps. However, one must also ensure that in addition to ignoring the zero elements of $\underline{\underline{A}}$ during the solution process, one never stores the zero elements in the first place.

Since the blocks are full, the most efficient scheme of storing the blocks would appear to be to simply move them to the left to form a rectangular array with N rows and B columns, as illustrated in Figure 10. This turns out to be rather simple to implement because simple indexing will transform the square array to the rectangular array: (using FORTRAN array terminology)

$$A(I,J) \longrightarrow A\big(I,\ J-INC(I)\big) \tag{178}$$

where INC(I) is an integer array indicating the horizontal indexing that must be done to move row I to the left side of the array. Thus the subroutines that compute matrix elements and which perform the LU decomposition and back substitutions can all be coded as if $\underline{\underline{A}}$ were square and then modified via the simple index change to accomodate the fact that $\underline{\underline{A}}$ is actually rectangular (and is dimensioned as such).

A simple example illustrates why this concern for zero elements is actually very important. A typical transport problem with linear finite elements may involve 100 spatial nodes and 10 angular nodes, or 1000 unknowns. If the matrix $\underline{\underline{A}}$ were to be dimensioned as a square array, then 10^6 storage locations would be required just to hold $\underline{\underline{A}}$.

However, if the non-zero blocks were stored, then $\underline{\underline{A}}$ would be a rectangular array of size 1000 x 30 or 30,000 locations, which is 3% of the storage required for the square array.

But as discussed above, the savings in execution time is also a considerable advantage. An estimate of this savings for the factorization alone can be made with a simple ratio of the operation counts for the full matrix, Eq. (176), and the banded matrix, Eq. (177),

$$\mathcal{R} = \frac{N N_w^2 (2 I_x + 1)}{N^3}$$

$$= \frac{(1000)(10^2)(3)}{1000^3} = 3 \times 10^{-4}$$

which, while only a crude estimate, is the motivation behind expending the additional coding effort to account for the sparse structure of $\underline{\underline{A}}$. In fact, for the solution of even moderately sized problems, this becomes a necessity, not a luxury.

VII.C. Dynamic Allocation of Storage

In the last section the advantages of storing the N x N matrix $\underline{\underline{A}}$ as a rectangular N x B array were explained. However, the dimensions of N and B can vary widely for the types of problems to be solved in neutron transport. For example, for the sample problem discussed in the previous section, N = 1000 and B = 30, while a problem with cubic spatial elements on a 25x25 mesh would require N = 625 and B = 175. Or a small problem with 10 spatial nodes and 3 angular nodes with linear elements would require N = 30 and B = 9. Thus it would be impossible

to dimension a fixed array for $\underline{\underline{A}}$ that would allow for a wide variety of problems yet would not result in a large amount of wasted storage.

To overcome these disadvantages with fixed dimensional arrays, dynamic dimensioning was used. Dynamic dimensioning is simply a method by which arrays may be dimensioned within a subroutine in accordance with parameters passed with the argument list. Thus the driver subroutine (in this case, START) can decide on the correct array dimensions for the specific problem being solved, and then can pass on to the appropriate subroutine(s) the size of the array which is to be dimensioned in that subroutine. The only restriction is that there is sufficient storage dimensioned so that all of the variably dimensioned arrays can fit within the allowed storage for that run.

The specific details will now be given for the dynamic dimensioning of real arrays in FTRAN. As noted earlier, the MAIN program dimensions a vector V with sufficient length to handle the specific problem. For example, V may be dimensioned V(20000). MAIN then calls START with the statement

```
CALL START (V(1), 20000, IV(1), 400)
```

where the last two arguments are for the integer array, which is treated in a similar manner and will not be discussed any further.

The above statement passes the first element of the vector V and the length of V. The key point is that START (or any other called subroutine) only cares where the initial element of the argument array is, because the length of the array within START will be determined by an appropriate dimension statement. For example, the first line in START is

```
SUBROUTINE START (V(1), IDIM, ...)
```

and the dimension statement in START,

```
DIMENSION V(IDIM), ...
```

will dimension the vector V with length IDIM, which is the second parameter passed by the calling statement. Thus in this case the vector V is identical within MAIN and START.

Now, however, START processes the input data and decided the correct dimensions for 28 separate 1-D, 2-D, and 3-D arrays, all of which are contained in the vector V. START then calls the various subroutines with the location of the element of V which is the first element in the appropriate array to be dimensioned in the called subroutine. START also passes the dimensions which the subroutine is to use. For example, the calling statement in START,

```
CALL INTGRL (V(IP1), M1, ..., V(IP28), M28, N28, L28, ...)
```

and the corresponding parameter list in INTGRL,

```
SUBROUTINE INTGRL (XP, M1, ...., ASX, M28, N28, L28, ..)
```

with the dimension statement

```
DIMENSION XP(M1),...., ASX (M28, N28, L28), ....
```

imply that within the subroutine INTGRL, the vector XP will have length M1 and the 3-D array ASX will be dimensioned M28 x N28 x L28, and these dimensions are determined by START from the input data. Also, the array XP begins with element V(IP1) and the array ASX begins with element V(IP28), although this knowledge is never needed by the user (once the code is debugged). Note that the effect of this is to COMMON the vector V with the 28 arrays of various sizes and shapes, thus requiring only one fixed dimension statement--the one for V in the MAIN program.

The advantage of this is that one can compile the MAIN program with V of sufficient length to contain the 28 arrays into which it is partitioned (an error message alerts the user if the arrays will not fit in V). Since MAIN is so short, this cost is negligible ($<$ 10 cents).

VIII.D. Algorithm for Calculation of Matrix Elements

Section III.E discussed the calculation of the matrix elements A_{ij} by means of a transformation to the standard interval [-1,1], which results in a considerable reduction in the number of unique integrals to be calculated. However, no mention was made of the actual algorithm for computing the matrix elements. Because of the large number of matrix elements that have to be computed, care must be taken to ensure that the algorithm is relatively efficient.

The principle behind the algorithm used in FTRAN is to make all decisions at the earliest point possible during the process of computing the A_{ij}. That is, the A_{ij} depend on the location of the nodes i and j, the region in which the nodes reside if a multi-region problem is being solved, whether or not nodes i and j are principal or internal nodes, etc. Thus if one were to naively set up a loop which cycled through all nodes i, and for each i, cycled through all the nodes j, and then made the decisions noted above to determine the matrix element A_{ij}, the total time to calculate the A_{ij} could easily be prohibitive. This would be an example of making decisions concerning the nodes i and j at the worst time--during the main loop which can easily include 10^5-10^6 cycles, even when only the non-zero elements of $\underline{\underline{A}}$ are computed. The solution is to make the decisions at a lower level and then when the main loop is cycled, very few decisions have to be made during each

cycle. The specific way in which FTRAN was constructed to do this will now be discussed.

Since each A_{ij} is made up of four or five (or more) integrals corresponding to the transport term, collision term, scattering term, (or sum of scattering terms for anisotropic scattering), boundary term, and interface term (if discontinuous spatial elements are used), consider a typical integral term

$$T_{ij} = \int_0^1 dx\, \Sigma_t(x) \int_{-1}^1 d\mu\, \psi_i^N(x,\mu)\, \psi_j^N(x,\mu)$$

or, using the tensor product basis functions,

$$T_{ij} = \left[\int_0^1 dx\, \Sigma_t(x)\, \psi_{ix}(x)\, \psi_{jx}(x)\right]\left[\int_{-1}^1 d\mu\, \psi_{i\mu}(\mu)\, \psi_{j\mu}(\mu)\right]$$

where ix is the x-node corresponding to node i, $i\mu$ is the μ-node corresponding to node i, etc.

Defining the factors of T_{ij} as

$$TX_{ij} = \int_0^1 dx\, \Sigma_t(x)\, \psi_{ix}(x)\, \psi_{jx}(x)$$

$$TU_{ij} = \int_{-1}^1 d\mu\, \psi_{i\mu}(\mu)\, \psi_{j\mu}(\mu)$$

the problem of calculating the 2-D integral T_{ij} is reduced to two separate 1-D integrals TX_{ij} and TU_{ij}. This in itself significantly reduces the effort because the spatial and angular integrals are done separately over their respective axes. For example, with 100 spatial nodes and 10 angular nodes, the number of integrals T_{ij} is ~9000 (assuming linear elements, each node interacts with 8 other nodes) while the number of

TX_{ij} integrals is ~300 and the number of TU_{ij} integrals is ~20. Thus the effort is reduced from computing 9000 integrals to computing 320 integrals, which are simply multiplied together to obtain

$$T_{ij} = TX_{ij} * TU_{ij}$$

Another important consequence is that different material properties (e.g. $\Sigma_t(x)$) can be factored into the integral TX_{ij}. However, the calculation of the integrals TX_{ij} and TU_{ij} can also take advantage of lower level calculations. Using the notation of Sec. III.E, a typical integral TX_{ij} (or TU_{ij}) can be expressed in terms of the generic basis functions over the standard interval [-1,1]: (assume j = i + 1)

$$TX_{ij} = \int_{X_{ix}}^{X_{i+1}} dx\, \psi_{ix}(x)\psi_{i+1}(x)$$

$$= \frac{\Delta X}{2} \int_{-1}^{1} d\xi\, \psi_1(\xi)\psi_2(\xi)$$

where $\psi_1(\xi)$ and $\psi_2(\xi)$ are as shown in Figure 3.

Since there are only two generic basis functions for linear elements, every TX_{ij} or TU_{ij} can be obtained from the three integrals

$$\int_{-1}^{1} d\xi\, \psi_1(\xi)\psi_1(\xi)\,,\quad \int_{-1}^{1} d\xi\, \psi_1(\xi)\psi_2(\xi)\,,\quad \int_{-1}^{1} d\xi\, \psi_2(\xi)\psi_2(\xi)\,.$$

Thus the computation of the 9000 integrals for the example above has been reduced to the computation of 3 integrals.

Once the TX_{ij} and TU_{ij} are computed, the calculation of T_{ij} is done in an efficient loop which now only needs to identify the x-nodes and μ-nodes corresponding to i and j, so that the TX_{ij} and TU_{ij} can be retrieved. Once these are known, T_{ij} is formed by taking the pro-

duct of the two array elements TX_{ij} and TU_{ij}. No other decisions need be made in this main loop to calculate the T_{ij}. Of course, symmetry (and anti-symmetry) of the contributions to A_{ij} can be used, as can the finite interaction distance between nodes to cut down the number of cycles in the main loop.

For example, as indicated in Table I, a problem involving 41 spatial nodes and 22 angular nodes will result in 902 unknowns and 59,320 non-zero matrix elements A_{ij} or $\sim$250,000 integrals to compute. The actual time required to do this was .671 sec for the main loop, or~3 μsec per integral. The time required in the intermediate calculation loops (for TX_{ij}, TU_{ij}, etc.) was .005 sec, which is negligible. If this earlier calculation had not been done, the time in the main loop might have been doubled or tripled.

CHAPTER VIII

NUMERICAL RESULTS

The previous chapters of this dissertation have been devoted to a thorough examination of the finite element method and its application to transport problems. This examination has included theoretical discussions concerning convergence rates as well as practical aspects such as the method by which the resultant system of equations is solved. So far, though, no numerical results have been presented to support a claim that the method is even successful, let alone substantiating the various claims and assertions concerning convergence rates as a function of mesh refinement or the improvement to be realized if one employs discontinuous finite elements in both space and angle. Thus, this chapter is intended to serve as the vehicle to substantiate the findings and results obtained in the earlier chapters.

This chapter includes confirmation that the method is applied and implemented correctly as well as examples to demonstrate the effectiveness of the method for solving eigenvalue problems, problems with strong heterogeneities, and time-dependent problems. In addition, numerical results are given to substantiate the claims made in Chapter II concerning the solution convergence rates and eigenvalue convergence rates as a function of mesh refinement.

VIII.A. Verification of Code

A necessary task in the development of any analytical procedure or computer code is to ensure that correct results are obtained when the code or procedure is applied to a problem with a known solution.

However, this is not conveniently done in neutron transport because exact solutions are known for only a few problems, and these problems generally are very simplified, such as semi-infinite mediums, pure absorption, etc. Even in these cases, the solution may exist in the form of an expansion that can only be evaluated numerically, thus introducing possible errors in the exact solution. If one wanted to examine the correct operation of a code that could handle multiple-region anisotropic scattering with arbitrary boundary conditions, then comparisons would have to be made with existing codes that people have confidence in or comparisons made with experimental data. If the goal is to obtain precise numerical results concerning convergence to an exact solution, then these methods would not suffice because of the errors inherent in each approach.

However, there is a method available that allows exact solutions to any general integro-differential, non-linear, inhomogeneous problem to be computed. The method is quite simple, although in practice it may require a more flexible code to use the method than would be required for the application the code was intended for. This method had been suggested earlier for the neutron transport equation[46] and has been found to be of extreme usefulness, both for debugging the code and for obtaining precise numerical convergence results. The method consists of assuming a solution to the transport equation and then computing the source necessary to balance the equation. Since the solution is known, the incoming boundary conditions are known. The code is then run with this source/boundary condition configuration as input, with the goal of obtaining the assumed angular flux (or some approximation to it) as output.

The observation that makes this method useful for a code that involves a numerical approximation technique is that if a solution is assumed that the method is able to approximate exactly, then a correctly written code should reproduce the assumed solution exactly (within machine precision). For example, if a bilinear angular flux (in x and μ) is assumed then the finite element method with linear or higher order elements should yield the assumed flux exactly. Any discrepancies in the solution which cannot be attributed to machine round-off probably indicate errors in the coding or in the formulation of the numerical approximation.

For example, the following test problem was used to verify several capabilities of FTRAN. The assumed flux was

$$Q(x,\mu) = -(x-11)(x+1)(5-\mu^2) \quad , \quad 0 \le x \le 10$$

which is a biquadratic polynomial in x and μ. Reflecting boundary conditions are automatically satisfied at x = 0 although x = 0 could also be a specified incoming flux boundary. The incoming flux at the right boundary is

$$Q(10,\mu) = 55 - 11\mu^2 \quad , \quad \mu \le 0$$

and if the assumed flux is substituted into the 1-D transport equation with isotropic scattering, the resultant source is easily calculated to be

$$S(x,\mu) = -\mu(2x-10)(5-\mu^2) - \Sigma_t (x-11)(x+1)(5-\mu^2) + \Sigma_s (x-11)(x+1)(14/3)$$

When FTRAN was run with $S(x,\mu)$ above as the source (via FUNCTION SORCE) and the above incoming boundary conditions at x = 10, the computed solution agreed with the assumed solution (within 8-10 decimal places) everywhere when quadratic or cubic finite elements were used. This close agreement must be obtained because quadratic or cubic elements can approximate the assumed flux exactly and since the transport equation in this case reduces to an algebraic identity with a unique solution, this exact solution must be the only solution.

The simple assumed flux

$$Q(x,\mu,t) = 1-t \quad , t \geq 0 \quad , 0 \leq x \leq 1,$$

is capable of testing the time-dependent portion of FTRAN, as well as the steady state portion, because the Crank-Nicholson scheme is capable of approximating this linear function of time exactly. This assumed flux leads to the source

$$S(x,\mu,t) = (\Sigma_t - \Sigma_s)(1-t)$$

and boundary conditions

$$Q(0,\mu,t) = 1-t, \quad \mu \geq 0$$
$$Q(1,\mu,t) = 1-t, \quad \mu \leq 0$$

plus initial conditions

$$Q(x,\mu,0) = 1$$

When FTRAN was run with the above as input, the angular flux at t = 1 was identically zero, independent of the time step chosen.

A more elaborate test problem, called ULT-TEST, is capable of testing all of the capabilities of FTRAN simultaneously, and thus is a useful tool for checking the code after changes have been made. ULT-TEST consists of a three-region slab with different anisotropic scattering coefficients in each region. The assumed flux is similar to the biquadratic flux above except time-dependence has been added:

$$Q(x,\mu,t) = -(x+1)(x-11)(5-\mu^2)(1-t)$$

$$\text{for} \quad 0 \le x \le 10$$

$$t \ge 0$$

which generates the initial conditions

$$Q_0(x,\mu) = -(x+1)(x-11)(5-\mu^2)$$

and time-dependent boundary conditions (reflecting at x = 0)

$$Q(0,\mu,t) = (55-11\mu^2)(1-t) \quad , \mu \le 0 .$$

When this assumed flux is substituted into the time-dependent trnasport equation with anisotropic scattering (L_{max} = 4), Eq. (154), the following source is calculated:

$$S(x,\mu,t) = \frac{1}{v}(x-11)(x+1)(5-\mu^2)$$

$$-(2x-10)\mu(5-\mu^2)(1-t) - \Sigma_t(x-11)(x+1)(5-\mu^2)(1-t)$$

$$+\left[\frac{14}{3}b_0 - (\mu^2-\tfrac{1}{3})b_2\right](x+1)(x-11)(1-t) .$$

Depending on which region of the slab is being considered, appropriate values of Σ_t, b_0, and b_2 are substituted into the above to obtain the specific source for each region. Note that the source

$S(x,\mu,t)$ is independent of the Legendre expansion coefficients b_1, b_3, and b_4. This is clearly true because the assumed flux consists of P_0 and P_2 Legendre polynomials, and P_1, P_3, and P_4 are orthogonal to P_0 and P_2 over the interval [-1, 1]. Hence these terms do not contribute to the scattering expansion.

When this source is input via FUNCTION SORCE and the time-dependent incoming flux is input via FUNCTION BDYFLX, FTRAN will yield a flux at t = 1 which is identically zero, which is the correct solution. Thus the test is easily run and checked at a glance.

Also, the correct operation of discontinuous spatial and angular elements can be verified because the computed solution should be continuous even if discontinuous elements are used. One could also choose an assumed angular flux with a discontinuity at $\mu = 0$ to check the correct operation of the discontinuous angular elements, which was done. However, the same could not be done with the discontinuous spatial elements because in fact continuity of the angular flux is a natural interface condition and if the assumed flux is analytically discontinuous (which it is not for real problems) the step during the calculation of the interface matrix elements where the flux is assumed continuous in the direction of neutron motion is no longer valid. The point is that the discontinuous spatial elements can approximate a rapidly varying angular flux but not a discontinuous angular flux.

To summarize, the above method is admittedly of no use for calculating actual solutions to the transport equation, but it has resulted in considerable savings in time and effort for checking out and debugging FTRAN, both initially and after the multitude of options were

added. However, the assumed flux method also contributed significantly to the study of the numerical convergence of the finite element method, as is discussed in the next section.

VIII.B. Angular Flux Convergence Results

The assumed solution method also allows for the performance of precise numerical convergence tests, because the exact solution is known and one can compute the error in the approximate method as the mesh is refined (as long as exact agreement is not possible). This was done with a single region problem with the following assumed angular flux:

$$Q(x,\mu) = 100\, e^{-x} \cos(\pi\mu/2), \quad 0 \le x \le 1$$

with

$$\Sigma_t = 2$$

$$\Sigma_s = 0, 1, 2 \quad \text{(to examine dependent of convergence rate on the scattering ratio } C = \Sigma_s/\Sigma_t)$$

This results in the following source and incoming boundary conditions:

$$S(x,\mu) = \left[100\, e^{-x} \cos(\pi\mu/2)\right]\left[(-\mu+\Sigma_t) - \frac{2}{\pi}\Sigma_s\right]$$

$$Q_0(0,\mu) = 100\cos(\pi\mu/2), \quad \mu \ge 0$$

$$Q_0(1,\mu) = 100\, e^{-1} \cos(\pi\mu/2), \quad \mu \le 0$$

The L_2-error in the approximate solution $Q_h(x,\mu)$ was calculated as

$$\|Q-Q_N\|_0 \equiv \left[\int_0^1 dx \int_{-1}^{1} d\mu \, |Q(x,\mu)-Q_h(x,\mu)|^2\right]^{1/2}$$

and was done with double 4, 5, or 6 point Gaussian quadrature over each mesh element, when linear, quadratic, or cubic finite elements were used, respectively. (The normal quadrature for the integrals in the matrix elements is double 2, 3, or 4 point on each element). The mesh spacing h was defined as $\frac{1}{N}$ where N was the number of mesh intervals which was the same in both the spatial direction and the angular direction. (A mesh interval is defined as an interval between any two adjacent nodes, whether interior nodes or principal nodes.)

The results were obtained with several mesh spacings for each of the finite elements (linear, quadratic, cubic) and with C = 0, 1/2, 1 and with the boundary conditions treated as essential boundary conditions and as natural boundary conditions. The numerical results have been plotted in Figure 12 and they motivate the following remarks:

1. The observed convergence rate (in the L_2-norm) is clearly

$$\|Q-Q_N\|_0 = O(h^{k+1})$$

where k is the degree of the finite element.

2. The observed convergence rate is independent of the scattering ratio C and is independent of how the boundary conditions were treated (natural vs. essential).
3. Although the error convergence rates are the same, the absolute error is less with the use of the natural boundary condition formulation versus the essential boundary condition formulation of the integral law. This might

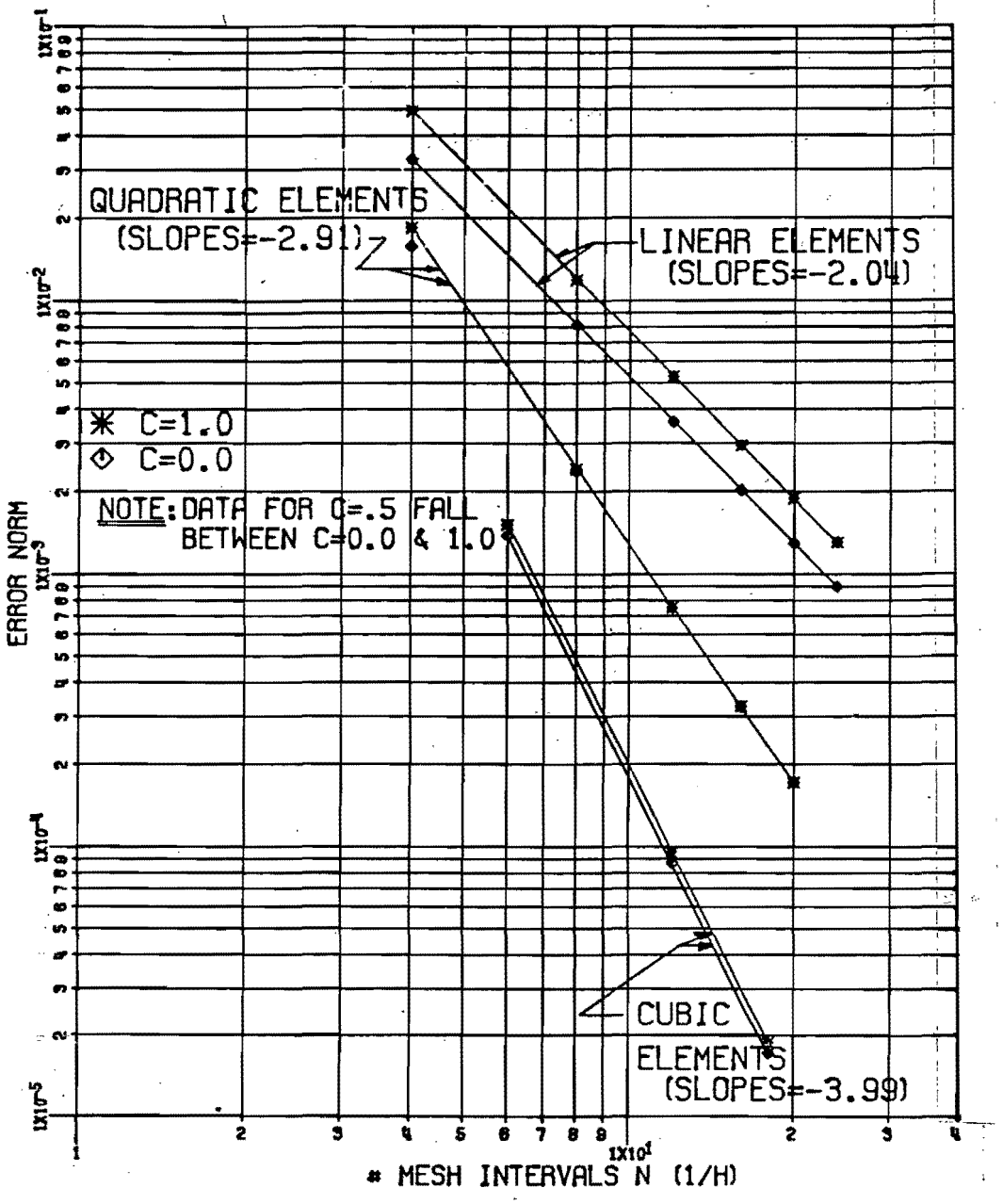


Figure 12. L_2 Error vs. Mesh Spacing

be expected because with the natural boundary conditions, the system is not restricted as to its choice for the solution on the incoming boundary, thereby allowing for a better overall solution to minimize the error throughout the region.

4. Again, the error rates are identical, but the absolute error is less with more absorption (small c).

VIII.C. Eigenvalue Problems

(1) Critical Slab Problem (Isotropic Scattering)

This application of our method concerns the classical eigenvalue problem of neutron transport theory--the calculation of the number of secondary neutrons, c, required per collision to achieve criticality in a slab of given half-width (measured in mean free paths). The specific eigenvalue equation to be solved is then

$$\mu \frac{\partial Q}{\partial x} + Q(x,\mu) = \frac{c}{2} \int_{-1}^{1} d\mu' \, Q(x,\mu') \qquad (159)$$

with boundary conditions

$$Q(0,\mu) = Q(0,-\mu)$$
$$Q(a,\mu) = 0, \quad \mu \leq 0$$

where a is the half-width in mfp.

Proceeding in an analogous manner with the derivation of the 1-D integral law in Sec. III.B., the equivalent integral for the eigenvalue problem is

Find $Q(x,\mu) \in H^1$ and C such that for all $\Psi(x,\mu) \in H^1$

$$-\left(Q, \mu \frac{\partial \Psi}{\partial x}\right) + \langle Q, \Psi \rangle_B + (Q, \Psi) = c\left(\frac{1}{2}\int_{-1}^{1} d\mu' Q(x, \mu'), \Psi\right)$$

where $\langle Q, \Psi \rangle_B$ is the surface term after the boundary conditions have been explicitly included (in the direction of neutron travel for the reflecting b.c.).

Now apply the finite element approximation to solve Eq. (160), as was done in Sec. III.C., to obtain the generalized matrix eigenvalue problem:

$$\underline{\underline{A}}\,\underline{\Phi} = c\,\underline{\underline{M}}\,\underline{\Phi} \tag{161}$$

where A_{ij} is the same as in Eq. (125) except the scattering term M_{ij} has been subtracted. Equation (161) is then solved using the inverse power iteration method in accordance with Sec. VI.C.

The following discussion contains the results of several different eigenvalue calculations, including comparisons with the production-level 1-D discrete ordinates, code, ANISN.[46]

Four different slabs were analyzed with various order finite elements and mesh spacings. Discontinuous angular elements at $\mu = 0$ were used for all FTRAN runs, and uniform spatial and angular meshes were used. The eigenvalue problems were also solved with ANISN to allow comparison with a discrete ordinates code. The ANISN runs were performed with $\Sigma_s = .5, \nu\Sigma_f = .5$ and the k-search option[47], with c calculated as $c = \frac{1}{2}(\frac{1}{k} + 1)$.

All runs were performed with single precision (32 bit) arithmetic on The University of Michigan AMDAHL 470V/6 computer. The eigenvalue error was chosen to be 10^{-6}, which is approaching the smallest error realistically obtainable with single precision arithmetic.

Table II summarizes the FTRAN and ANISN results for the four slabs and includes benchmark eigenvalues reported by Kaper, Leaf, and Lindeman[49] for the different slabs. Table II also includes actual solution times for FTRAN and ANISN for the slabs of half-width 1.0 mfp and 100.0 mfp. For FTRAN, the time is the CPU time required to factor $\underline{\underline{A}}$ and perform the inverse iterations, while for ANISN the time is that required to perform the necessary outer-inner iterations to converge to the eigenvalue. Because these codes are so dissimilar, these timing results should be viewed as providing only an approximate comparison of the relative efficiencies of the two methods for a specific application. The comparison in Table II stimulates several observations.

First, the convergence of the finite element method is extremely rapid and excellent agreement with the benchmark eigenvalues of Kaper, et al[48] was obtained for even the very coarse mesh of two spatial and two angular intervals. For the higher order elements and/or the refined meshes (which are still fairly coarse), the agreement approaches the accuracy possible in single precision arithmetic. In fact, this rapid convergence forced a restriction to cases with fairly coarse meshes when single precision arithmetic was used. Secondly, the use of higher order elements on the same mesh generally results in better agreement; however, this increased accuracy results in significantly longer running times due to the increased coupling between the spatial

$\frac{1}{2}$ width (mfp)	Benchmark[a] (Ref.48)	FTRAN Results-- L(Linear),Q(Quadratic),C(Cubic) Number of mesh intervals in x,μ					ANISN	
		LL (2,2)	LL(4,4)	LL(6,6)	QQ(4,4)	CC(6,6)	S_6, N_x=6	S_{16}, N_x=16
.5	1.615379	1.60319 (.75%)	1.61140 (.25%)	1.61462 (.05%)	1.61217 (.2%)	1.61539 (.0007%)	1.65302 (2.3%)	1.61554 (.01%)
1.0	1.277102	1.27193 (.4%)	1.27676 (.027%)	1.27706 (.003%)	1.27708 (.002%)	1.27712 (.001%)	1.28417 (.55%)	1.27724 (.01%)
5.0	1.024879	1.02479 (.008%)	1.02488 (.0001%)	1.02493 (.005%)	1.02490 (.002%)	1.02489 (.001%)	1.02627 (.14%)	1.02496 (.008%)
10.0	1.007135	1.00717 (.003%)	1.00721 (.007%)	1.00718 (.004%)	1.00720 (.006%)	1.00717 (.003%)	1.00884 (.17%)	1.00726 (.01%)
CPU Time[b] (sec)	1.0 mfp	.032	.10	.22	.15	.46	.065	.26
	10.0 mfp	.15	.37	.95	.50	1.7	.34	1.9

[a]Rounded to 7 figures

[b]AMDAHL 470V/6

TABLE II. Eigenvalues (Isotropic Scattering)

Position	Benchmark[a] (Ref. 48)	FTRAN Results (Same notation as Table II.)					ANISN	
		LL (2,2)	LL(4,4)	QQ(4,4)	LL(8,8)	CC(12,12)	$S_6(N_x=4)$	$S_{16}(N_x=16)$
0.0	1.00[c]	1.00[c]	1.00	1.00	1.00	1.00	1.00	1.00
0.25	.9660	.9660[b] (.6%)	.9827 (1.7%)	.9691 (.3%)	.9658 (.02%)	.9663 (.03%)	.9645 (.16%)	.9660 (0%)
0.50	.8651	.9200 (6%)	.8665 (.16%)	.8666 (.17%)	.8646 (.06%)	.8645 (.07%)	.8587 (.74%)	.8652 (.01%)
0.75	.6982	.6728 (3.6%)	.7186 (2.9%)	.7037 (.8%)	.6986 (.06%)	.6985 (.04%)	.6840 (2%)	.6984 (.03%)
1.00	.4341	.4256 (2.0%)	.4415 (1.7%)	.4412 (1.6%)	.4405 (1.5%)	.4368 (.6%)	.4401 (1.4%)	.4346 (.12%)

[a] Rounded to 4 figures
[b] Linear extrapolation
[c] Center fluxes normalized to unity

TABLE III. Scalar Fluxes (Isotropic Scattering)

nodes. Finally, a comparison of the ANISN results with the FTRAN results indicate that the finite element method yields more accurate eigenvalues on relatively coarser meshes, with at least comparable execution times.

The eigenfluxes were also compared for the slab of half-width 1.0 mfp. Table III contains a comparison of the nodal scalar flux corresponding to the converged angular flux eigenfunction for each mesh spacing. These results include different order finite elements as well as two different ANISN runs. The comparison of the results with the benchmark fluxes of Ref. 48 should be viewed with some caution; however, because the normalization used to compare fluxes was to set the flux at the center of the slab ($x = 0$) equal to unity for all cases. In particular for the finite element results, where the best approximation to the actual flux may not involve equality at $x = 0$, this normalization is suspect. However, a more satisfactory normalization, such as multiplying the approximate solution by a constant which is chosen to minimize the least square error (L_2 error) between the benchmark flux and approximate flux, would involve more effort than justified for the limited conclusions that may be drawn.

In any event, it would appear from Table III that reasonable agreement is obtained with the coarse FTRAN runs although the best results were obtained with the S_{16} ANISN run. The S_6 ANISN run compared favorably with the LL(4,4) FTRAN run, which is unexpected on the basis of the eigenvalue results for these cases.

(2) Critical Slab Problem (Linear Anisotropic Scattering)

The second application was to repeat the above criticality problem including anisotropic scattering. Furthermore, the limiting cases of very thin and very thick slabs were considered with slab half-widths of .01 mfp, 1.0 mfp, and 100 mfp being analyzed. FTRAN was run with anisotropic scattering coefficients b_1 = 0, +1/3, and -1/3 for each of the three slabs (see Eq. (120) for definition of the b_i). Table IV compares the results of the FTRAN runs with earlier results reported by Kschwendt[49], who used the SP_N-P_L method to calculate the eigenvalues. Note that different mesh spacings and element types were used for the individual slabs, since a thin slab requires a highly refined angular mesh while a thick slab requires a relatively more refined spatial mesh. A comparison of the results for the thin slab (.01 mfp) would indicate that the eigenvalues reported in Ref. 49 are more accurate, since the FTRAN eigenvalues appear to be approaching the Ref. 49 eigenvalues as the angular mesh is refined. A comparison of the LL(1,24) and LL(2,24) results for the .01 mfp slab clearly illustrates the insensitivity to refinement in the spatial mesh. It is interesting to note that the LC(1,24) results are significantly more accurate than the LL(1,24) results, although the number of unknowns and the execution times are identical. This is due to the fact that the angular nodes already being fully coupled by the scattering, cannot be coupled any more strongly for the cubic angular elements; hence the increased accuracy is free (neglecting the insignificant increase in time to assemble the matrix).

As noted by Kschwendt[49], the thin slab (.01 mfp) cannot be treated by the S_N method unless a large number of angular quadrature points are

$\frac{1}{2}$ width	Anisotropy (b_1)	Ref. 49	FTRAN Results			
.01 (mfp)	Element & Mesh		LL[a](1,24)	LL(2,24)	LC(1,24)	LL(i,50)
	$+\frac{1}{3}$	20.9031	21.9064	21.8913	21.6447	21.0512
	0	20.6219	21.5839	21.5700	21.3334	20.7634
	$-\frac{1}{3}$	20.3544	21.2810	21.2655	21.0405	20.4919
1.0	Element & Mesh		LL(2,2)	LL(6,6)	QQ(4,4)	CC(6,6)
	$+\frac{1}{3}$	1.32147	1.31491	1.32141	1.32145	1.32149
	0	1.27710	1.27193	1.27706	1.27708	1.27712
	$-\frac{1}{3}$	1.24468	1.24045	1.24465	1.24466	1.24471
100.0	Element & Mesh		LL(3,2)	LL(25,2)	CL(24,2)	LL(50,2)
	$+\frac{1}{3}$	1.00012	1.00020	1.00040	1.00042	1.00035
	0	1.00008	1.00014	1.00035	1.00035	1.00033
	$-\frac{1}{3}$	1.00006	1.00011	1.00032	1.00030	1.00027

TABLE IV. Eigenvalues (Anisotropic Scattering)

EIGENVALUES FOR 1.0 mfp SLAB (ISOTROPIC SCATTERING) (For comparison of discontinuous vs. continuous angular elements at μ=0)		
MESH	FTRAN (Continuous)	FTRAN (Discontinuous)
LL(2,2)	1.2538 (1.8%[a])	1.2719 (.49%)
LL(4,4)	1.2738 (.25%)	1.2768 (.03%)
LL(6,6)	1.2761 (.08%)	1.27706 (.003%)
LL(8,8)	1.2766 (.04%)	1.27708 (.002%)
LL(10,10)	1.2768 (.02%)	1.27709 (.0009%)

EIGENVALUES FOR .01 mfp ISOTROPIC SLAB (Indicates sensitivity of angular mesh & insensitivity of spatial mesh)		
MESH	FTRAN (Discontinuous)	% Difference[b]
QQ(6,6)	28.46	38%
QQ(6,8)	24.87	21%
QQ(6,16)	22.42	9%
LL(14,14)	22.86	11%
LL(6,2)	34.22	66%
LL(4,2)	34.22	66%
LL(1,2)	34.22	66%

[a]Compared with c= 1.27710 (Reference 48)
[b]Compared with c= 20.62 (Reference 49)

TABLE V. Miscellaneous Eigenvalue Results

used, such as S_{32}. Again, the reason for this is that the thin slab necessitates a good angular approximation and the typical S_N quadrature sets (up to S_{16}) do not suffice. Use of arbitrarily fine angular meshes pose no problem with FTRAN, which is somewhat of an advantage for the finite element method.

For the intermediate width slab (1.0 mfp), the FTRAN and Ref. 49 results are in excellent agreement, even though the most refined FTRAN mesh had only 6 angular and 6 spatial mesh intervals. The results for the thick (100 mfp) slab, however, do not agree as well; this might be expected because the eigenvalues are very close to unity and a 10^{-5} error effectively changes the eigenvalue considerably. However, we believe that the FTRAN results with the refined spatial meshes are more accurate because the Ref. 49 results were obtained with a spatial truncation equivalent to a coarse spatial mesh and it was observed with FTRAN that a coarse spatial mesh resulted in very small eigenvalues (see Table V), similar to Ref. 49. In addition, the FTRAN results with the finer spatial meshes compared favorably with an ANISN S_6 run with 25 spatial intervals which yielded C = 1.00041. Note that the use of cubic spatial elements on the same mesh as linear elements does not appear to yield any significant difference, although the computational effort is increased substantially due to the stronger spatial coupling.

It is interesting to note that the execution times were independent of the inclusion of anisotropic scattering. That is, the number of eigenvalue iterations was independent of the anisotropy, and since the direct solution time only depends on the total number of nodes (which was constant for the isotropic and anisotropic cases) the total execution was relatively constant with a given mesh.

Several calculations were also performed using continuous angular elements for comparison with the discontinuous elements. The resultant eigenvalues were consistently poorer with the continuous elements, espeically for the .01 mfp slab where the eigenvalue is extremely sensitive to the angular approximation. These results are shown in Table V, along with several other miscellaneous eigenvalue results.

VIII.D. Eigenvalue Convergence Rates

The isotropic scattering eigenvalue problem considered above was also analyzed with the intention of obtaining precise numerical results for the eigenvalue convergence rate as a function of mesh spacing (h) for linear, quadratic, and cubic elements with both continuous and discontinuous angular elements at $\mu = 0$. Therefore, FTRAN was modified to incorporate double precision arithmetic for all calculations and a 10^{-10} error criterion was placed on the eigenvalue convergence. Although a limit on the maximum angular flux error was not imposed, this error was calculated and was within 10^{-7} for the last two iterations for all of the cases considered. Figure 13 illustrates the results for a series of FTRAN runs with different mesh spacings, different finite elements, and continuous and discontinuous angular elements. The results clearly indicate the improvement in the convergence rate with discontinuous angular elements, since $\sim O(h^3)$ convergence was achieved with discontinuous linear elements, while $\sim O(h^{2.5})$ convergence was obtained with continuous linear elements.

Although the results are not shown, the observed convergence rates for quadratic and cubic continuous elements was nearly identical to the

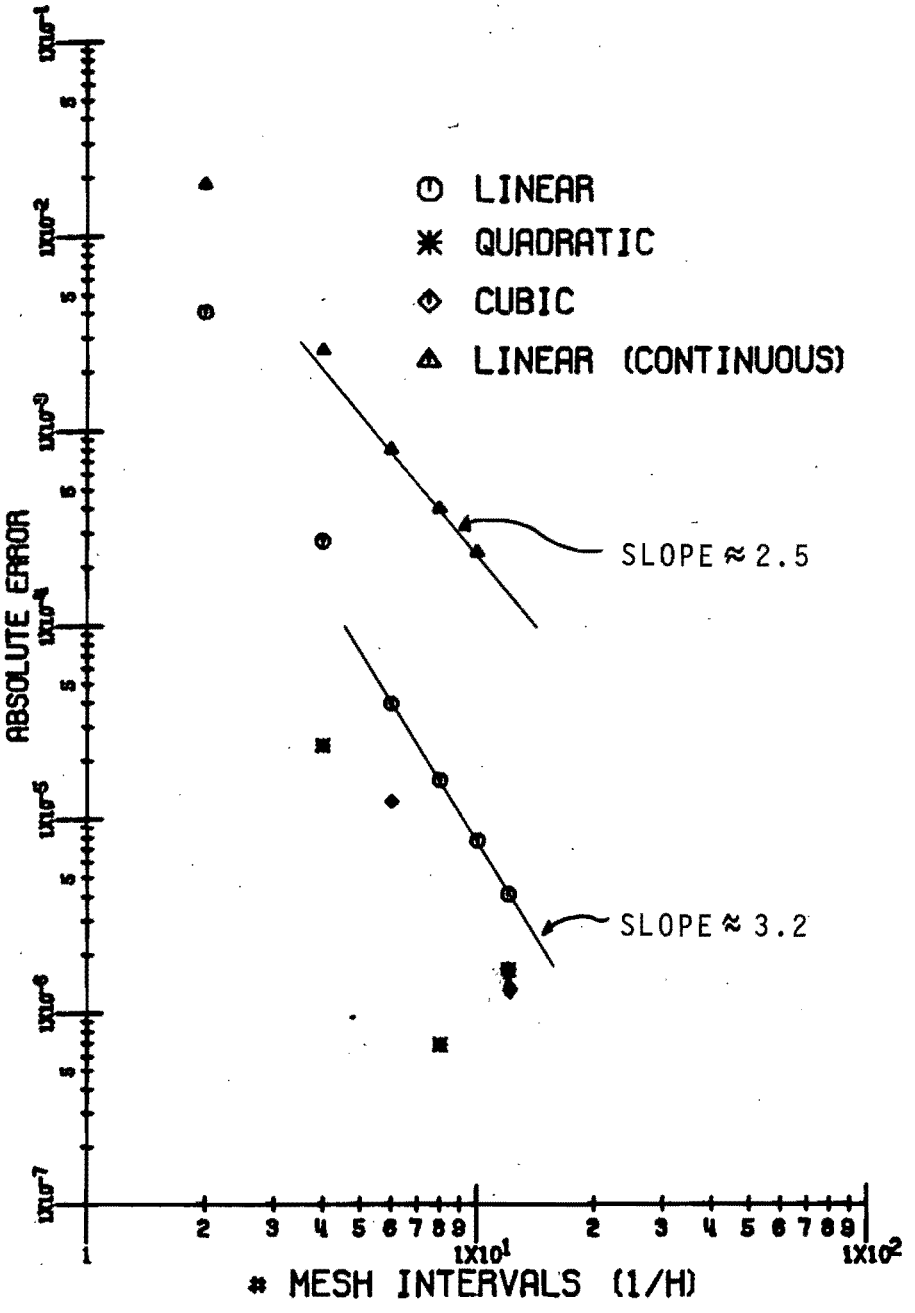


Figure 13. Eigenvalue Error vs. Mesh Spacing

observed rate for linear continuous elements, which indicates that the effect of the singularity at $\mu = 0$ is dominating the error.

For the quadratic and cubic discontinuous elements, the error was apparently too small to be resolved. That is, with quadratic elements the smallest mesh configurations, 4 x 4 and 8 x 8, yielded errors of 2.4×10^{-5} and 6.5×10^{-7}, respectively, which indicates better than $O(h^6)$ convergence. However, the 12 x 12 result was poorer than the 8 X 8, probably due to the fact that pivoting was not performed which became important for small errors. For cubic elements, the smallest mesh, 6 x 6, yielded an error of 1.2×10^{-5} while the error for the next mesh, 12 x 12, which might have been near 10^{-10} on the basis of the linear and quadratic results, was only nominally better at 1.7×10^{-6}. Also, the linear elements with the 12 x 12 mesh resulted in a higher than expected error which supports the contention that other factors became important for extremely small ($< 10^{-6}$) errors. Regardless of the specific results for the higher order elements, though, we feel that the important conclusions here are that discontinuous angular elements effectively match the singularity at $\mu = 0$, and linear elements appear to yield $O(h^3)$ convergence.

As discussed earlier in the section on the theoretical error analysis, the observed eigenvalue convergence rates, which suggest $O(h^{2k+1})$ convergence for finite elements of degree k, are $O(h)$ faster than predicted for the second-order, elliptic, self-adjoint problem. Thus the results may be viewed with some skepticism. However, as the results of Sec. II.G indicate, the predicted convergence rate for linear elements is

$$\epsilon = C_1 h^3 + C_2 h^4 = O(h^3) \quad \text{as } h \to 0$$

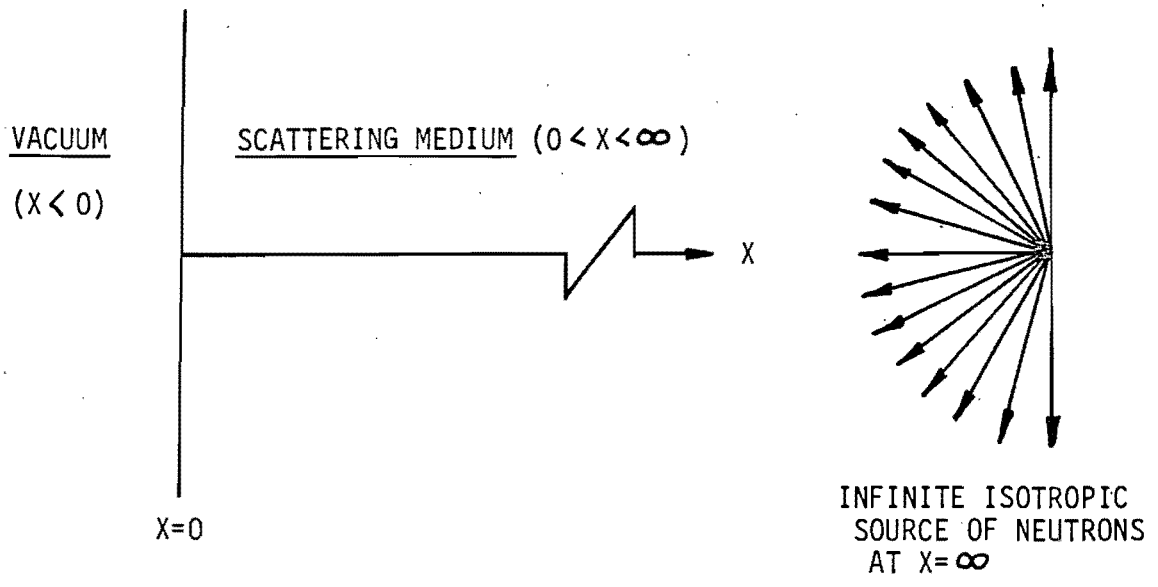
where the $O(h^4)$ term will, of course, be negligible as the mesh if refined. But a look at Figure 13 reveals the interesting fact that the initial dependence of the error is indeed $O(h^4)$, and as h becomes smaller, the $O(h^4)$ term becomes negligible compared with the $O(h^3)$ term. Thus the numerical results, at least for the linear elements, would appear to strongly support the predicted convergence rate

$$\epsilon \sim O(h^{2k+1}) \quad .$$

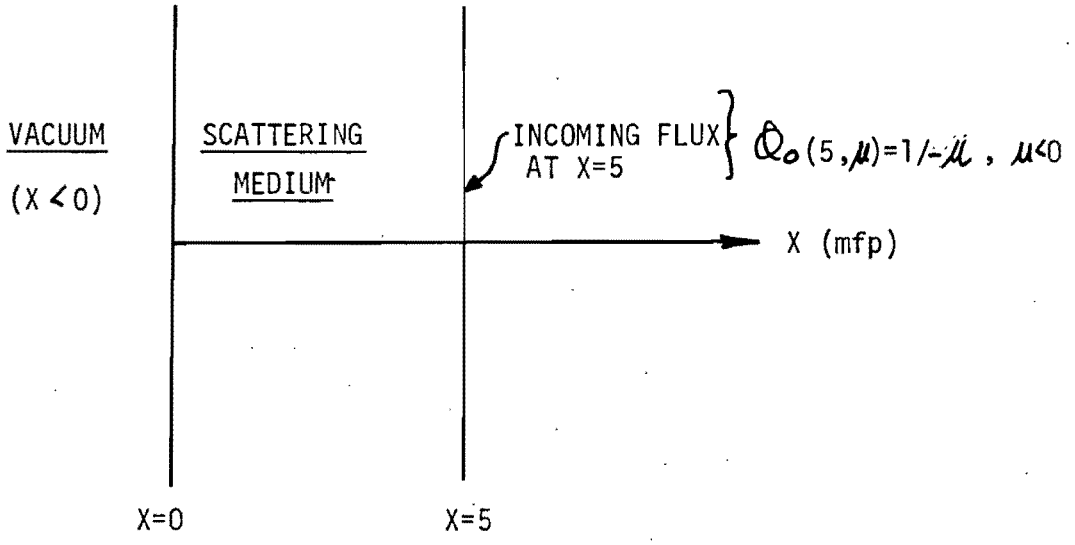
VIII.E. Milne Problem

The Milne problem is a classic problem in transport theory and is one of the few transport problems that is amenable to an exact solution. Therefore it is a typical test problem for transport codes in order to compare numerical results with exact results. Originally the Milne problem was formulated and solved to obtain the observed angular distribution of radiation emitted by a star. However, the problem may also be posed in terms of neutron transport, which will now be done.

The physical configuration is a vacuum for $x < 0$ and a pure scattering medium for $x > 0$, as indicated in Figure 14. Deep within the pure scattering region ($x \to +\infty$) is an infinite source of neutrons and these neutrons traverse the scattering medium until they reach the vacuum, where they simply stream to $-\infty$. The problem is to find the angular distribution of neutrons emerging from the scattering medium at $x = 0$. Also, the spatial dependence of the scalar flux near the vacuum



ACTUAL CONFIGURATION



EQUIVALENT CONFIGURATION

FIGURE 14. Configuration for Milne Problem

is of interest because the position in the vacuum where the asymptotic scalar flux extrapolates to zero is of considerable interest to nuclear analysis (extrapolated end point).

The infinite medium may be approximated satisfactorily by using a finite slab which is several mean free paths thick and a plane isotropic source of neutrons at the right boundary. Neutronically this finite slab behaves like an infinite medium because if it is thick enough the angular flux will attain an asymptotic form away from the source but before the effect of the vacuum boundary at x = 0 is felt. Specifically, FTRAN was used to solve the Milne problem assuming the slab was 5 mfp thick with a plane isotropic source of unit magnitude at x = 5. Although FTRAN is not specifically set up to handle a plane source, it may be accommodated in the following manner.

Case, de Hoffman, and Placzek[50] note that for a problem where all neutrons are due to a surface source (the present configuration), the equivalent problem consists of specifying the value of the incoming flux. That is, if S_0 is the strength of the plane isotropic source then the equivalent problem consists of an incoming flux with magnitude

$$\Psi(\underline{r}_s, \hat{\Omega}) = \frac{S_0}{|\hat{\Omega}\cdot\hat{n}|} \quad , \quad \hat{\Omega}\cdot\hat{n} < 0$$

where $\hat{n}$ is the unit outward normal to the surface $\underline{r}_s$.

For a 5 mfp slab with a unit isotropic plane source at x = 5, the equivalent problem consists of a source-free slab and an incoming flux

$$Q_0(5,\mu) = \left|\frac{1}{\mu}\right| \quad , \mu < 0 .$$

FTRAN was run with the above incoming flux distribution at x = 5 and vacuum boundary conditions at x = 0. The absorption cross-section was zero and the scattering cross-section was unity, therefore x is measured in mfp. The spatial mesh was uniform with spacing .05 mfp, resulting in 101 spatial nodes. The angular mesh was also uniform with $\Delta\mu = .20$, resulting in 11 nodes for continuous angular elements and 12 nodes for discontinuous angular elements. Thus the number of unknowns was 1111 or 1212, depending on the choice of angular elements. To give some information concerning the efficiency of FTRAN, the CPU time required to factor the matrix $\underline{\underline{A}}$ and backsolve for the solution was < 2.5 sec, while the time required to assemble $\underline{\underline{A}}$ was $< .6$ sec. This is consistent with most FTRAN runs with this approximate angular mesh, ~1-2 milli-sec per unknown to solve the system of equations $\underline{\underline{A}}\,\underline{\Phi} = \underline{S}$. (Data is for 1212 unknowns.)

The numerical results for the emergent angular distribution of neutrons are tabulated in Table VI for both the continuous and discontinuous angular elements, along with exact values from Case, et al.[50] The FTRAN results are normalized to $\int_{-1}^{0} du Q(0,\mu) = 1$, consistent with the normalization used for the exact results. Note that even with the relatively coarse angular mesh ($\Delta\mu = .20$), the FTRAN results agree quite well with the exact results for the discontinuous angular elements. Even the continuous results indicate good agreement, except near $\mu = 0$ where the forced continuity results in a numerical value for the $\mu = 0$ angular flux which is nearly half of the actual discontinuity. In addition, the values for the angular flux on the vacuum boundary for the discontinuous elements are much closer to zero (factor of 10) than with the continuous elements.

$-\mu$	Exact[a]	FTRAN[b]	FTRAN[c]
0.0	.500	.514 (2.8%)	.209 (58%)
0.2	.725	.727 (.28%)	.791 (9%)
0.4	.915	.915 (0%)	.929 (1.6%)
0.6	1.097	1.099 (.18%)	1.127 (2.7%)
0.8	1.276	1.278 (.16%)	1.308 (2.5%)
1.0	1.454	1.446 (.6%)	1.48 (1.8%)

[a]Reference 50. All angular fluxes normalized to unit emergent flux,

$$\int_{-1}^{0} d\mu \, Q(0,\mu) = 1$$

[b]Discontinuous angular elements at $\mu=0$
$\Delta\mu=.2$ (linear), $\Delta X=.05$ (linear)

[c]Continuous angular elements at $\mu=0$, same mesh

TABLE VI. Nodal Emergent Angular Flux (Milne)

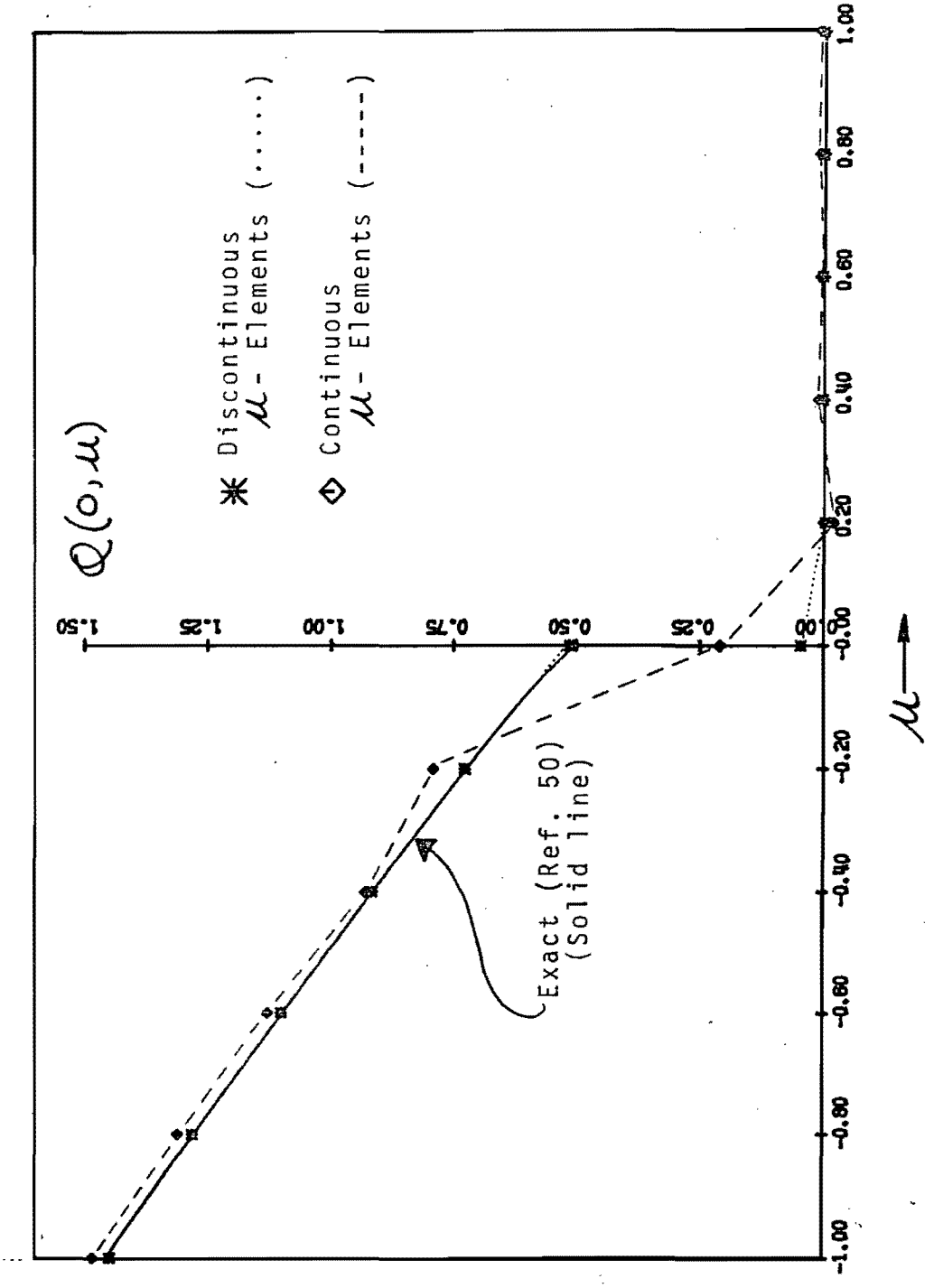


Figure 15. Emergent Angular Distribution (Milne)

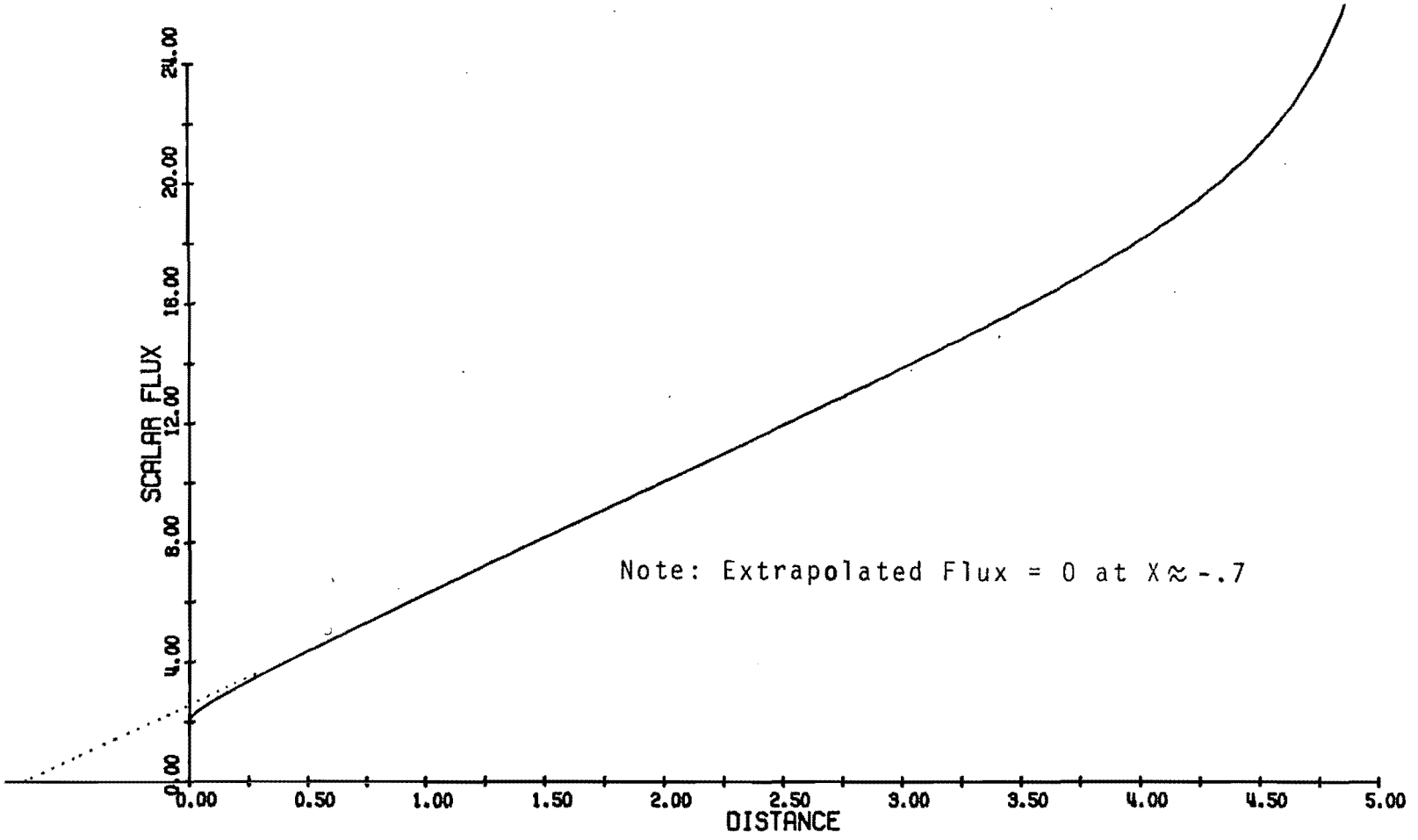


FIGURE 16. Scalar Flux (Milne)

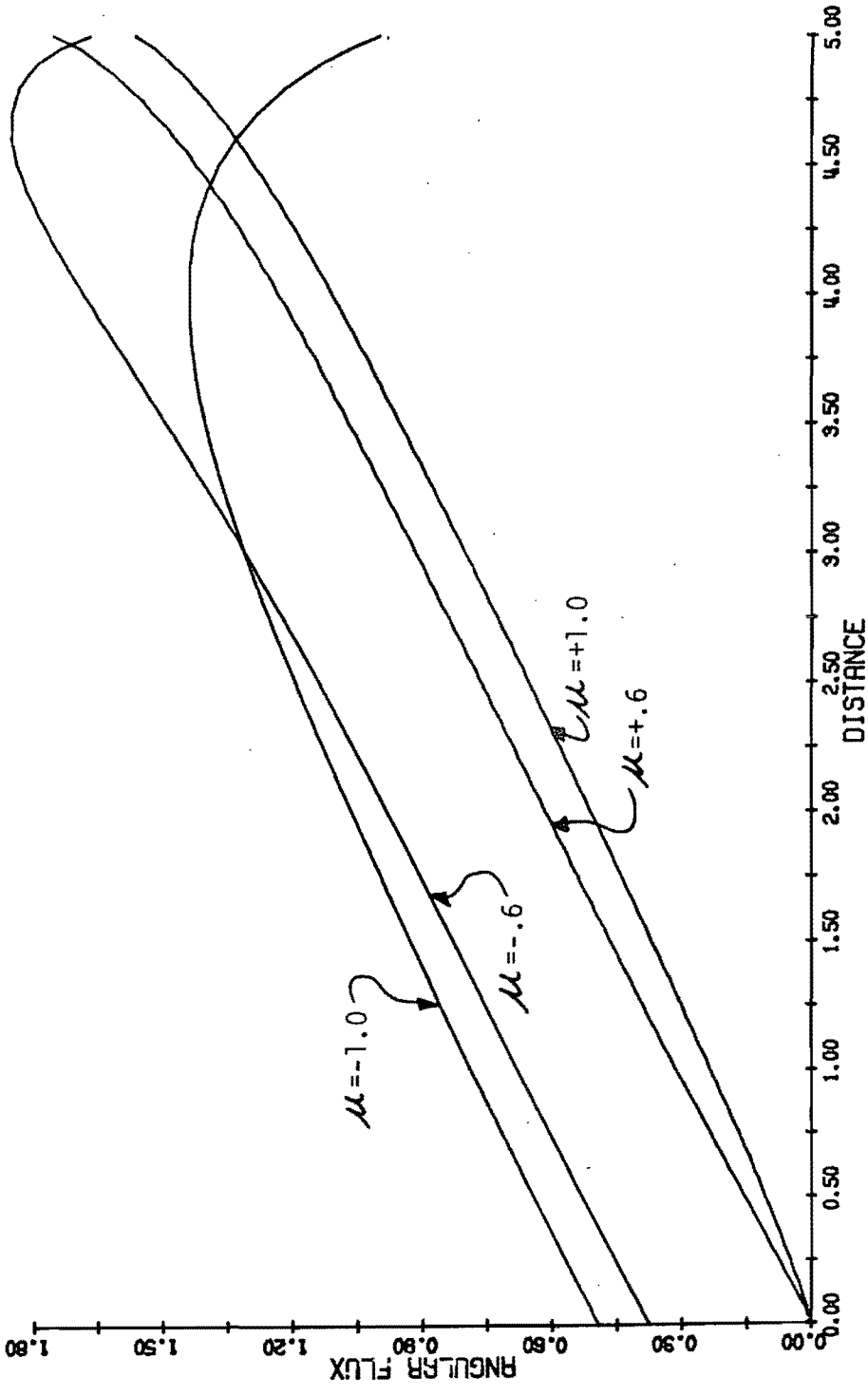


FIGURE 17. Angular Fluxes (Milne)

Figure 15 illustrates the results for the FTRAN results versus the exact values[50]. It can be seen that the discontinuous elements approximate the discontinuity at $\mu = 0$ quite well, which also allows an excellent approximation to the incoming flux, which should be zero. It should be noted that the results are almost identical to results from two different discrete ordinates solutions[51], one with a typical S_{16} quadrature and another with a quadrature (DP_7) which allows better resolution of the angular flux near $\mu = 0$.

Figure 16 is a plot of the scalar flux results from FTRAN (discontinuous elements) versus position. Note that the extrapolated scalar flux becomes zero at $x \approx -.7$, which is consistent with the exact value of .7104[52]. Figure 17 is a plot of several angular fluxes obtained with the FTRAN run with discontinuous elements and it can be seen that the fluxes do achieve their asymptotic forms away from the boundaries, which indicates that the choice of a 5 mfp slab was satisfactory.

Thus FTRAN yields excellent results for the solution of the Milne problem, which is further evidence for the successful application of the finite element method to problems in neutron transport.

VIII.F. Source Problem With Severe Heterogeneities

Since realistic problems in reactor analysis involve multi-region problems with strong heterogeneities and localized sources, any viable numerical method for solving the transport equation should be capable of treating such problems. This section contains a detailed examination of the application of the finite element method to a problem with

strong heterogeneities, including the use of discontinuous spatial and angular finite elements, which are shown to result in a marked improvement in the solution.

This particular problem was chosen because it has been considered previously by other investigators to test their numerical methods for treating heterogeneous problems. The initial treatment of the problem was performed by Reed[53], who applied various spatial difference schemes within the discrete ordinates approach to solve the problem. Pitkaranta[26] subsequently used the problem to test the application of a finite element method which effectively solved both the even and odd parity forms of the second order transport equation by using a non-self-adjoint variational principle. The following is then a third solution of this problem, and care will be taken to compare the FTRAN results with the results obtained by Reed[53] and Pitkaranta[26]. Also, a solution of this problem has been obtained with the ONETRAN[23] code, which allows a direct comparison of the treatment of spatial discontinuities by FTRAN and ONETRAN.

The physical problem consists of 4 regions (see Figure 18)--a strong absorber in region 1, a moderate absorber in region 2, a void in region 3, and a predominantly scattering medium in region 4. Region 1 also contains a strong isotropic source of neutrons and part of region 4 contains a weaker isotropic source. Reflecting boundary conditions are imposed on the left boundary and vacuum boundary conditions on the right.

The initial attempt to solve the problem employed continuous linear finite elements in space and angle. The mesh consisted of 40 spatial intervals and 8 angular intervals, which is consistent with the mesh used by Reed[53] (S_8, 40 spatial intervals) and Pitkaranta[26]. The results

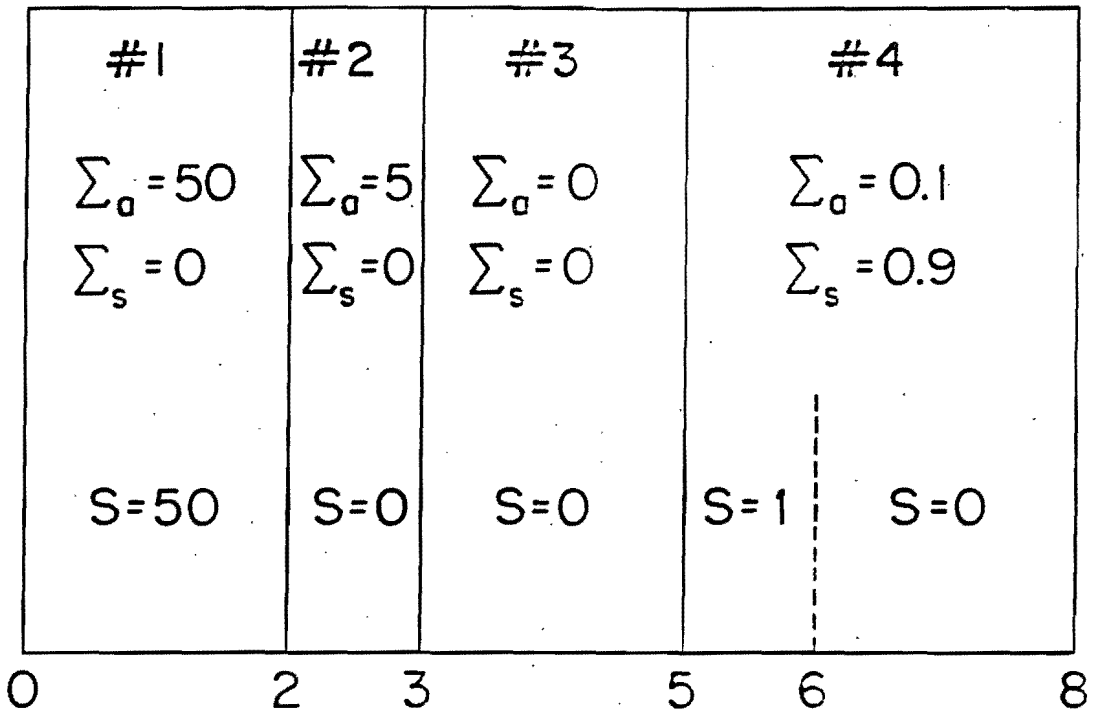


FIGURE 18. Geometry for Heterogeneous Problem

were extremely poor, as illustrated in Figure 19, where an extremely large scaled plot had to be used to contain the huge oscillations. Attempts to improve these results by using higher order finite elements in both space and angle did not result in any improvement of these results. (Note that the scalar flux has been plotted, and generally an integrated quantity will smooth out oscillations.)

Pitkaranta[26] also observed significant oscillations in the scalar flux when a uniform spatial mesh was used, although his oscillations were considerably smaller in magnitude than the oscillations in Figure 19. However, Pitkaranta noted that refining the mesh on either side of the vacuum region resulted in a considerable decrease in the magnitude of the oscillations, due to the fact that the rapidly varying flux in these regions could be approximated better.

Therefore, following the lead of Pitkaranta[26], this mesh refinement was made (adding nodes at x = 2.9, 2.95, 2.975, 2.99, 2.995, 5.005, 5.01, 5.025, 5.05, and 5.1) and FTRAN was run again with the continuous linear elements in space and angle. The results, which are also plotted in Figure 19, still exhibit gross oscillations and do not represent an improvement in the solution. Experimenting with higher order elements in space and angle did not improve these results.

The next attempt was to use discontinuous angular elements (at μ=0) although continuous spatial elements were used.* The result was a significant improvement--the oscillations, though still objectionable, were orders of magnitude less than before, as can be seen in Figure 20, which

---

* At the time this problem was being investigated, the discontinuous spatial elements had not been considered, let alone incorporated into FTRAN.

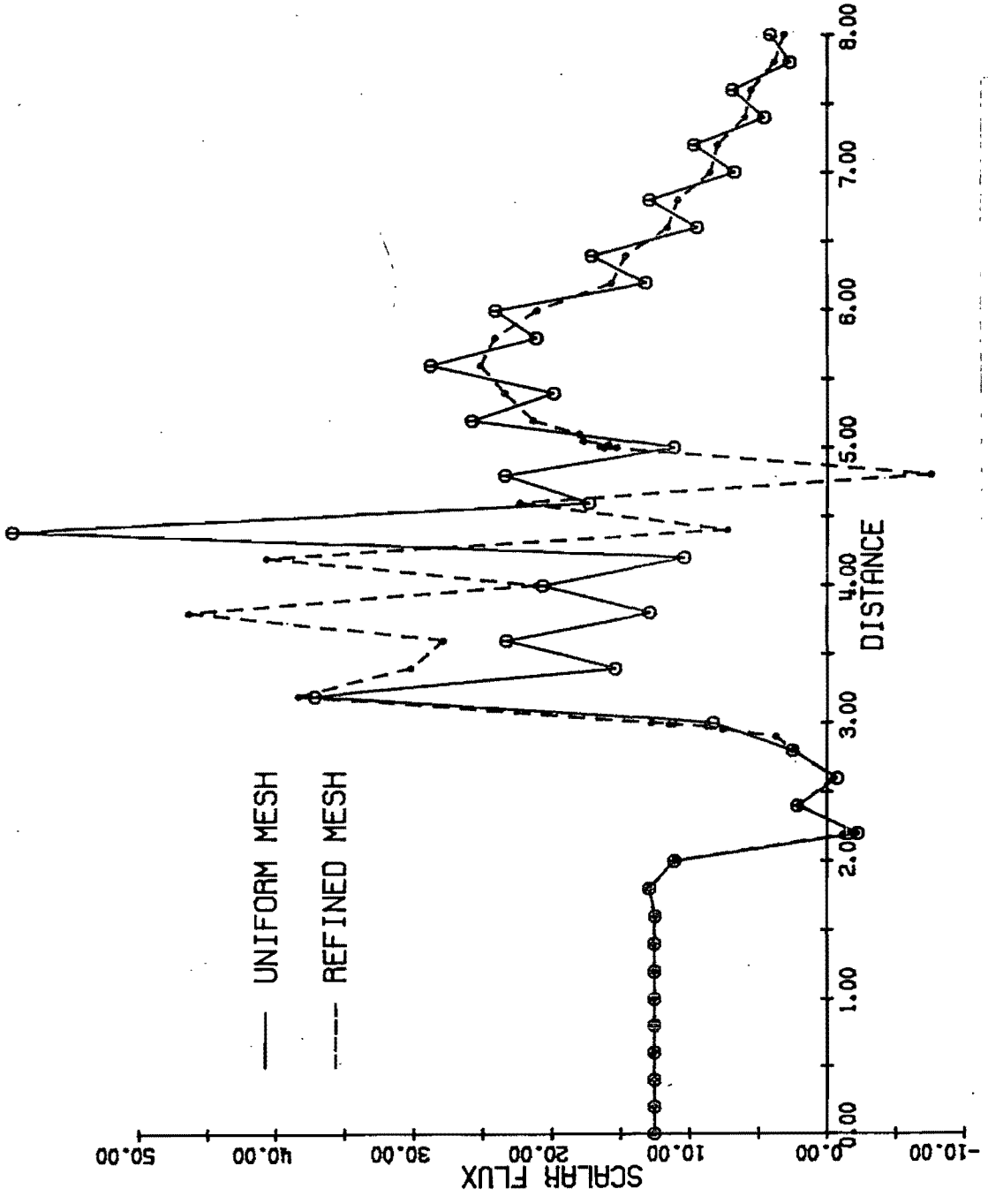


Figure 19. Scalar Flux (Continuous Elements)

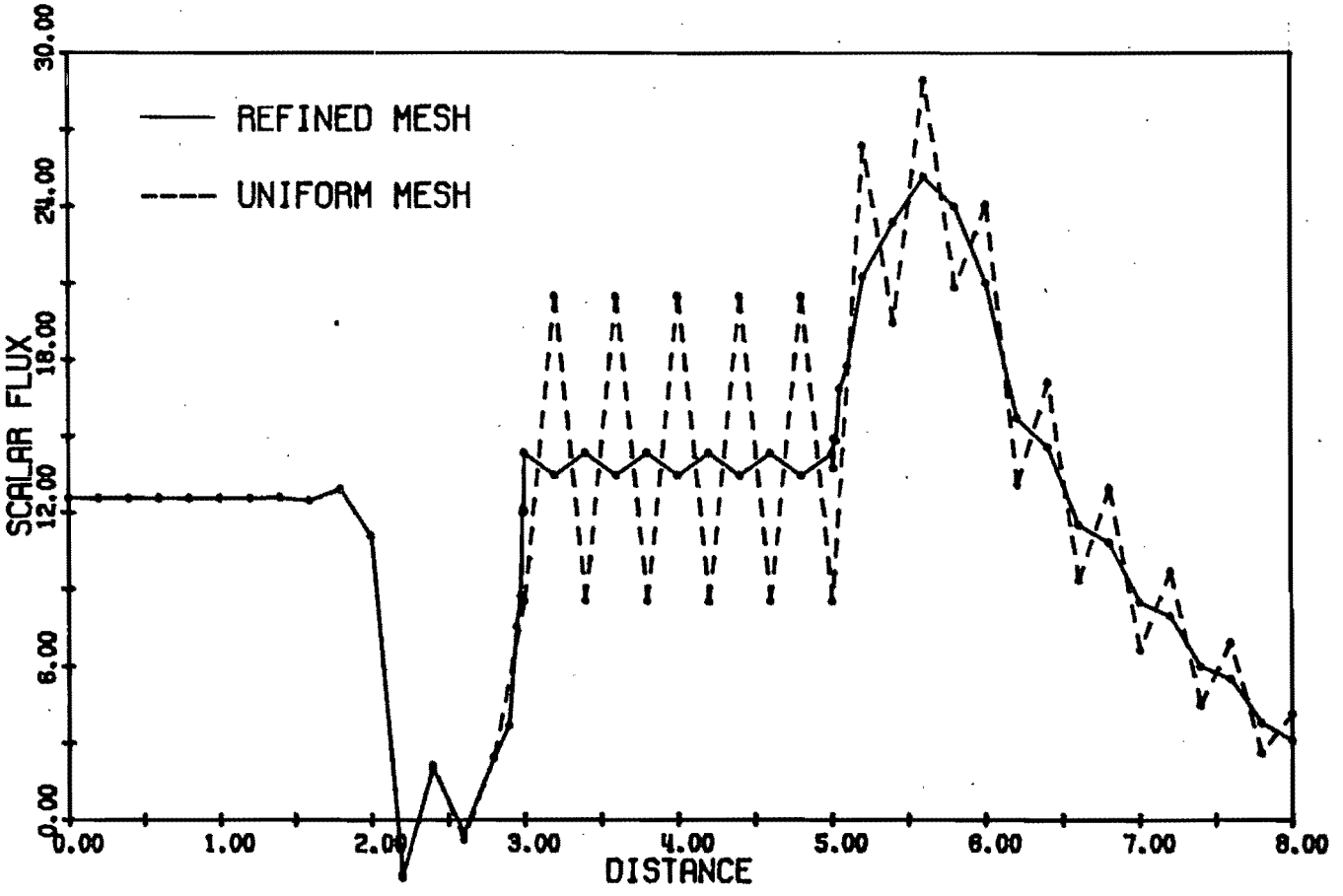


Figure 20. Scalar Flux (Discontinuous μ Elements)

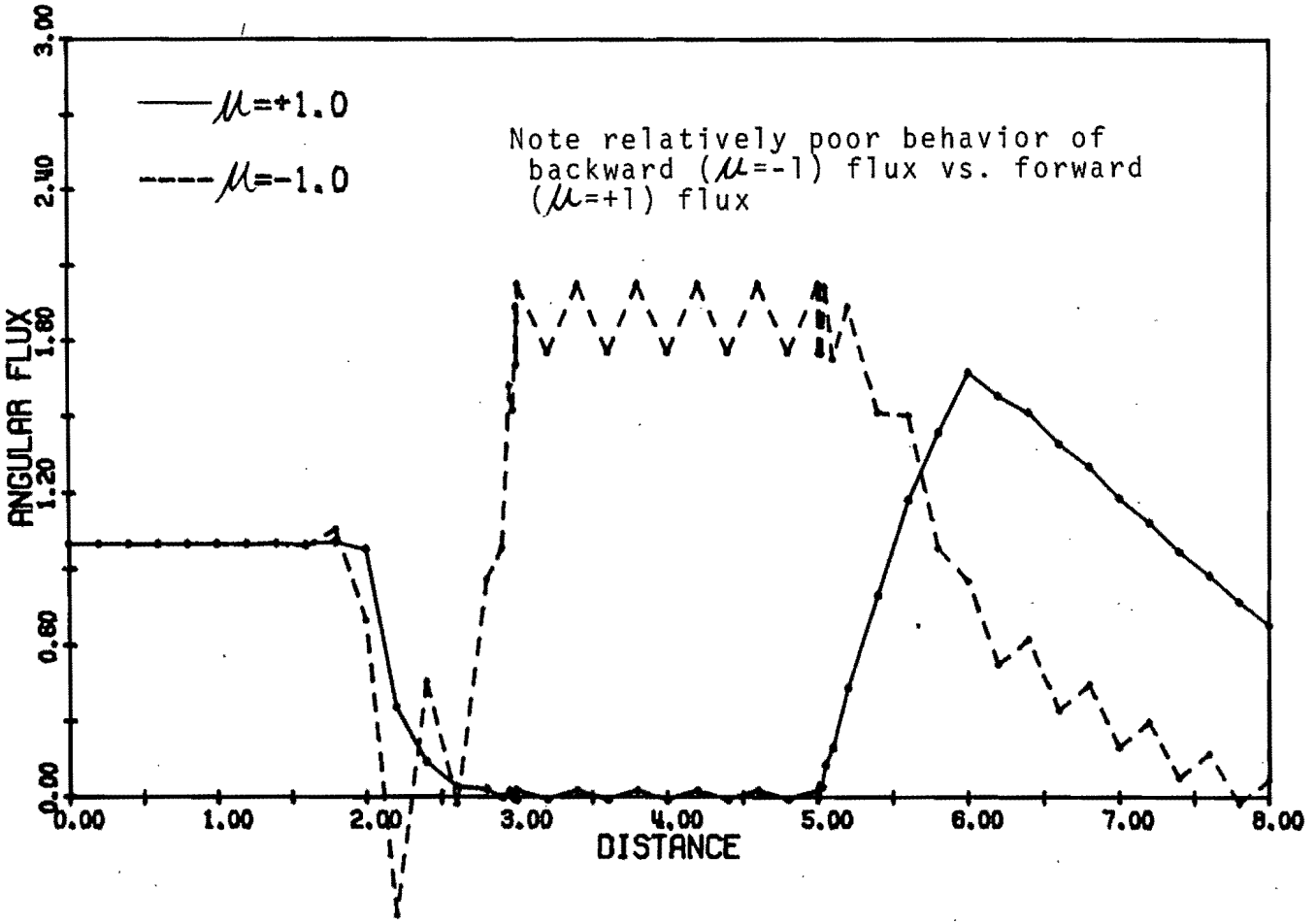


Figure 21. Angular Fluxes (Discontinuous μ Elements)

is drawn on a different scale than the two preceding plots. Now when the minor mesh refinement suggested by Pitkaranta was employed, as discussed above, the results were even better, as indicated in Figure 20. The magnitude of the oscillations was considerably reduced, similar to that reported by Pitkaranta. However, the oscillations on the void are still noticeable and the disturbing behavior of the flux at the interface at x = 2 is still present. In fact, a persual of Figures 19 and 20 indicates that this behavior of the flux near x = 2 was not affected by any of the above changes.

It is interesting to observe that the scalar flux oscillations and negative values are due to oscillations in the backward angular fluxes ($\mu < 0$) because the forward angular fluxes are quite well behaved. Figure 21, which clearly shows the relatively poor behavior of the μ=-1 angular flux compared with the μ= +1 flux, illustrates this remark. This particular point will be discussed in more detail below when the physical discontinuities contained within this problem are discussed.

In an attempt to improve the results even further, especially the behavior of the flux near the interface at x = 2, where the scalar flux becomes negative, FTRAN was modified to allow treatment of near-discontinuities in the angular flux at arbitrary spatial positions. The details on this modification have been discussed previously in Chapter V, but the key point is that the principal effect of this change is to allow continuity of the flux across the interface to be imposed as a natural interface condition in the direction of neutron travel. FTRAN was applied to the problem with the original uniform mesh, allowing discontinuities at x = 2, 3, 5, and 6, which are the interfaces for

the material and source discontinuities. The results indicated a substantial improvement in the solution, with complete elimination of the oscillations in the void and elimination of the negative flux behavior at $x \gtrsim 2$ (the scalar flux results are plotted in Figure 22 and a few angular flux profiles are plotted in Figures 23 and 24).

As the above discussion indicates, the use of discontinuous spatial and angular finite elements results in a vast improvement in the numerical results for the solution of the heterogeneous problem. As noted, this improvement was seen primarily in two regions--the void and the interface at x = 2. The numerical phenomena in these two regions are unrelated, although the use of the discontinuous spatial elements resulted in improvement in each region. The following discussion examines this point in more detail.

First we note that the backward ($\mu < 0$) angular flux exhibits a near-discontinuity at x = 2 because its behavior to the left is determined almost entirely by the source in region 1 while its behavior to the right is determined almost entirely by the source in region 4. Consequently the backward flux is nearly zero for $x \gtrsim 2$ due to absorption in region 2 and is a constant (unity) to within a few mean free paths (mfp) to the left of x = 2. Since the mesh spacing in region 1 is 10 mfp (due to the large absorption), the numerical solution cannot resolve the rapid variation of the flux near x = 2 and the net result of imposing continuity on the backward flux at x = 2 is to force the outgoing flux of region 2 to meet a boundary condition (unity) at x = 2 which is a non-physical condition. Based on these arguments, the forward flux at x = 5 will exhibit this effect although it is reduced somewhat by the

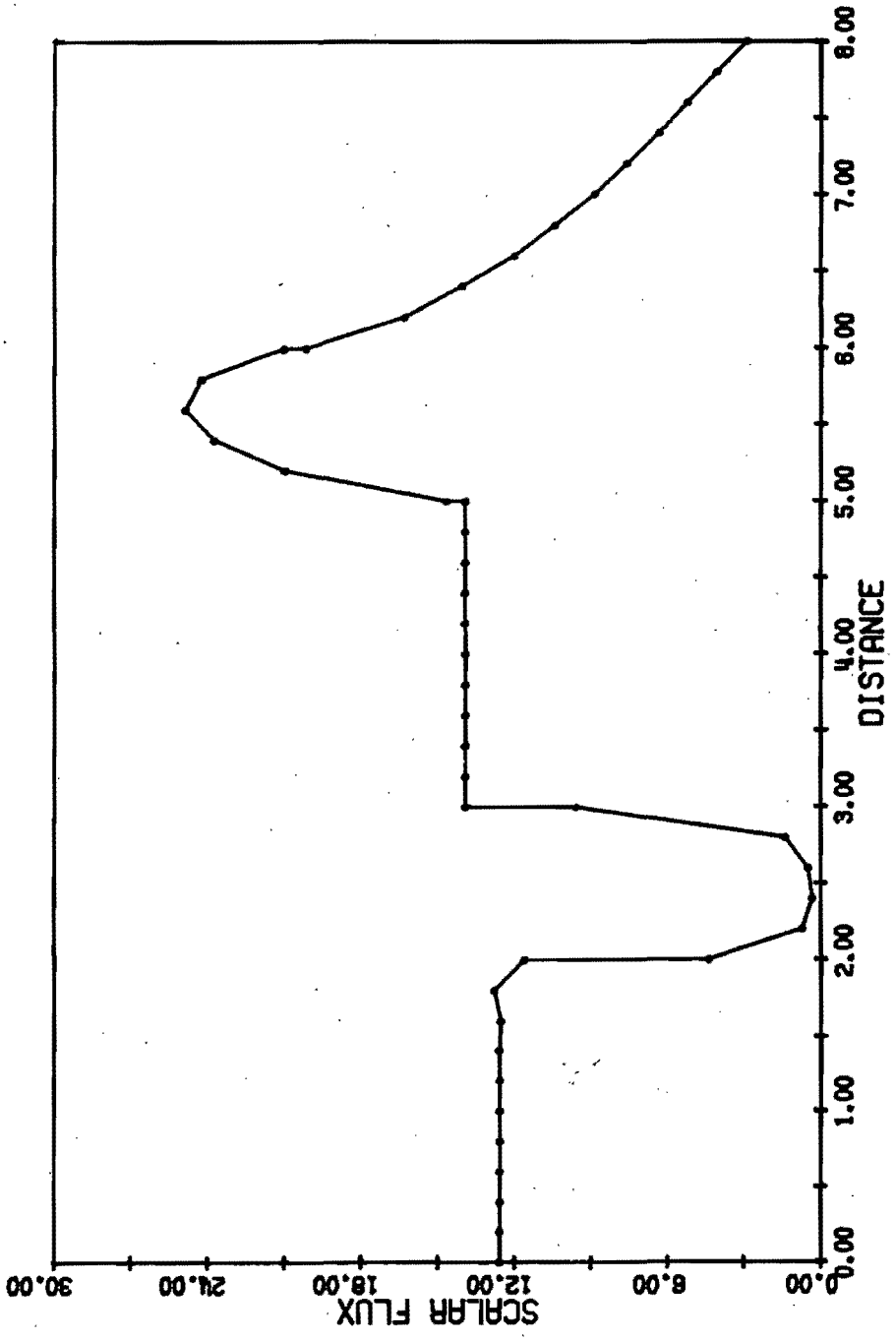


Figure 22. Scalar Flux (Discontinuous X and μ Elements)

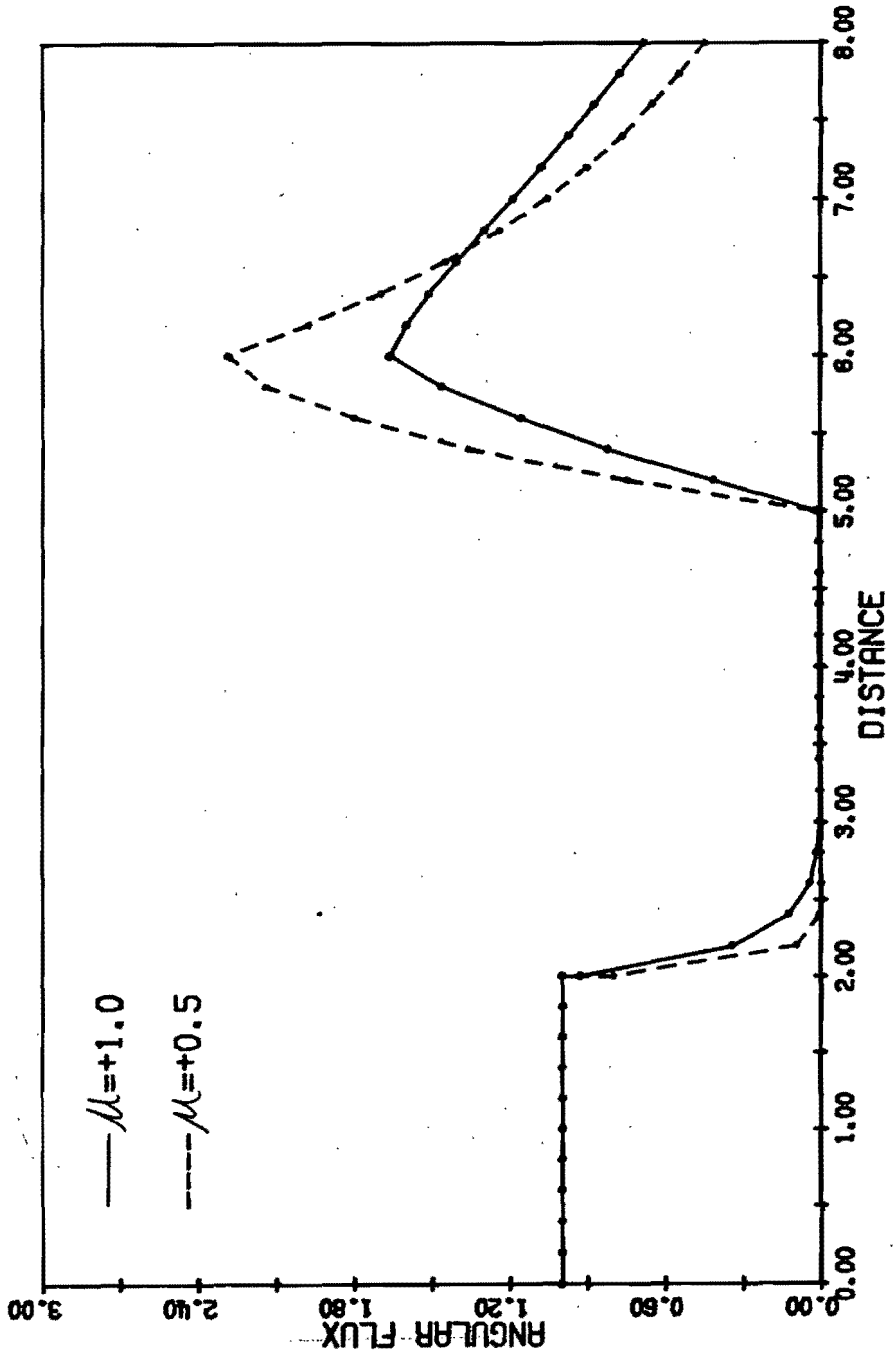


Figure 23. Angular Fluxes (Discontinuous X and μ Elements)

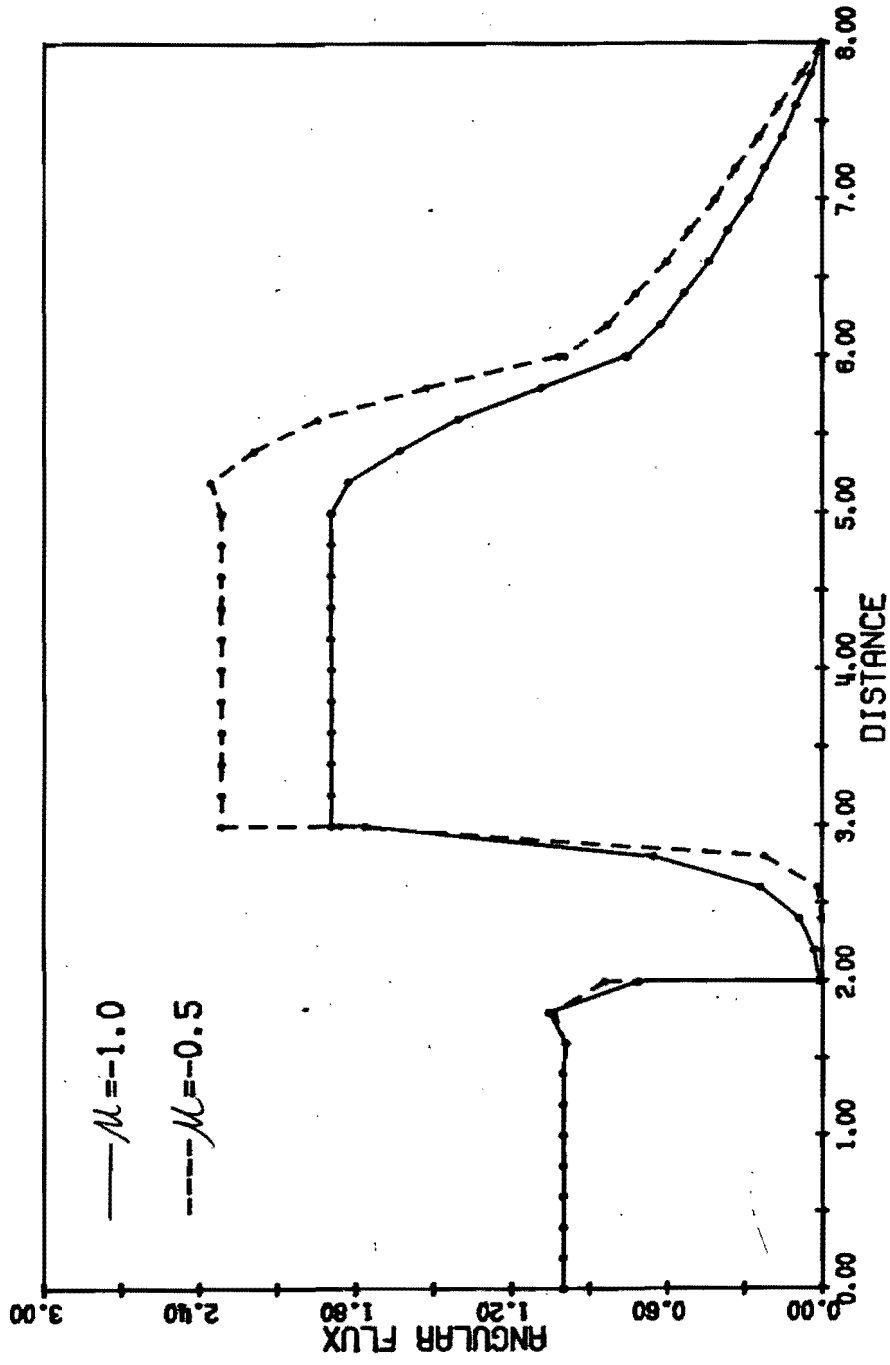


Figure 24. Angular Fluxes (Discontinuous X and μ Elements)

relatively finer mesh (in mfp) in region 3. Also, the backward flux at x = 6 is similarly affected, but to a much lesser extent due to the substantial amount of scattering which couples the forward and backward fluxes.

Thus the use of discontinuous spatial elements should relieve the solution of meeting the restrictive conditions at x = 2 for the outgoing flux of region 2 and the other interfaces as noted above, because now this flux may be discontinuous in order to approximate the actual solution. However, continuity in the direction of neutron travel is still a natural boundary condition to which the numerical solution will tend.

The coupling of one region to another in the direction of neutron flow is clearly observed at x = 2 for the $\mu = 1$ angular flux, which is plotted in Figure 23. The flux for x = 2+ tends to be the same as its value for x = 2-; however, this condition is a natural interface condition, not forced, and therefore continuity will in general not be achieved. For the $\mu = -1$ flux at this point, this natural interface condition results in an effective vacuum boundary condition because the backward flux at $x \gtrsim 2$ has traversed 5 mfp and is nearly zero. A look at Figure 24 illustrates the advantage of the natural interface condition for the $\mu = -1$ flux at this point, in that the solution is well-behaved compared with the negative oscillations obtained with continuous elements which force continuity.

The separate problem of the presence of the large oscillations in the void region, which are significantly damped by the mesh refinement on either side of the void, will now be considered. In order to examine the cause of these oscillations, we will need to consider the finite

difference relations which are equivalent to the linear finite element scheme on a uniform mesh. These equivalent finite difference relations were derived in Sec. III.F. For the vacuum region, where only the transport term is non-zero, the relation coupling one node to another is

$$\mu \frac{\partial Q}{\partial X} \approx \left(\frac{1}{6}\right)\left(\frac{\mu_j + \mu_{j+1}}{2}\right)\left(\frac{Q_{i+1,j+1} - Q_{i-1,j+1}}{2\Delta X}\right)$$

$$+ \left(\frac{2}{3}\right)\mu_j\left(\frac{Q_{i+1,j} - Q_{i-1,j}}{2\Delta X}\right)$$

$$+ \left(\frac{1}{6}\right)\left(\frac{\mu_j + \mu_{j-1}}{2}\right)\left(\frac{Q_{i+1,j-1} - Q_{i-1,j-1}}{2\Delta X}\right)$$

$$= 0$$

Note that i and j as used here do not correspond to the i and j as used in the global numbering for A_{ij}. This scheme is seen to be a weighted "leap-frog" scheme which has the well-known consequence[54] of perpetuating an initial error across the mesh. In particular, a flux difference between nodes immediately adjacent to the void will be propagated through the void in a regular (leap-frogging) manner. Thus a rapidly varying flux adjacent to the void will result in a large oscillation in the void. Therefore, the reason for the partial success of the mesh refinement discussed above is obvious--the difference between the nodal fluxes adjacent to the void is reduced, thereby decreasing the magnitude of the resultant oscillation in the void.

The key observation for this application is that the discontinuous elements decouple the void region and the adjacent regions except for the surface terms at each interface to account for transport of neutrons across the interface. Thus the natural interface conditions on each side of the vacuum region completely eliminate any trace of oscillations in the angular flux because the flux within the vacuum now depends only on the incoming flux at its boundaries, not on the angular fluxes adjacent to its boundaries. In this case the "leap-frog" scheme propagates only the incoming flux, rather than some combination of fluxes near the boundaries, across the void.

The absolute necessity of using discontinuous angular elements is clearly illustrated in Figure 25, which is a plot of the μ=0+ and μ=0- angular fluxes, which are quite different throughout most of the region. Clearly any numerical scheme which imposes continuity at μ=0 will fail, as illustrated by the earlier results with continuous elements.

There remains the numerical difficulty of the anticipation by the numerical solution of the effective vacuum boundary condition at x = 2 for the backward flux in region 1. Since the mesh spacing in region 1 is 10 mfp, the numerical solution cannot resolve this behavior, and its attempt to do so results in the "bump" near x = 1.7. One obvious remedy is to insert additional nodes near $x \lesssim 2$, which was done. As expected, when this minor mesh refinement was performed, the "bump" was eliminated.

The heterogeneous problem considered in this section was also solved with the 1-D discrete ordinates code ONETRAN[23], which employ discontinuous finite elements in space. Although the actual problem solved

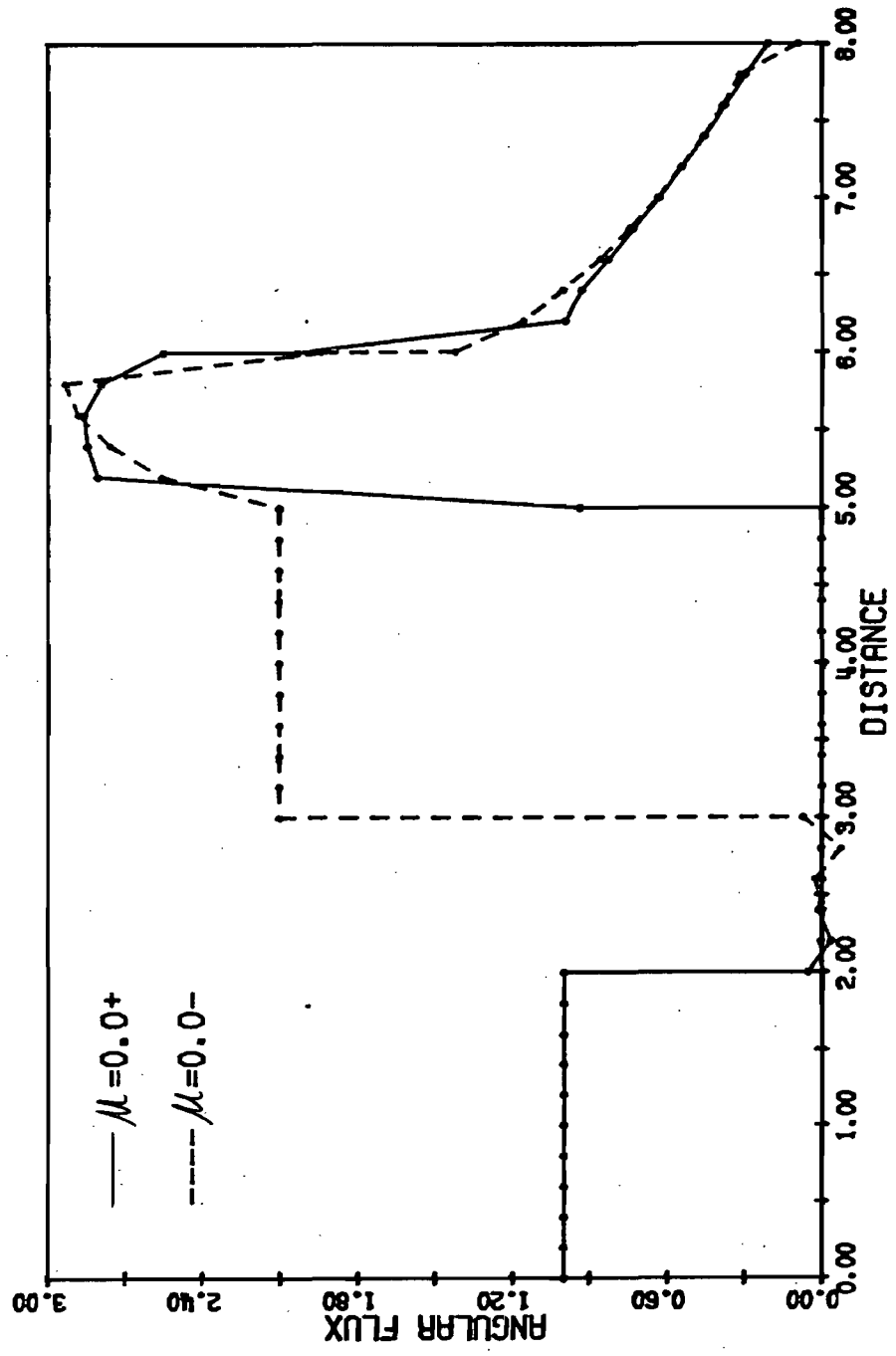


Figure 25. Angular Fluxes (Discontinuous X and μ Elements)

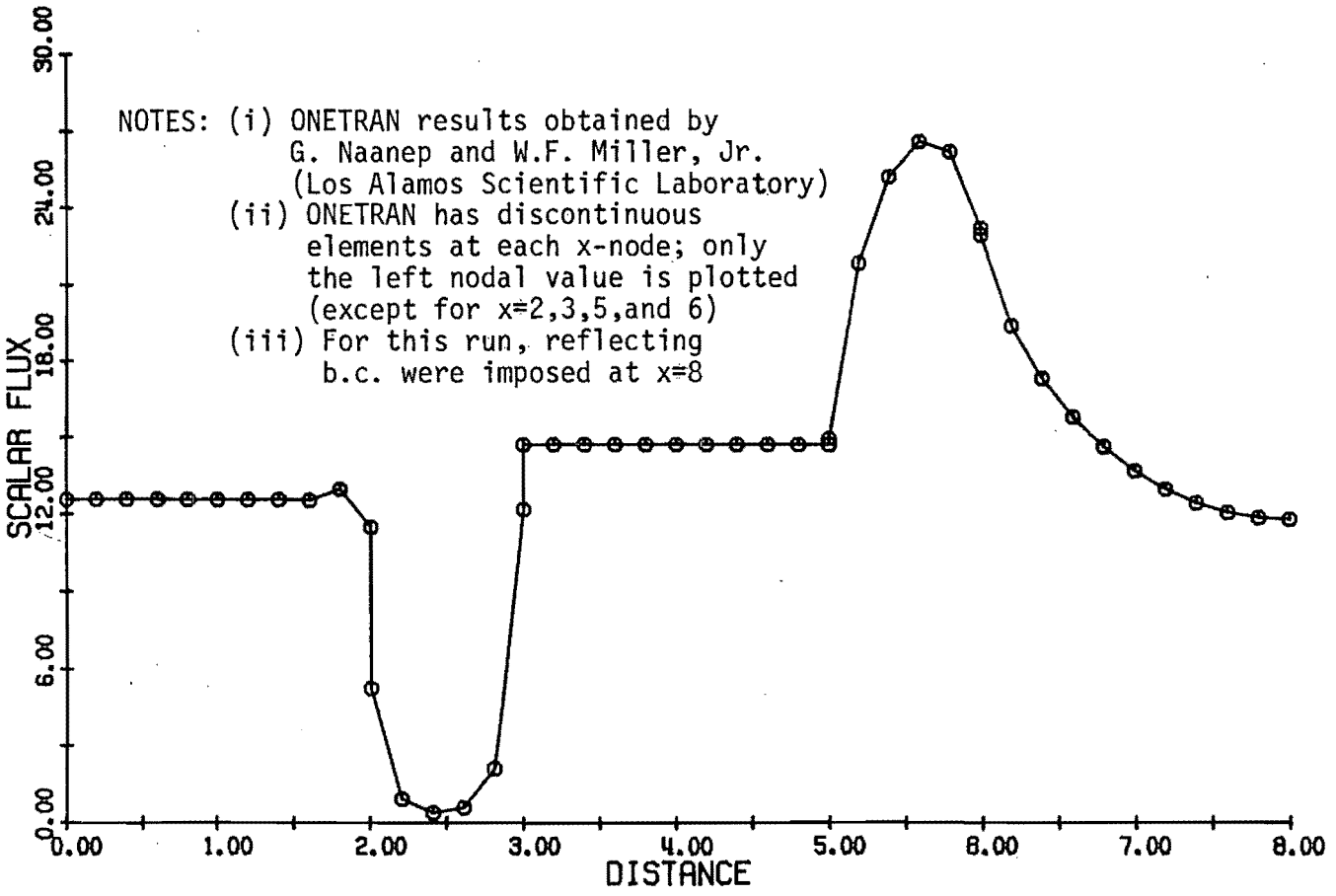


Figure 26. Scalar Flux (ONETRAN)

was slightly different because reflecting boundary conditions were imposed on the right boundary, the ONETRAN results may still be compared with the FTRAN results because the dependence of the solution at the interface at x = 2 and within the void are not affected. Figure 26 is a plot of the scalar flux results from ONETRAN, and it can be seen that they are nearly identical with the FTRAN results plotted in Figure 22, including the magnitude of the numerical discontinuity at x = 2, disregarding the changes due to the use of reflecting boundary conditions on the right boundary.

Thus the remarks made earlier in Chapter IV, concerning the equivalence of the FTRAN and ONETRAN approaches for incorporating spatial discontinuities, would appear to be substantiated by these results.

VIII.G. Time-Dependent Problems

This section contains two examples in time-dependent neutron transport--the decay of an equilibrium neutron distribution and the propagation of a neutron wave in a vacuum.

1. Decay of Equilibrium Distribution

The physical configuration for this example, which is illustrated in Figure 27, resembles an assembly of fuel, clad, and moderator. The equilibrium flux distribution, which is present for $t \leq 0$, is established by an isotropic source of neutrons in Region 1, which is a pure absorber. Region 2 is a void and Region 3 is a pure scatterer. Reflecting boundary conditions are imposed on the left boundary and vacuum boundary conditions are imposed on the right boundary.

At t = 0 the source in Region 1 is removed and the neutron distribution then decays to its new equilibrium distribution, which is zero.

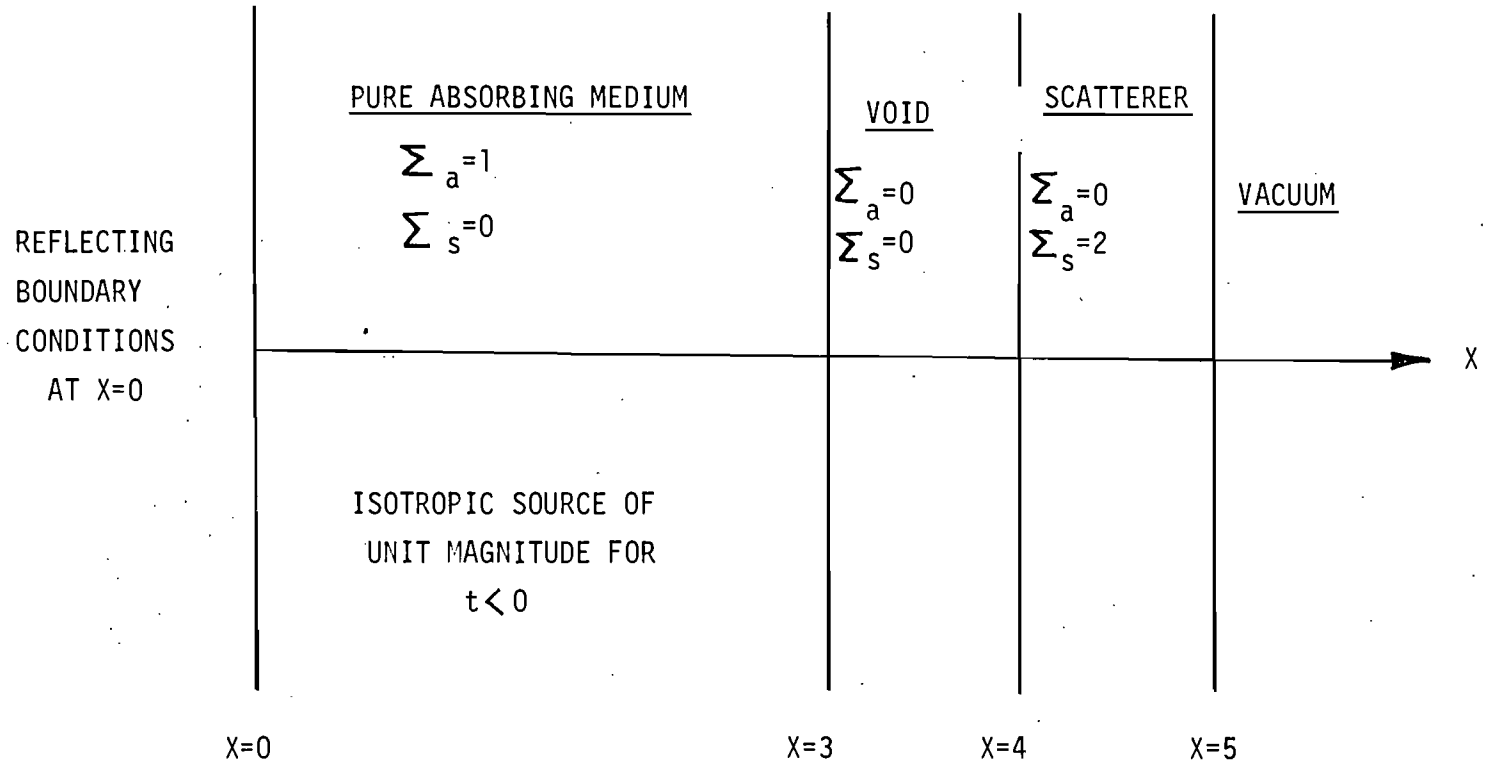


FIGURE 27. Geometry for Equilibrium Decay Problem

FTRAN was used to solve this problem with linear elements in space and angle. In addition, discontinuous angular elements at $\mu = 0$ were used and discontinuous spatial elements were employed at x = 3 and 4. The spatial and angular meshes were uniform with $\Delta x = .25$ and $\Delta\mu = .25$, which results in 23 spatial nodes and 10 angular nodes, or 230 nodes total.

Figure 28 is a plot of the resultant scalar flux distribution at various times during the transient and Figures 29-34 are plots of selected angular fluxes at the various times.

On the basis of the numerical results, the following remarks may be made. First, the dominant decay mechanism is the absorption in Region 1 although leakage from the right boundary also contributes to the loss of neutrons. Neutron wave propagation is clearly illustrated by the motion of the disturbance in the scalar flux (or angular fluxes) across the void, which reaches Region 3 at t = 1 second. Since the neutron velocity is 1 cm/sec and the void is 1.0 cm thick, this is the correct time. Also, the flux in Region 3 remains a constant until this disturbance reaches it, which is also predicted because there is no absorption in Region 3 and its boundary conditions remain constant until the disturbance arrives.

Figures 29-34 illustrate various angular modes of the decaying neutron distribution and the wave motion across the void is clearly seen. Also, it is observed that the numerical solutions are fairly well behaved, even at later times where oscillations are prone to occur. As will be noted in the next example, oscillations in time-dependent problems occur quite readily, especially at long times, and one needs to be careful with the choice of the angular mesh.

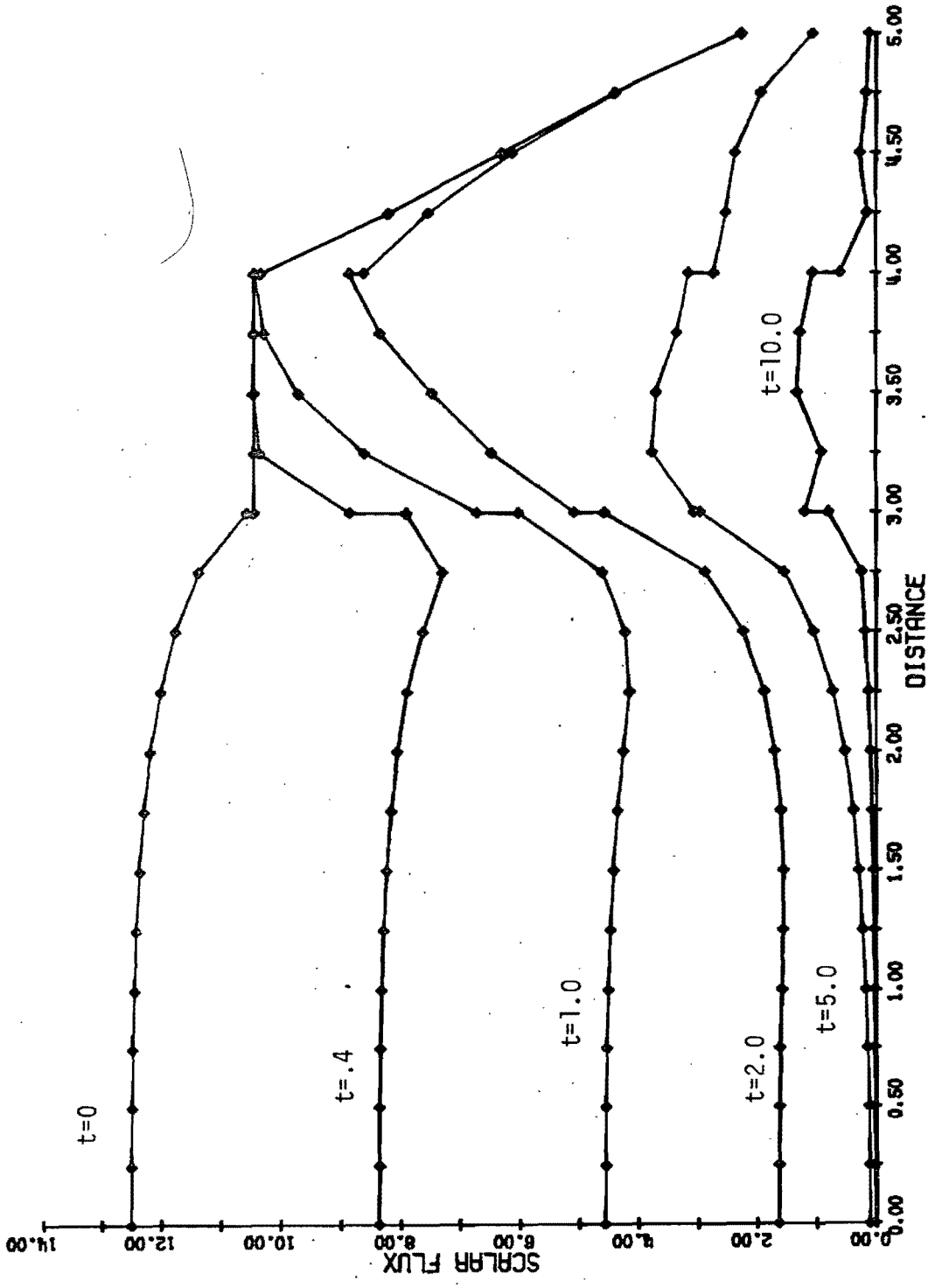


FIGURE 28. Decay of Scalar Flux

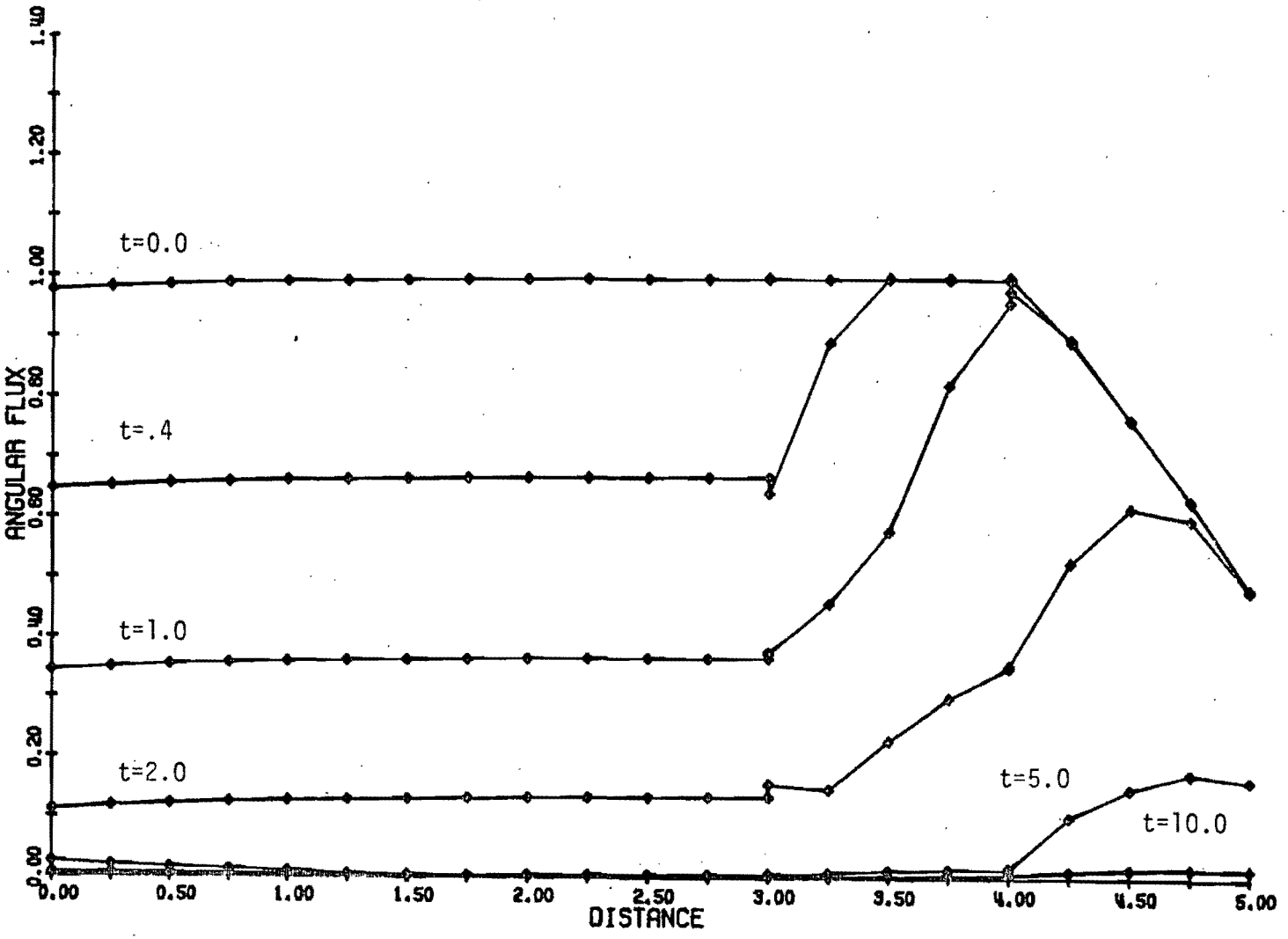


FIGURE 29. Decay of Angular Flux (μ=+1.0)

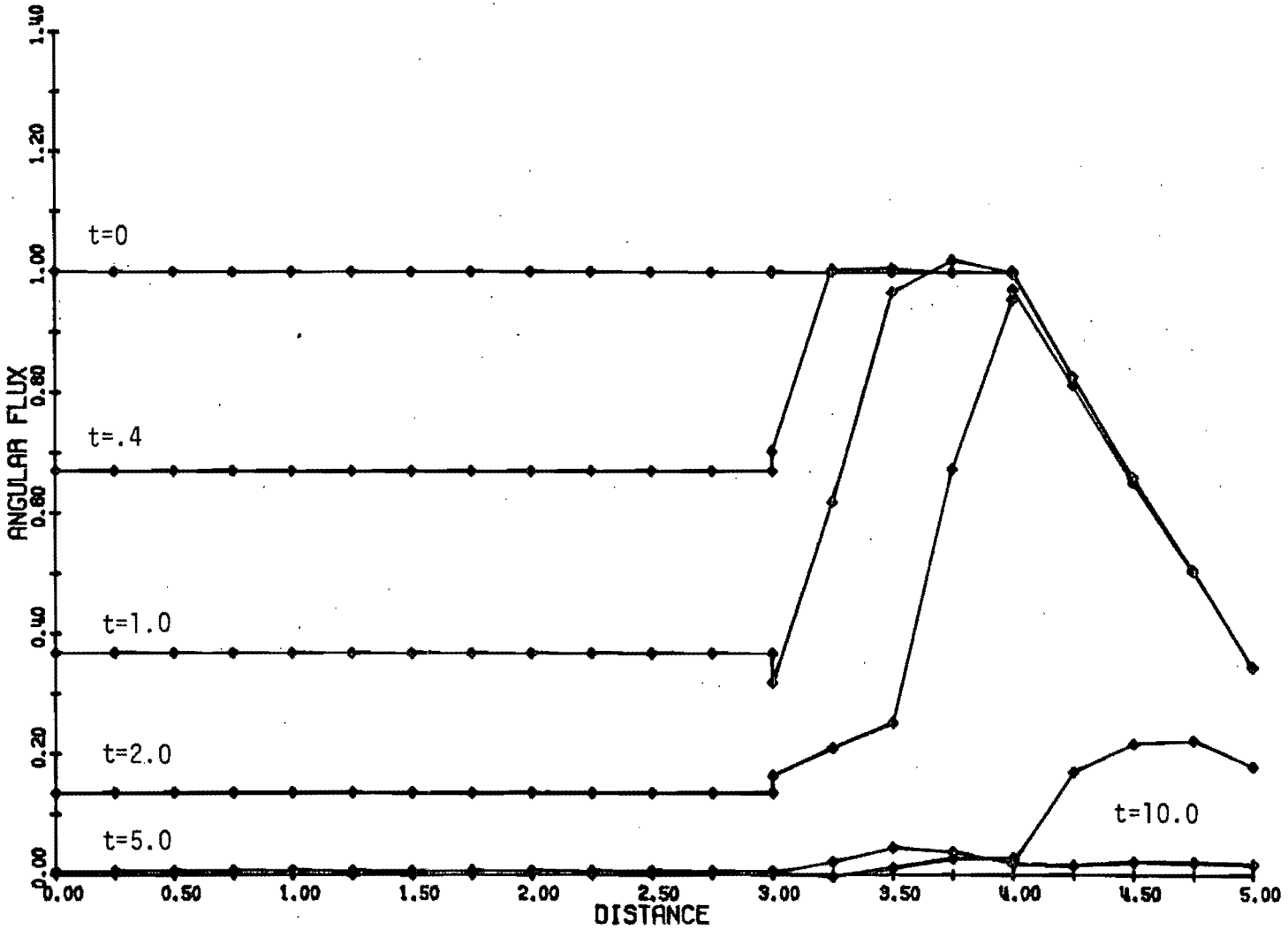


FIGURE 30. Decay of Angular Flux (μ=+.5)

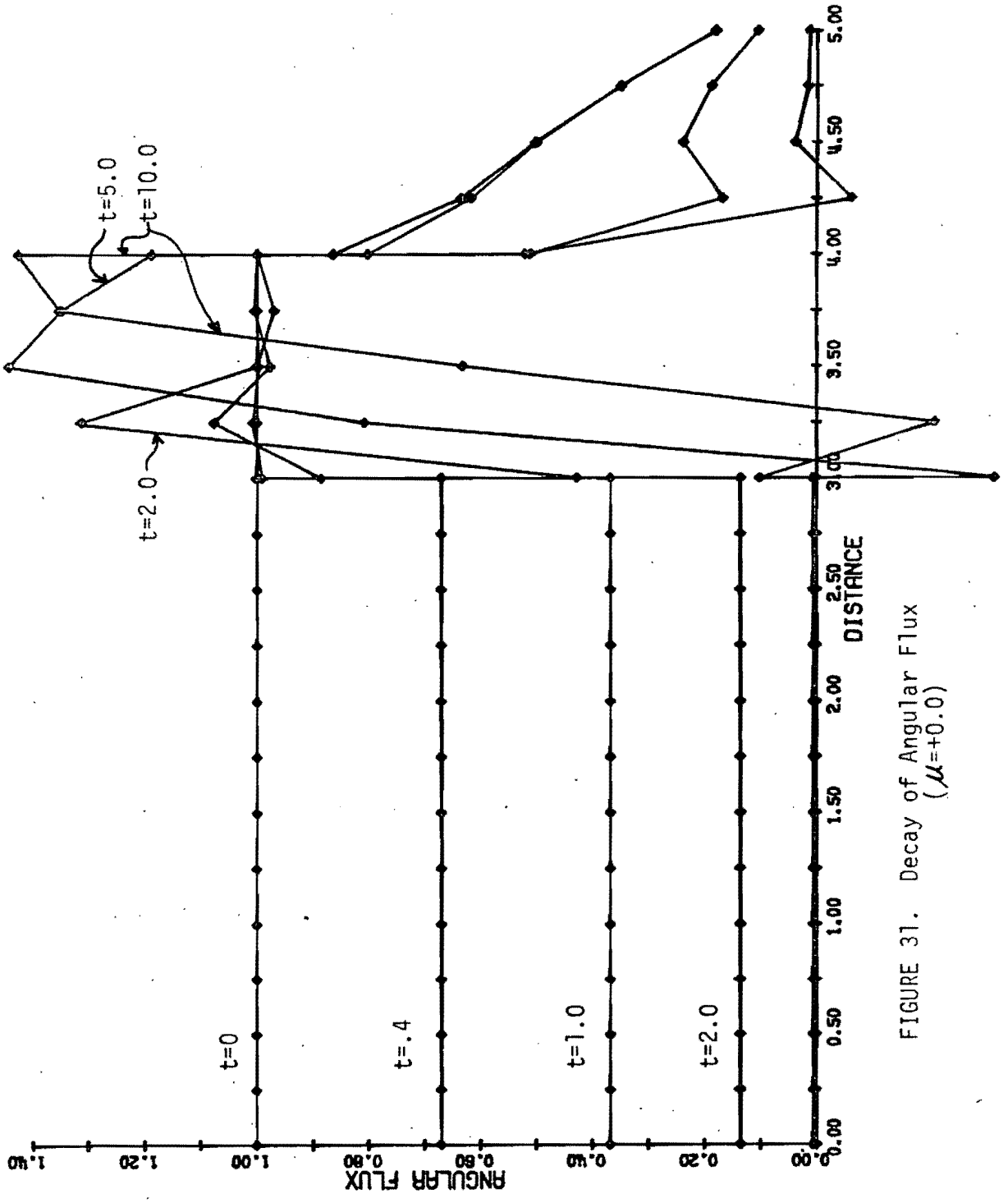


FIGURE 31. Decay of Angular Flux (μ=+0.0)

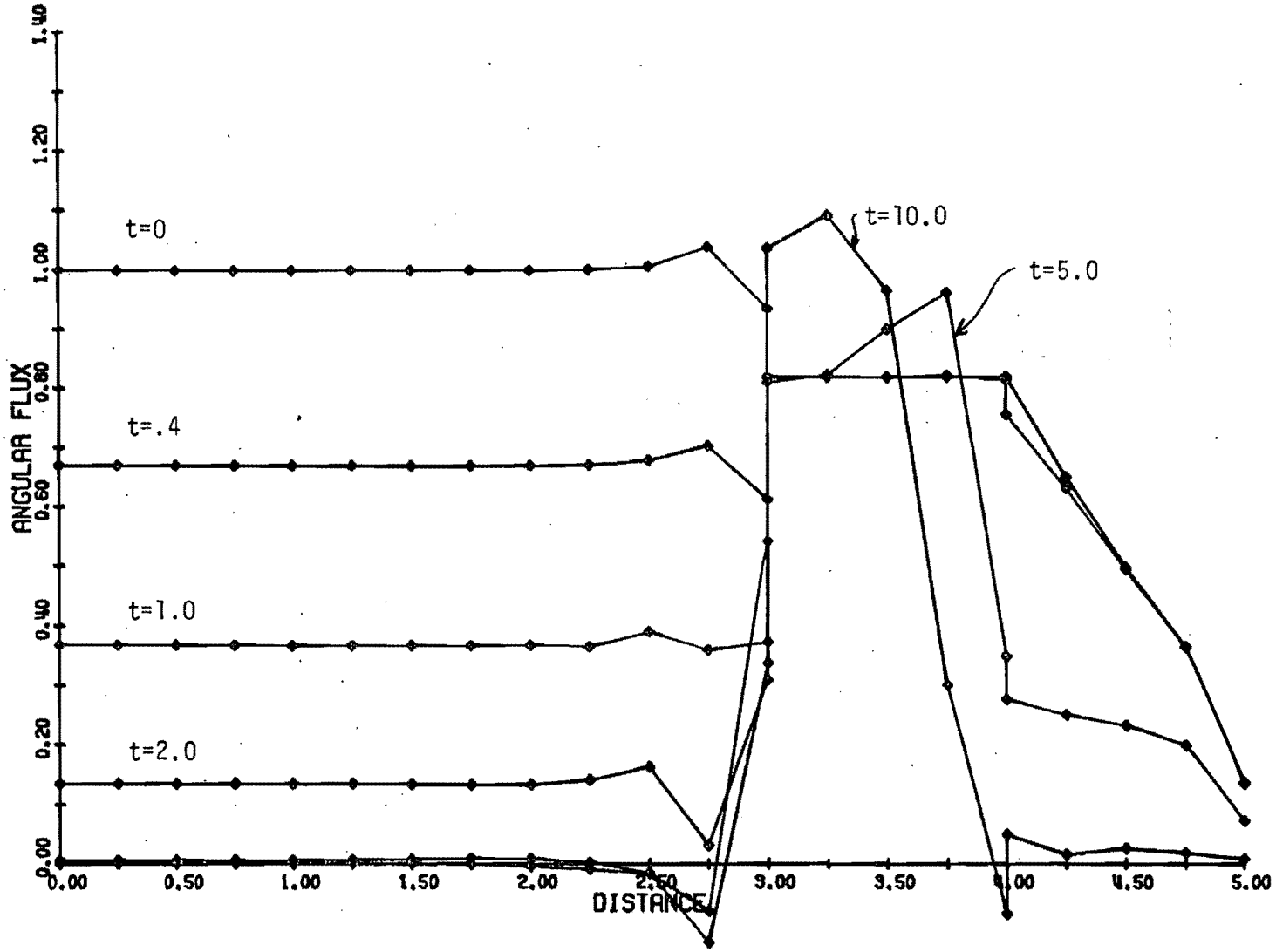


FIGURE 32. Decay of Angular Flux (μ=-0.0)

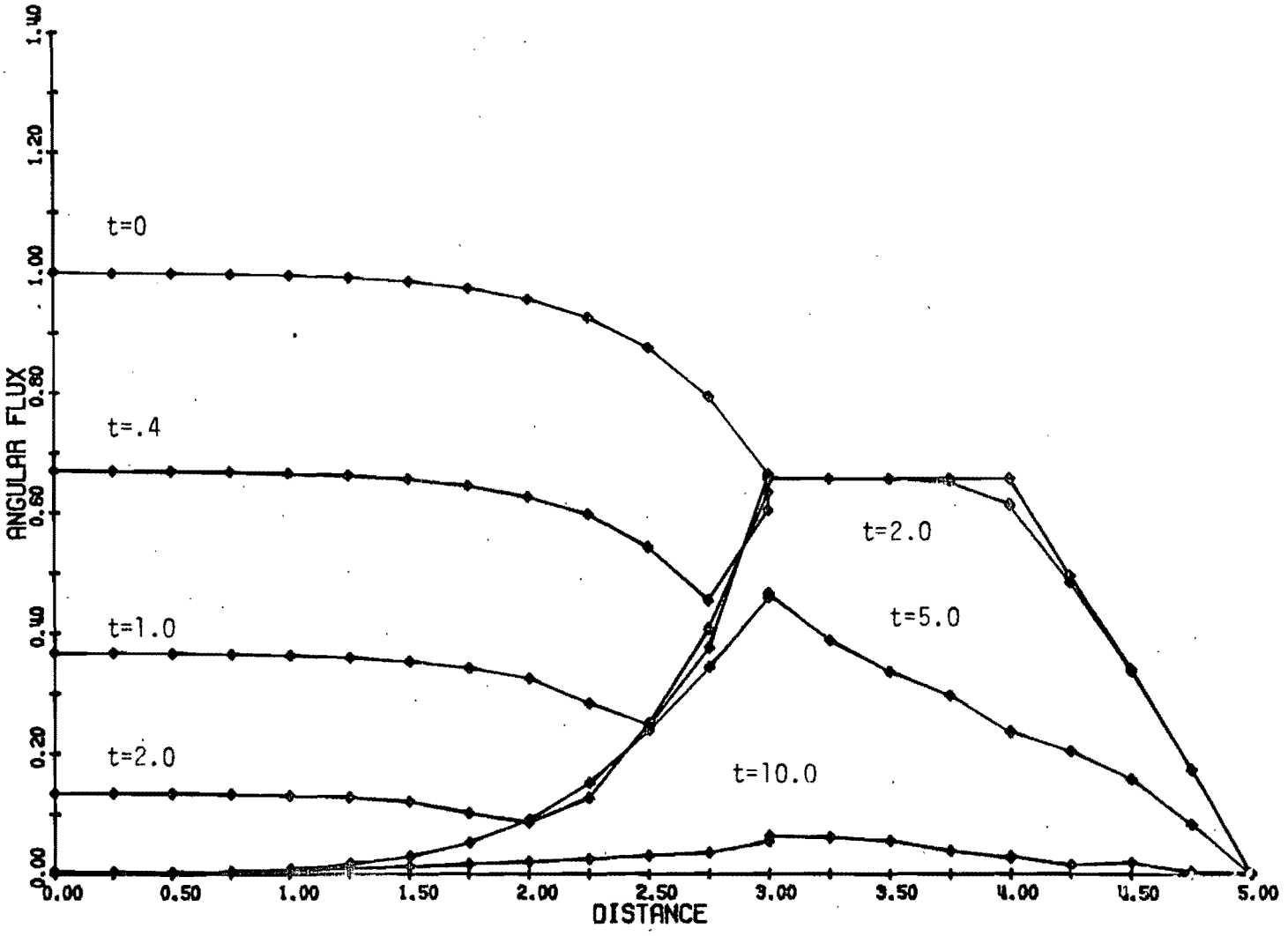


FIGURE 33. Decay of Angular Flux (μ=-.5)

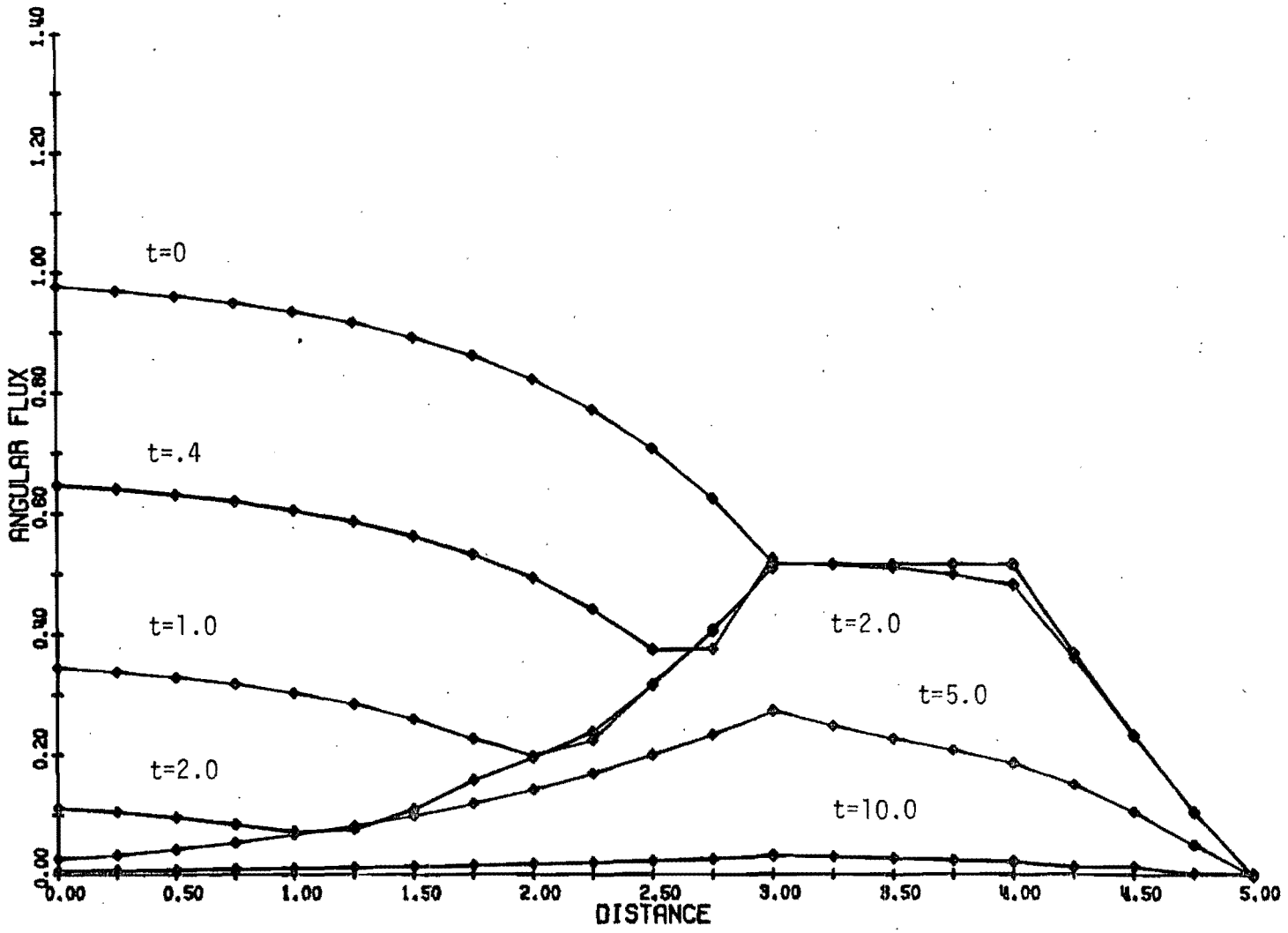


FIGURE 34. Decay of Angular Flux (μ=-1.0)

2. Pulse Propagation in a Vacuum

The simple problem of the propagation of a neutron pulse in a vacuum will now be considered. This problem will illustrate the phenomenon of angular mode separation with the advancement of time, and the numerical difficulties that are subsequently encountered.

The physical configuration is a Gaussian (in space) distribution of neutrons traveling in the +x direction. The initial angular distribution is isotropic in the forward ($\mu > 0$) directions and zero in the backward ($\mu < 0$) directions. The specific initial angular flux is

$$Q(x,\mu,0) = Q_0(x,\mu) = \begin{cases} e^{-(x-1)^2/.25}, & \mu \geq 0 \\ 0, & \mu \leq 0 \end{cases} \qquad (162)$$

which is a Gaussian centered about x = 1 cm with a half-width of .5 cm. The neutron velocity is 1 cm/sec and the spatial domain considered is $0 \leq x \leq 5$. In order to resolve the pulse, the spatial mesh was chosen to be

$$\Delta x = .25$$

and the time step was .1 sec, which restricts the movement of the pulse to less than one mesh cell per time step. Although their use would appear redundant for this problem, vacuum boundary conditions are imposed at x = 0 and x = 5 cm.

Since the initial distribution is not continuous at $\mu = 0$, discontinuous angular elements were used. Three different angular meshes were used, keeping the spatial mesh and element type (linear) constant. The angular meshes were uniform in the forward directions ($\mu > 0$) with

the only nodes in the backward directions ($\mu < 0$) being the nodes at $\mu = -1.0$ and $\mu = 0.0-$. The angular flux should be identically zero for the $\mu < 0$ nodes since there is no initial distribution of neutrons with a velocity in the -x direction and no mechanism (i.e., no scattering) to transfer neutrons from forward directions to backward directions.

The specific angular meshes chosen were (for $\mu > 0$) $\Delta\mu = 1.0$, .25, and .1. The results offer convincing evidence that the angular mesh has a dominant effect on obtaining acceptable results. That is, with the extremely coarse angular mesh ($\Delta\mu = 1.0$), the scalar flux, which is plotted in Figure 35, and the angular fluxes at $\mu = 1.0$ and 0.0+, which are plotted in Figures 36 and 37, contain severe oscillations which are amplified as time advances. With the finer mesh of $\Delta\mu = .25$, the scalar flux (Figure 38) and angular fluxes (Figures 39-41) are somewhat better but also develop severe oscillations within a few seconds into the transient. However, as will be discussed in more detail shortly, the oscillations do not appear as soon as with the $\Delta\mu = 1.0$ case. The finest angular mesh examined, $\Delta\mu = .1$, gave good results for the scalar flux (Figure 42) and the angular fluxes (Figures 43-45) over the entire transient ($0 \leq t \leq 5$), although the forward peaked angular fluxes (μ=1.0) were beginning to develop negative oscillations toward the later times. However, the scalar flux behaves quite well, especially in comparison with Figures 35 and 38 for the coarser meshes, in that the oscillatory behavior cannot be observed for any time. Note that the $\mu = 0.0+$ flux (Figure 45) illustrates quite well the stationary behavior of this particular angular mode, which represents streaming perpendicular to the x-axis.

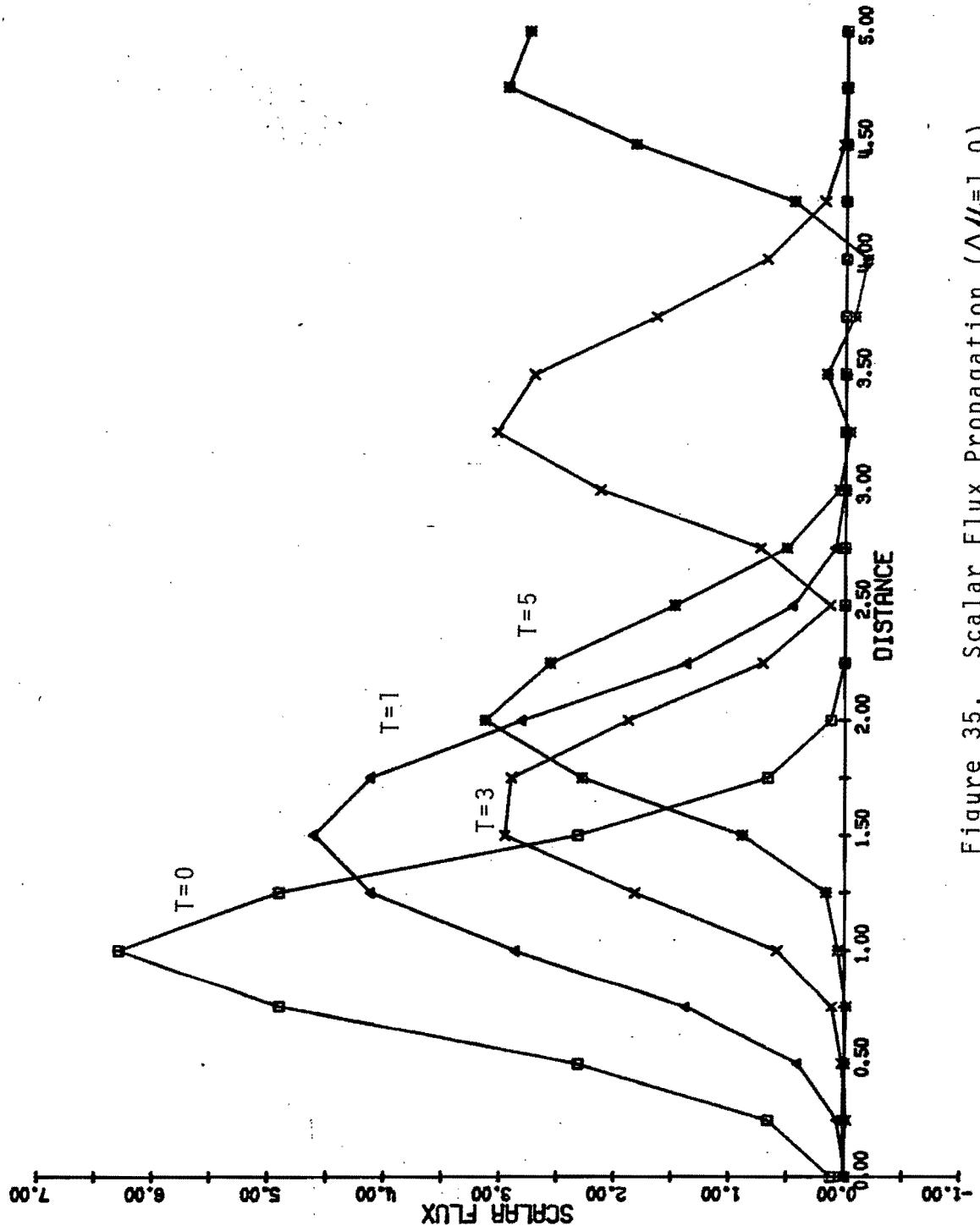


Figure 35. Scalar Flux Propagation ($\Delta\mu$=1.0)

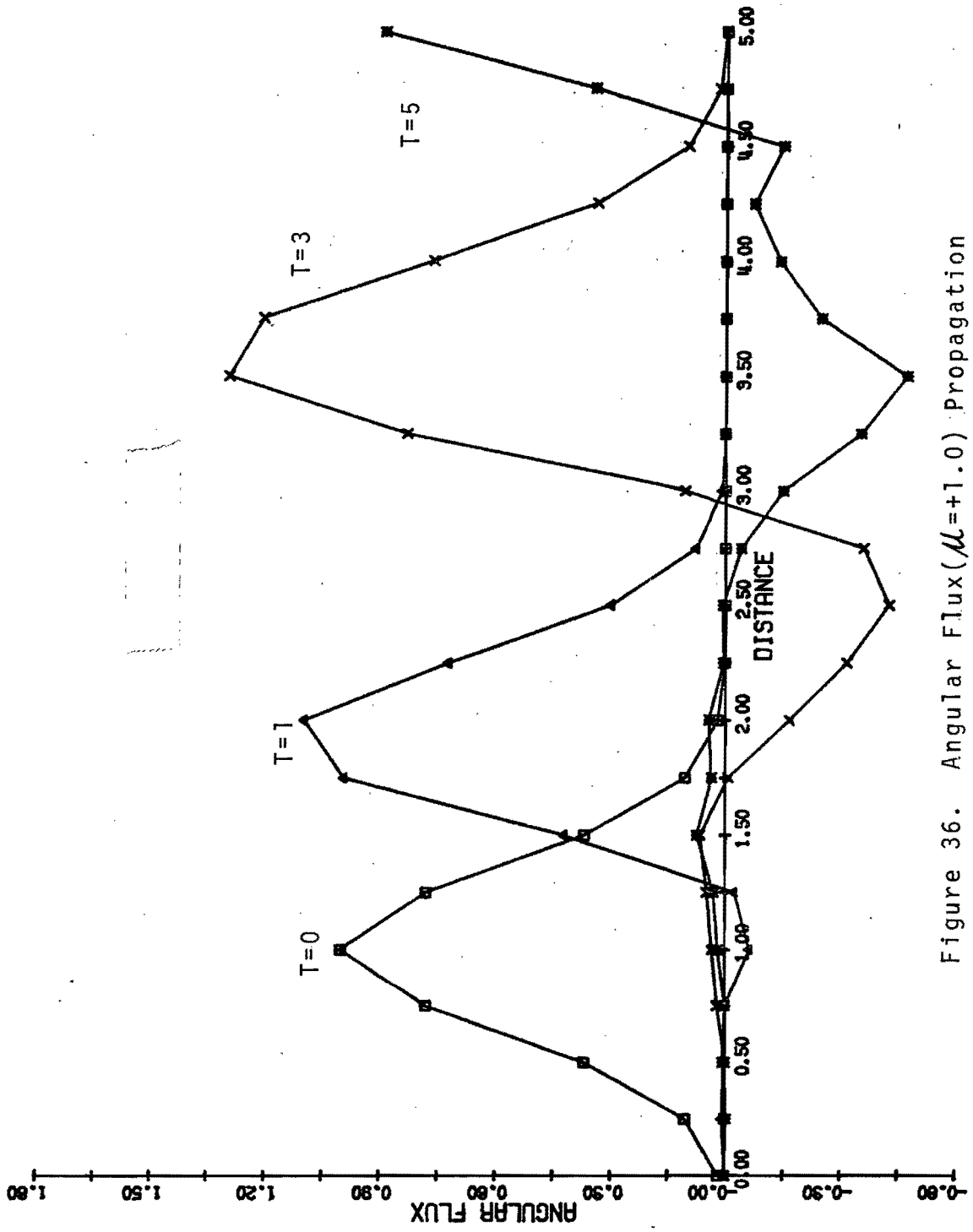


Figure 36. Angular Flux(μ=+1.0) Propagation

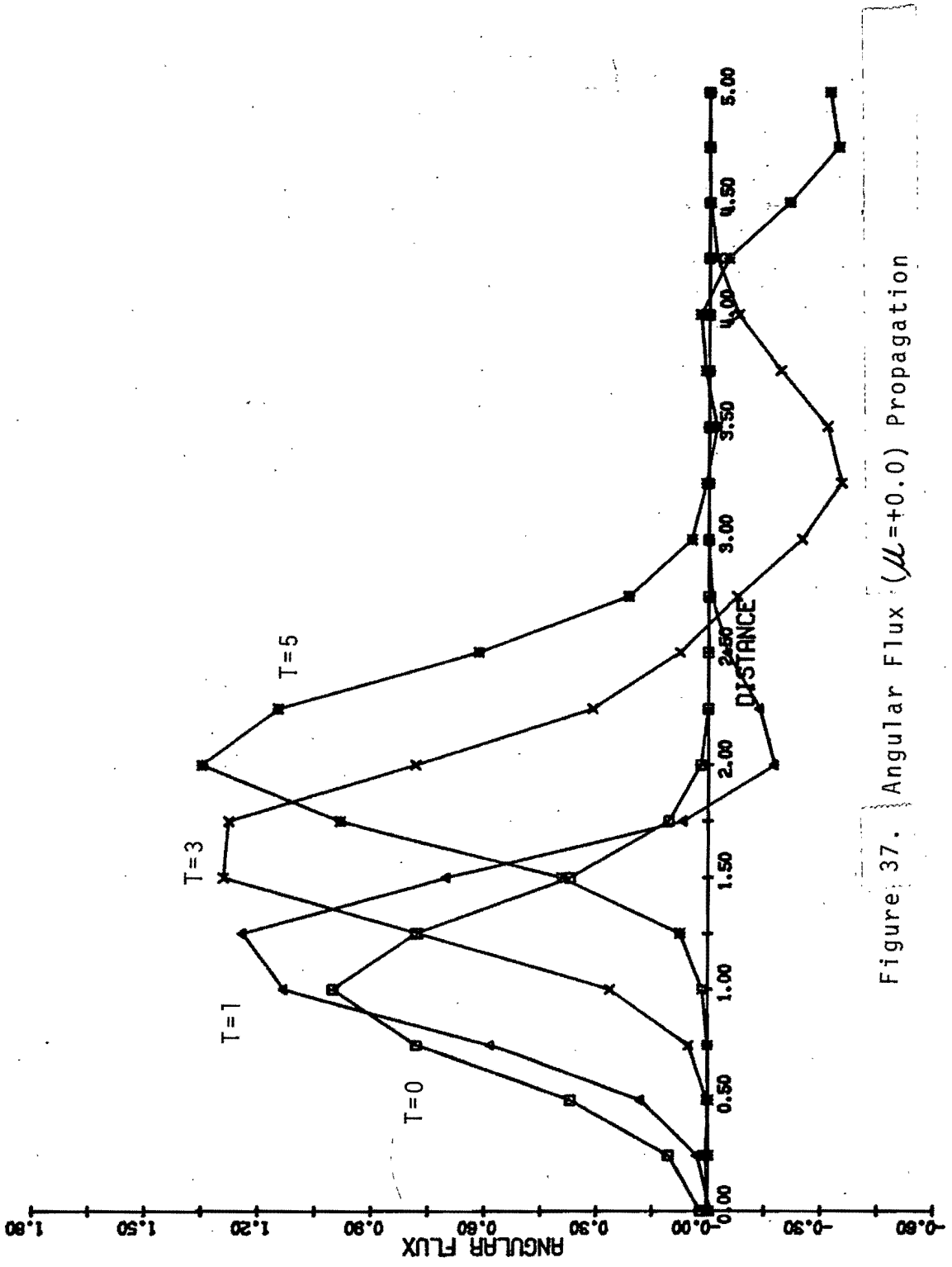


Figure 37. Angular Flux (μ=+0.0) Propagation

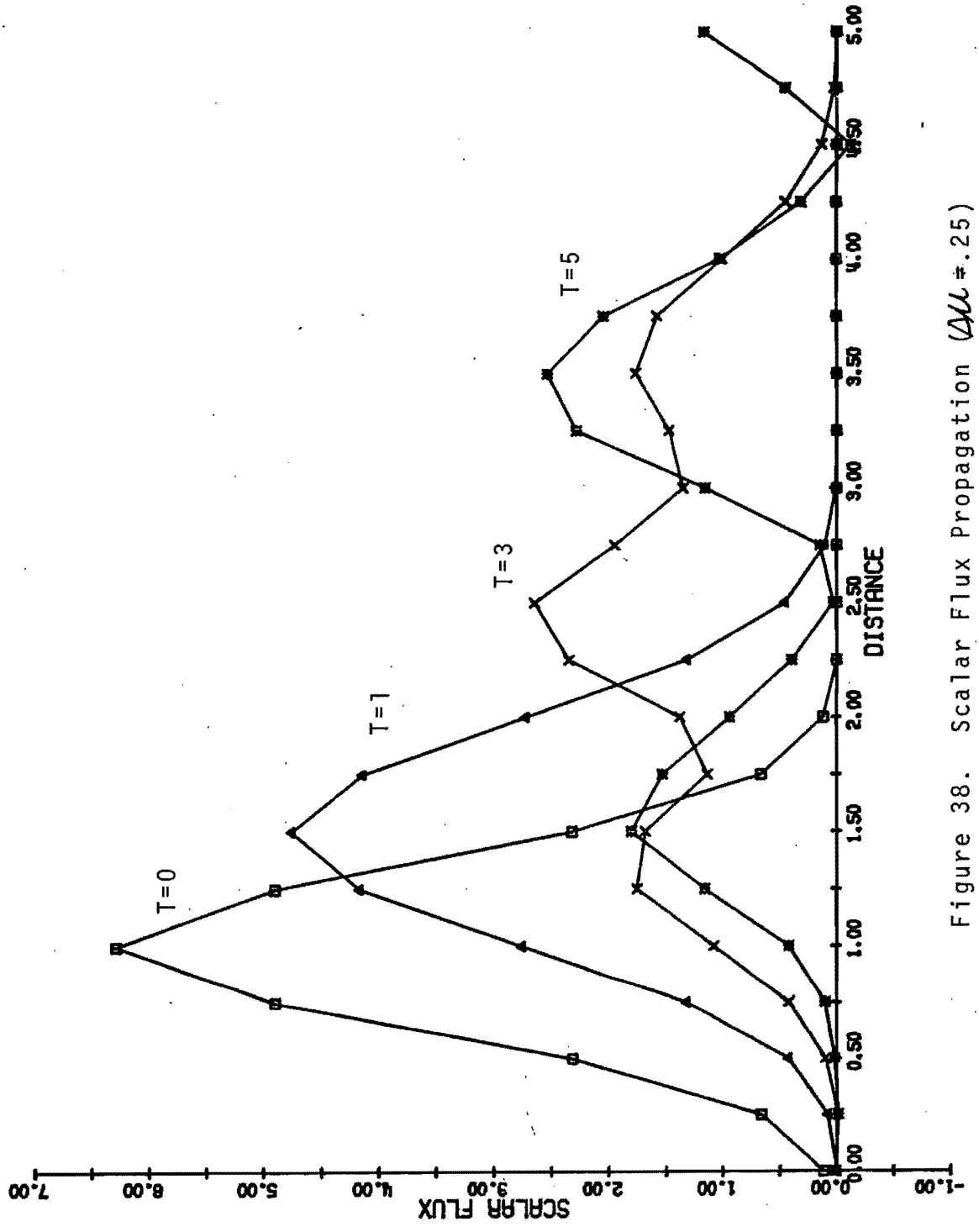


Figure 38. Scalar Flux Propagation ($\Delta\mu = .25$)

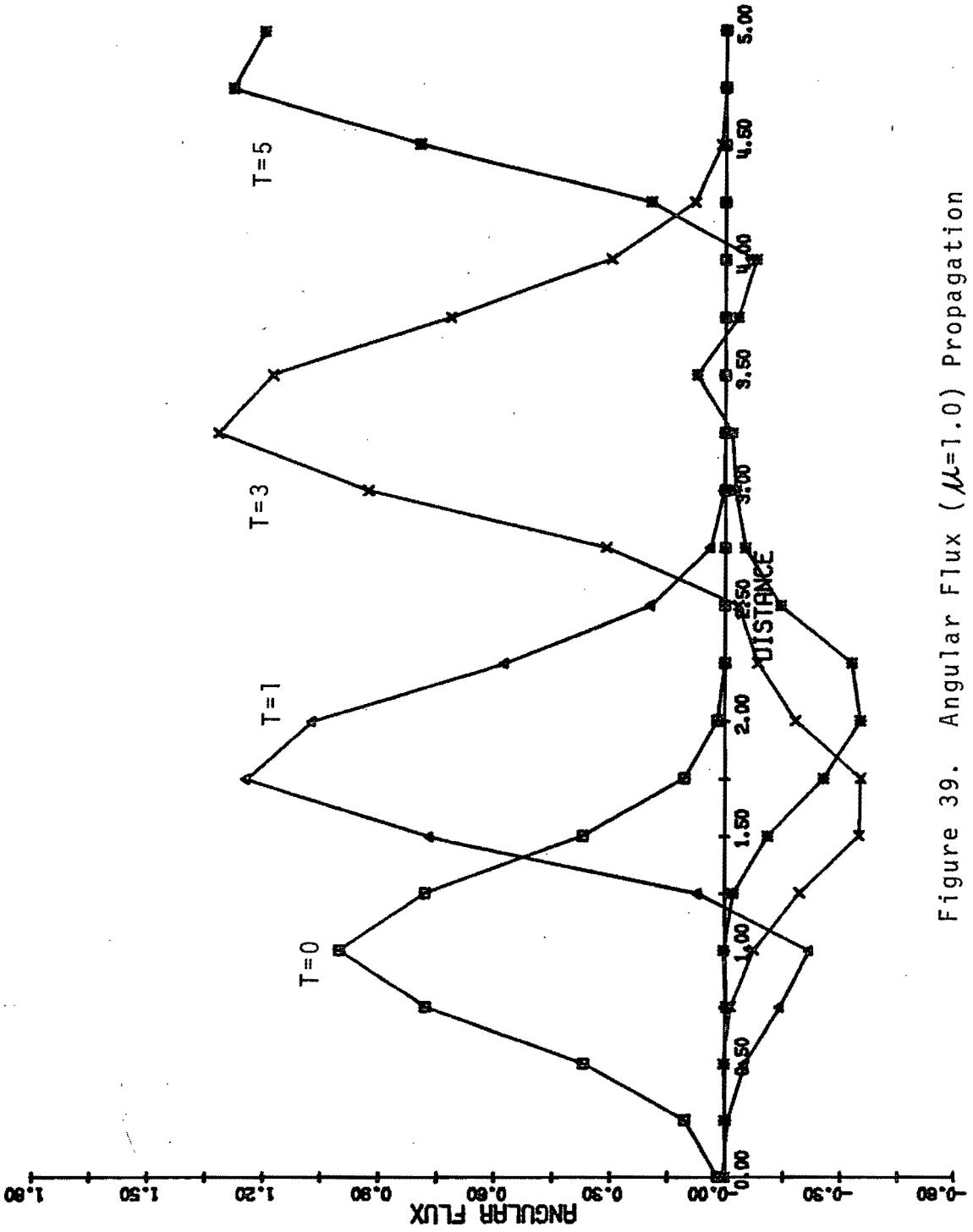


Figure 39. Angular Flux (μ=1.0) Propagation

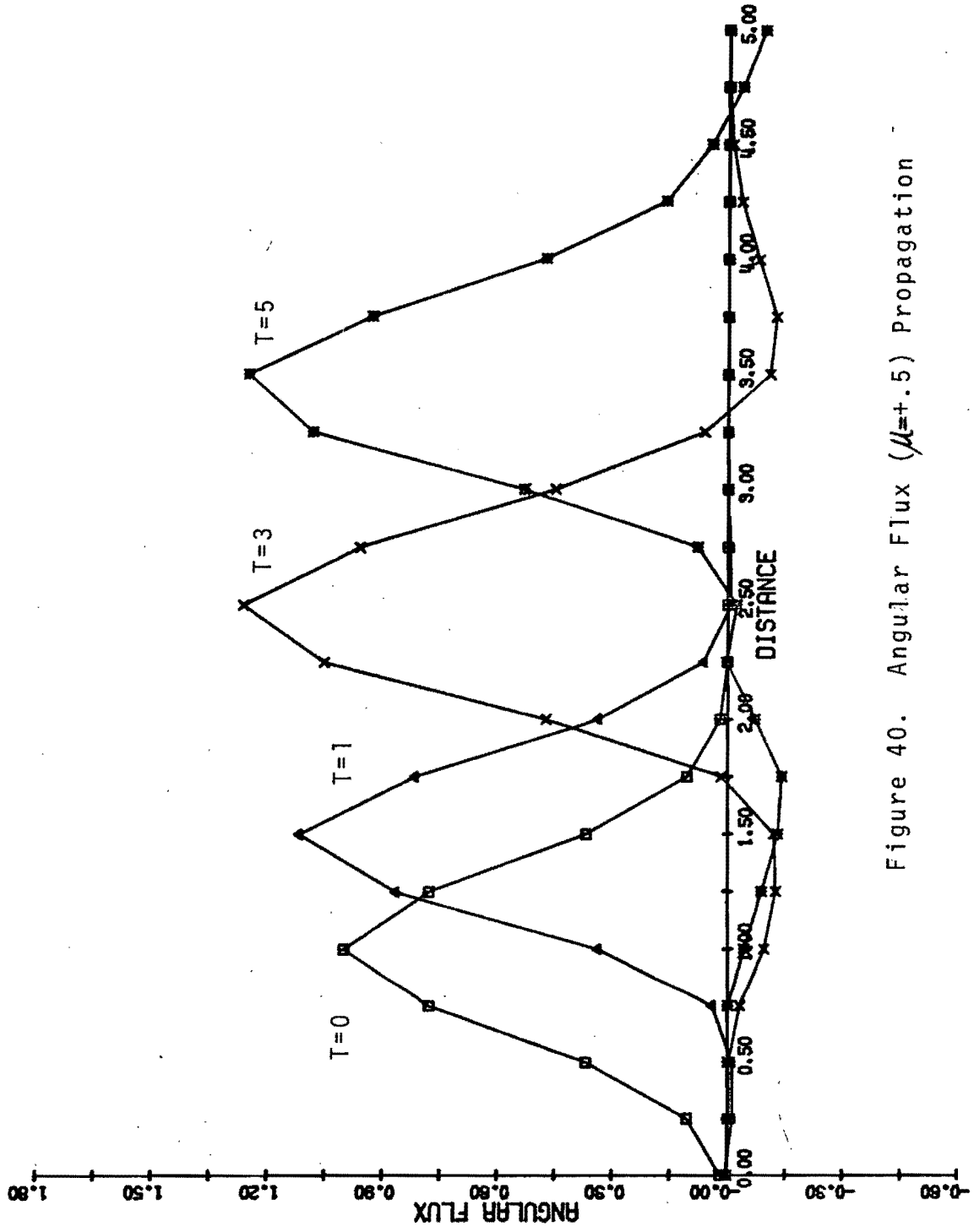


Figure 40. Angular Flux (μ=+.5) Propagation

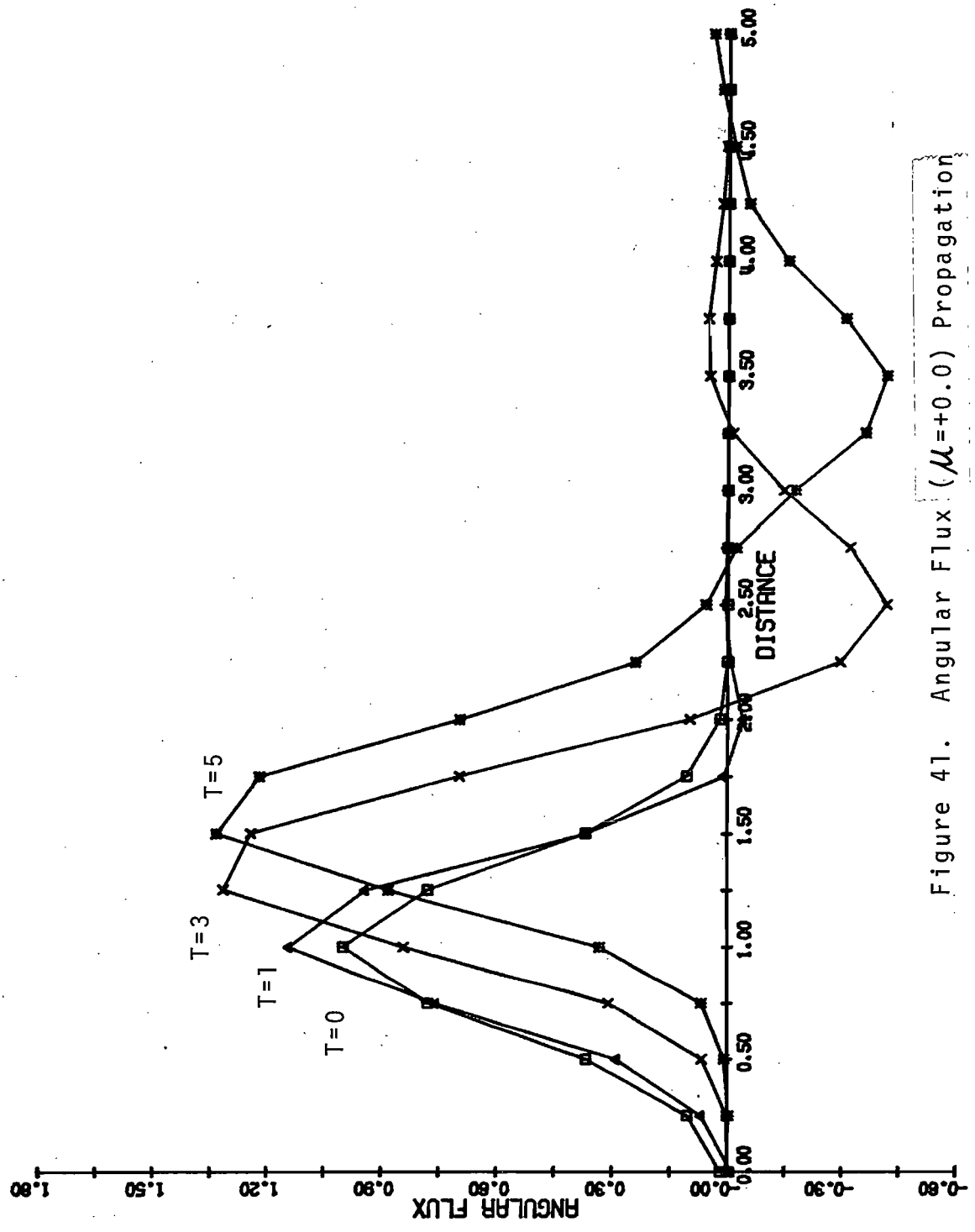


Figure 41. Angular Flux (μ=+0.0) Propagation

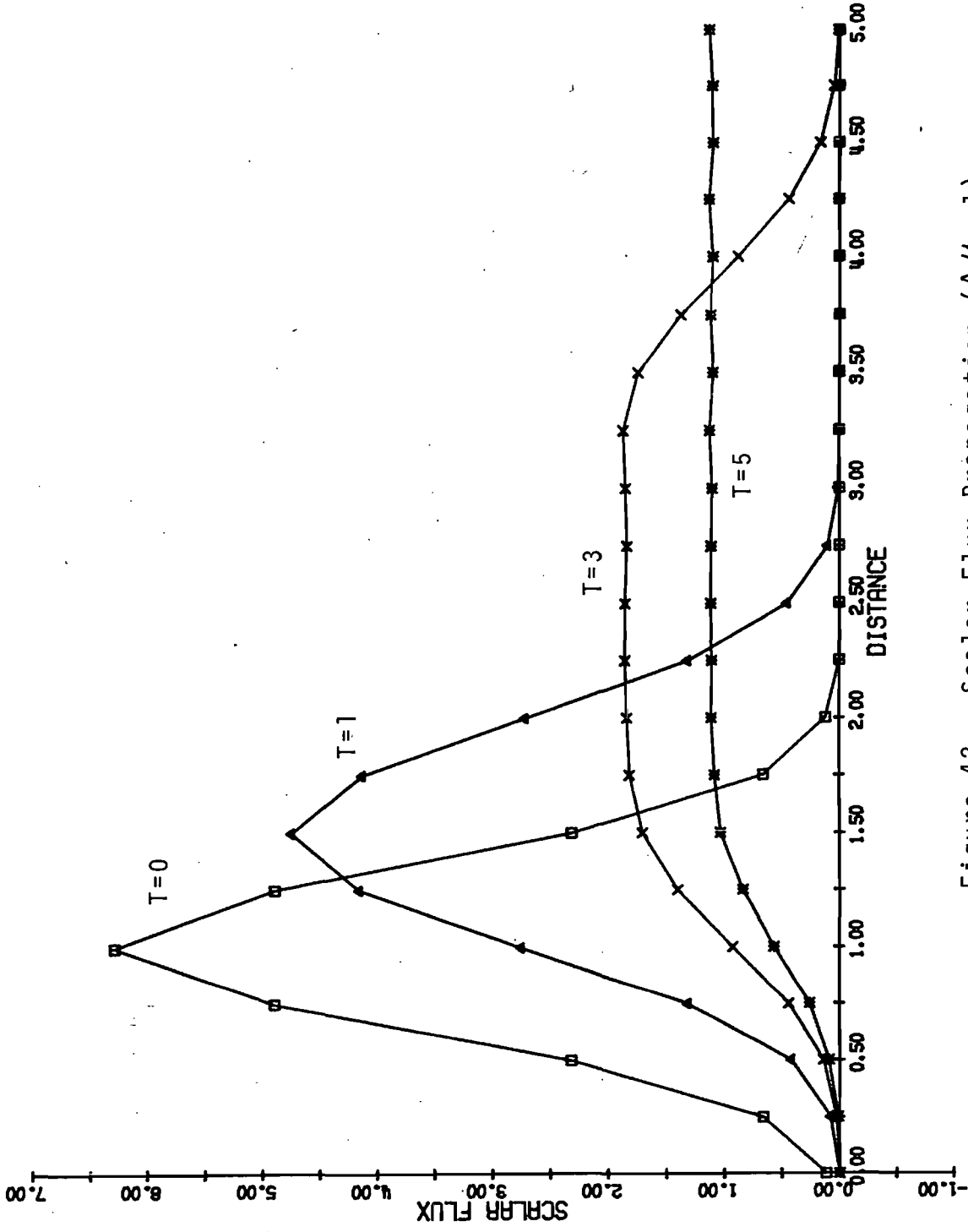


Figure 42. Scalar Flux Propagation ($\Delta\mu$=.1)

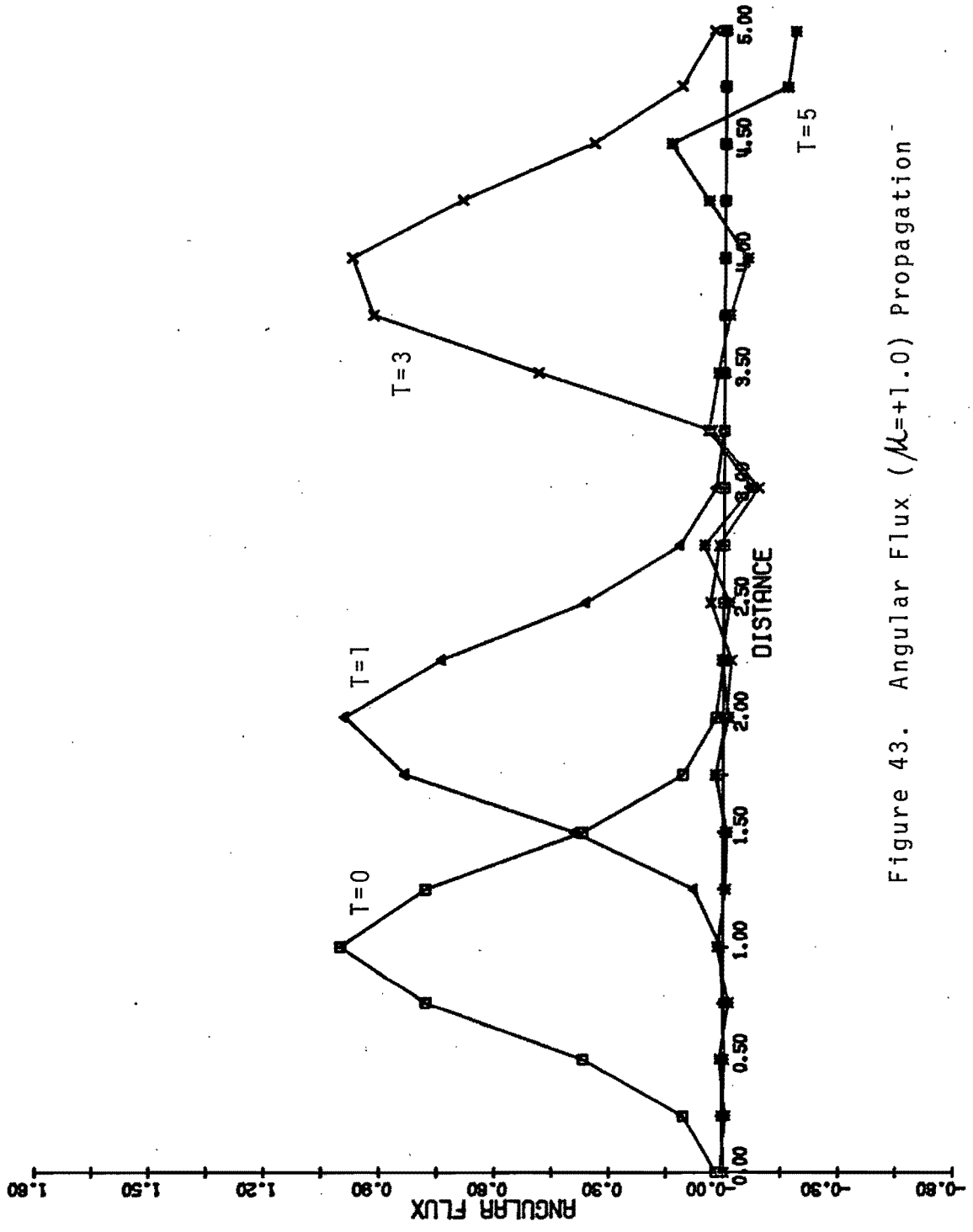


Figure 43. Angular Flux (μ=+1.0) Propagation

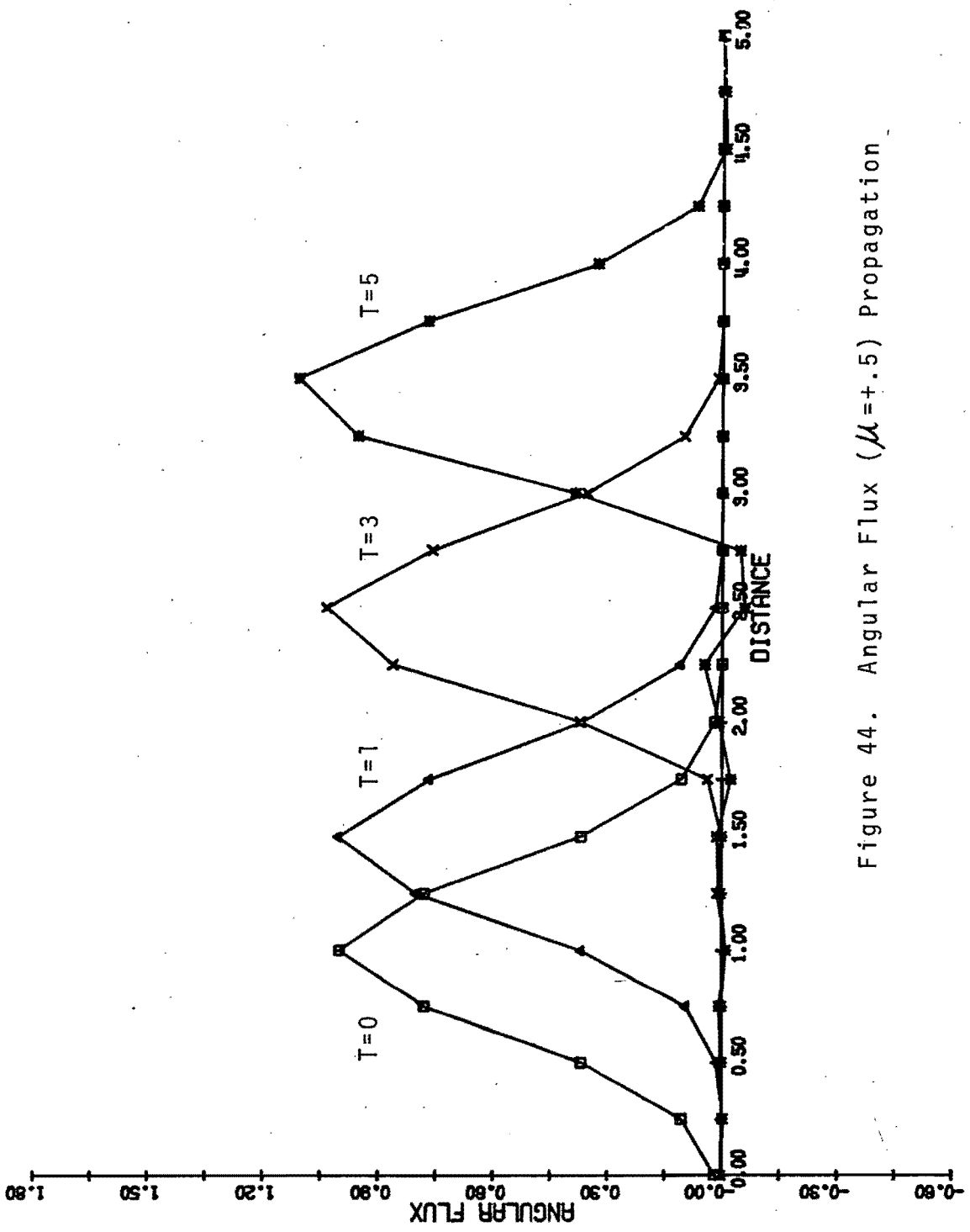


Figure 44. Angular Flux (μ=+.5) Propagation

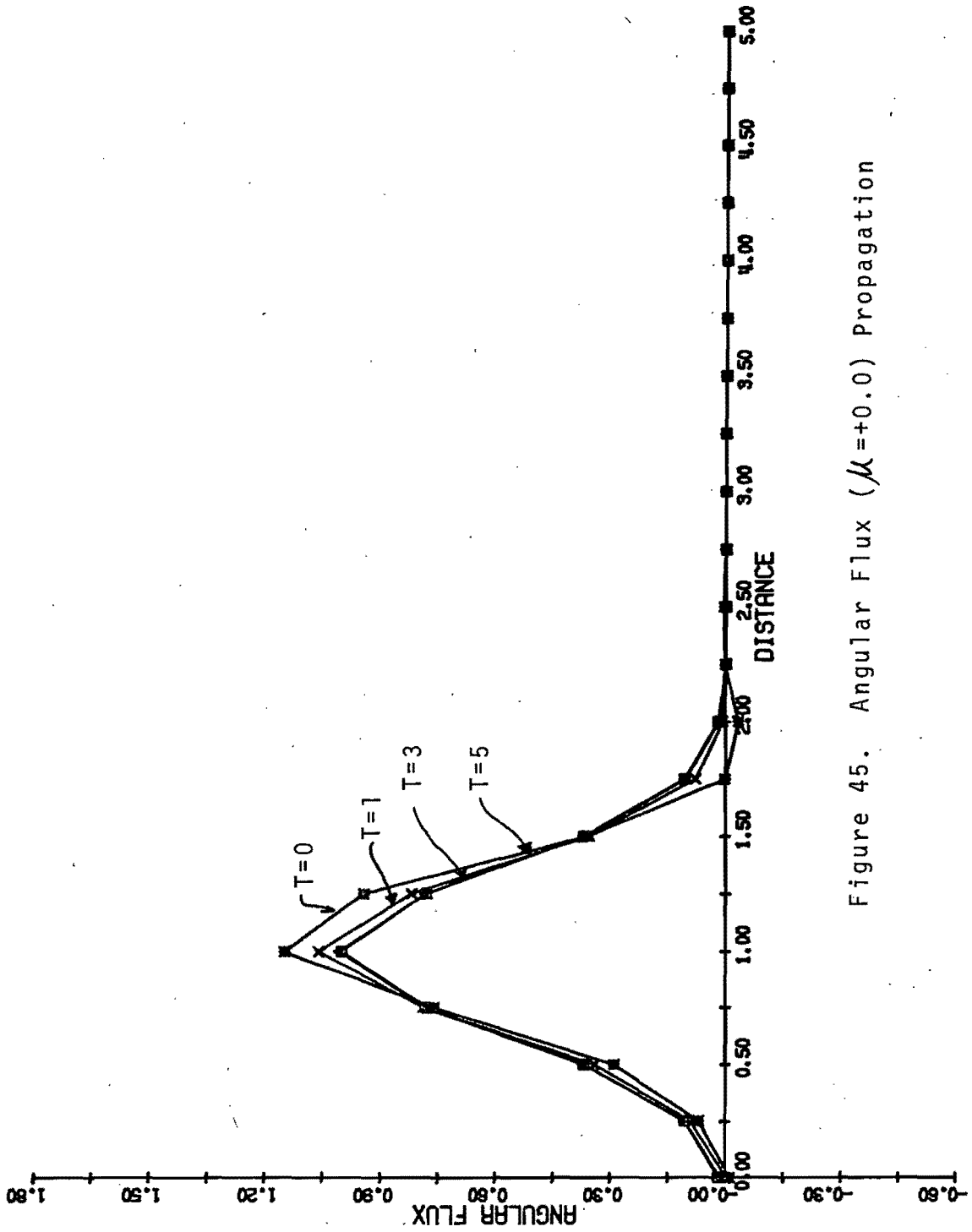


Figure 45. Angular Flux (μ=+0.0) Propagation

It should be noted that even for the coarser angular meshes, the individual angular fluxes propagate along the x-axis with the correct velocity, which is μv for the flux traveling in the direction $\theta = \cos^{-1}\mu$. However, it is exactly this propagation of angular modes with different velocities which results in the severe oscillatory behavior, especially with the coarser meshes. This separation of the angular modes (which is a numerical problem) is what causes the regular oscillations (or waves) in the scalar flux, because these waves correspond exactly to the location at that time of the individual angular modes along the x-axis. The following discussion will now consider this in more detail.

Let us consider the above pulse propagation problem in more generality, by assuming the initial angular flux is a Gaussian with half-width ξ,

$$Q_0(x,\mu) = e^{-x^2/\xi^2} \tag{162}$$

Now the exact solution to the time-dependent transport problem for this situation,

$$\frac{1}{v}\frac{\partial Q}{\partial t} + \mu\frac{\partial Q}{\partial x} = 0 \tag{163}$$

is simply

$$Q(x,\mu,t) = Q_0(x-\mu v t,\mu) \tag{164}$$

which is readily verified by direct substitution into Eq. (163).

Using Eq. (164), the solution to the initial Gaussian pulse problem is

$$Q(x,\mu,t) = e^{-(x-\mu v t)^2/\xi^2} \tag{165}$$

which is valid for $t \geq 0$.

Now define an equivalent temporal half-width τ corresponding to the spatial half-width ξ,

$$\tau \equiv \xi/v$$

and express Eq. (165) in the form

$$Q(x,\mu,t) = e^{-(\mu - \frac{x}{vt})^2/(\tau/t)^2} \tag{166}$$

which is now seen to be a Gaussian distribution in μ centered about x/vt with half-width τ/t. Physically, one expects the pulse to be centered at μ= x/vt since the neutrons propagate at a velocity μv along the x-axis. However, the half-width, τ/t, is interesting because it decreases with time, which implies that the angular distribution is becoming narrower as time advances.

This poses a severe challenge for a numerical method because in order to resolve the solution satisfactorily, the mesh must be finer than the width of the pulse. However, the above analysis indicates that no mesh is capable of treating this simple transport problem because if one waits long enough, the half-width will certainly becomes less than the mesh. Thus the angular mesh will always fail to resolve the pulse if sufficiently long times are considered.

An estimate may be made concerning the maximum time that a given angular mesh will resolve the pulse by requiring

$$\Delta\mu < \tau/t$$

i.e., requiring the angular mesh to be less than the half-width of the pulse in the angular variable. For the problem solved above,

$$v = 1.0 \text{ cm/sec}$$
$$\xi = .5 \text{ cm}$$

which results in

$$\tau = \xi/v = .5 \text{ sec}$$

or

$$t < .5/\mu$$

for satisfactory resolution of the angular dependence of the neutron pulse. For the three cases solved, $\Delta\mu = 1.0$, .25, and .1, thus results in

$$t < .5 \text{ sec}, 2.0 \text{ sec}, \text{ AND } 5.0 \text{ sec}$$

respectively. Interestingly, the numerical results agree reasonably well with these estimates in that for $\Delta\mu = 1.0$, the oscillatory behavior occurs within 1 sec while for $\Delta\mu = .25$, the oscillations do not begin until t = 2-3 seconds. Finally, for the $\Delta\mu = .1$ case, the oscillations are insignificant until t = 5, and then only for the forward peaked angular modes.

The significance of this phenomenon is not restricted to pulses of the type considered here. Since any distribution may be expressed in terms of a superposition of such pulses, this problem will occur

for any time-dependent transport problem. This discussion is limited of course to orthogonal geometries where it is possible for a pulse to be localized in space even after long times (e.g., the angular modes near $\mu = 0$ will remain near $x = 0$). For spherical geometries, this will not occur because the pulses will move away from the origin for any angle, although the distribution may become forward peaked for large radius. However, the separation of angular modes will not occur because the geometry does not allow streaming in specific directions (i.e., μ changes as the neutron streams in spherical geometry).

CHAPTER IX

CONCLUSIONS AND RECOMMENDATIONS

IX.A. Conclusions

On the basis of the analytical and numerical results presented in the preceding chapters, we make the following conclusions concerning the applicability of the finite element method to neutron transport, as applied via the integral law approach to the first order form of the transport equation.

The results of the eigenvalue calculations (with both isotropic and anisotropic scattering) and the Milne problem indicate that the finite element method yields extremely accurate results with reasonable size meshes. At the same time, the computational effort required to attain a given accuracy would appear to be at least comparable to conventional methods, and significantly less for those applications which can be solved using repeated back substitutions.

As the solution of the heterogeneous problem in Sec. VIII.F. indicates, when discontinuous spatial elements (arbitrary positions) and discontinuous angular elements (at $\mu = 0$) are used, the finite element method is capable of analyzing problems involving severe heterogeneities. In particular, the use of discontinuous spatial elements can eliminate the need for local mesh refinement to approximate a rapidly varying solution near an interface or other discontinuity. For problems that involve a large amount of angular uncoupling (such as the problem considered in Sec. VIII.F), the use of discontinuous angular elements at $\mu = 0$ is necessary to obtain reasonable results. Moreover,

the use of discontinuous angular elements at $\mu = 0$ has generally resulted in superior results for all applications, including eigenvalue problems, uniform source problems, etc., and should probably be used as a matter of course.

Since one has considerable choice with the finie element method when constructing the angular mesh, the only restrictions being a node (preferably a double node) at $\mu = 0$ and nodes at $\mu = \pm 1.0$, the finite element method is capable of treating extremely anisotropic problems with few "wasted" nodes away from the angular region of interest. If one were to employ triangular elements in the $x - \mu$ phase space, this would probably allow an excellent treatment of deep penetration problems. However, it would probably be necessary in this case to restrict triangles from crossing the line $\mu = 0$, so that discontinuous basis functions could be used.

The finite element method works equally well for problems characterized by pure absorption, pure scattering, voided regions, or problems with arbitrarily high orders of anisotropy, with no increase in execution time for comparable size meshes. This is in sharp contrast to discrete ordinates methods where solution accuracy and efficiency are quite sensitive to the physical problem being solved. Theoretically there is no problem associated with its application to multi-dimensional problems, even when anisotropic scattering is included.

If desired, higher order finite element approximations may easily be used. For the angular approximation this presents no increase in execution time since all angular nodes are coupled by scattering. However, for the spatial approximation, this results in more spatial

coupling which leads to greater storage requirements and greater execution times. Thus the increased accuracy may not justify the increase in cost to perform the calculation. This particular point has not been examined carefully, however, and it is possible that for reasonably homogeneous problems the use of higher order spatial elements may allow the use of a coarse enough mesh to result in a decrease in the execution time required to achieve a given accuracy.

For time-dependent transport, the finite element method in conjunction with the Crank-Nicholson discretization of the time dependence would appear to yield acceptable results. This is true despite the fact that transport in plane geometry is seen to result in unique numerical difficulties due to the separation of angular modes or, equivalently, the sharpening of the neutron distribution in phase space. The time-dependent solution method is seen to be extremely attractive because only back substitutions were required once the coefficient matrix is factored at the beginning of the solution process.

As far as the convergence of the finite element method with mesh refinement is concerned, the numerical results indicate $O(h^{k+1})$ convergence for the L_2 error in the approximate solution and $O(h^{2k+1})$ convergence for the error in the smallest eigenvalue. Theoretically there are difficulties with obtaining these estimates; however, plausible arguments have been made to support these observed convergence rates. In addition, the $O(h^{k+1})$ convergence was explicitly demonstrated for linear elements ($k = 1$) by using the truncation error associated with the finite element approximation and a numerical estimate for the norm of the inverse of the coefficient matrix. Finally, it was shown that

the observed convergence rates for the solution and the eigenvalue were consistent, which is more evidence to support the conclusion that the observed convergence rates are valid.

To summarize, the finite element method is an efficient, reliable numerical technique for obtaining stable and accurate solutions to both steady-state (including eigenvalue formulations) and time-dependent transport equations. The main objection to the finite element method, as noted by others[20,21] would appear to be the necessity to compute and store the coefficient matrix and then to solve the corresponding system of equations directly, at least for the first order approach. For large problems, especially multi-dimensional applications, the storage requirement may become prohibitive for many computing installations. However, as advances are made in data management techniques and computer fast memory development, this drawback may be mitigated to some extent.

IX.B. Recommendations for Further Study

Since this entire investigation has been concerned with 1-D plane geometry as far as the practical application of the finite element method is concerned, the obvious generalizations are to examine 2-D or even 3-D geometries and curvilinear coordinates. The generalization to 1-D spherical geometry should be straightforward, since only a few of the integrals will change, as long as tensor product basis functions in Λ-μ space are used. 1-D cylindrical geometry poses more of a challenge because two angles are required even though only one spatial coordinate is needed; however, the generalization to treat the two angles is a logical step on the way to a code capable of analyzing transport in multi-dimensional geometries.

However, the application to multi-dimensional geometries poses a substantial challenge because of the size of the problems encountered. For example, a 3-D transport problem involves five variables, ignoring time and energy. Even a coarse mesh with 10 nodes along each axis involves 10^5 unknowns, and matrices requiring 10^6-10^7 storage locations at least. Also, the structure of the matrices will be block-banded along the diagonal with bands of blocks in the interior, hence care will need to be taken during the elimination process to avert fill-in of zeroes (this was not a concern with the 1-D geometry).

Time-dependent transport could be examined with the multi-dimensional geometries, since the method is quite general. However, the cost of the solution could become quite exorbitant.

Energy dependence can easily be included since standard source iteration techniques could be used to solve the multi-group transport problems. It is possible that their solution could be made quite efficient if the within-group matrix problems are solved by the LU method, and the LU matrices are stored on relatively efficient peripheral storage. In this way, when the particular group equation is being solved, the $\underline{\underline{L}}$ and $\underline{\underline{U}}$ matrices could be recalled and the solution obtained by simple back substitutions. Of course, this would require storage of an $\underline{\underline{L}}$ and $\underline{\underline{U}}$ for each group; however, the increase in efficiency may justify the cost and effort to store the matrices. Since the acceptance of the finite element method as an acceptable tool in reactor analysis will probably not occur until a code is developed with the multi-group capability of standard codes such as ANISN or ONETRAN, this generalization to a multi-group treatment may be warranted.

In summary, the recommendations for future effort would be to allow treatment of

(1) spherical geometry

(2) cylindrical geometry

(3) 2-D and 3-D geometries

(4) multi-group treatment

In addition, from a theoretical standpoint, a rigorous proof of the $O(h^{k+1})$ solution error and $O(h^{2k+1})$ eigenvalue error should be given.

REFERENCES

1. Zienkiewicz, O.C., The Finite Element Method in Engineering Science, McGraw-Hill, London (1971).

2. Zienkiewicz, O.C., "Finite Elements - The Background Story," in The Mathematics of Finite Elements and Applications, J.R. Whiteman, ed., Academic Press (1973).

3. Strang, G. and Fix, G.J., An Analysis of the Finite Element Method, Prentice-Hall (1973).

4. Norrie, D.H., and de Vries, G., The Finite Element Method, Academic Press (1973).

5. Kaper, H.G., Leaf, G.K., and Lindeman, A.J., "A Timing Comparison Study for Some High Order Finite Element Approximation Procedures and a Low Order Finite Difference Approximation Procedure for the Numerical Solution of the Multigroup Neutron Diffusion Equation," Nuclear Science and Engineering, 49, 27-48 (1972).

6. Hansen, K.F. and Kang, C.M., "Finite Element Methods in Reactor Physics Analysis," in Advances in Nuclear Science and Technology, Vol. 8, 173-253, Academic Press (1975).

7. Kang, C.M., and Hansen, K.F., "Finite Element Methods for Reactor Analysis," Nuclear Science and Engineering, 51, 456-495 (1973).

8. Semenza, L.A., Lewis, E.E., and Rossow, E.C., "The Application of the Finite Element Method to the Multigroup Neutron Diffusion Equation," Nuclear Science and Engineering, 47, 302-310 (1972).

9. Lathrop, K.D., "Transport Theory Numerical Methods," Proc. Conf. on Mathematical Models and Computational Techniques for Analysis of Nuclear Systems, USAEC Report CONF-730414-P1, I1-19, (1973).

10. Froehlich, R., "Current Problems in Multi-Dimensional Reactor Calculations," Proc. Conf. on Mathematical Models and Computational Techniques for Analysis of Nuclear Systems, USAEC Report CONF-730414-P2, VII-1 - VII-66 (1973).

11. Lathrop, K.D., "Remedies for Ray Effects," Nuclear Science and Engineering, 45, 255 (1971).

12. Kaper, H.G., Leaf, G.K., and Lindeman, A.J., "Application of Finite Element Techniques for the Numerical Solution of Neutron Transport and Diffusion Equations," Proc. Conf. on Transport Theory, Second CONF-710107, Los Alamos, New Mexico (1971).

24. Pitkaranta, J. and Silvennoinen, P., "Computational Experimentation on the Finite Element Method in Bare Slab Criticality Calculations," Nuclear Science and Engineering, 50, 297-300 (1973).

25. Pitkaranta, J. and Silvennoinen, P., "Finite Element Analysis of Some Critical Fast Assemblies," Nuclear Science and Engineering, 52, 447-453 (1973).

26. Pitkaranta, J., "A Non-Self-Adjoint Variational Procedure for the Finite-Element Approximation of the Transport Equation," Transport Theory and Statistical Physics, 4, 1-24 (1975).

27. Bell, G.I. and Glasstone, S., Nuclear Reactor Theory, Van Nostrand Reinhold Co. (1970).

28. Duderstadt, J.J. and Hamilton, L.J., Nuclear Reactor Analysis, John Wiley and Sons (1976).

29. Zlamal, M., "Some Recent Advances in the Mathematics of Finite Elements," in The Mathematics of Finite Elements and Applications, J.R. Whiteman, ed., 59-81, Academic Press (1973).

30. Mitchell, A.R., "An Introduction to the Mathematics of the Finite Element Method," in The Mathematics of Finite Elements and Applications, J.R. Whiteman, ed., 37-58, Academic Press (1973).

31. Oden, J.T., "Finite Element Applications in Mathematical Physics," in The Mathematics of Finite Elements and Applications, J.R. Whiteman, ed., 239-282, Academic Press (1973).

32. Case, K.M. and Zweifel, P.F., Linear Transport Theory, Addison-Wesley Publishing Co. (1967).

33. Hellwig, G., Differential Operators of Mathematical Physics, Addison-Wesley Publishing Co. (1964).

34. Fix, G.J., "Eigenvalue Approximation by the Finite Element Method," Advances in Mathematics, 10, 300-316 (1973).

35. Fix, G.J., Lecture Notes for Mathematics 571-572, Course in Numerical Analysis, University of Michigan (1973-1974).

36. Martin, W.R., "Convergence of the Finite Element Method in Neutron Transport," Transactions of the American Nuclear Society Winter Meeting, 22, 251-252 (1975).

37. Nering, E.D., Linear Algebra and Matrix Theory, John Wiley and Sons, Inc. (1963).

38. Lasaint, P. and Raviart, P.A., "On a Finite Element Method for Solving the Neutron Transport Equation," Mathematical Aspects of Finite Elements in Partial Differential Equations, Proc. Univ. of Wisconsin Symposium, 89-123, Academic Press (1974).

13. Ohnishi, T., "Application of Finite Element Solution Technique to Neutron Diffusion and Transport Equations," Proc. Conf. on New Developments in Reactor Mathematics and Applications, USAEC Report CONF-710302 (1971).

14. Ohnishi, T., "Finite Element Solution Technique for Neutron Transport Equations (F_N Approximation)," Proceedings Symposium on Numerical Reactor Calculations, 629-638, International Atomic Energy Agency, Vienna (1972).

15. Ukai, S., "Solution of Multi-Dimensional Transport Equation by Finite Element Method," Journal of Nuclear Science and Technology, 9, 366-373 (1972).

16. Miller, W.F., Lewis, E.E., and Rossow, E.C., "The Application of Phase-Space Finite Elements to the One-Dimensional Neutron Transport Equation," Nuclear Science and Engineering, 51, 148-156 (1973).

17. Miller, W.F., Lewis, E.E., and Rossow, E.C., "The Application of Phase-Space Finite Elements to the Two-Dimensional Neutron Transport Equation in X-Y Geometry," Nuclear Science and Engineering, 52, 12-22 (1973).

18. Lewis, E.E., Miller, W.F., and Henry, T.P., "A Two-Dimensional Finite Element Method for Integral Neutron Transport Calculations," Nuclear Science and Engineering, 58, 203-212 (1975).

19. Briggs, L.L., Miller,W.F., and Lewis, E.E., "Ray-Effect Mitigation in Discrete Ordinate-Like Angular Finite Element Approximations in Neutron Transport," Nuclear Science and Engineering, 57, 205-217 (1975).

20. Kaper, H.G., Leaf, G.K., and Lindeman, A.J., "Applications of Finite Element Methods in Reactor Mathematics. Numerical Solution of the Neutron Transport Equation," ANL-8126, Argonne National Laboratory (1974).

21. Yuan, Y.C., Lewis, E.E., and Miller, W.F., "Iterative Solution Methods for Two-Dimensional Finite Element Approximations in Neutron Transport," Proceedings of ANS Conference on Computational Methods in Nuclear Engineering, CONF-750413, p. III-85, Charleston, S.C. (1975).

22. Reed, W.H., Hill, T.R., Brinkley, F.W., and Lathrop, K.D., "TRIPLET: A Two-Dimensional, Multigroup, Triangular Mesh, Planar Geometry, Explicit Transport Code," LA-5428-MS, Los Alamos Scientific Laboratory (1973).

23. Hill, T.R., "ONETRAN: A Discrete Ordinates Finite Element Code for the Solution of the One-Dimensional Multigroup Transport Equation," LA-5990-MS, Los Alamos Scientific Laboratory (1975).

39. Cavendish, J.C., "Local Mesh Refinement Using Rectangularly Blended Finite Elements," Journal of Computational Physics, 19, 211 (1975).

40. Roache, P.J., Computational Fluid Dynamics, Hermosa Publishers, Albuquerque, N.M. (1972).

41. Fried, I., "The l_2 and l_∞ Condition Numbers of the Finite Element Stiffness and Mass Matrices, and the Pointwise Convergence of the Method," in The Mathematics of Finite Elements and Applications, J.R. Whiteman, ed., 163-174, Academic Press (1973).

42. Kang, C.M. and Hansen, K.F., Letter to the Editor, "Response to "Comments on 'Finite Methods for Reactor Analysis'"" , Nuclear Science and Engineering, 56 , 226-228 (1974).

43. Dahlquist, G. and Bjorck, A., Numerical Methods, Prentice-Hall (1974).

44. Steinke, R.G., "A Review of Direct and Iterative Strategies for Solving Multi-Dimensional Finite Difference Problems," University of Michigan Nuclear Engineering Report (1971).

45. Lingus, C., "Analytical Test Cases for Neutron and Radiation Transport Codes," Proc. Conf. on Transport Theory, Second CONF-710107, Los Alamos, New Mexico (1971).

46. Engle, W.W., Jr., ANISN, A One Dimensional Discrete Ordinates Transport Code with Anisotropic Scattering," K-1693, Oak Ridge National Laboratory (1967).

47. "ANISN-ORNL, A One Dimensional Discrete Ordinates Transport Code," CCC-254, Oak Ridge National Laboratory (1973).

48. Kaper, H.G., Lindeman, A.J., and Leaf, G.K., "Benchmark Values for the Slab and Sphere Criticality Problem in One-Group Neutron Transport Theory," Nuclear Science and Engineering, 54, 94-99 (1974).

49. Kschwendt, H., "The SP_N-P_L Method for Neutron Transport in Homogeneous Slabs with Anisotropic Scattering," Nuclear Science and Engineering, 44, 423-434 (1971).

50. Case, K.M., de Hoffman, F., and Placzek, G., Introduction to the Theory of Neutron Diffusion, Los Alamos Scientific Laboratory (1953).

51. Engle, W.W., Boling, M.A., and Colston, B.W., "DTF-II, A One-Dimensional, Multigroup Neutron Transport Program," NAA-SR-10951, Atomics International (contained within Reference 47 above), (1966).

52. Glasstone, S. and Edlund, M.C., The Elements of Nuclear Reactor Theory, D. Van Nostrand Co., Inc. (1952).

53. Reed, W.H., "New Difference Schemes for the Neutron Transport Equation," Nuclear Science and Engineering, 46, 309-314 (1971).